
# DEVELOPMENT AND VALIDATON OF MULTIVARIATE STRATEGIES FOR FOOD QUALITY CONTROL

**Glòria Rovira Garrido**

UNIVERSITAT ROVIRA i VIRGILI

# Development and validation of multivariate strategies for food quality control

Glòria Rovira Garrido

**DOCTORAL THESIS**

**2024**

Glòria Rovira Garrido

# Development and validation of multivariate strategies for food quality control

Doctoral Thesis

Supervised by:

Prof. María Pilar Callao Lasmarías

Prof. Itziar Ruisánchez Capelástegui

Department of Analytical Chemistry and Organic Chemistry

UNIVERSITAT
ROVIRA i VIRGILI

Tarragona

2024

UNIVERSITAT ROVIRA i VIRGILI
Dept. de Química Analítica
i Química Orgànica

Prof. María Pilar Callao Lasmarías and Prof. Itziar Ruisánchez Capelástegui, both Professor of Analytical Chemistry at the Department of Analytical and Organic Chemistry at Universitat Rovira i Virgili,

CERTIFY,

that the Doctoral Thesis entitled *"Development and validation of multivariate strategies for food quality control"*, submitted by Glòria Rovira Garrido to receive the degree of Doctor with International Mention by Universitat Rovira i Virgili has been carried out under our supervision at the Department of Analytical and Organic Chemistry of this University, and all the results presented in this thesis were conducted by the above mentioned student.

Tarragona, 28 May 2024,

María Pilar Callao Lasmarías

Itziar Ruisánchez Capelástegui

The research work presented in this doctoral thesis is the outcome of a project funded by both institutions under the collaboration framework agreement between the Diputació de Tarragona and the Universitat Rovira i Virgili for the period 2020–2023, with the reference number 21/22/23PIN-DIPTA-URV01.

The research work has been developed with the Martí i Franquès (2020PMF-PIPF-22) grant from the Universitat Rovira i Virgili.

# Agraïments

Farà quatre anys en una època de pandèmia mundial, vaig acabar un màster amb moltes ganes de menjar-me el món laboral. Però degut a les circumstàncies externes no va ser possible, i vaig decidir no quedar-me estancada i tirar endavant per una altra via, que els que em coneixen molt bé saben que era una opció que no m'havia plantejat mai. Soc una mica l'exemple de la dita "no diguis mai d'aquesta aigua no en beuré". Vist en perspectiva és una de les millors decisions que vaig prendre, tot i que ha sigut una aventura difícil i tediosa en algun moment, però molt satisfactòria i en la que he estat sempre molt ben acompanyada.

Tengo que empezar primero dando las gracias a mis dos supervisoras, Pilar y Itziar. Gracias por darme la oportunidad de formar parte y crecer tanto personal como profesionalmente dentro del grupo de Chemosens. Por confiar, por el apoyo y por haberme enseñado tantas cosas a lo largo de estos tres años. No puc deixar de donar les gràcies a tots els integrants del grup d'investigació Chemosens, per les calçotades, les paelles a l'estiu i per les paraules d'ànim. Especialment, volia agrair al Santi per totes les ajudes amb el Matlab.

Eu quero também agradecer minha orientadora na Universidade Federal de Minas Gerias (UFMG). Scheilla, muito obrigada por me receber quase como uma filha durante os três meses em Belo. Obrigada por todas as palavras de apoio e motivação, pelos abraços nos momentos más difíceis e por todos os momentos vividos dentro e fora do laboratório. Você tem um grupo de pesquisa maravilhoso. Carolina, muito obrigada pela sua ajuda com o NIR e o RMN e pela sua calma com os resultados. Foi um prazer trabalhar com você e compartir tempo fora do laboratório. Ronália, minha querida amiga e companheira de forro. Muito obrigada por estar sempre disposta a me ajudar dentro e fora do laboratório. Pelos nossos passeios, botecos e karaokê! Marina, você estava no meio da redação da sua tese más também foi um grande apoio. Muito obrigada, pelas palavras de ânimo nos

seminários fazendo um cafezinho e muito obrigada pelo passeio em Ouro Preto com Ronália, foi ótimo! Laura, muito obrigada pela sua ajuda com as amostras, eu acho que não teria conseguido preparar e medir todas as amostras sem você. Bruna, muito obrigada pela sua ajuda também e por sempre ter um sorriso no laboratório. Marcão, muito obrigada por sempre estar disposto a ajudar.

Não posso esquecer do Marcelo e do seu grupo de pesquisa. Marcelo muito obrigada por o primeiro passeio na Lagoa, por me ensinar a história de Belo. Muito obrigada pela sua ajuda durante os três meses, os conselhos, ideias e palavras de apoio. Poliana, muito obrigada pela sua ajuada com os RMN e os decaimentos. Ana, muito obrigada pela sua ajuda com o NIR, eu tenho que voltar para comer a melhor coxinha de Belo Horizonte. Muito obrigada a todos por querer formar parte dessa experiência comigo.

Dins de la universitat hi ha tres persones que també han format part d'aquesta aventura, tot i no formar part del meu grup d'investigació, i de ben segur que sense ells no hagués sigut el mateix. Uri, gràcies per seguir formant part de la meva vida després de tants anys. Gràcies pel recolzament i els ànims durant tot el doctorat. Adrià, gràcies per fer-me riure cada dia, per compartir els enfados i blasfemar amb mi. Et trobem a faltar una mica per aquí, però en res estem a Ghent tocant la pera. Javi, gracias por tus bromas y por estar siempre dispuesto a dar un abrazo. Als tres, no tinc paraules suficients per agrair tot el que hem compartit, fa 10 anys que formeu part de la meva vida i us heu convertit en família. Gràcies pels dinars, les festes, els viatges, les nits de jocs i els riures, gràcies de veritat. Us estimo molt.

Fora de la universitat hi ha molta gent que m'ha acompanyat al llarg d'aquest viatge. Magda, Aina i Cristina, tota la vida juntes i seguim aquí. Gràcies per estar sempre i seguir compartint la vida. Sou imprescindibles. A les meves JEFAS: Ana, Andrea, Clara, Profi, Conxi, Marisa i Irene, des de batxillerat donant guerra. Actualment, casi que estem cada una a un punt del mapa

diferent i es fa una mica més difícil coincidir. Gràcies per ser-hi, per compartir amb mi la indignació i per les cerveses curatives.

Gràcies a l'Andrea per les nostres trucades infinites que animen a qualsevol, per ser llum i força. Marisa i Irene, gràcies per ser calma i alegria en tot moment. Gràcies pels vins, les cerveses i les converses que curen l'ànima. Gràcies per ser-hi sempre. Moltes gràcies a totes, us estimo.

Marina, gràcies per ser la meva companya de ball, de purpurina i de barra. Per les converses, els xismes i riures al cotxe. Gràcies per acompanyar-nos dalt i baix de l'escenari, per no deixar-nos mai la mà en les nostres aventures de la vida i sostenir-nos sempre. T'estimo molt, ets indispensable a la meva vida.

No puc deixar de donar les gràcies a la meva família: mama, papa, Marta, Jordi. Moltes gràcies per confiar i recolzar-me en tot moment. Per calmar-me en moments d'angoixa i animar-me. Arnau, ets l'alegria de la família, segueix sent el nen més simpàtic del món, la teva padrineta t'adora. Moltes gràcies per estar sempre, us estimo molt!

Por último, pero no menos importante, gracias Derrick. Por confiar en mí, por ser mi gran apoyo, por calmarme en mis momentos de oscuridad y hacerme ver la luz al final del túnel. Gracias por tus ánimos y mimos en momentos de bajón, por hacerme reír cada día, por hacerme desconnectar, por darme fuerza y sobre todo por tu paciencia a lo largo de esta tesis. Gracias por seguir a mi lado cumpliendo metas, otra más conseguida juntos de la mano. Te quiero mucho.

*"If you never try you'll never know just what you're worth"*

*Coldplay*

## Abbreviations list

ACC: Accuracy
ACR: Absolute centered residual
ANN: Artificial neural networks
ATR: Attenuated total reflectance
ATR-FTIR: Attenuated total reflectance Fourier transform infrared
AUC: Area under the curve
BN: Brazilian nut
CC: Centering correction
CCα: Decision limit
CCβ: Detection capability
CPGM: Carr-Purcell-Meiboom-Gill
CVCE: Cross-validation classification error
DG: Directorate General
DO: Denomination of origin
DS: Direct Standardization
EC: European Commission
EEFS: Excitation-emission fluorescence spectroscopy
EFF: Efficiency
Em: Emission
EUC: Eucalyptus honey
EVOO: Extra virgin olive oil
Ex: Excitation
FDA: Food and Drug Administration
FIR: Far infrared spectroscopy
FN: False negative
FP: False positive
FT: Fourier transform
GC: Gas chromatography
GC-FID: Gas chromatography flame ionization detector
GLSW: Generalized least squares weighting
HCA: Hierarchical clustering analysis
HPLC: High-performance liquid chromatography
ICP-MS: Inductively coupled plasma mass spectrometry
ILT: Inverse Laplace transform
IPC-OES: Inductively coupled plasma optical emission spectroscopy
IR: Infrared
IRMS: Isotope ratio mass spectrometry
ISO: International Organization for Standardization
KNN: K-nearest neighbor
LDA: Linear discriminant analysis
LF-NMR: Low-field nuclear magnetic resonance
LG: Les Garrigues
LV: Latent variables
LVF: Linear variable filter
M: Macadamia nut
MAPA: Ministerio da agriultura, pecuária e abastecimento
MIR: Mid-infrared spectroscopy
MSC: Multiplicative scatter correction
NA: Non-adulterated
NIR: Near-infrared spectroscopy

NMR: Nuclear magnetic resonance

OCPLS: One-class partial least squares

OCURR: Occurrence

OLAF: European anti-fraud office

P: Peanut

PC: Principal component

PCA: Principal component analysis

PCC: Performance characteristic curves

PCR: Principal component regression

PDO: Protected designation of origin

PDS: Piecewise direct standardization

PGI: Protected geographical indication

PLS: Partial least squares

PLS-DA: Partial least squares discriminant analysis

PN: Pecan nut

PR: Precision

QDA: Quadratic discriminant analysis

RF: Radio frequency

RMSE: Root mean square of errors

ROC: Receiver Operating Characteristic

S: Siurana

SBC: Slope/bias correction

SD: Score distance

SEN: Sensitivity

SFS: Synchronous fluorescence spectroscopy

SIMCA: Soft independent modelling of class analogies

SPC: Specificity

SVM: Support vector machine

SWS: Single wavelength standardization

TN: True negative

TP: True positive

TSG: Traditional specialties guaranteed

UNEQ: Unequal dispersed classes

UR: Uncertainty region

UV-Vis: Ultraviolet-Visible

W: Wild honey

## Abstract

The most common forms of food fraud are adulteration and authentication. This ongoing issue affects the integrity of the food supply chain and it might imply serious risks to consumer health and safety, as well as economic damage to businesses. Olive oil, honey, and nuts are highly consumed products, and due to their economic value, they become a target for different fraudulent practices. Although there are various official control protocols for nuts, olive oil, and honey analysis, new analytical methodologies are being developed using spectroscopic techniques coupled with multivariate analysis. Those methodologies offer numerous advantages over the reference protocols, including speed, sustainability, and the ability to preserve samples.

This thesis aims to develop and validate multivariate methodologies for the detection of adulteration or authentication of olive oil, honey, and nuts including the next issues: i) using different molecular spectroscopic instruments such as Near-Infrared (NIR), Attenuated Total Reflectance Fourier Transform Infrared (ATR-FTIR), Fluorescence, and Low-Field Nuclear Magnetic Resonance (LF-NMR); ii) defining an uncertainty region from the semi-quantitative information in adulteration cases; iii) optimizing the performance parameters obtained from the multivariate analysis, and iv) applying a transfer technique to keep the performance of models if measurement conditions change.

Three different adulteration problems are presented in this work; olive oil adulterated with sunflower oil, cashew nuts adulterated with other nuts (Brazilian nut, Pecan nut, Macadamia nut, and Peanut), and honey adulterated with syrups (Inverted sugar, corn and rice syrup). In all cases, adulteration was considered at several levels. So, different strategies were proposed to obtain semi-quantitative information.

For **olive oil** samples, the definition of a cut-off to determine the adulteration of a sample was proposed and Performance Characteristic Curves (PCC) were applied to obtain the semi-quantitative performance parameters.

In the case of **cashew nuts**, two different strategies were proposed, the first one is to use Receiver Operating Characteristic (ROC) curves for the optimization of the class limit value and data fusion to improve the performance parameters. The second strategy was to propose two class limits instead of just one, in a way to define an uncertainty region. For **honey** adulteration, the PCC were used to characterize the model, and an uncertainty region was defined setting two class limits. It has been proven that defining two class limits decreases the error in the sample assignment because the samples present in the uncertainty region should be submitted to a confirmatory analysis.

In the authentication problem with **extra virgin olive** oils from two different Protected Denominations of Origin (PDO), a standardization approach was proposed to manage the variations between harvests. It has been demonstrated the usefulness of the standardization techniques to maintain the performance of a multivariate classification model.

**Table of contents**

**List of Figures**

Paper 2.



Paper 3.

**List of tables**

# Chapter 1. Scope and Objectives



## 1.1. Scope

Food fraud is defined by Codex Alimentarius as "any deliberate action of businesses or individuals to deceive others regarding the integrity of food to gain undue advantage. Types of food fraud include but are not limited to adulteration, substitution, dilution, tampering, simulation, counterfeiting, and misrepresentation" [1]. The two principal problems related to food fraud are food authentication and food adulteration.

Food authentication is the process that confirms whether a food product complies with the description provided on its label. This can include the origin, production method, or processing technologies [2]. In Europe, one of the main issues in food authenticity is the origin. That is why the European Union legislation [3,4,5] defines specific names for geographical indications for food products: Protected Designation of Origin (PDO), Protected Geographical Indication (PGI), and Traditional Specialties Guaranteed (TSG). Recently, optional quality terms such as "mountain product" and "product of island farming" were introduced [6].

Food adulteration is defined as the deliberate addition of a non-declared substance to a food to alter its chemical composition and/or appearance, typically motivated by economic considerations. As this act can have significant consequences on public health, government institutions established specific commissions and laws to prevent this practice. Over the years, several food adulteration incidents were encountered.

One of the most known cases was the contamination of powdered infant milk with melamine in China in 2008, which caused illness in 294,000 children. Between January 2022 and March 2022, the Food and Drug Administration (FDA) interjected and tested 144 imported honey samples, ten percent of which were adulterated with undeclared added sweeteners. In April 2022, the Spanish Agency for Food Safety and Nutrition alerted of the production of fake olive oil and extra-virgin olive oil adulterated with other vegetable oils [7].

These are just a few examples of all the incidents that have occurred over the years and are still occurring today. As a result of these incidents, institutions around the world have initiated activities to combat food fraud. In Europe, numerous initiatives focusing on detecting, preventing, and mitigating food fraud have recently begun. These include initiatives at the regulatory body level, research institutions, not-for-profit organizations, and private legal entities. The European Commission (EC) has different departments named Directorate Generals (DG) which actively participates in preventive measures and food fraud detection through different approaches. Recently, DG SANTE has established a Food Fraud Unit and launched the Food Fraud Network, to facilitate the collaboration between the competent authorities of EU Member States and Europol [8].

The DG Joint Research Centre has established a unit known as the Knowledge Centre for Food Fraud and Quality (KC-Food), which is developing new analytical methods to detect food fraud, offering intelligence support to EU Member States and to the general public, and conducting analytical measurements for the European Anti-Fraud Office (OLAF) anti-fraud investigations [8]. The European Union has already developed methods of analysis for honey [9], wine [10], and olive oil [11].

Figure 1.1. shows the number of cases of food fraud over the last five years collected by the European Commission's food fraud newsletter [12]. It can be seen that 873 are related to other types of fraud like food contraband or illegal sale of food products. Cases related to food adulteration are slightly higher than food authentication, mainly caused by adding other substances. For the authentication problem, generally, the cases are related to products labeled with a geographical indication to which they do not belong.

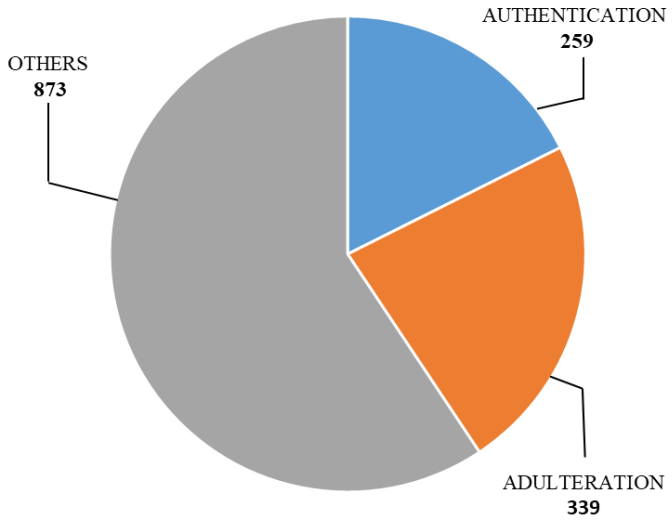



**Figure 1.1.** Number of food fraud cases collected by the European Commission from 2019-2024 (Beginning of 2024).

Thanks to the collective efforts and stringent controls that have been put in place, it is common to carry out controls on foods in which adulteration is not expected. Therefore, the objective of the analysis is to determine whether or not a food is adulterated, and not to obtain a concentration value of the possible adulterant. On the other hand, when it comes to an authentication problem, a binary response is also required (yes or no it is authentic). Withal, qualitative analysis appears to be an effective approach for conducting routine analyses and a good tool to implement in laboratories for food regulation and quality control.

Whether the approach used is qualitative or quantitative [13], it cannot be solved in many cases with just one value (variable) [14]. So, multivariate data analysis techniques are required to process the data and to obtain the relevant information from the samples. For quantitative analysis through the application of multivariate calibration techniques and qualitative analysis through the application of classification techniques.

The concern about food fraud is reflected in the scientific papers addressing this issue. Figure 1.2. shows the number of papers reporting the use of

quantitative or qualitative multivariate methods to solve problems related to food adulteration and food authentication in the last five years. The bibliographic research has been done in the Web of Science database using two principal keywords *"food adulteration"* and *"food authentication"* together with *"data analysis", "multivariate calibration", "classification",* and *"multivariate analysis"* between 2019 and the beginning of 2024.

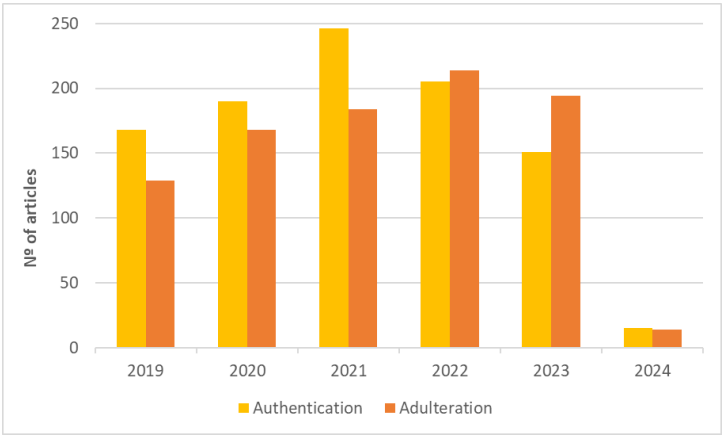



**Figure 1.2.** The number of papers related to food adulteration and food authentication coupled with multivariate analysis published in the last five years.

Today many analytical techniques provide multivariate data. Among them, spectroscopic techniques should be highlighted since, usually, they do not require the use of chemical reagents, and sample manipulation is minimal. Nowadays, the use of handheld spectrophotometers has increased. Because offers the benefit of portability enabling in situ measurements, saving time in obtaining analytical results, and enhancing the analysis frequency. This type of spectrophotometer has been used in the development of this thesis and will be properly explained later [15]. The methods utilizing these techniques align with green chemistry principles, making them more respectable to the environment [16].

Like any analytical method, multivariate methods must be validated. Whenever reference is made to method validation, it must be considered whether it refers to univariate/multivariate and quantitative/qualitative.

The validation of univariate quantitative methods is extensively studied and regulated both at national and european levels. Among international organizations, Eurachem has the objective of establishing an international system for traceability of chemical measurements and promoting food quality practices. In 2012, they published a guide named *"Quantifying Uncertainty in Analytical Measurement"* which offers detailed guidance for the evaluation and expression of uncertainty in univariate quantitative chemical analysis [17].

The validation of qualitative methods, both univariate and multivariate, is not as widely developed. In 2014, Eurachem published another guide named *"The Fitness for purpose of analytical methods: A laboratory guide to method validation and related topics"* [18]. It focuses on the validation of quantitative methods although some of the principles can be applied to qualitative methods. For univariate methods, introduce the cut-off reference, which is established in univariate qualitative methods as a concentration parameter based on false negative rates. For multivariate methods just describe the identification of unknown compounds by matching absorbance signals (i.e. 'NIR peaks') in the analyte spectrum with those of reference spectra stored in a spectral library.

More recently (2021), the Eurachem organization published a guide named *"Assessment of performance and uncertainty in qualitative chemical analysis"* which focuses on the fundamental principles for evaluating the performance of qualitative analysis, reports on the uncertainty of qualitative analytical results, and provides practical examples of the described theories in use [19]. This guide is devoted to univariate qualitative methods, and as above mentioned multivariate just deals with matching spectra bands. In any case, some of the principles can be also applied to multivariate qualitative

methods like the definition of contingency tables and the performance parameters derived from them.

This leads us to say that the validation of qualitative methods is scarcely developed and implemented and that there is still much work to be done. It should be noted that one of the main difficulties is due to the not numerical response obtained from these types of methods. Therefore instead of calculating the performance parameters from the well-known statistics, probabilistic calculations should be considered. Another issue to be contemplated is the particular case of adulteration problem. If that is the case, an approach can be made between quantitative and qualitative methods called semi-quantitative methods. Semi-quantitative methods have similar output as qualitative methods but the performance parameters can be related to concentration values. In this case, we can say that there are studies that address these methods but there is no guide that refers to them and defines the validation parameters.

Given the above, in this Thesis, the following objectives are proposed for the development of multivariate qualitative methods and their validation.

## 1.2. Objectives

The main objective of this thesis is to develop and validate multivariate qualitative and semi-quantitative analytical methodologies based on spectroscopic techniques and chemometric data analysis for food fraud analysis. Within this general framework, the following sub-objectives have been considered:

a) Develop multivariate qualitative methodologies based on molecular spectroscopy measurement and classification techniques, for different types of foods in authentication and adulteration problems.
b) In the case of adulteration, propose tools associated with semi-quantitative information, that is, analysis with a binary response (yes/no it is adulterated) for different levels of adulteration.

c) Propose strategies to optimize the performance parameters of the developed methods.
d) Propose strategies such as multivariate transfer techniques, to increase the usefulness of models developed under certain conditions.

## 1.3. Thesis structure

This thesis is structured in four chapters.

**Chapter 1: Scope and Objectives.** In this chapter, it is presented the framework that give rise to implement multivariate analytical methods in different food fraud problems. Then the objectives and the structure of the thesis are described.

**Chapter 2. Theoretical and practical aspects.** This chapter has three sections. The first section contains a brief description of the different samples studied (honey, nuts and olive oil). The second section presents an introduction to the theoretical aspects of the instrumental analytical techniques used. Finally, the third section contains the theoretical background of data treatment and an explanation of all chemometric tools applied in the experimental part of this thesis.

**Chapter 3. Results.** This chapter contains the results of the experimental work that has been carried out and is divided into two sections. Section 3.1. is devoted to studying several adulteration problems. Each of them uses different spectroscopic techniques as well as multivariate classification techniques. Four papers present the results: *"Multivariate qualitative methodology for semi-quantitative information. A case study: Adulteration of olive oil with sunflower oil"* published in *Analytica Chimica Acta 1206 (2022) 339785; "In-depth chemometric strategy to detect up to four adulterants in cashew nuts by IR spectroscopic techniques"* published in *Microchemical Journal 181(2022) 107816; "One-class model with two decision thresholds for the rapid detection of cashew nuts adulteration by*

*other nuts"* published in *Talanta 253 (2023) 123916.* The last one named *"A semi-quantitative one-class modelling method for detecting honey adulteration using two-class limits"* is submitted for publication. In each studied case, the chemometric treatment required for each type of multivariate signal was evaluated. Also are evaluated, the different performance parameters applied regarding whether it is a semi-quantitative or a qualitative food adulteration case. Section 3.2. is centered on food authentication. It contains one paper in which an approach of a multivariate transfer method technique is applied to study and correct the effect of seasonality on agricultural products: *"Data standardization strategy to correct the effect of seasonality in the authentication of virgin olive oil"* published in *Microchemical Journal 195 (2023) 109520.*

**Chapter 4. Conclusions.** This chapter presents the general conclusions of the thesis.

## References

[1] Discussion paper on food integrity and food authenticity. In *Codex Committee on food import and export inspection and certification systems*; Codex Alimentarius Commission (CX/FICS 18/24/7): Brisbane, **2018**.

[2] Danezis, G.P.; Tsagkaris, A.S.; Camin, F.; Brusic, V.; Georgiou, C. A. Food authentication: Techniques, trends & emerging approaches. *TrAC - Trends Anal. Chem*. **2016**, 85, 123-132, doi: 10.1016/j.trac.2016.02.026.

[3] Commission Regulation (EC) Nº 1898/2006. In *Official Journal of the European Union*, **2006**.

[4] Commission Regulation (ECC) Nº 2081/1998. In *Official Journal of the European Communities*, **1998**.

[5] Council Regulation (EC) Nº 510/2006. In *Official Journal of the European Union*, **2006**.

[6] Regulation (EU) Nº 1151/2012 of the European Parliament and of the council. In *Official Journal of the European Union*, **2012**.

[7] Momtaz, M.; Bubli, S.Y.; Khan, M.S. Mechanisms and Health aspects of food adulteration: A comprehensive review. *Foods* **2023**, 12, 199, doi: 10.3390/foods12010199.

[8] Popping, B.; Buck, N.; Bánáti, D.; Brereton, P.; Gendel, S.; Hristozova, N.; Chaves, S.M.; Saner, S.; Spink, J.; Willis, C.; Wunderlin, D. Food inauthenticity: Authority activities, guidance for food operators, and mitigation tools. *Compr. Rev. Food Sci. Food Saf.* **2022**, 21, 4776-4811, doi: 10.1111/1241-4337.13053.

[9] Harmonised methods of the International Honey Commission. In *International Honey Commission*, **2001**.

[10] International Organisation of Vine and Wine (OIV), available at: https://www.oiv.int/index.php/what-we-do/standards.

[11] International Olive Council (IOC), available at: https://www.internationaloliveoil.org/what-we-do/chemistry-standardisation-unit/standards-and-methods/.

[12] The Food Fraud Monthly Report is available at: https://knowledge4policy.ec.europa.eu/food-fraud-quality/monthly-food-fraud-summary-reports_en.

[13] Muñoz-Olivas, R. Screening analysis: an overview of methods applied to environmental, clinical and food analyses. *TrAC - Trends Anal. Chem*. **2004**, 23, 203–216, doi: 10.1016/S0165-9936(04)00318-8.

[14] Callao, M.P.; Ruisánchez, I. An overview of multivariate qualitative methods for food fraud detection. *Food Control* **2018**, 86, 283-293, doi: 10.1016/j.foodcont.2017.11.034.

[15] Rovira, G.; Miaw, C.S.W.; Martins, M.L.C.; Sena, M.M.; De Souza, S.V.C.; Callao, M.P.; Ruisánchez, I. One-class model with two decision thresholds for the

rapid detection of cashew nuts adulteration by other nuts. *Talanta* **2023**, 253, 123916, doi: 10.1016/j.talanta.2022.123916.

[16] Ruisánchez, I.; Jiménez-Carvelo, A.M.; Callao, M.P. ROC curves for the optimization of one-class model parameters. A case study: Authenticating extra virgin olive oil from a Catalan protected designation of origin. *Talanta* **2021**, 222¸ 121564, doi: 10.1016/j.talanta.2020.121564.

[17] Ellison, S.L.R.; Williams, A (Eds.). Eurachem/CITAC guide: Quantifying Uncertainty in Analytical Measurement, 3rd Edition, **2012**. ISBN 978-0-948926-30-3.

[18] Magnusson, B.; Örnemark, U. (Eds.). Eurachem Guide: The Fitness for Purpose of Analytical Methods – A Laboratory Guide to Method Validation and Related Topics, 2nd Edition, **2014**. ISBN 978-91-87461-59-0.

[19] Bettencourt da Silva, R; Ellison, S.L.R. (Eds.) Eurachem/CITAC Guide: Assessment of performance and uncertainty in qualitative chemical analysis, 1st Edition, **2021**. ISBN 978-0-948926-39-6.

# Chapter 2. Theoretical and practical aspects

## 2.1. Samples

There are several reasons to commit food fraud, some of them are the high product demand during the off-season, the elevated product prices in specific periods, the lack of surveillance systems, and the poor enforcement of regulations [1-3]. So, there are many foods subject to fraud, being one of the main reasons the economics behind.

The most studied foods susceptible to fraud are shown in Figure 2.1. This information has been obtained from a search in the Web of Science database over a short period (5 years). Honey, olive oil, species, and milk, are the most common with percentages between 10% and 14%.

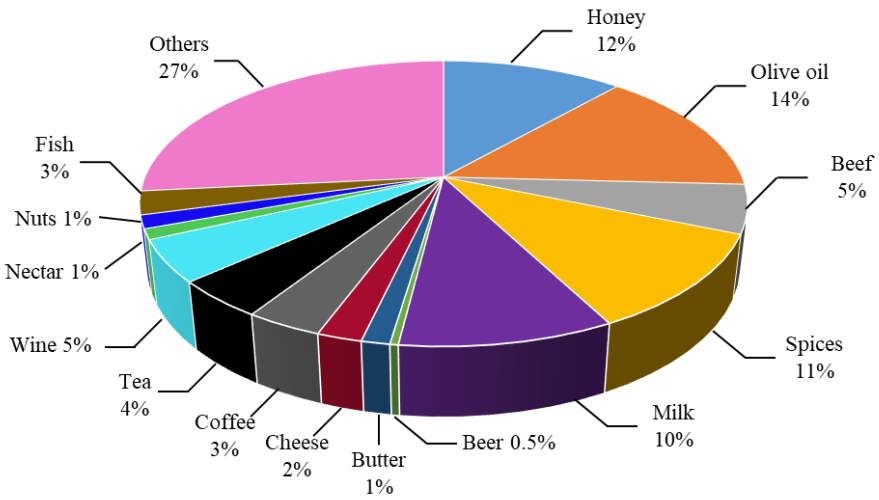



**Figure 2.1.** Percentage of published papers of the most common food subject to fraud. Source Web of Science from 2019 to the beginning of 2024.

This thesis focuses on the analysis of food fraud in honey, nuts, and olive oil. All these foods have high health benefits and are highly consumed by the population.

### *2.1.1. Honey*

The honey samples were obtained directly from traceable producers under Ministerio da Agricultura, Pecuária e Abastecimento (MAPA) of Brazil in Minas Gerais State.

In Brazil, the production value was 957.811 million Brazilian Reals, producing 60.966.305 kg of honey, in 2022. A significant increase compared to 2021, when the production value was 851.354 million Brazilian Reals [4]. Brazil is the fifth main world exporter of honey, exporting between January and June of 2023 around 15.000 tons [5]. Principally to the United States, Germany, Canada, Belgium, and the Netherlands [6].

In 2022 and 2023, the MAPA developed the "Operações do Mel 2022/2023" to fight against honey fraud. In this operation, honey adulteration was observed in 14.14% and 8.95% of the samples analyzed in 2022 (99 samples) and 2023 (67 samples), respectively [7].

The European Anti-Fraud Office (OLAF) in 2023, provided support to a European action against honey adulteration led by the European Commission. During this action, 133 businesses (70 importers and 63 exporters) were involved in shipments of suspected adulterated honey, claiming that 46% of the honey imported to the European Union was not authentic [8,9].

As can be seen, honey fraud is a current and recurrent problem worldwide. Therefore, the scientific community is investigating different solutions to try to detect easily and stop this problem. Figures 2.2. and 2.3., show the published papers related to honey fraud in the last 5 years, according to the instrumental and chemometric techniques most used.

As can be observed in Figure 2.2., the most used instrumental techniques to detect honey fraud are the spectroscopic techniques with 69% of the total followed by the chromatographic ones with 23%. Figure 2.3. shows that PCA, which is an exploratory technique, is the most used in studies of honey fraud detection. Discriminant classification techniques such as PLS-DA and LDA are more implemented in comparison with modelling techniques such as SIMCA.

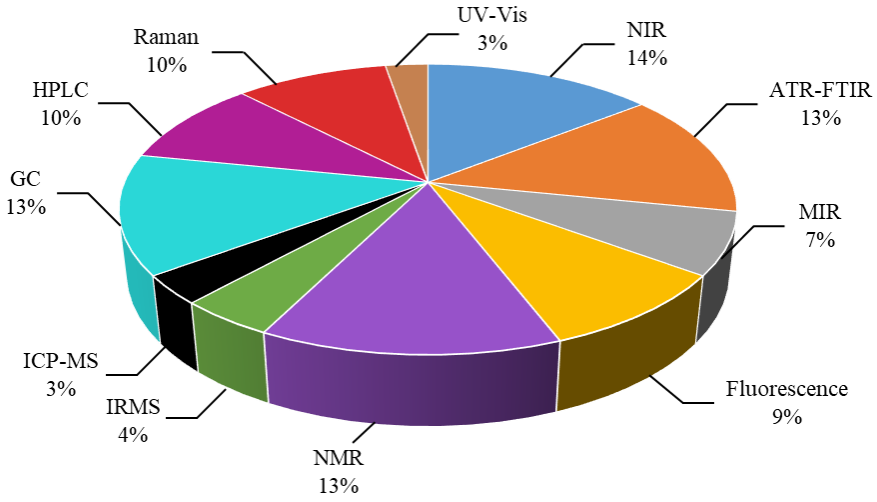



**Figure 2.2.** Papers in the last 5 years using different analytical techniques to detect honey fraud. Source Web of Science from 2019 to the beginning of 2024.

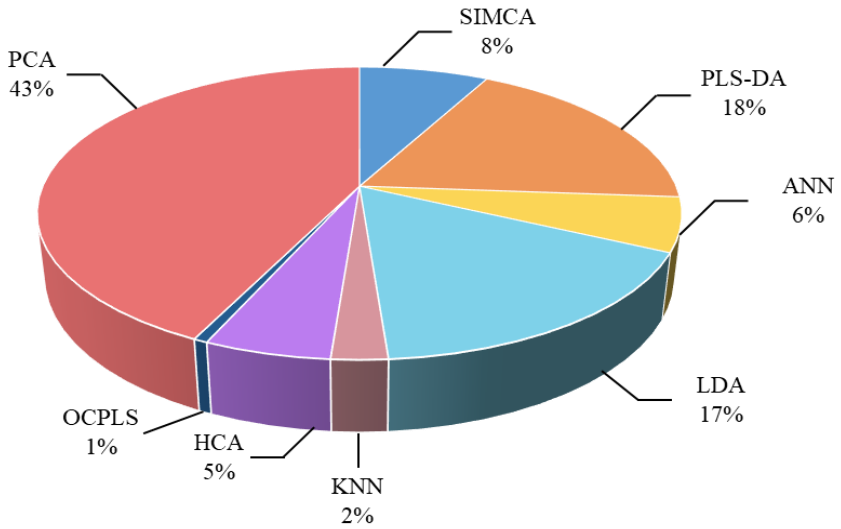



**Figure 2.3.** Papers in the last 5 years using various classification techniques in honey fraud detection. Source Web of Science from 2019 to the beginning of 2024.

### *2.1.2. Nuts*

Commercial nuts samples (Cashew nuts, Brazilian nuts, Macadamia nuts, Peanuts, and Pecan nuts) were acquired from certified producers from Brazil. This country is one of the major world producers of cashew nuts, producing an estimated 116.0 thousand tons in 2023 [10]. Comparing the actual market prices cashew nuts are sold for R$40-55, while peanuts, Pecan, and Brazil nuts are sold for R$5-10, R$25-50, and R$30-45, Macadamia nuts are sold for the same price as cashew nuts, because most of it is imported [11].

The MAPA of Brazil, in August 2022, published in the Diário Oficial da União, the technical regulation defining the minimum requirements of identity and quality for almonds, chestnuts, walnuts, and dried fruits [12]. The principal objective of this standard is to fill the lack of a specific Official Classification Standard for each product and enable these products to be controlled and offered to consumers while respecting a minimum level of quality and sanitary conditions.

Since 2002, in Europe, the Regulation (EC) Nº 178/2002 includes the general principles and requirements of food law, the establishment of the European Safety Authority, and the procedures in matters of food safety. The last one includes the prevention of fraudulent practices concerning food safety [13].

Figures 2.4. and 2.5. illustrate the papers on nut fraud published over the past 5 years, categorized by the instrumental and chemometric techniques employed. Among spectroscopic techniques, infrared spectroscopy techniques are by far the most widely used (NIR, MIR, and ATR-FTIR). Again, the application of PCA is generalized as an exploratory technique of the data. In that case, the modelling approach (SIMCA) is as implemented as the discriminant approach (PLS-DA, LDA) and the published papers use similar discriminant or modelling classification techniques.

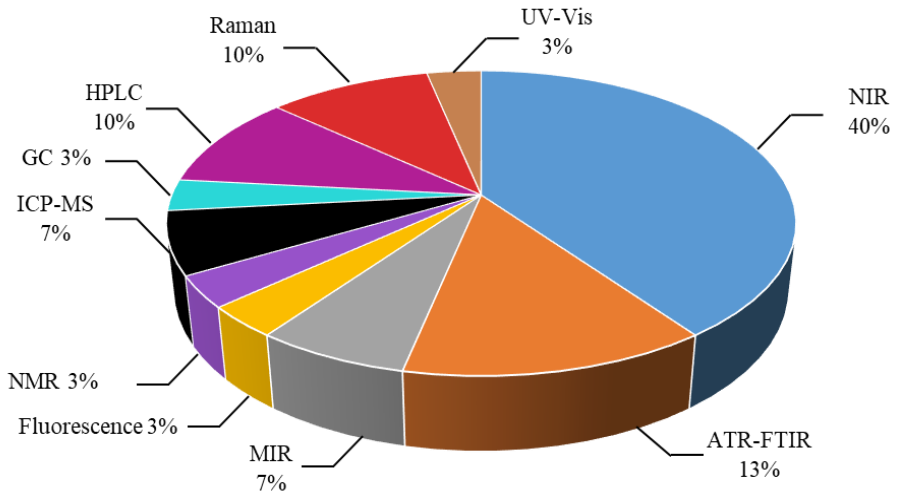



**Figure 2.4.** Papers in the last 5 years using different analytical techniques to detect nut fraud. Source Web of Science from 2019 to the beginning of 2024.

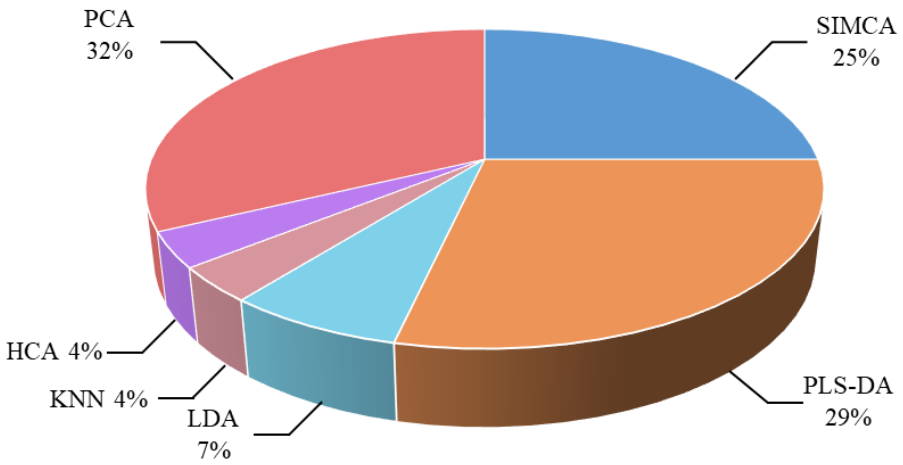



**Figure 2.5.** Papers in the last 5 years using various classification techniques in nut fraud detection. Source Web of Science from 2019 to the beginning of 2024.

### *2.1.3. Olive oil*

Olive oil is one of the bases of the Mediterranean diet. Catalonia is one of the olive oil producing areas of Spain. There are five Catalan protected designations of origin (PDO) for olive oils such as Siurana, Empordà, Terra Alta, Les Garrigues, and Baix Ebre-Montsià [14]. The olive oil samples used in this thesis are from Siurana and Les Garrigues PDOs supplied by the Catalan Government's Tasting Panel of Virgin Olive Oils of Catalonia. In 2022, the production between these two regions was 6772 tons. Nowadays, due to climate change, harvests are affected here in Catalonia by the lack of rainfall. This increases the market price and makes it susceptible to fraud.

The European Union produces 67% of the world's olive oil and exports 65% mainly to the United States, Brazil, and Japan. As in the case of nuts, the general food law covers all phases of olive oil production, processing, and distribution [13]. Legislation related to olive oil covers a range of areas including marketing standards, properties of olive and olive-pomace oils, organizations of producers, support initiatives, private storage, and price notifications.

There are two international organizations, the International Olive Council and Codex Alimentarius in which there are standard methods for the analysis of olive oil and specific standards for olive oils and olive pomace oils [15].

Figures 2.6. and 2.7. shows the percentage of research papers related to the fraud of olive oils according to the instrumental and classification technique used. In this case, spectroscopic techniques are the most used, within the group of chromatographic techniques, the use of gas chromatography stands out. Discriminant techniques are more widely used than modelling techniques.

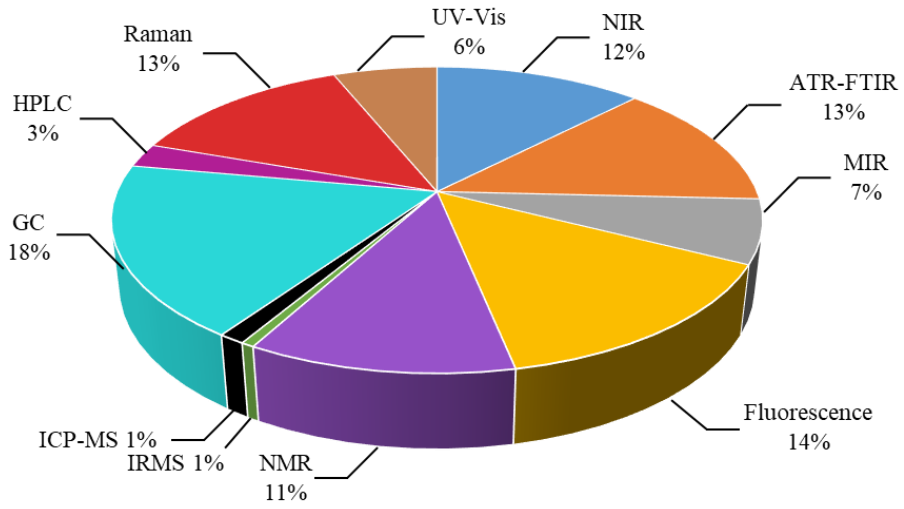



**Figure 2.6.** Papers in the last 5 years using different analytical techniques to detect olive oil fraud. Source Web of Science from 2019 to the beginning of 2024.

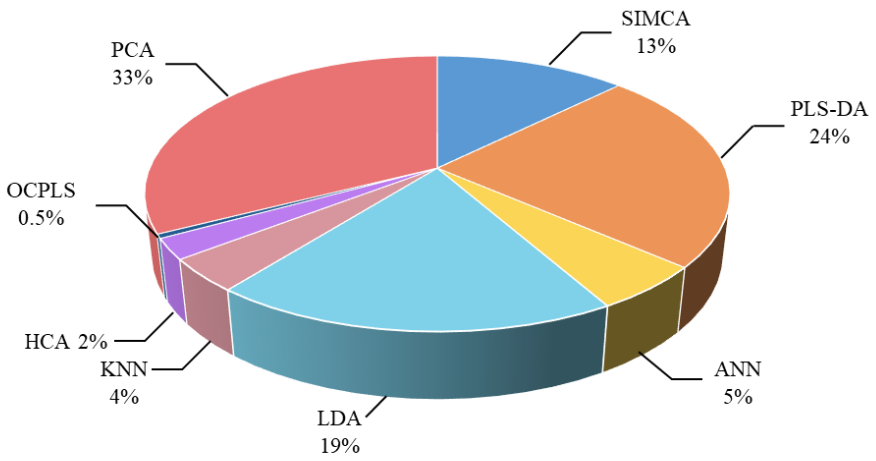



**Figure 2.7.** Papers in the last 5 years using various classification techniques in olive oil fraud detection. Source Web of Science from 2019 to the beginning of 2024.

## 2.2. Analytical Techniques

Over the past few decades, analytical techniques have made significant progress thanks to great technological advances. Among them, spectroscopies have undergone rapid development thanks to further developments in laser technology, the miniaturization of optical and electronic components, etc.

One of the main areas of research for analytical chemists is the development of new methods or the improvement of existing ones in most areas, including the food sector. In that scenario, the application of modern, advanced, and innovative spectroscopic techniques plays a fundamental role, in increasing overall knowledge in food analysis.

Spectroscopy can be defined as the study of the interaction between electromagnetic radiation and matter [16]. The electromagnetic spectrum (Figure 2.8.) is divided into different specific ranges of wavelengths such as ultraviolet (UV), visible (VIS), infrared (IR), and radio (nuclear magnetic resonance, NMR). Different information can be acquired based on the specific range of the spectrum used for the analysis.

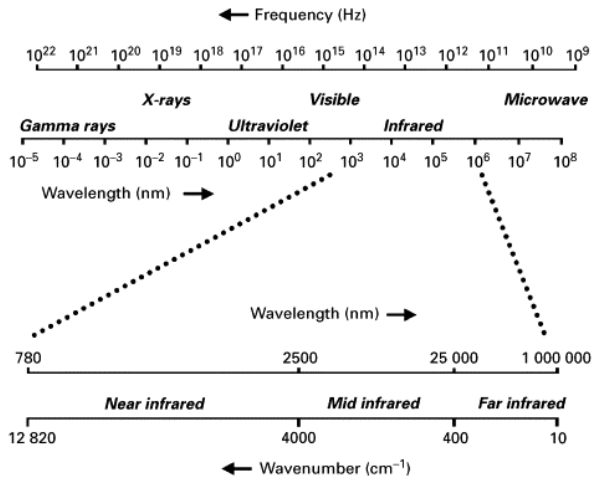



**Figure 2.8**. Regions of the electromagnetic spectrum. Adapted from Shawn, 1999 [17].

The spectroscopic techniques can be classified as molecular or atomic depending on the type of species analyzed. Also, according to the type of radiation-matter interaction such as absorption, emission, diffraction, stroke shift, or fluorescence [18]. In this section, only the techniques used in developing the thesis will be discussed.

### *2.2.1. Infrared spectroscopy*

The infrared region is divided into three subregions, the near-infrared region (NIR) (800-2500 nm), the mid-infrared region (MIR) (2500nm-25μm), and the far infrared region (FIR) (25μm-1mm) [19].

NIR and MIR spectroscopies are associated with the vibrations of molecular bonds of functional groups that include a hydrogen atom such as O-H, C-H, and N-H (S-H is very weak) and from the C=O groups. Specifically, in MIR the spectra show the fundamental vibrations of the functional groups mentioned above [19]. In NIR spectroscopy, absorption bands correspond to combinations and the three principal overtones of the fundamental bands. The combination bands arise from vibrational combinations of these chemical groups. A more detailed explanation of the theory of NIR absorption bands can be found in the references [18,20-22].

Infrared instruments can be categorized into two groups dispersive spectrometers and Fourier Transform (FT) spectrometers. The instrumental components consist of a radiation source, a monochromator in dispersive spectrometers or an interferometer in FT spectrometers, a sample holder, and a detector usually linked to an amplifier system for spectrum recording [16,19]. The FT spectrometers allow us to obtain higher sensitivity, superior wavelength resolution, and wavelength accuracy [19,23,24].

In infrared spectroscopy, there are different measurement configurations (Figure 2.9.) that give rise to two ways of obtaining absorbance measurements; transmittance (configuration (a)) and reflectance (configurations (b), (c), and (d)).

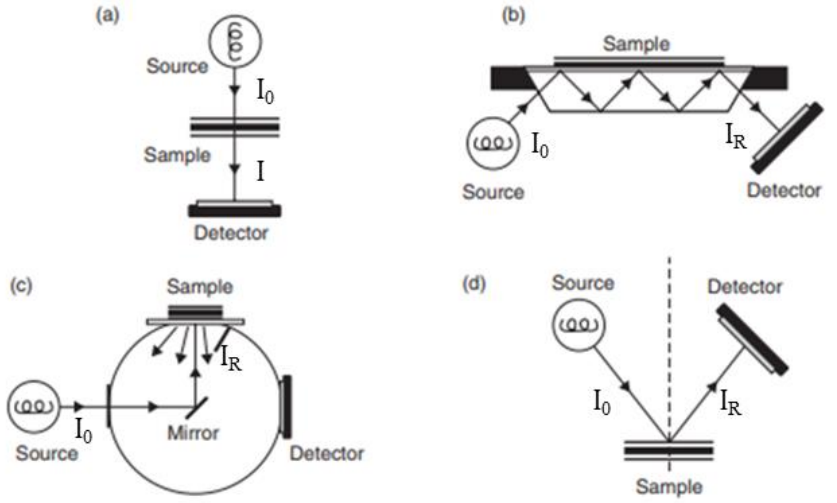



**Figure 2.9.** IR measurement configurations: (a) transmission, (b) attenuated total reflectance (ATR), (c) diffuse reflectance, and (d) specular reflectance. Adapted from Subramanian, and Rodriguez-Saona, 2009 [25].

As the absorbed light cannot be measured directly, transmittance and diffuse reflectance are correlated to absorbance following Equations 1 and 2 [20,26].

$$A = -\log T = \log \frac{I_0}{I} \qquad \text{Equation 1}$$

$$A = \log\left(\frac{1}{R}\right) = \log\left(\frac{I_{Rstandard}}{I_R}\right) \qquad \text{Equation 2}$$

Where *A* is absorbance, *T* is transmittance, *R* is reflectance, $I_0$ is the intensity of the light before it falls on the sample, *I* is the intensity of the light after it has passed through the sample, $I_R$ is the reflected intensity after it has interacted with the sample, and $I_{Rstandard}$ is the reflected intensity of a highly reflective material such as Teflon or Spectralon [26].

The reflectance configurations can be categorized into three distinct types: attenuated total reflectance (ATR) (b), diffuse reflectance (c), and specular reflectance (d) [19,27].

Specular reflectance (Figure 2.9. (d)) is a technique where the angle of the incident IR radiation and its reflection angle are identical. The reflected IR radiation from the sample surface is directed to the detector using a separate mirror. This type has limited applications in analyzing food products. Diffuse reflectance (Figure 2.9. (c)) happens when the incident IR radiation is reflected from the sample's surface in various directions. This type is particularly suitable for solid and powdered materials. The ATR (Figure 2.9. (b)) is widely used in food analysis and will be deeply explained in section 2.2.1.2 [25].

In the transmittance method (Figure 2.9. (a)) the IR radiation passes directly through the sample to the detector on the opposite side. It is suitable for analyzing solid, liquid, and gaseous samples, but is limited by the sample thickness, while reflectance methods are not affected by the sample thickness [19,27]. In this thesis, all IR measurements have been carried out in reflectance mode.

#### 2.2.1.1. Near-Infrared Spectroscopy (NIR)

NIR spectroscopy has very good characteristics, principally as a non-destructive and for its ability to carry out in situ analysis through the use of optical fibers. It can be applied to samples in different states, shapes, and thicknesses. All these characteristics make NIR a good option for determining multiple types of samples and parameters, acquiring many scans in a short period [21,28].

The instrumentation of NIR spectroscopy can be categorized into two main types: benchtop spectrometers, which are restricted to a laboratory, and autonomous spectrometers, which can be used on-site. Within the autonomous spectrometers, they can be classified according to the weight of the equipment as follows: transportable (>20 kg), ‘suitcase’ (1-20 kg), and handheld (<1 kg). In recent years, a new class of spectrometers has emerged, the miniaturized NIR devices that can weigh less than 100 g [29].

The principal requirement of miniaturizing an analytical instrument is to guarantee that the decrease in size does not compromise measurement accuracy or performance [30]. Addressing this problem, recent studies that compared handheld and benchtop spectrometers, have shown that equivalent qualitative and quantitative results can be achieved [30-33]. In the scientific publication of Kranenburg et al., a comparative table of characteristics of NIR spectrometers is presented [33].

The use of handheld NIR spectrometers in the food field is continuously increasing, and many studies appeared in the literature, some of which can be seen in Table 2.1.

**Table 1.** Examples of different applications of NIR handheld spectrometers in food analysis.

| Sample | Mode | Study | Reference |
|---|---|---|---|
| Milk | Reflectance | Authentication | [34] |
| Butter<br>Cheese | Diffuse reflectance | Adulteration | [35] |
| Meat | Diffuse reflectance | Adulteration | [36] |
| Olive oil | Transmission and Diffuse reflectance | Adulteration | [37] |
| Fish | Diffuse reflectance | Authentication | [38] |
| Pineapple | Reflectance | Authentication | [39] |
| Tea | Diffuse reflectance | Authentication | [40] |
| Paprika | Diffuse reflectance | Adulteration | [41] |

In this thesis, a handheld miniaturized NIR instrument (MicroNIR 1700) (Figure 2.10.) was used in diffuse reflection mode. It is more appropriate than transmission mode to analyze ground and solid samples such as nuts and honey.

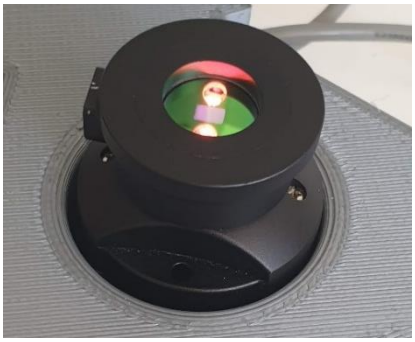


**Figure 2.10.** MicroNir 1700 from Viavi Solutions.

### 2.2.1.2. Attenuated Total Reflectance Fourier Transformed Infrared (ATR-FTIR) (Figure 2.9. (b))

This technique is based on the principle of infrared light attenuation when it is directed at an interface between an internal reflection element (crystal), characterized by high refractive index properties, and the sample, usually with a lower refractive index. When the light interacts with the surface an evanescent wave is produced and penetrates the sample, which absorbs specific wavelengths reducing the intensity of the light reflected. The attenuated light that emerges from the crystal, having been altered by the sample, is then measured [42].

Since the radiation does not pass through the sample, it ensures that the spectrum collected is consistent without considering the sample's volume [19]. ATR minimizes the need for sample preparation, eliminates the variations in cell path lengths, and facilitates the acquisition of consistent spectra. Table 2.2. summarizes some examples of the application of ATR-FTIR spectroscopy as an analytical technique in foods.

**Table 2.** Examples of different applications of ATR-FTIR spectroscopy in food analysis.

| Sample | Study | Reference |
|---|---|---|
| Edible oils | Authentication and Adulteration | [43] |
| Milk | Adulteration | [44] |
| Lentils | Authentication | [45] |
| Coconut oil | Adulteration | [46] |
| Saffron | Adulteration | [47] |
| Honey | Adulteration | [48] |

### *2.2.2. Fluorescence*

The Jablonski diagram shown in Figure 2.11a, explains how the phenomenon of fluorescence occurs. Initially, a molecule known as a fluorophore is excited to an electronic singlet state. In this excited state, an internal conversion takes place to the lowest vibrational level of the excited state. Finally, the fluorophore returns to its ground state, emitting light of a longer wavelength [49,50].

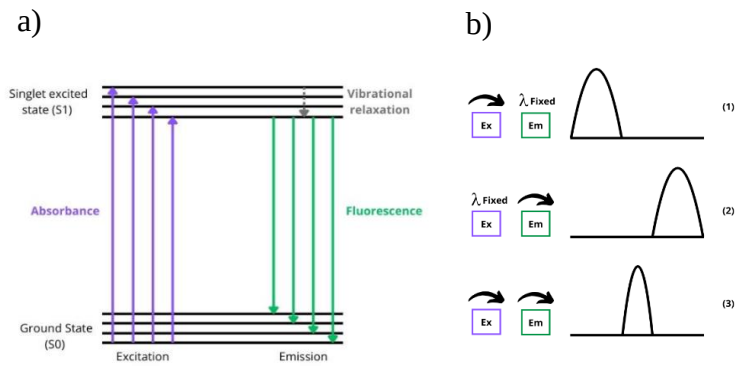



**Figure 2.11. (a)** Jablonski diagram. The horizontal black lines indicate the vibrational levels within an electronic state. Purple vertical arrows correspond to the excitation to an electronic stage and green vertical arrows correspond to the emission of light. **(b)** Fluorescence spectra depend on the excitation (Ex) and emission (Em) measurement: (1) excitation spectrum, (2) emission spectrum, and (3) synchronous spectrum. Adapted from Li et al. 2010 [50].

In Figure 2.12., a scheme of a Fluorescence spectrophotometer is presented, and as can be seen, the detector is placed at 90º to the incident light beam to reduce the risk of incident light that is transmitted or reflected reaching the detector [42,49].

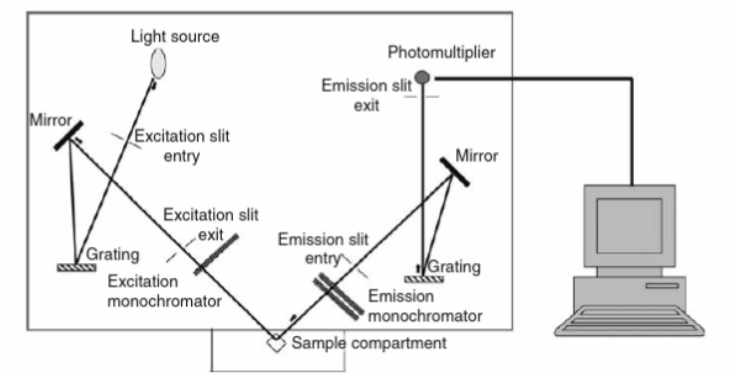



**Figure 2.12.** Scheme of a Fluorescence spectrophotometer. Adapted from Andersen et al. 2008 [49].

Figure 2.11b, shows different types of spectra obtained in different measurement modes. The excitation spectrum (Figure 2.11b (1)) is determined by monitoring fluorescence emission at the wavelength of maximum intensity while the sample is excited through a wavelength interval. Conversely, to obtain the emission spectrum (Figure 2.11b (2)), the emission monochromator is scanned across different wavelengths, with the excitation monochromator held at an optimal excitation wavelength [50]. This measurement mode corresponds to the conventional one in fluorescence spectrometry.

Currently, other modes of fluorescence measurements are developed and applied giving rise to synchronous fluorescence spectroscopy (SFS) (Figure 2.11b (3)) and excitation-emission fluorescence spectroscopy (EEFS) [42]. The SFS (Figure 2.11b (3)) measurement mode involves simultaneously scanning the excitation and emission monochromators while maintaining a consistent interval of wavelength difference between them ($\Delta\lambda$). There are different types of SFS procedures: constant-energy SFS, maintaining a

constant frequency difference; and variable-angle SFS, allowing simultaneous variation of excitation and emission wavelengths at different rates. The first one is the most applied. Among the remarkable advantages of SFS compared to conventional fluorescence spectroscopy, can be mentioned the simplification of the spectra, the reduction of light scattering, and the improvement of the selectivity. EEFS mode consists of measuring both the excitation and emission spectra for a given sample, obtaining a 3D excitation-emission data matrix as a result. Unlike conventional fluorescence spectroscopy, EFFS provides simultaneous information about different fluorophores in a sample. In the development of this thesis, both conventional and synchronous fluorescence spectroscopy have been used. Table 2.3. shows some recent papers that use the conventional mode, SFS, and EEFS for detecting adulterants and authenticating the geographical origin of different samples.

**Table 3.** Examples of applications of Fluorescence spectroscopy using conventional, SFS, and EEFS measurement modes in foods.

| Sample | Mode | Study | Reference |
|---|---|---|---|
| Olive oil | SFS | Adulteration | [51] |
| Cumin | SFS | Adulteration | [52] |
| Peanut oil | SFS | Adulteration | [53] |
| Tea | SFS | Authentication | [54] |
| Olive oil | SFS | Authentication | [55] |
| Olive oil | EEFS | Adulteration | [56] |
| Saffron | EEFS | Adulteration | [57] |
| Green tea | EEFS | Authentication | [58] |
| Grape seed oil | EEFS | Authentication | [59] |
| Almond | EEFS | Authentication | [60] |
| Olive Oil | Conventional | Authentication | [61] |
| Honey | Conventional | Authentication | [62] |
| Milk | Conventional | Adulteration | [63] |
| Turmeric | Conventional | Adulteration | [64] |
| Honey | Conventional | Adulteration | [65] |

### *2.2.3. Low-Field Nuclear Magnetic Resonance (LF-NMR)*

The principle of Nuclear Magnetic Resonance spectroscopy (NMR) is nuclear magnetism which arises from the spins of nucleons, such as protons and neutrons. In the request to achieve a net nuclear magnetization moment, the nucleus should have an odd number of nucleons, such as $^{1}H$, $^{31}P$, $^{15}N$, and $^{23}Na$ being the $^{1}H$ the nuclei most applied [66].

NMR informs about the molecular structure and their environment. So, the spectrum obtained can be considered as a molecular "fingerprint" of the sample under study. Various types of NMR spectra have been used in food analysis, including high-resolution and low-resolution [67,68]. In this thesis, low-resolution NMR was used, which gives information about relaxation times, both intramolecular and intermolecular movements [69].

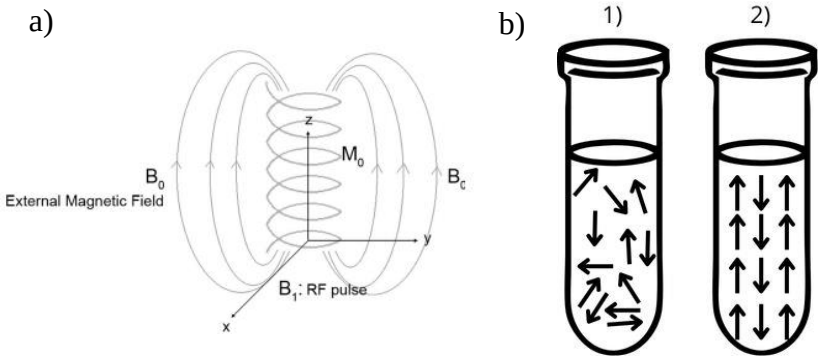



**Figure 2.13.** (a) Scheme of the external magnetic field in the z-direction. Adapted from Kirtil et al. 2016 [70]. (b) Illustration showing the orientation of protons in samples: 1) Without an external magnetic field present. 2) Under the influence of an external magnetic field $B_0$.

To obtain a signal from a sample, it is introduced into an external magnetic field ($B_0$), as shown in Figure 2.13a. Once in the magnetic field, the protons in the sample align themselves to match with the external magnetic field in the z-direction. The protons are not always oriented towards the external magnetic field (Figure 2.13b (1)), each proton has a magnetic moment,

moment, which can align with the external magnetic field in the same direction or opposite direction (Figure 2.13b (2)) [70,71].

The number of protons aligned in the same direction as $B_0$ slightly exceeds those aligned in the opposite direction. This results in a net magnetic field in the sample along the z-axis, which is known as longitudinal magnetization. Those protons do not always face the external magnetic field, but instead exhibit a spin movement named precession. The magnetization in the xy plane resulting from this movement is named transverse magnetization. Which becomes zero because the precessing protons do not occupy the same spatial position simultaneously and cancel each other out [70,72].

A radio frequency (RF) pulse is applied to shift the net magnetization away from the z-axis towards the xy plane. This pulse causes some protons to align themselves opposite to $B_0$. When the RF pulse is turned off, the protons revert to their original state, this process is known as relaxation. The measurement of the relaxation of longitudinal and transverse magnetization provides valuable information about the samples. The longitudinal relaxation time ($T_1$) is defined as the time required for spins to realign along the direction of the external magnetic field. The transverse relaxation time ($T_2$) is the duration required for the transverse magnetization to decay back to its equilibrium value of zero [70,71]. Table 2.4. recaps some examples of the application of LF-NMR spectroscopy as an analytical technique in foods.

**Table 4.** Examples of applications of LF-NMR spectroscopy in foods.

| Sample | Study | Reference |
|---|---|---|
| Honey | Adulteration | [73] |
| Rapeseed oil | Adulteration | [74] |
| Olive oil | Adulteration | [75] |
| Hibiscus | Authentication | [76] |
| Coconut oil | Authentication | [77] |

## 2.3. Multivariate analysis

Multivariate analysis is a set of tools to manage extensive data sets effectively. It is considered an essential tool in analytical chemistry since it enables the extraction of the most relevant information from complex data matrices. Different steps must be followed to develop a multivariate analytical method. Figure 2.14. shows the sequential steps, as well as the details required for the implementation of each stage that has been used in the development of this Doctoral Thesis.

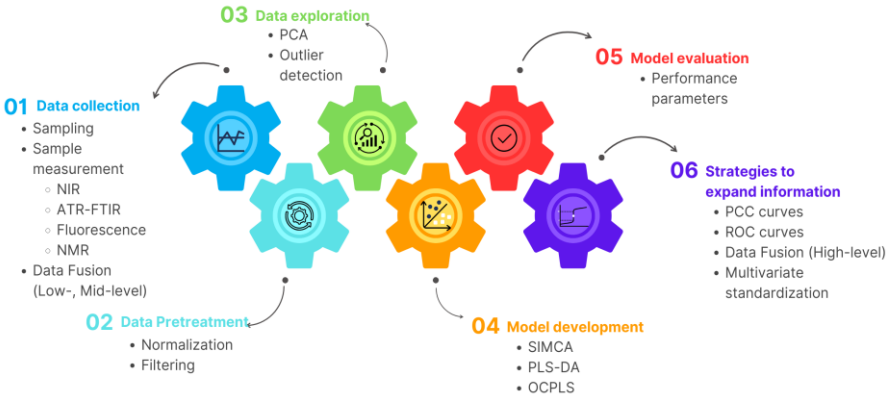



**Figure 2.14.** Diagram of the steps and tools used in each of them for multivariate qualitative analysis.

As can be seen in Figure 2.14. the first step consists of collecting representative samples of the problem to be solved, analyzing them by the appropriate instrumental technique, and finally ensemble the data in a data matrix. From the spectroscopic techniques described in the previous section,

a high amount of data is generated for each sample analyzed, corresponding to the interaction of radiation at multiple wavelengths with the sample. As a result, data are organized/ensembled in two-dimensional matrices (2D **X**) obtained from the spectroscopic measurements in which each row (*i*) corresponds to the spectrum of each sample, and the columns (*j*) correspond to each variable for all samples [78].

### *2.3.1. Data pretreatment*

Data obtained by the spectroscopy techniques may contain unwanted variation due to measurement mode, sample state, and other external physical, chemical, and environmental factors. In addition, can be corrupted with inconsistencies such as outliers and missing values [79]. Therefore, most of the time, data pre-processing is required which includes the steps we need to follow to transform or encode data so that the multivariate techniques may easily parse it. Following the diagram (Step 02, Figure 2.14.), this section will discuss the data pre-treatments applied in this Thesis. Depending on the spectroscopic techniques and the objective, sometimes a combination of them has to be used.

Given the matrix **X**, the following nomenclature is used in all pretreatments: each value of **X** is coded as $x_{ij}$, $\mathbf{x}_i$ is the spectrum for a sample *i*, *j* is the number of variables in the spectrum and $\mathbf{x}_{i,\text{corrected}}$ is the spectrum corrected according to the pretreatment applied in each case.

#### 2.3.1.1. Sample Normalization

Normalization is a pre-processing technique used to reduce spectral variability ensuring that all samples have an equal influence on the model [80,81]. In the development of the thesis, **normalization by the maximum value** has been used. To correct the spectrum of each sample, first, the variable (*j*) with the maximum value of the spectrum (*i*) is identified (Equation 3).

$$w_{ij} = Max(\mathbf{x}_i) \qquad \text{Equation 3}$$

The entire spectrum is then divided by this value (Equation 4).

$$\mathbf{x}_{i,\text{corrected}} = \frac{\mathbf{x}_i}{w_{ij}} \qquad \text{Equation 4}$$

As a result of this pretreatment, the normalized spectrum takes values between 0 and 1. Figure 2.15. shows the application of the normalization by the maximum value to the LF-NMR spectra for the eucalyptus honey (EUC) data set. In this case, all the spectra start at 1 since the first variable is the one with the maximum value.

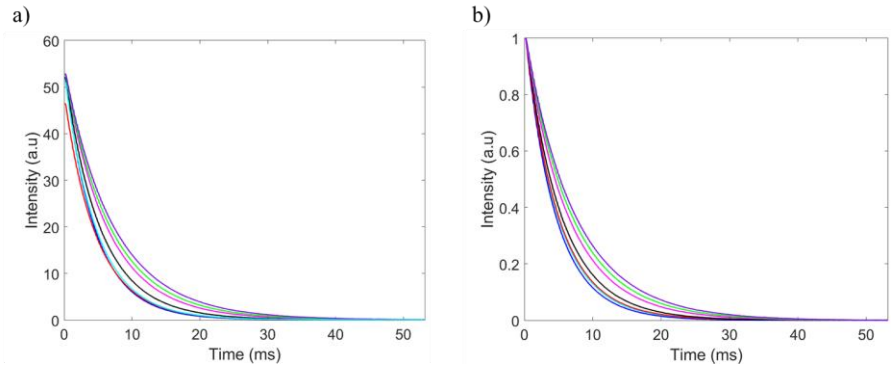



**Figure 2.15.** LF-NMR spectra of Eucalyptus honey before (a)) and after (b)) applying the normalization pretreatment.

***Multiplicative scatter correction (MSC)*** corrects for scale and baseline effects. A linear regression is applied between each spectrum $\mathbf{x}_i$ and a reference spectrum (Equation 5). In this Thesis, the reference spectrum considered is the mean spectrum ($\bar{\mathbf{x}}$) of the data matrix **X** [82].

$$\mathbf{x}_i = a_i + b_i\bar{\mathbf{x}} + e_i \qquad \text{Equation 5}$$

Where $a_i$ and $b_i$ are the coefficients calculated for each sample *i* obtained from the regression adjusted by least squares. To obtain the corrected spectrum, the intercept ($a_i$) is subtracted from each spectrum, and it is divided by the slope ($b_i$) (Equation 6) [81,83-85]:

$$\mathbf{x}_{i,\text{corrected}} = \frac{[\mathbf{x}_i - a_i]}{b_i} \qquad \text{Equation 6}$$

Figure 2.16. shows the application of MSC normalization to ATR-FTIR spectra for the authentic cashews data set. As a result, while the scale remains roughly the same, the variability between samples is reduced.

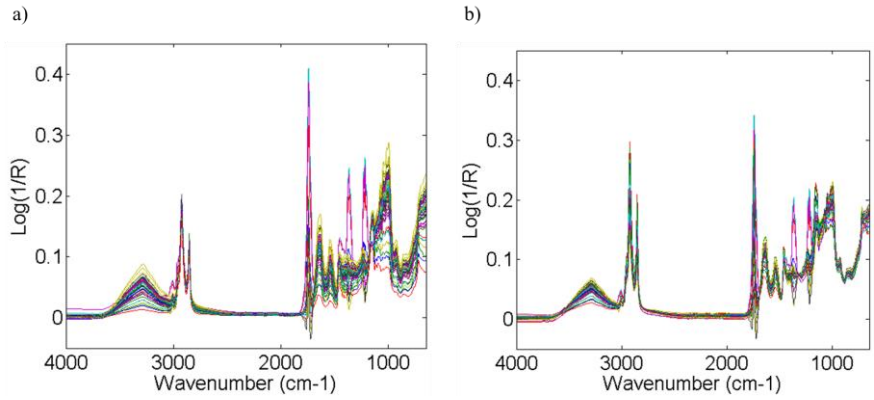



**Figure 2.16**. ATR-FTIR spectra of authentic cashew nuts before (a)) and after (b)) applying the MSC pretreatment.

### 2.3.1.2. Filtering

Filters remove high- or low-frequency, such as noise, baseline distortions, positive or negative slopes in spectra, or other baseline effects. **Smoothing** is a low-pass filter employed to eliminate high-frequency signals in samples, attributed to noise. Among other smoothing algorithms, the most used is the *Savitzky-Golay*. The spectrum is divided into successive subsets of points (windows) which are fitted to a polynomial function that minimizes the fitting error. This procedure is done for each defined window through all the data points (spectra) obtaining the corrected spectra [81,86,87].

For the sake of clarity, Figure 2.17. is presented showing the process, in this case for a window size of 6 points (blue circles). The fitting process can be observed in Figure 2.17a, which shows the raw data (solid blue line), the window size to be smooth (6 blue circles), the polynomial function, and the fitted line in red. Figure 2.17b shows the final step with the smoothed spectra indicated with blue points.

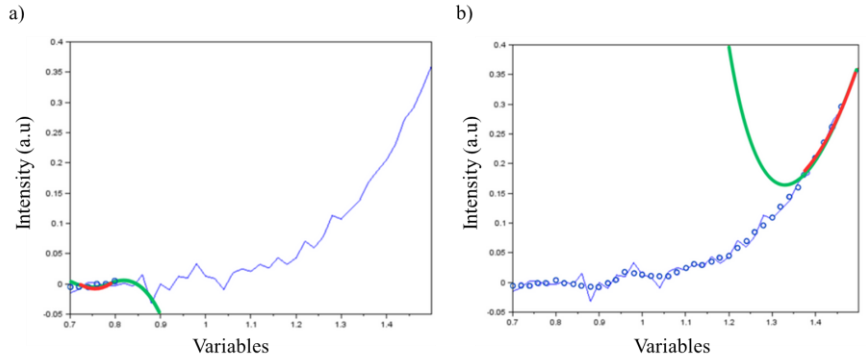



**Figure 2.17.** Example of smoothing by Savitzky-Golay method. The solid blue line is the raw data, blue circles are the points after the smoothing pretreatment, the green curve is the polynomial function, and the red curve is the polynomial function restricted to the window defined. Figures a) and b) correspond to two different windows.

Figure 2.18. shows the ATR-FTIR spectra of cashew nuts before (2.18a) and after the smoothing (2.18b). As a result, the bands in Figure 2.18b are better defined, with noise removed.

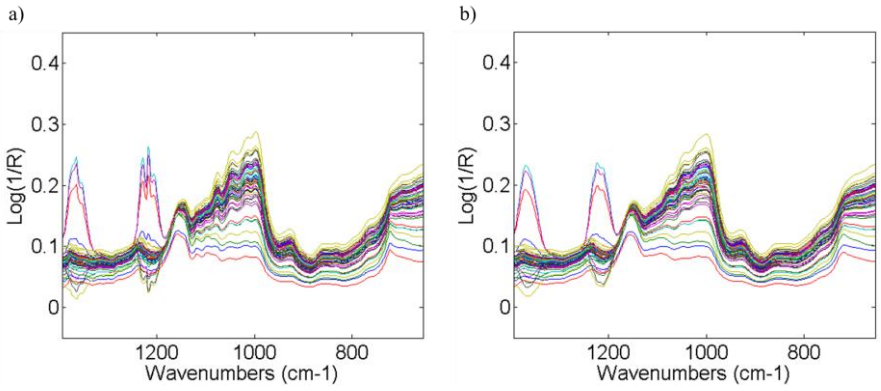



**Figure 2.18.** ATR-FTIR spectra of authentic cashew nuts before (a)) and after (b)) applying the Smoothing pretreatment.

Other filters widely used are the **derivative methods**, which amplify small peaks, highlight differences in shapes between very similar peaks, and remove unimportant baseline signals [88,89]. It is common to obtain the first or the second derivative, being the first derivative the one that has been applied in this thesis. To obtain the first derivative of a spectrum $\mathbf{x}_i$, the

difference between two adjacent points (*j*, *j-1*) has been calculated, for all the variables. The corrected value for sample *i* and variable *j* is calculated according to Equation 7.

$$x_{ij,corrected} = x_{ij} - x_{ij-1} \qquad \text{Equation 7}$$

Before obtaining the first derivative, the spectra are normally smoothed [90].

Figure 2.19. shows the application of the first derivative to the NIR spectra of authentic cashew nuts dataset. As can be seen in Figure 2.19b, the $\mathbf{x}_{i,corrected}$ spectra are centered around 0 and the baseline has been corrected.

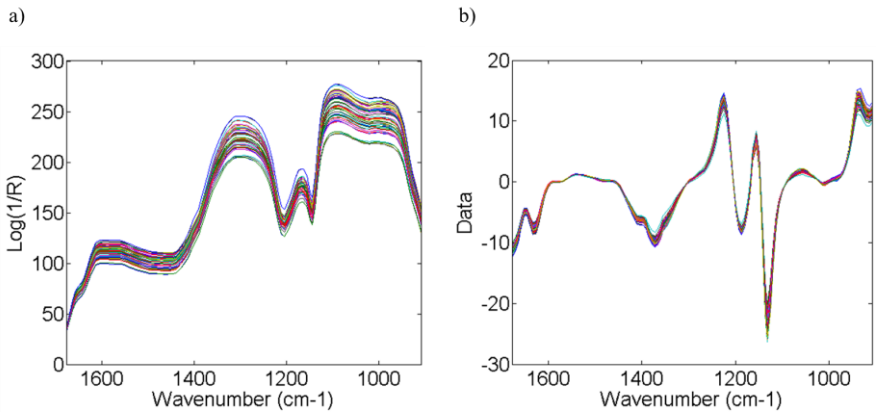



**Figure 2.19.** NIR spectra of authentic cashew nuts before (a)) and after (b)) applying the First Derivative pretreatment.

**Generalized least squares weighting** (**GLSW**) is a filter applied to minimize the within-class variance as much as possible without reducing the distance between classes [81,91]. This filter generates a matrix that accounts for the differences between pairs or groups of samples that are expected to be similar [92]. As can be observed in Figure 2.20b, in the corrected spectra all samples are centered in 0 and the variability between samples has been reduced. This reduction is not evident in the graph due to the change in scale on the Y-axis.

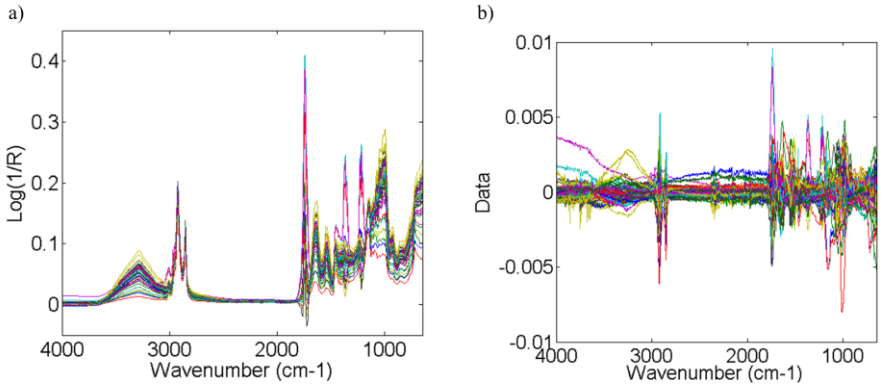



**Figure 2.20.** ATR-FTIR spectra of authentic cashew nuts before (a)) and after (b)) applying the GLSW pretreatment.

**Baseline correction using specified points**. The baseline of a spectrum is defined as those regions of the spectrum where there are no significant instrumental responses (ex. zero absorbance units). Due to various factors, the baseline may shift vertically or curve. When this occurs, the baseline has to be corrected.

For this purpose n number of variables (*j* to *j*+*n*) are selected, which are considered baseline points that ought to be set at zero. To obtain the corrected spectra for a sample *i* the mean value of the *n* selected variables is subtracted from the value of each variable *j*, according to Equation 8:

$$x_{ij,corrected} = x_{ij} - \bar{x}_{n,i} \qquad \text{Equation 8}$$

Where $\bar{x}_{n,i}$ is the average value of the *n* selected spectrum points, being *n* the same variables for the whole data set, although each sample will have a different value. In Figure 2.21. can be seen in the effect of the baseline correction in NIR spectra of the cashew nuts dataset, where in Figure 2.21b the corrected spectra start at the same point [81,93].

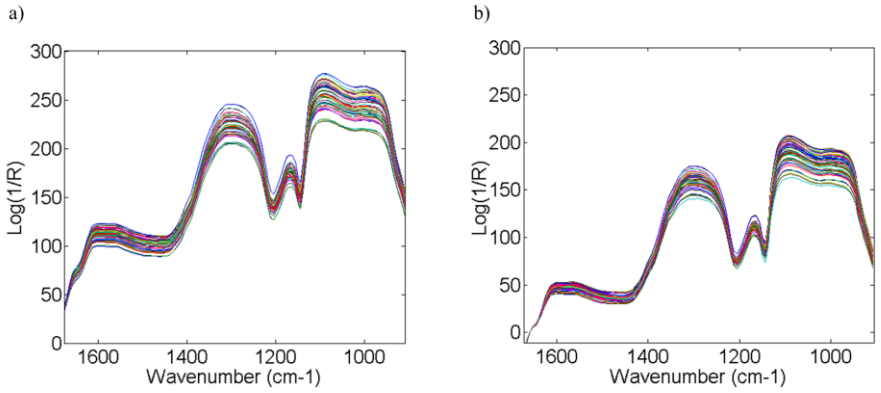



**Figure 2.21.** NIR spectra of authentic cashew nuts before (a)) and after (b)) applying the baseline correction pretreatment.

**Mean centering** involves adjusting a dataset to relocate the centroid to the origin of the coordinate system, according to Equation 9:

$$\mathbf{x}_{i,corrected} = \mathbf{x}_i - \bar{\mathbf{x}} \qquad \text{Equation 9}$$

Where the $\bar{\mathbf{x}}$ is the average spectra. This approach retains the differences in spectral shapes while equalizing the magnitude of peaks across the spectrum as can be seen in Figure 2.22. [93].

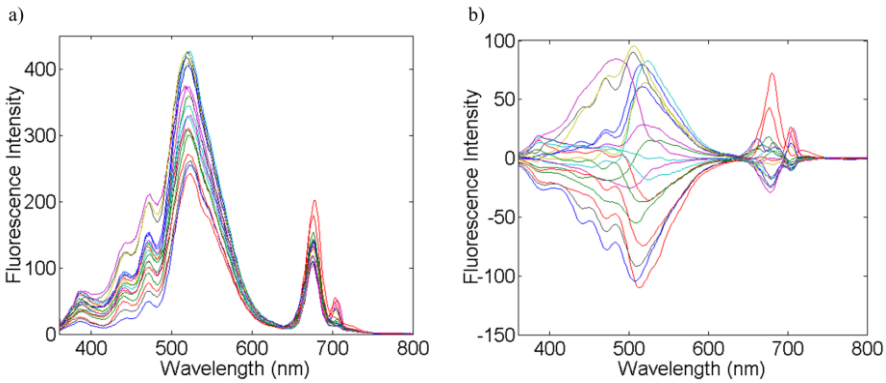



**Figure 2.22.** Fluorescence spectra of DO Siurana extra virgin olive oil before (a)) and after (b)) applying the mean center pretreatment.

### *2.3.2. Data exploration. Principal Component Analysis (PCA)*

The next step in the development of a multivariate model is data exploration (Step 03, Figure 2.14.). One of the most used exploratory methods is the representation of Principal Components (PCs). PCA is a bilinear decomposition or projection technique, that can condense datasets containing multiple variables into a few new variables known as Principal Components (PCs).

PCA decomposes the data matrix **X** following Equation 10:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^{\mathbf{T}} + \mathbf{E} \qquad \text{Equation 10}$$

Where **T** is the score matrix, which is the projection of samples into the new coordinate system, **P** is the loadings matrix indicating the weight coefficients for the original variables and **E** is the residuals matrix, which has the information not captured by the new variables [94,95].

In this way, the first component PC1 explains the highest variance, that is it contains the maximum information. The second PC2 explains the next highest variance and the maximum information not explained by the first PC. This pattern continues accordingly [79,94,96].

PCA is a versatile unsupervised technique that can be used as a data visualization tool for exploratory analysis, as a data reduction technique, and as a support for various classification and calibration strategies. The most known application is to get an initial idea about the pattern of the data analyzed. With the generated scores plot, PCA enables a visual representation that demonstrates how data is distributed along the PCs, and it is useful for detecting outliers and identifying connections and trends within the data.

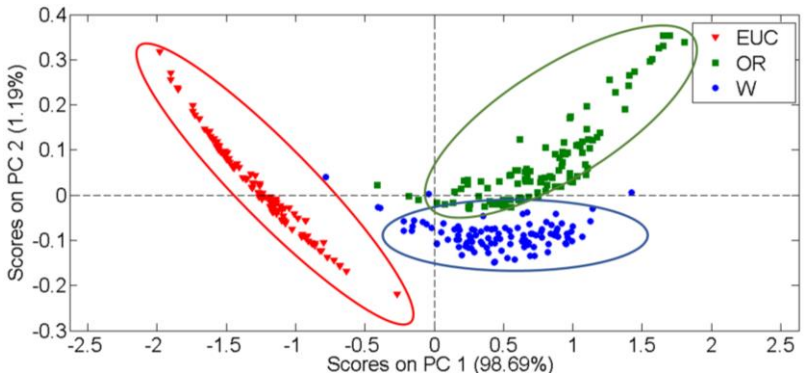



**Figure 2.23.** Example of a PCA plot. Red triangles are for the eucalyptus (EUC) type of honey, green squares are for the orange (OR) type of honey, and blue circles are for the wild (W) type of honey.

As an example, in Figure 2.23. a PCA score plot obtained from data corresponding to three types of honey is shown. As can be observed there are three differentiated groups of samples, one for each variety of honey, eucalyptus (red triangles), orange (green squares), and wild (blue circles).

### *2.3.3. Model development. Classification techniques*

The next step (Step 04, Figure 2.14.) is the development of the model using classification techniques, that can be grouped into two main types: discriminant and class-modelling techniques.

**Discriminant techniques** divide the data space into the same number of regions as defined classes; therefore, their focus is on the differences between samples from different classes. In these techniques, every sample is classified into one of the predefined classes if only there are two predefined classes. If there are more than two classes, the samples can be assigned to more than one class. Some of the main discriminant techniques are linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), k nearest neighbors (kNN), partial least squares discriminant analysis (PLS-DA), and one-class partial least squares (OCPLS) [96-98]. Although traditional discriminant techniques need at least two classes,

nowadays one-class discriminant techniques such as OCPLS have been developed.

**Class-modelling techniques** establish an independent model for each class. In contrast to the discriminant ones, they are focused on the similarities between samples of the same class. If a sample falls within the limits of the modelled class it is accepted by this category, while if it falls outside, the class model rejects it. Some of the main modelling techniques are soft independent modelling of class analogy (SIMCA) and unequal dispersed classes (UNEQ) [96-98].

The discriminant approach ensures every sample is assigned to a class, which complicates the detection of outliers. On the other hand, modelling techniques can assign samples to one class, more than one class, or none, leading to results that can be ambiguous or inconclusive [96].

In Figure 2.24. can be seen that in the last five years, both discriminant and modelling techniques have been used in the food fraud field, being the discriminant ones slightly higher.


**Figure 2.24.** The number of papers in the last five years using discriminant or class-modelling techniques. Source Web of Science from 2019 to the beginning of 2024.

The classification techniques implemented in the development of this thesis have been: SIMCA, PLS-DA, and OCPLS. In the next sub-sections, the fundamentals of them will be shortly discussed.

#### 2.3.3.1. Soft independent modelling of class analogy (SIMCA)

In SIMCA, the model for each class is determined using a principal component decomposition of appropriate dimensionality. As an example, the model for class 1 could be described by a Z principal component, according to Equation 11:

$$\mathbf{X_1} = \mathbf{T}_Z \cdot \boldsymbol{P}_Z^T + \mathbf{E} \qquad \text{Equation 11}$$

Where $\mathbf{X_1}$ is the matrix of the original data set formed only from the samples of class 1, $\mathbf{T_Z}$ and $\mathbf{P^T_Z}$ are the matrices containing the first Z scores and loading vectors, and **E** is the residual matrix [98]. Similarly, models are obtained for as many classes are pre-defined.

For each sample, values of two statistics are obtained: *Hotelling* $T^2$ which is the distance within the model space, and *Q residual* which is the orthogonal distance to the model space. A class limit value is selected for both statistics to define the model limits ($T^2_{lim}$ and $Q_{lim}$), at a specific confidence level (usually 95%). Therefore, for a sample *"i"* to fit the model, it should have both parameters lower than the class limit ($T^2_i < T^2_{lim}$ and $Q_i < Q_{lim}$). For the sake of simplicity, both statistics, are usually expressed as a reduced value ($T^2_{i,r}$ and $Q_{i,r}$) calculated as the ratio between each sample statistic and the corresponding class limit (Equations 12 and 13).

$$T^2_{i,r} = \frac{T_i^2}{T_{lim}^2} \qquad \text{Equation 12}$$

$$Q_{i,r} = \frac{Q_i}{Q_{lim}} \qquad \text{Equation 13}$$

Therefore, for a sample to fit the model, it should have its reduced parameters lower than 1 ($T^2_{i,r} < 1$ and $Q_{i,r} < 1$).

From the reduced statistics value ($T^2_{i,r}$ and $Q_{i,r}$) another parameter can be obtained, the reduced distance ($d_{ik}$) obtained from equation 14:

$$d_{ik} = \sqrt{\left(Q_{i,r}\right)^2 + \left(T^2_{i,r}\right)^2} \qquad \text{Equation 14}$$

Where, $d_{ik}$ is the distance of a sample to the class *"k"*, "*r*" stands for the reduced values of sample "*i*". To assign a sample to a model, different class distance limits can be used, the most common are $d_{ik}<1$ and $d_{ik}<\sqrt{2}$ [99,100]. Accordingly, if a sample has a $d_{ik}<1$ (or $d_{ik}<\sqrt{2}$) it will be accepted by the modelled class, otherwise will be rejected.

2.3.3.2. Partial Least Squares-Discriminant Analysis (PLS-DA)

PLS-DA is based on the regression model of the PLS algorithm, which relates an independent data matrix **X** and a dependent matrix **Y.** Those matrices are decomposed into scores and loadings according to Equations 15 and 16:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P}^{\mathbf{T}} + \mathbf{E}_{\boldsymbol{x}} \qquad \text{Equation 15}$$

$$\mathbf{Y} = \mathbf{U} \cdot \mathbf{Q}^{\mathbf{T}} + \mathbf{E}_{y} \qquad \text{Equation 16}$$

Where **X** is a matrix, whose rows are the spectra of the samples of the training set and **Y** is a matrix with as many columns as classes and whose elements respond to a binary code (0 and 1) where 1 indicates that the sample belongs to the class and 0 that do not belong. **T** and **U** are the score matrices, respectively, **P** and **Q** are the loading matrices, and $\mathbf{E_x}$ and $\mathbf{E_y}$ are the residual matrices of the **X** and **Y** matrices, respectively.

The predicted value for each sample will be either close to 0 or 1. A class limit ($Y_{lim}$) is set by assuming that the predicted values adhere to a Gaussian distribution, estimated by utilizing the mean and standard deviation of the predicted values for each class [99,100]. If the predicted value ($Y_{predicted}$) is greater than the class limit set ($Y_{lim}$) will be classified as belonging to the class. If $Y_{predicted}$ is lower than the $Y_{lim}$ will be assigned as not belonging to the class.

#### 2.3.3.3. One-Class Partial Least Squares (OCPLS)

OCPLS is based, as in the case of PLS-DA, on PLS but using only the authentic class in the training stage.

This technique, similar to PLS-DA, employs an **X** matrix with the spectra and **y** a vector with all values equal to 1 [24].

Similarly as in SIMCA, when the OCPLS model is built two statistics are obtained: Hotelling $T^2$, which is based and named score distance (*SD*), and the absolute centered residual (*ACR*). A class limit value is selected for both statistics to define the model limits ($SD_{lim}$ and $ACR_{lim}$), at a specific confidence level (usually 95%) [101-103].

Therefore, for a sample to fit the model, it should have both parameters lower than the class limit ($SD_i<SD_{lim}$ and $ACR_i<ACR_{lim}$). Those statistics can be used to calculate the reduced distance, using Equation 17:

$$d_{r,i} = \sqrt{\left(\frac{SD_i}{SD_{lim}}\right)^2 + \left(\frac{ACR_i}{ACR_{lim}}\right)^2} \qquad \text{Equation 17}$$

where $SD_i$ and $ACR_i$ are the statistical parameters of a sample *"i"* and the $SD_{lim}$ and $ACR_{lim}$ are the corresponding statistical class limits at a determinate level of significance.

### *2.3.4. Model evaluation*

The next step (Step 05, Figure 2.14.) of the flowchart presented at the beginning of this section is the model evaluation. This implies estimating its performance parameters which give information about the sample's assignation accuracy, the capacity of prediction, and the statistical significance of the obtained results [104].

The most basic way of quantifying the performance of a qualitative analysis method implies determining in a set of known membership samples whether the model assigns as belong (are within the limits of the class) or does not belong (are outside the limits of the class). Consequently, true positive (TP) and false positive (FP) or true negative (TN) and false negative (FN) rates

will be obtained, comparing the prediction outputs with the real situation of the samples concerning the class. From these outputs, different qualitative performance parameters can be calculated as can be seen in Table 2.5. [104-106].

**Table 5.** Performance parameters, from Eurachem guide 2021 (Ellison S.L.R et al.) [107].

| Performance characteristics | Expression |
|---|---|
| True positive rate, *TP* (Sensitivity, *SS*) | $tp/pc = tp/(tp + fn) = 1 - FN$ |
| False positive rate, *FP* | $fp/nc = fp/(tn + fp) = 1 - TN$ |
| True negative rate, *TN* (Specificity, *SP*) | $tn/nc = tn/(tn + fp) = 1 - FP$ |
| False negative rate, *FN* | $fn/pc = fn/(tp + fn) = 1 - TP$ |
| 'Precision' or 'Positive predictive value', *PPV* | $tp/p = tp/(tp + fp)$ |
| 'Negative predictive value', *NPV* | $tn/n = tn/(tn + fn)$ |
| Efficiency, *E* | $(tp + tn)/(p + n)$ |
| Youden Index, *Y* | $SS(\%) + SP(\%) - 100$ |
| Likelihood ratio of positive results, $LR(+)$ | *TP*/*FP* |
| Likelihood ratio of negative results, $LR(-)$ | *TN*/*FN* |
| Posterior probability | See Annex A |

*tp* – number of true positive results; *fp* – number of false positive results; *tn* – number of true negative results; *fn* – number of false negative results; *p* – number of positive results (*tp* + *fp*); *n* – number of negative results (*tn* + *fn*); *pc* – number of positive cases and *nc* – number of negative cases.

It should be emphasized, that one peculiarity of the qualitative analysis is that the positive and negative results can be established in different ways and, hence, involve different interpretations of the performance characteristic values (Table 2.5.). Therefore, unambiguous definitions of positive and negative results should be specified and/or adapted to the terminology used in the calculation of the performance parameters of Table 2.5.

In this thesis, in any adulteration problems studied, the positive result is defined for contaminated samples (presence of a contaminant). Therefore the negative output remains for the non-adulterated class. Both for one-class and multiclass approaches.

Along this thesis, sensitivity, specificity, and efficiency have been reported which are calculated for each class according to Table 2.5. and are defined as:

- Sensitivity: The percentage of samples that are properly assigned as they fit the class model.

- Specificity: The percentage of samples that are properly assigned as they do not fit the class model.
- Efficiency: Corresponds to the global prediction ability. It is calculated as the geometric mean of sensitivity and specificity values.

As has been stated previously, these performance parameters arise from the model outputs which are obtained in the prediction step considering a defined class limit. Nowadays, some papers propose the definition of two class limits instead of just one, defining an uncertainty region (UR) between them [108,109]. If that is the case, when samples fall in the UR, inconclusive assignations are reported since the results do not provide sufficient confidence in their classification. Inconclusive results require further study, usually by confirmatory analysis, to report results as "conclusive" [107]. This approach permits less error in the assignment of unknown samples because allows to detection of which samples should be taken to a confirmatory analysis. Both criteria (one and two class limits) will be applied in this thesis.

Some authors refer to this error region as uncertainty and others as unreliability region. The main difference is related to the response type that is obtained. The Eurachem Guide defines that uncertainty pertains to quantitative results and is represented by a range of concentrations where the results are expected to fall. In contrast, the unreliability region is related to binary responses (yes/no) and refers to a range where errors are produced. The principal differences between uncertainty and unreliability regions are summarized in Table 2.6. [107,110].

**Table 6.** Similarities and differences between unreliability and uncertainty, from Ríos, A. et al. 2003 [110].

| Metrological term | PROPERTY OF | | RANGE | INTERVAL WHERE | |
|---|---|---|---|---|---|
| | QUANTITATIVE RESULTS | BINARY YES/NO RESPONSES | | RESULTS CAN BE EXPECTED | ERRORS ARE PRODUCED |
| UNRELIABILITY | | X | X | | X |
| UNCERTAINTY | X | | X | X | |

In this thesis, we have used uncertainty since we determine it in a semi-quantitative model and it is related to the concentration range.

### *2.3.5. Strategies to expand information*

The next step (Step 06, Figure 2.14.) is the implementation of several strategies that allow to optimize the multivariate model or improve the classification results.

One of the thesis goals was the optimization of the model parameters by adjusting the class limit value. For this purpose, ROC curves and PCC have been applied.

Whenever working with spectroscopic data, sometimes, an improvement of the model or in the performance parameters can be obtained by the combinations of more than one instrumental technique. With this idea in mind, data fusion has been applied combining the spectra obtained from two instrumental techniques and combining the classification results obtained from models built individually with each technique.

Finally, another situation that has been considered is how to address the loss of validity of the model, in the prediction stage of new samples due to some change of condition with respect to the condition in which the model was developed and validated. If that is the case, transfer techniques are a good option to ensure that a developed model is still useful.

#### 2.3.5.1 Performance Characteristic Curves (PCC)

Performance Characteristic Curves (PCC) are a plot of the probability of having a positive result versus the concentration level of the analyte. Ideally,

the curve has a sigmoidal shape even though another type of curve shape might appear (i.e. exponential) [105,111].

PCC allow the estimation of semi-quantitative performance parameters and provides semi-quantitative information about a qualitative analytical method. Figure 2.25 shows a theoretical PCC, indicating the semi-quantitative information that can be obtained.

The decision limit (*CCα*) is the minimum concentration of the analyte (adulterant) that can be reliably detected or identified in a sample with low statistical certainty, usually 5%. The detection capability (*CCβ*) is the concentration of the analyte (adulterant), from which its presence can be reliably detected or identified in a sample with statistical certainty, usually 95%. Those parameters are obtained from the intersection with the two horizontal lines, which represent the probabilities of false positive and false negative errors. Those probabilities are usually set at P(X) = 95% or 5% in the upper and lower line, respectively [105,109].

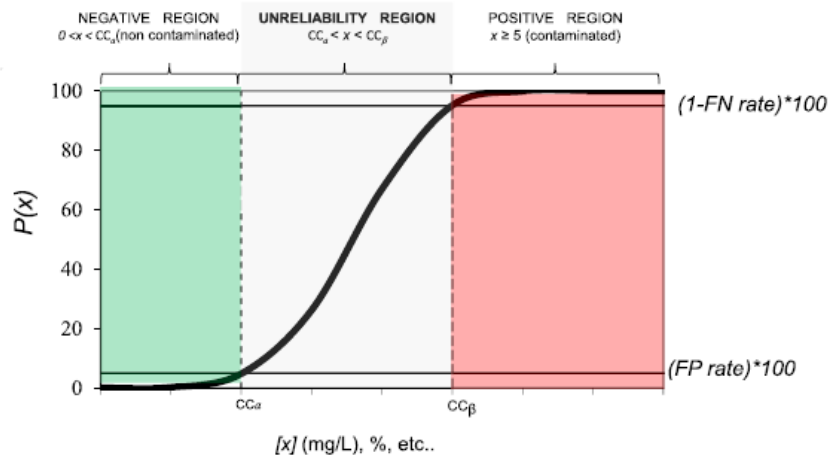



**Figure 2.25.** A theoretical example of a Performance Characteristic Curve (PCC), from López et al. 2015 [105].

The uncertainty region is the one between the *CCα* and *CCβ* class limits, in this region, false errors are probable [109,112].

### 2.3.5.2. Receiver Operating Characteristic (ROC) curves

The Receiver Operating Characteristic (ROC) curve is a graph showing the performance of a classification model when the value of a parameter is changed, such as the PC number, the α fixed to establish the class limits, and the distance limit, among others. (Figure 2.26.). For each value of the studied parameter, the curve plot shows the sensitivity against 1-specificity [109,113].

Figure 2.26, shows five theoretical examples of ROC curves. Curve A (blue point) presents a perfect model with a sensitivity and specificity of 100%. Curves B and C represent suitable methods, where sensitivity ≥ 1-specificity. Curve D represents the chance diagonal where the probability of classifying a sample as positive is equal to the probability of classifying it as negative. Finally, Curve E represents a non-suitable method, where 1-specificity ≥ sensitivity.

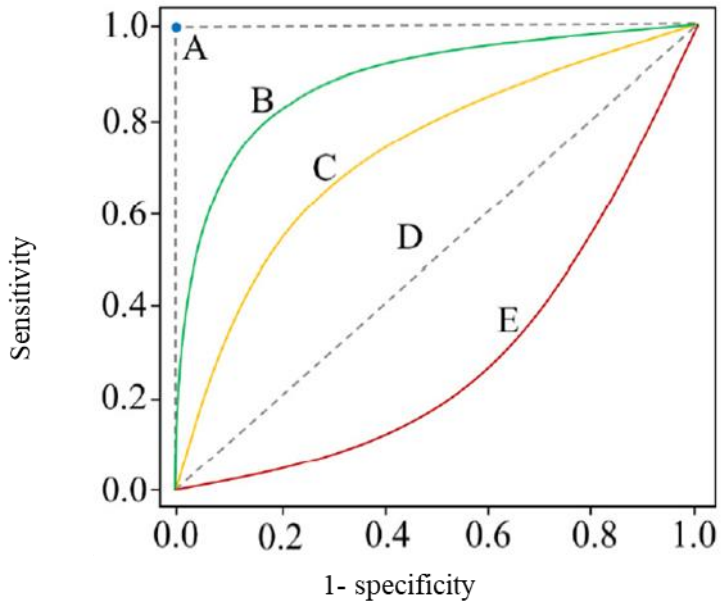



**Figure 2.26.** A theoretical example of a Receiver Operating Characteristic (ROC) Curve from EURACHEM 2021.

Such curves serve to compare the performance of the model and therefore select the parameter value which shows the best compromise between sensitivity and specificity. The area under the curve (AUC) is frequently used as a tool to assess the model performance. An AUC value of 0.5 indicates a minimal classification ability, while a value of 1 indicates a model operating at maximum efficiency [106,109,114].

#### 2.3.5.3. Data fusion

Data fusion combines data from diverse sources (instruments) to enhance the performance parameters beyond what could be achieved using individual sources alone [115].

The data can be combined at three levels: low-, mid-, and high-level. In low-level data fusion, raw data from multiple sources are directly combined, ensuring that all variables are on the same scale before integration. In mid-level data fusion, a subset of raw variables is selected and then combined. For these two approaches, a single classification model is developed once the spectra are combined. In high-level data fusion, classification models are individually calculated from each data source (instrument). Then the results from the individual models are combined [115,116].

In this thesis, the classification results (distances) have been fused by the fuzzy set theory. Four fuzzy aggregation connective operators were used: minimum, maximum, average, and product value [117]. To illustrate how the connective operators work and to understand how data fusion occurs, Table 2.7. shows an example of merging sample distances obtained from two classification models developed individually for two instrumental techniques.

**Table 7.** Example of high-level data fusion results. Code: In bold are shown the decisive values of the operators to obtain the ensemble decision.

| | *d* values | | Connective operators | | | | |
|---|---|---|---|---|---|---|---|
| Nº Sample | Instr. 1 Model | Instr. 2 Model | Min | Max | Prod | Avg | Ensamble decision |
| 1 | 0.63 | 1.10 | **0.63** | 1.10 | **0.69** | **0.86** | YES |
| 2 | 0.90 | 1.83 | 0.90 | **1.83** | **1.64** | **1.36** | NO |
| 3 | 1.07 | 0.90 | 0.90 | 1.70 | 0.96 | 1.30 | Inconclusive |

The second and third columns of Table 2.7. show the distance value of three different samples to the two models, developed from data measured with two instrumental techniques (Instr.). By way of example sample 1 is assigned as belonging to the model build with instrument 1 (Instr. 1) ($d$=0.63<1; YES), but with the model build with instrument 2 (Instr. 2) is assigned as not belonging to the model ($d$=1.10>1; NO).

The following columns show the result of applying the four fuzzy aggregation connective operators. The minimum (Min) and the maximum (Max) do not require mathematical calculations, it is simply to identify which is the maximum and minimum value between the distances. The product value (Prod) is the direct multiplication of the distance value. The average (Avg) is the sum of the distance values divided, in this example, by 2 because the results of the two models are fused.

Once the operators are applied, the sample is assigned following the majority vote criteria. Therefore, in Table 2.7, sample 1 is assigned to the model because 3 of the 4 operators have a value lower than 1 (shown in bold). Sample 2 is assigned as not belonging to the model because it has 3 operators with a value greater than 1 (shown in bold). And sample 3 remains assigned as inconclusive.

#### 2.3.5.4. Multivariate standardization

For the usefulness of a multivariate model, the performance parameters determined and validated must continue to be acceptable when predicting future samples. So, it is important to periodically test the model's validity to be sure its predictive capacity is sustained.

If a scenario occurs where the model might lose its effectiveness for the correct prediction of new samples, there are two approaches to solve it [118]. The first is creating a new model although this is the less practical solution due to the associated time and cost. The most beneficial solution is to apply chemometric methods to rectify variations in measurements obtained under different conditions, from which the model has been established. Once the spectra are corrected as if they were measured in the first conditions, they can be properly predicted by the model. With this strategy, the spectra are modified but not the model. These strategies are known as calibration transfer or standardization methods [118,119]. There are many standardization methods deeply explained in the literature [120-123]. In this thesis, an approach based on the standardization technique of Piecewise Direct Standardization (PDS) is used. This approach is developed in section 3.2, where we present the paper to which this strategy has given rise.

## References

[1] Gwenzi, W.; Makuvara, Z.; Marumure, J.; Simbanegavi, T.T.; Mukonza, S.S.; Chaukura, N. Chicanery in the food supply chain! Food fraud, mitigation, and research needs in low-income countries. *Trends Food Sci. Technol.* **2023**, 136, 194-223, doi: 10.1016/j.tifs.2023.03.027.

[2] Giannakas, K.; Yiannaka, A. Food Fraud: Causes, Consequences, and Deterrence Strategies. *Annu. Rev. Resour. Econ.* **2023**, 15, 85-104, doi: 10.1146/annurev-resource-101422-013027.

[3] Robson, K.; Dean, M.; Huaghey, S.; Elliot. C. A comprehensive review of food fraud terminologies and food fraud mitigation guides. *Food Control* **2021**, 120, 107516, doi: 10.1016/j.foodcont.2020.107516.

[4] Instituto Brasilerio de Geografia e Estatística, Produção Agropecuaria, available at: https://www.ibge.gov.br/explica/producao-agropecuaria/mel-de-abelha/br.

[5] European Commission, *Honey Market Presentation*, **2023**.

[6] Observatory of Economic Complexity (OEC), available at: https://oec.world/es/profile/bilateral-product/honey/reporter/bra.

[7] Araujo, P.H.L. Câmara Setorial de Mel. In *Ministério da Agricultura e Pecuária*, **2023**.

[8] (No) sugar for my honey: OLAF investigates honey fraud. In: *European Anti-Fraud Office (OLAF).* Press Release Nº 3/2023, **2023**.

[9] Questions and Answers surrounding Honey adulteration. Official controls food fraud. In *European Commission*, **2021**.

[10] Agência IBGE notícias, Governo de Brazil, available at: https://agenciadenoticias.ibge.gov.br/en/agencia-press-room/2185-news-agency/releases-en/38886-ibge-expects-record-harvest-of-306-5-million-tonnes-in-2024-with-a-drop-of-2-8-from-2023.

[11] M.F. Rural, Castanhas à venda com preço, Mercado Físico Rural, Marília, Brazil, available at: https://www.mfrural.com.br/produtos/3-2992/alimentos-castanhas.

[12] Portaria SDA Nº 635, In: *Diário Oficial Da União*, **2022**.

[13] Regulation (EC) Nº 178/2002 of the European Parliament and of the Council. In: *European Commission*, **2022**.

[14] Llei 14/2003 de qualitat agroalimentària. In: *Portal Jurídic de Catalunya*, **2003**.

[15] Standard for olive oils and olive pomace oils. In: *Codex Alimentarius* (CX-33-1981), **1981**.

[16] Abbas, O.; Dardenne, P.; Baeten, V. Near-Infrared, Mid-Infrared, and Raman Spectroscopy. In: *Chemical analysis of food: Techniques and applications*; Pico, Y., Ed; Academic Press, **2012**; pp. 59-89. ISBN: 9780123848628.

[17] Shawn, R.A.; Mantsch, H.H. Near-IR Spectrometers. In: *Encyclopedia of Spectroscopy and Spectrometry (2nd Edition)*; Lindon, J.C., Ed; Academic Press, **1999**; pp. 1738-1747. ISBN: 9780123744135.

[18] Penner, M.H. Basic Principles of Spectroscopy. In *Food Analysis*; Nielsen, S.S., Ed; Springer, **2017**; pp. 78-88. ISBN: 9783319457741.

[19] Rodriguez-Saona, L.; Ayvaz, H.; Wehling, R. L. Infrared and Raman Spectroscopy. In *Food Analysis*; Nielsen, S.S., Ed; Springer, **2017**; pp. 107-127. ISBN: 9783319457741.

[20] Kang. S. NIR Spectroscopy for chemical composition and internal quality in foods. In*: Emerging Technologies for Food Quality and Food Safety Evaluation*; Cho, Y-J., Kang. S., Sun, D-W., Ed; Taylor & Francis Group, **2011**; pp.113-148. ISBN: 9781138199132.

[21] Ozaki, Y.; Morisawa, Y. Principles and Characteristics of NIR spectroscopy. In: Near-Infrared Spectroscopy; Ozaki, Y., Huck, C., Tsuchikawa, S., Engelsen, S.B., Ed; Springer, **2021**. ISBN: 9789811586477.

[22] Fagan, C.C. Infrared Spectroscopy. In *Process Analytical Technology for the Food Industry*; O'Donell, C.P, Fagan, C., Cullen, P.J., Ed; Springer, **2014**; pp. 73-102. ISBN: 9781493903108.

[23] Karthika, B.R.; Nishad, V.M.; Prasobh, G.R. An overview on infrared spectroscopy. *Int. J. Res. Publ. Rev.* **2022**, 3, 526-552, ISSN: 2585-7421.

[24] Valand, R.; Tanna, S.; Lawson, G.; Bengtström, L. A review of Fourier Transform Infrared (FTIR) spectroscopy used in food adulteration and authenticity investigations. *Food Addit. Contam. Part A* **2020**, 37, 19-38, doi: 10.1080/19440049.2019.1675909.

[25] Subramanian, A.; Rodriguez-Saona, L. Fourier transform infrared (FTIR) spectroscopy. In *Infrared spectroscopy for food quality analysis and control*; Sun, D-W., Ed; Academic Press, **2009**; pp. 145-178. ISBN: 9780123741363.

[26] Agelet, L.E.; Hurburgh, C.R. A tutorial on near infrared spectroscopy and its calibration. *Crit. Rev. Anal. Chem.* **2010**, 40, 246–260, doi: 10.1080/10408347.2010.515468.

[27] Spragg, RA. IR Spectroscopy Sample Preparation Methods. In *Encyclopedia of Spectroscopy and Spectrometry*; Lindon J., Ed; Elsevier Science & Technology, **2016**; pp. 1058-1066. ISBN: 9780128032244.

[28] Cozzolino D. Advantages, Opportunities, and Challenges of Vibrational Spectroscopy as Tool to Monitor Sustainable Food Systems. *Food Anal. Methods* **2022**, 15, 1390–1396, doi: 10.1007/s12161-021-02207-w.

[29] Huck, C.W. New trend in instrumentation of NIR spectroscopy miniaturization. In *Near-Infrared Spectroscopy;* Ozaki, Y.; Huck, C.; Tsuchikawa, S.; Engelsen, S. B., Ed; Springer, **2021**; pp. 193-210. ISBN: 9789811586477.

[30] Yan, H.; Neves, M.G.; Noda, I.; Guedes, G.M.; Ferreira, A.C.S.; Pfeifer, F.; Chen, X.; Siesler, H.W. Handheld Near-Infrared Spectroscopy: State-of-the-Art Instrumentation and Applications in Material Identification, Food Authentication, and Environmental Investigations. *Chemosens.* **2023**, 11, 272, doi: 10.3390/chemosensors11050272.

[31] Yang, S.; Zhao, Z.-N.; Yan, H.; Siesler, H.W. Fast detection of cotton content in silk/cotton textiles by handheld near-infrared spectroscopy: A performance comparison of four different instruments. *Text. Res. J.* **2022**, 92, 147–153, doi: 10.1177/00405175221082324.

[32] Yan, H.; Siesler, H.W. Quantitative analysis of a pharmaceutical formulation: Performance comparison of different handheld near-infrared spectrometers. *J. Pharm. Biomed. Anal.* **2018**, 160, 179–186, doi: 10.1016/j.jpba.2018.07.048.

[33] Kranenburg, R.F.; Weesepoel, Y.; Alewijn, M.; Sap, S.; Arisz, P.W.F.; van Esch, A.; Keizers, P.H.J.; van Asten, A.C. The importance of wavelength selection in on-scene identification of drugs of abuse with portable near-infrared spectroscopy. *Forensic Chem.* **2022**, 30, 100473, doi: 10.1016/j.forc.2022.100437.

[34] Liu, N.; Parra, H.A.; Pustjens, A.; Hettinga, K.; Mongondry, P.; van Ruth, S.M. Evaluation of portable near-infrared spectroscopy for organic milk authentication. *Talanta* **2018**, 184, 128–135, doi: 10.1016/j.talanta.2018.02.097.

[35] Medeiros, M.L.S.; Lima, A.F.; Gonçalves, M.C.; Godoy, H.T.; Barbin, D.F. Portable near-infrared spectrometer and chemometrics for rapid identification of butter cheese adulteration. *Food Chem*. **2023**, 425, 136461, doi: 10.1016/j.foodchem.2023.136461.

[36] Silva, L.C.R.; Folli, G.S.; Santos, L.P.; Barros, I.H.A.S.; Oliveira, B.G.; Borghi, F.T.; dos Santos, F.D.; Filgueiras, P.R.; Romão, W. Quantification of beef, pork, and chicken in ground meat using a portable NIR spectrometer. *Vib. Spectrosc.* **2020**, 111, 103158, doi: 10.1016/j.vibspec.2020.103158.

[37] Borghi F.T.; Santos P.C.; Santos F.D.; Nascimento M.H.C.; Corrêa T.; Cesconetto M. et al. Quantification and classification of vegetable oils in extra virgin olive oil samples using a portable near-infrared spectrometer associated with chemometrics. *Microchem. J.* **2020**, 159, 105544, doi: 10.1016/j.microc.2020.105544.

[38] Grassi, S.; Casiraghi, E.; Alamprese, C. Handheld NIR device: A non-targeted approach to assess authenticity of fish fillets and patties. *Food Chem.* **2018**, 243, 382–388, doi: 10.1016/j.foodchem.2017.09.145.

[39] Amuah, C.L.Y.; Teye, E.; Lamptey, F.P.; Nyandey, K.; Opoku-Ansah, J.; Osei-Wusu Adueming, P. Feasibility study of the use of handheld NIR spectrometer for simultaneous authentication and quantification of quality parameters in intact pineapple fruits. *J. Spectrosc*. **2019**, 2019, 5975461, doi: 10.1155/2019/5975461.

[40] Wang, Y.-J.; Li, T.-H.; Li, L.-Q.; Ning, J.-M.; Zhang, Z.-Z. Micro-NIR spectrometer for quality assessment of tea: Comparison of local and global models. *Spectrochim. Acta A* **2020**, 237, 118403, doi: 10.1016/j.saa.2020.118403.

[41] Oliveira, M.M.; Cruz-Tirado, J.P.; Roque, J.V.; Teófilo, R.F.; Barbin, D.F. Portable near-infrared spectroscopy for rapid authentication of adulterated paprika powder. *J. Food Compos. Anal.* **2020**, 87, 103403, doi: 10.1016/j.jfca.2019.103403.

[42] García-González, D.L.; Baeten, V.; Pierna, J.A.F.; Tena, N. Infrared, Raman, and Fluorescence Spectroscopies: Methodologies and Applications. In *Handbook of Olive oil: Analysis and Properties*; Aparicio, R., Harwood, J., Ed; Springer, **2013**; pp; 335-334. ISBN: 9781461477761.

[43] Tachie, C.Y.E.; Obiri-Ananey, D.; Alfaro-Cordoba, M.; Tawiah, N.A.; Aryee, A.N.A. Classifications of oils and margarines by FTIR spectroscopy in tandem with Machine learning. *Food Chem.* **2024**, 431, 137077, doi: 10.1016/j.foodchem.2023.137077.

[44] Tan, E.; Julmohammad, N.B.; Koh, W.Y.; Sani, M.S.A.; Rasti, B. Application of ATR-FTIR incorporated with multivariate data analysis for discrimination and quantification of urea as an adulterant in UHT milk. *Foods* **2023**, 12, 2855, doi: 10.3390/foods12152855.

[45] Biancolillo, A.; Foschi, M.; Di Micco, M.M.; Di Donato, F.; D'Archivio, A.A. ATR-FTIR-based rapid solution for the discrimination of lentils from different origins, with special focus on PGI and Slow Food typical varieties. *Microchem. J.* **2022**, 178, 107327, doi: 10.1016/j.microc.2022.107327.

[46] Amit, Jamwal, R.; Kumari, S.; Dhaulaniya, A.S.; Balan, B.; Singh, D.K. Application of ATR-FTIR spectroscopy along with regression modelling for the detection of adulteration of virgin coconut oil with paraffin oil. *LWT-Food Sci. Tech.* **2020**, 118, 108754, doi: 10.1016/j.lwt.2019.108754.

[47] Foschi, M.; Tozzi, L.; Di Donato, F.; Biancolillo, A.; D'Archivio, A.A. A novel FTIR-based chemometric solution for the assessment of saffron adulteration with non-fresh stigmas. *Molec.* **2023**, 28, 33, doi: 10.3390/molecules28010033.

[48] Riswahyuli, Y.; Rohman, A.; Setyabudi, F.M.C.S; Raharjo, S. Indonesian wild honey authenticity analysis using attenuated total reflectance-Fourier transform infrared (ATR-FTIR) spectroscopy combined with multivariate statistical techniques. *Heliyon* **2020**, 6, e03662, doi: 10.1016/j.heliyon.2020.e03662.

[49] Andersen, C.M.; Wold, J.P.; Engelsen, S.B. Autofluorescence Spectroscopy in Food Analysis. In *Handbook of Food Analysis Instruments*; Ötles, S., Ed: Taylor & Francis Group, **2008**; pp: 347-364. ISBN: 9781420045666.

[50] Li, Y-Q.; Li, X-Y.; Shindi, A.A.G.; Zou, Z-X.; Liu, Q.; Lin, L.R.; Li, N. Synchronous fluorescence spectroscopy and its Applications in clinical analysis and food safety evaluation. *Rev. Fluoresc.* **2010**, 95-117, doi: 10.1007/978-1-4419-9828-6_5.

[51] Arslan, F.N.; Akin, G.; Karuk Elmas, Ş.N.; Yilmaz, I.; Janssen, H.-G.; Kenar, A. Rapid detection of authenticity and adulteration of cold-pressed black cumin seed oil: A comparative study of ATR–FTIR spectroscopy and synchronous fluorescence with multivariate data analysis. *Food Control* **2019**, 98, 323–332, doi: 10.1016/j.foodcont.2018.11.055.

[52] Tan, J.; Liu, J.-Y.; Su, H.; Yang, X.-H.; Li, H.-F. Detection of adulteration of cumin powder by front-face synchronous fluorescence spectroscopy: The influence of the natural variation of adulterants. *Food Control* **2024**, 158, 110228, doi: 10.1016/j.foodcont.2023.110228.

[53] Huyan, Z.; Ding, S.; Liu, X.; Yu, X. Authentication and adulteration detection of peanut oils of three flavor types using synchronous fluorescence spectroscopy. *Anal. Methods* **2018**, 1, 327–3214, doi: 10.1039/c8ay00837j.

[54] Dankowska, A.; Kowalewski, W. Tea types classification with data fusion of UV–Vis, synchronous fluorescence and NIR spectroscopies and chemometric analysis. *Spectrochim. Acta A* **2019**, 211, 195-202, doi: 10.1016/j.saa.2018.11.063.

[55] Fort, A.R.; Ruisánchez, I.; Callao, M.P. Chemometric strategies for authenticating extra virgin olive oils from two geographically adjacent Catalan protected designations of origin. *Microchem. J.* **2021**, 169, 106611, doi: 10.1016/j.microc.2021.106611.

[56] Meng, X.; Yin, C.; Yuan, L.; Zhang, Y.; Ju, Y.; Xin, K.; Chen, W.; Lv, K.; Hu, L. Rapid detection of adulteration of olive oil with soybean oil combined with chemometrics by Fourier transform infrared, visible-near-infrared and excitation-emission matrix fluorescence spectroscopy: A comparative study. *Food Chem.* **2023**, 405, 134828, doi: 10.1016/j.foodchem.2022.134828.

[57] Chen, Y.; Wu, H.; Wang, T.; Wu, J.; Liu, B.; Ding, Y.; Yu, R. Rapid detection and quantification of adulteration in saffron by excitation–emission matrix fluorescence combined with multi-way chemometrics. *J. Sci. Food Agric.* **2024**, 104, 1391–1398, doi: 10.1002/jsfa.13028.

[58] Hu, X-C.; Yu, H.; Deng, Y.; Chen, Y.; Zhang, X-H.; Gu, H-W.; Yin, X-L. Rapid authentication of green tea grade by excitation-emission matrix fluorescence spectroscopy coupled with multi-way chemometric methods. *Eur. Food Res. Technol.* **2023**, 290, 767–775, doi: 10.1007/s00217-022-04174-w.

[59] Rahmani, N.; Mani-Varnosfaderani, A. Excitation-emission fluorescence spectroscopy and sparse chemometric methods for grape seed oil classification and authentication. *Chemom. Intell. Lab. Syst.* **2023**, 241, 104939, doi: 10.1016/j.chemolab.2023.104939.

[60] Mani-Varnosfaderani, A.; Masroor, M.J.; Yamini, Y. Designating the geographical origin of Iranian almond and red jujube oils using fluorescence spectroscopy and l1-penalized chemometric methods. *Microchem. J.* **2020**, 157, 104984, doi: 10.1016/j.microc.2020.104984.

[61] Ríos-Reina, R.; Salatti-Dorado, J.A.; Ortiz-Romero, C.; Cardador, M.J.; Arce, L.; Callejón, R. A comparative study of fluorescence and Raman spectroscopy for discrimination of virgin olive oil categories: Chemometric approaches and evaluation against other techniques. *Food Control* **2024**, 158, 110250, doi: 10.1016/j.foodcont.2023.110250.

[62] Truong, H.T.D.; Reddy, P.; Reis, M.M.; Archer, R. Internal reflectance cell fluorescence measurement combined with-multi-way analysis to detect fluorescence signatures of undiluted honeys and a fusion of fluorescence and NIR to enhance predictability. *Spectrochim. Acta. A* **2023**, 290, 122274, doi: 10.1016/j.saa.2022.122274.

[63] Boukria, O.; Boudalia, S.; Bhat, Z.F.; Hassoun, A.; Aït-Kaddour, A. Evaluation of the adulteration of camel milk by non-camel milk using multispectral image, fluorescence, and infrared spectroscopy. *Spectrochim. Acta A* **2023**, 300, 122932, doi: 10.1016/j.saa.2023.122932.

[64] Hao, S.; Yuan, J.; Wu, Q.; Liu, X.; Cui, J.; Xuan, H. Rapid identification of corn sugar syrup adulteration in wolfberry honey based on fluorescence spectroscopy coupled with chemometrics. *Foods* **2023**, 12, 2309, doi: 10.3390/foods12122309.

[65] Ali, Z.; Saleem, M.; Atta, B.M.; Khan, S.S.; Hammad, G. Determination of curcuminoid content in turmeric using fluorescence spectroscopy. *Spectrochim. Acta A* **2019**, 213, 192-198, doi: 10.1016/j.saa.2019.01.028.

[66] McCarthy, M.J.; McCarthy, K.L. Magnetic Resonance Imaging and Nuclear Magnetic Resonance Spectroscopy. In *Process Analytical Technology for the Food Industry*; O'Donell, C.P, Fagan, C., Cullen, P.J., Ed; Springer, **2014**; pp. 135-156. ISBN: 9781493903108.

[67] Sobolev, A.P.; Thomas, F.; Donarski, J.; Ingallina, C.; Circi, S.; Marincola, F.C.; Capitani, D.; Mannina, L. Use of NMR application to tackle future food fraud issues. *Trend. Food Sci. Technol.* **2019**, 91, 347-353, doi: 10.1016/j.tifs.2019.07.035.

[68] Belmonte-Sánchez, E.; Romero-González, R.; Fenich A.G. Applicability of high-resolution NMR combination with chemometrics for the compositional analysis and quality control of spices and plant-derived condiments. *J. Sci. Food Agric.* **2020**, 101, 3541-3550, doi: 10.1002/jsfa.11051.

[69] Santos, A.D.C.; Fonseca, F.A.; Lião, L.M.; Alcantara, G.B., Barison, A. High-resolution magic angle spinning nuclear magnetic resonance in foodstuff analysis. *TrAC Trend. Anal. Chem.* **2015,** 73, 10-18, doi: 10.1016/j.trac.2015.05.003.

[70] Kirtil, E.; Oztop, M.H. $^{1}$H nuclear magnetic resonance relaxometry and magnetic resonance imaging and Applications in food science and processing. *Food Eng. Rev.* **2016**, 8, 1-22, doi: 10.1007/s12393-015-9118-y.

[71] Hashemi, R.H.; Bradley, W.G.; Lisanti, C.J. MRI: the basics. Ed; Lippincott Williams & Wilkins, **2010**; pp. 16-40. ISBN: 9781608311156.

[72] Bernstein, M.A.; King, K.F.; Zhou, X.J. Handbook of MRI pulse sequences. Ed; Elsevier Science, **2004**; pp. 29-34. ISBN: 9780080533124.

[73] Ribeiro, R. de O.R.; Mársico, E.T.; Carneiro, C. da S.; Monteiro, M.L.G.; Júnior, C.C.; Jesus, E.F.O. Detection of honey adulteration of high fructose corn syrup by Low Field Nuclear Magnetic Resonance (LF $^{1}$H NMR). *J. Food Eng.* **2014**, 135, 39-43, doi: 10.1016/j.jfoodeng.2014.03.009.

[74] McDowell, D.; Defernez, M.; Kemsley, E.K.; Elliott, C.T.; Koidis, A. Low vs high field $^{1}$H NMR spectroscopy for the detection of adulteration of cold pressed rapeseed oil with refined oils. *Food Sci. Technol.* **2019**, 111, 490-499, doi: 10.1016/j.lwt.2019.05.065.

[75] Ok, S. Detection of olive oil adulteration by low-field NMR relaxometry and UV-Vis spectroscopy upon mixing olive oil with various edible oils. *Grasas y Aceites (Sevilla)* **2017**, 68, 173-173, doi: 10.3989/gya.0678161.

[76] Tahir, H.E.; Arslan, M.; Mahunu, G.K.; Mariod, A.A.; Wen, Z.; Xiaobo, Z.; Xiaowei, H.; Jiyong, S.; El-Seedi, H. Authentication of the geographical origin of Roselle (Hibiscus sabdariffa L) using various spectroscopies: NIR, low-field NMR

and fluorescence. *Food Control* **2020**, 114, 107231, doi: 10.1016/j.foodcont.2020.107231.

[77] Bao, R.; Tang, F.; Rich, C.; Hatzakis, E. A comparative evaluation of low-field and high-field NMR untargeted analysis: Authentication of virgin coconut oil adulterated with refined coconut oil as a case study. *Anal. Chim. Acta* **2023**, 1237, 341537, doi: 10.1016/j.aca.2023.341537.

[78] Varmuza, K.; Filzmoser, P. Introduction to Multivariate Statistical Analysis in Chemometrics. Ed: CRC Press, **2009**; pp: 31-32. ISBN: 9781420059496.

[79] Vigni, M.L.; Durante, C.; Cocchi, M. Exploratory Data Analysis. In *Chemometrics in Food Chemistry;* Marini, F, Ed; Elsevier, **2013**; pp: 55-126. ISBN: 9780444595287.

[80] Oliveri, P.; Malegori, C.; Simonetti, R.; Casale, M. The Impact of Signal Pre-Processing on the Final Interpretation of Analytical Outcomes - A Tutorial. *Anal Chim. Acta* **2018**, 1058, 9–17, doi: 10.1016/j.aca.2018.10.055.

[81] Wise, B.M.; Gallagher, N.B.; Bro, R.; Shaver, J.M.; Windig, W.; Koch, R.S. Chemometrics Tutorial for PLS_Toolbox and Solo. Es; Eigenvector Research, Inc, **2006**. ISBN: 0976118416.

[82] Dhanoa, M.S.; Lister, S.J.; Sanderson, R.; Barnes, R.J. The link between Multiplicative Scatter Correction (MSC) and Standard Normal Variate (SNV) transformations of NIR spectra. *J. Near Infrared Spectrosc.* **1994**, 2, 43-47, doi: 10.1255/jnirs.30.

[83] Maleki, M.R.; Mouazen, A.M.; Ramon, H.; De Baerdemaeker, J. Multiplicative scatter correction during on-line measurement with Near Infrared Spectroscopy. *Biosyst. Eng.* **2007**, 96, 427-433, doi: 10.1016./j.biosystemseng.2006.11.014.

[84] Martens, H.; Jensen, S.A.; Geladi, P. Multiplicative linearity transformation for near infrared reflectance spectra of meat. Application Spectroscopy, Proceedings of the Nordic Symposium, Applied Statistics. Stockholm Forlag Publication, Stanger, Norway, **1983**, 235-267.

[85] Huang, F.; Song, H.; Guo, L.; Guang, P.; Yang, X.; Li, L.; Zhao, H.; Yang, M. Detection of adulteration in Chinese honey using NIR and ATR-FTIR spectral data fusion. *Spectrochim. Acta A* **2020**, 235, 118297, doi: 10.1016/j.saa.2020.118297.

[86] Tahir, H. E.; Xiaobo, Z.; Zhihua, L.; Jiyong, S.; Zhai, X.; Wang, S.; Mariod, A. A. Rapid prediction of phenolic compounds and antioxidant activity of Sudanese honey using Raman and Fourier transform infrared (FT-IR) spectroscopy. *Food Chem.* **2017**, 226, 202–211, doi: 10.1016/j.foodchem.2017.01.024.

[87] Savitzky, A.; Golay, M.J.E. Smoothing and differentiation by data simplified least squares procedures. *Anal. Chem.* **1964**, 1627, doi: 10.1021/ac6021a047.

[88] Chuen Lee, L.; Liong, C.-Y.; Aziz Jemain, A. A Contemporary Review on Data Preprocessing (DP) Practice Strategy in ATR-FTIR Spectrum. *Chemometr. Intell. Lab.* **2017**, 163, 64–75, doi: 10.1016/j.chemolab.2017.02.008.

[89] Ciao, Y.; Cai, H.; Ni, K. Identification of geographical origin and adulteration of Northeast China soybeans by mid-infrared spectroscopy and spectra augmentation, *J. Consumer Prot. Food Saf.* **2024**, 19, 99-111, doi: 10.1007/s00003-023-01471-8.

[90] Rinnanm, Å.; van den Berg, F.; Engelsen, S.B. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends Anal. Chem.* **2009**, 28, 1201-1222, doi: 10.1016/j.trac.2009.07.007.

[91] Nespeca, M.G.; Pavini, W.D.; de Oliveira, J.E. Multivariate filters combined with interval partial least square method: A strategy for optimizing PLS models developed with near infrared data of multicomponent solutions. *Vib. Spectrosc.* **2019**, 102, 97-102, doi: 10.1016/j.vivspec.2019.05.001

[92] Rozenstein, O.; Paz-Kagan, T.; Salbach, C.; Karnieli, A. Comparing the effect of preprocessing transformations on methods of land-use classification derived from spectral soil measurements. *J. Sel. Top. Appl. Earth Obs. Sens.* **2015**, 8, 2393-2404, doi: 10.1109/JSTARS.2014.2371920.

[93] Rinnanm Å.; Nørgaard, L.; van den Berg, F.; Thygesen, J.; Bro, R.; Engelsen, S.B. Data Pre-processing. In *Infrared Spectroscopy for Food Quality Analysis and Control;* Sun, D-W, Ed; Academic Press, **2009**; pp: 29-50. ISBN: 9780123741363.

[94] Esbensen, K.H.; Geladi, P. Principal Component Analysis: Concept, Geometrical interpretation, Mathematical Background, Algorithms, History, Practice. In: *Comprehensive Chemometrics Chemical and Biochemical Data Analysis*. Ed: Elsevier, **2009**; pp: 211-226, doi: 10.1016/B978-044452701-1.00043-0.

[95] Geladi, P. Chemometrics in Spectroscopy. Part 1. Classical Chemometrics. *Spectrochim Acta B* **2003**, 58, 767–782, doi:10.1016/s0584-8547(03)00037-5.

[96] Callao, M.P.; Ruisánchez, I. An overview of multivariate quality methods for food fraud detection. *Food Control* **2018**, 86, 289-293, doi: 10.1016/j.foodcont.2017.11.034.

[97] Bevilaqua, M.; Bucci, R.; Magrì, A.D.; Magrì, A.L.; Nescatelli, R.; Marini, F. Classification and Class-Modelling. In *Chemometrics in Food Chemistry;* Marini, F, Ed; Elsevier, **2013**; pp: 171-233. ISBN: 9780444595287.

[98] Marini, F. Classification methods in chemometrics. *Current Anal. Chem.* **2010**, 6, 72-79, doi: 10.2174/157341110790069592.

[99] Miaw, C.S.M.; Sena, M.M.; de Souza, S.V.C.; Callao, M.P.; Ruisánchez, I. Detection of adulterants in grape nectars by attenuated total reflectance Fourier-transform mid-infrared spectroscopy and multivariate classification. *Food Chem.* **2018**, 266, 254–261, doi: 10.1016/j.foodchem.2018.06.006.

[100] Durante, C.; Bro, R.; Cocchi, M. A classification tool for N-way array based on SIMCA methodology. *Chemom. Intell. Lab. Syst.* **2011**, 106, 73–85, doi: 10.1016/j.chemolab.2010.09.004.

[101] Xu, L.; Goodarzi, M.; Shi, W.; Cai, C-B.; Jiang, J-H. A MATLAB toolbox for class modeling using one-class partial least squares (OCPLS) classifiers. *Chemom. Intell. Lab. Syst.* **2014**, 139, 58-63, doi: 10.1016/j.chemolab.2014.09.005.

[102] Xu, L.; Yan, S-M.; Cai, C-B.; Yu, X-P. One-class partial least squares (OCPLS) classifier. *Chemom. Intell. Lab. Syst.* **2013**, 126, 1-5, doi: 10.1016/j.chemolab.2013.04.008.

[103] Gagneten, M.; Buera, M.P, Rodríguez, S.D. Evaluation of SIMCA and PLS algorithms to detect adulterants in canola oil by FT-IR. *Inst. Food Sci. Technol.* **2021**, 56, 2596-2603, doi: 10.1111/ijfs.14866.

[104] Ellison, S.L.R.; Fearn, T. Characterising the performance of qualitative analytical methods: Statistics and terminology. *TrAC Trends Anal. Chem.* **2005**, 24, 468-476, doi: 10.1016/j.trac.2005.03.007.

[105] López, M.I.; Callao, M.P.; Ruisánchez, I. A tutorial on the validation of qualitative methods: From the univariate to the multivariate approach. *Anal. Chim. Acta* **2015**, 891, 62-72, doi: 10.1016/j.aca.2015.06.032.

[106] Oliveri, P.; Downey, G. Multivariate class modeling for the verification of food-authenticity claims. *TrAC Trends Anal. Chem.* **2012**, 35, 74-86, doi: 10.1016/j.trac.2012.02.005.

[107] Ellison, S.L.R.; Williams, A. Quantifying uncertainty in analytical measurement. Eurachem/CITAC Guide, **2021**.

[108] Lemyre, F.C.; Desharnais, B.; Laquerre, J.; Morel, M.A.; Côté, C.; Mireault, P.; Skinner, C. D. Qualitative threshold method validation and uncertainty evaluation: a theoretical framework and application to a 40 analytes LC-MS/MS method. *Drug Test. Anal.* **2020**, 12, 1287–1297, doi: 10.1002/dta.2867.

[109] Gondim, C.S.; Junqueira, R.G.; de Souza, S.V.C.; Callao, M.P.; Ruisánchez, I. Determining performance parameters in qualitative multivariate methods using probability of detection (POD) curves. Case study: two common milk adulterants. *Talanta* **2017**, 168, 23–30, doi:10.1016/j.talanta.2016.12.065.

[110] Ríos, A.; Barceló, D.; Buydens, L.; Cárdenas, S.; Heydorn, K.; Karlberg, B.; Klemm, K.; Lendl, B.; Milman, B.; Neidhart, B.; Stephany, R.W.; Townshend, A.; Zschunke, A.; Valcárcel, M. Quality assurance of qualitative analysis in the framework of the European project 'MEQUALAN'. *Accred. Qual. Assur.* **2003**, 8, 68-77, doi: 10.1007/s00769-002-0556-x.

[111] Trullols, E.; Ruisánchez, I.; Rius, F.X. Validation of qualitative analytical methods. *TrAC Trends Anal. Chem.* **2004**, 23, 137-145, doi: 10.1016/S0165-9936(04)00201-8.

[112] López, M.I.; Colomer, N.; Ruisánchez, I.; Callao, M.P. Validation of multivariate screening methodology. Case study: Detection of food fraud. *Anal. Chim. Acta* **2014**, 827, 28-33, doi: 10.1016/j.aca.2014.04.019.

[113] Ruisánchez, I.; Jiménez-Carvelo, A.M.; Callao, M.P. ROC curves for the optimization of one-class model parameters. A case study; authenticating extra virgin olive oil from a Catalan protected designation of origin. *Talanta* **2021**, 222, 121564, doi: 10.1016/j.talanta.2020.121564.

[114] Fawcett, T. An introduction to ROC analysis, *Pattern Recogn. Lett.* **2006**, 27, 861–874, doi: 10.1016/j.patrec.2005.10.010.

[115] Borràs, E.; Ferré, J.; Boqué, R.; Mestres, M.; Aceña, L.; Busto, O. Data fusion methodologies for food beverage authentication and quality assessment-A review. *Anal. Chim. Acta* **2015**, 891, 1-14, doi: 10.1016/j.aca.2015.04.042.

[116] Márquez, M.; López, M.I.; Ruisánchez, I.; Callao, M.P. FT-Raman and NIR spectroscopy data fusion strategy for multivariate qualitative analysis food fraud. *Talanta* **2016**, 161, 80-86, doi: 10.1016/j.talanta.2016.08.003.

[117] Ramos, P.M.; Callao, M.P.; Ruisánchez, I. Data fusion in the wavelet domain by means of fuzzy aggregation connectives. *Anal. Chim. Acta* **2007**, 584, 360-369, doi: 10.1016/j.aca.2006.11.051

[118] Di Anibal, C.V.; Ruisánchez, I.; Fernández, M.; Forteza, R.; Cerdà, V.; Callao, M.P. Standardization of UV-visible data in a food adulteration classification problem. *Food Chem.* **2012**, 134, 2326-2331, doi: 10.1016/j.foodchem.2012.03.100.

[119] Feudale, R.N.; Woody, N.A.; Tan, H.; Myles, A.J.; Brown, S.D.; Ferré, J. Transfer of multivariate calibration models: a review. *Chemom. Intell. Lab. Syst.* **2002**, 64, 181-192, doi: 10.1016/S0169-7439(02)00085-0.

[120] Norgaard, L. Direct standardisation in multi wavelength fluorescence spectroscopy. *Chemom. Intell. Lab. Syst.* **1995**, 29, 283–293, doi: 10.1016/0169-7439(95)80103.

[121] Bouveresse, E.; Massart, D. L.; Dardenne, P. Calibration transfer across near-infrared spectrometric instruments using Shenk's algorithm: Effects of different standardization samples. *Anal. Chim. Acta* **1994**, 297, 405–416, doi: 10.1016/0003-2670(94)00237-1.

[122] Walczak, B.; Bouveresse, E.; Massart, D. L. Standardization of near-infrared spectra in the wavelet domain. *Chemom. Intell. Lab. Syst.* **1997**, 36, 41–51, doi: 10.1016/S0169-7439(96)00075-5.

[123] Macho, S.; Rius, A.; Callao, M.P.; Larrechi, M.S. Monitoring ethylene content in heterophasic copolymers by near-infrared spectroscopy: Standardisation of the calibration model. *Anal. Chim. Acta* **2001**, 445, 213–220, doi: 10.1016/S0003-2670(01)01281-8.

# Chapter 3. Results

# Chapter 3

This chapter presents the experimentation carried out and the results obtained in developing this thesis to achieve the objectives proposed in Chapter 1. Figure 3.1. presents, schematically, the content of this chapter. First, the type of fraud being addressed is indicated. Secondly, the type of samples we have worked with is represented. Next, the main objective to be developed is outlined indicating the paper where the results are presented. Finally, the instrumentation and classification techniques used in the experimentation that have produced significant results.

Results are presented in a publication format and divided into two sections; Section 3.1. contains four scientific publications on food adulteration problems, and Section 3.2. contains one scientific publication on food authentication.

The specific objective a) is addressed in all the experimentation carried out since it deals with developing multivariate qualitative analysis using spectroscopy measurement and classification techniques, both for adulteration and authentication problems.

In Section 3.1. the specific objectives b) and c) are developed. b) proposes different tools to obtain semi-quantitative information if different levels of adulteration have been considered, and c) proposes strategies to optimize the model performance parameters by selecting the limit of the class(es).

Specifically, in Paper 1 a strategy was proposed to set a cut-off to obtain semi-quantitative information using PCC curves for olive oil adulteration (Figure 3.1. Paper 1). The second and third publications presented (Figure 3.1. Paper 2, and Paper 3) address the adulteration of cashew nuts. In Paper 2, the optimization of the model performance parameters has been carried out from two independent strategies, by applying ROC curves and high-level data fusion. In this work, only one class limit was considered. In Paper 3, two class limits were defined, so an uncertainty region was set. The last publication (Figure 3.1. Paper 4) is in progress and addresses the

adulteration of honey. It is based on the characterization of the one-class model using PCC curves and defining two-class limits.

In Section 3.2. the specific objectives d) is developed. Strategies to extend the usefulness of a classification model has been proposed to deal with changes in the experimental conditions. This is fulfilled with Paper 5 (Figure 3.1.) in which a multivariate transfer technique is proposed to correct the effect of seasonality variability in olive oil.

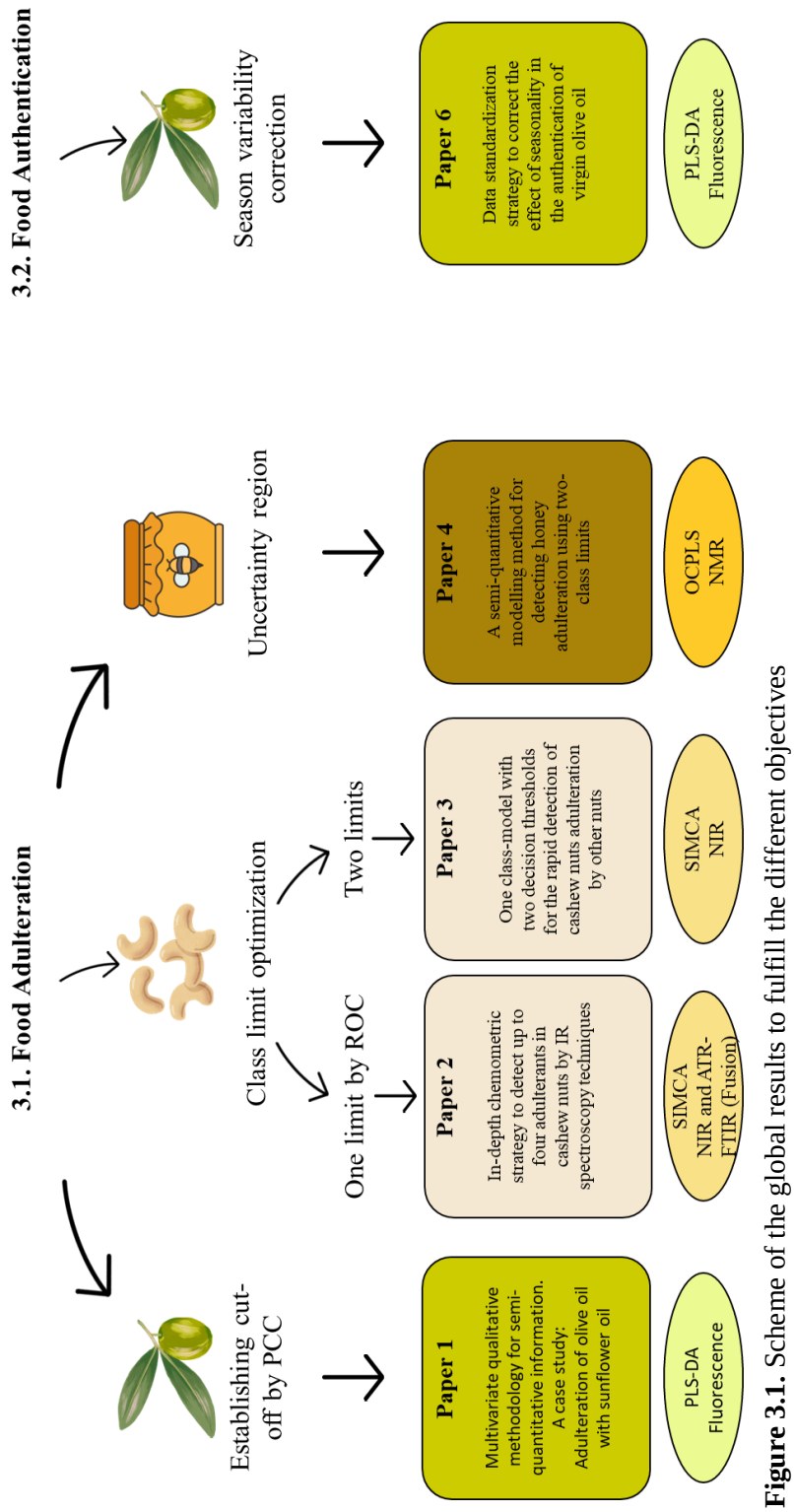



**Figure 3.1.** Scheme of the global results to fulfill the different objectives

# Section 3.1. Food Adulteration

## Section 3.1.

As stated in the Eurachem guide, whenever decisions are based on analytical results, it is important to have some indication of the quality of the results. Qualitative analysis, like quantitative analysis, needs to be demonstrably reliable. Keeping this idea in mind, the objectives of this thesis are related to theoretical and practical problems of uncertainty evaluation in qualitative analysis.

In quantitative analysis, uncertainty is clearly defined (ISO, EURACHEM), and it is represented by a numerical value that indicates the range where the measured value is expected to fall, with a certain level of confidence. In qualitative analysis because of the response obtained is binary (Yes/No) and not numerical, it is not possible to assign a numerical value to the result. So another term is proposed. The uncertainty is described as a region of potential error probabilities, in which the false responses (positive or negative) will be obtained. This region is scarcely implemented in qualitative analysis.

The common thread of this section is the determination of adulteration in foods. As summarized in Figure 3.1, the different works developed have been applied to three types of foods: oil, nuts, and honey. Two aspects have been the focus of the works presented, the consideration of food adulteration from a semi-quantitative point of view and the establishment of the optimal class limit(s).

Nowadays, there are many studies that establish multivariate qualitative methods to determine possible adulteration in food, but very few of them determine parameters related to the concentration of the adulterant. In this thesis, we have implemented PCC curves as a tool to calculate figures of merit of qualitative methods related to concentration (decision limit-CCα, detection capability-CCβ), and uncertainty region-UR). PCC are extensively used in univariate qualitative methods, but not in multivariate.

Regarding the second item, our work has evolved from an initial proposal of establishing a single optimized limit to obtain the best quality parameters (sensitivity and specificity), jointly considered. This study is carried out by implementing ROC curves. Secondly, it is proposed to establish an uncertainty region by defining two class limits: a lower and an upper limit. The uncertainty region is the one that lies between the two limits, the samples that fall in this region will be assigned as inconclusive and therefore subjected to confirmatory analysis.

It is considered that the use of the two limits and consequently the establishment of an uncertainty region, results in having less error in the assignment of unknown samples as either adulterated or non-adulterated. It allows the detection of which samples should be taken for confirmatory analysis. Establishing two class limits is an innovative approach because it is not as commonly used as establishing only one.

## Paper 1

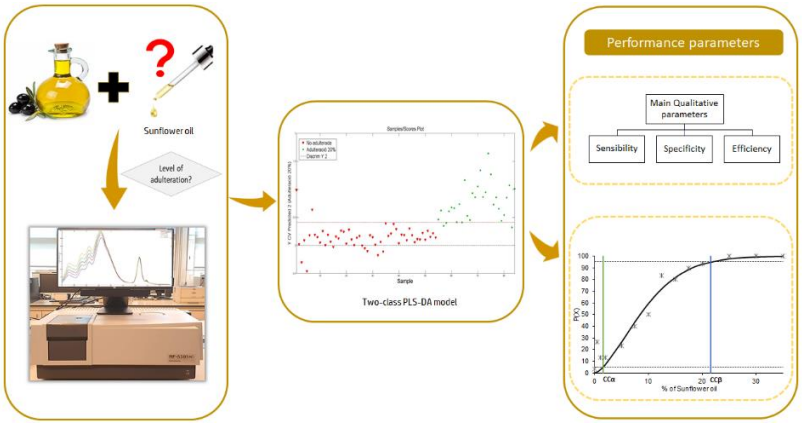
**Multivariate qualitative methodology for semi-quantitative information. A case study: Adulteration of olive oil with sunflower oil**

Itziar Ruisánchez, Glòria Rovira, M. Pilar Callao

*Analytica Chimica Acta,* **2022**, 1206, 339785

https://doi.org/10.1016/j.aca.2022.339785

### Graphical Abstract




**Keywords:** Olive oil adulteration, Multivariate screening, PLS-DA, Semi-quantitative performance parameters, Performance characteristic curve.



# Multivariate qualitative methodology for semi-quantitative information. A case study: Adulteration of olive oil with sunflower oil

Itziar Ruisánchez, Glòria Rovira, M. Pilar Callao*

Chemometrics, Qualimetric and Nanosensors Group, Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo S/n, 43007, Tarragona, Spain.

Analytica Chimica Acta 1026 (2022) 339785

**Abstract**

This paper proposes a strategy to assess the performance of a multivariate screening method for semi-quantitative purposes. The adulteration of olive oil with sunflower oil was considered as a case study using fluorescence spectroscopy and two-class Partial Least Squares Discriminant Analysis (PLS-DA). Building the proper screening methodology based on a two-class multivariate classification model involves setting the cut-off value for the adulterated class (class 2). So, four classification models were established for four levels of adulterant (cut-off). Model validation involved calculating the main quality parameters (sensitivity, specificity, and efficiency) and three additional semi-quantitative parameters (limit of detection, detection capability, and unreliability region).

The probability of successfully recognizing non-adulterated samples as such was set by the main performance parameters of the two-class model. However, the probability of successfully recognizing adulterated samples as such was more accurately extracted from the performance characteristic curves (PCC) curves instead of just from the sensitivity of the adulterated class.

The main performance parameters of the PLS-DA models increased as the cut-off level increased although, after a particular value, the increase was less pronounced. As an example, when the cut-off was changed from 5% to 20%, sensitivity changed from 70 to 93%, specificity changed from 87 to 97%, and efficiency changed from 78 to 95%. The same can be stated for the semi-quantitative parameter's decision limit and detection capability, which moved from 0 to 1.6 and from 17.7 to 21.6 (% of adulterant), respectively.

## 1. Introduction

There are different qualitative approaches that provide a binary response as output (for example, yes/no, accepted/not accepted, etc.). Qualitative methods can be classified in several ways depending on the information sought and the nature of the data used. Related to the information sought, it can be differentiated between non-quantifiable and semi-quantifiable methods. An example is when a qualitative method is developed to check whether the samples fit an overall property (i.e., geographical origin). Qualitative methods that give semi-quantifiable information determine whether a compound is present in a sample above or below a certain threshold or if a sample contains or does not contain it (i.e., samples have been adulterated or not). Related to data used, can be either a specific signal or multiple non-specific variables known as univariate and multivariate qualitative methods, respectively. To develop a multivariate qualitative method, a chemometric approach is required by applying multivariate classification techniques.

Qualitative methods based on multivariate classification have proven to be faster, more economic, and environmentally friendly, resulting in more sustaining methods. Consequentially they are currently in high demand. Examples are numerous in many fields, food science being one of them [1,2]. As in all newly developed methods, one of the steps is its validation. Considerable effort has been made in this matter although there is still a lot to be done [3,4]. Nowadays the quality parameters sensitivity (SEN), specificity (SPC), efficiency (EFF) or accuracy (ACC) and precision (PR) are well established [3,5,6]. Recently, a new quality parameter called occurrence (OCURR) [7] and a work dealing with time stability of the classification models [8] have been reported. These performance parameters can be calculated by all of the classification approaches mentioned above. If the classification method is approached for semi-quantitative purposes, there are other parameters of great interest such as the detection limit ($CC\alpha$),

the detection capability (*CCβ*), and the unreliability region (UR). These parameters have hardly been addressed in the literature [3,9-13] although they are of the utmost importance when the qualitative method is to be used for screening purposes.

All the quality parameters mentioned are calculated from the probabilities that arise from the four well-known binary responses [14]: True positive (TP) and True Negative (TN) when the qualitative method rightly considers a sample to be positive or negative that is indeed positive or negative, respectively; and False positive (FP) and False negative (FN) when the qualitative method wrongly gives a positive output for a sample that is negative or a negative output for a sample that is positive. Depending on the classification technique used, in a multi-class approach, some responses can be inconclusive, that is, a sample is assigned to no class or is assigned to more than one [15-17].

The present study proposes a strategy to assess the performance of a multivariate screening method for semi-quantitative purposes. It includes the calculation of the main quality parameters and the three additional semi-quantitative parameters. As a case study, an adulteration of olive oil with sunflower oil has been considered using fluorescence measurements.

To develop a multivariate screening, a decision must be made on using class modelling or discriminating classification techniques. Discriminant methods require at least two classes and are applied to multi-class problems, while class modelling methods can be developed for just one-class or for more than two classes (multiclass). A review [18] has recently been published that describes and critically compares the main multivariate qualitative strategies. Deciding which one is the most appropriate in each case is not straight forward. Dealing with adulteration problem, some authors [19,20] prefer class modelling methods than discriminant ones considering the difficulties in acquiring a sample set representative of all possible types of adulterations. But, in case the adulterant under study is

known, the use of a two-class discriminant technique has the advantage of making unambiguous assignments [21]. In such cases, in the prediction of a sample, it belongs or not to the adulterated class.

Given this and considering that the adulterant is known (sunflower oil), the proposed methodology consists of building a two-class PLS-DA model. Class 1 consists of the non-adulterated samples and class 2 of the adulterated samples. The key point of the two-class strategy is that the adulteration level used to define the adulteration class must be specified. We will refer to this level as the cut-off level since it is related to the threshold concentration below which the sample is acceptable (compliant) and above which it is not acceptable (non-compliant). Two situations can be addressed, when there is a reference or threshold value (concentration, %, etc.) and the goal is to verify whether the compound exceeds it or not. In the other, there is no reference or threshold value. Therefore, the goal is to determine from what level of adulteration the method is able to determine if the sample is adulterated, with predetermined probabilities of success. In the food area, this cut-off value is not usually known a priori. So, different values have to be tested so that the one that best fits the purpose of the application under consideration can be selected.

The decision will also depend on considerations related to the nature of the problem ahead, whether the compound is a contaminant with an impact on health or the economy, etc.

Once the cut-off has been established, the discriminant model can be developed and validated on the basis of the main quality parameters (sensitivity, specificity, and efficiency). If the main performance parameters are adequate, performance characteristic curves (PCC) will be established to determine the parameters related to the concentration ($CC\alpha$, $CC\beta$, and unreliability region) by analyzing samples with adulteration percentages above and below the cut-off value. PCCs are widely used in the validation of univariate qualitative methods [11,22,23] but not to the same extent as

multivariate qualitative methods. The main drawback of PCC is the need for a large number of experiments for each level of adulterant. Despite this, PCC have proven to be a well-studied and tested tool in practice for estimating the parameters of qualitative methods, although their use for multivariate methods is still scarce. Recently, new ways to estimate the three semi-quantitative parameters have been proposed although, as the authors stated, they have not yet been fully investigated and validated in practice [4].

Nowadays, there are many studies, particularly on food, that establish multivariate qualitative methods to determine possible adulteration in food but very few of them determine parameters related to the concentration of the adulterant. This paper is a contribution in this sense, and it also demonstrates the usefulness of taking advantage of the information from the PCC curves in the unreliability range. Additionally, it shows, from a practical point of view, the benefits of allowing the analyst to set the cut-off value in each case.

## 2. Samples, instrumentation, and software

A total of 60 virgin and extra virgin Arbequina oil samples from Catalonia were supplied by the Catalan Government's Official Tasting Panel of Virgin Olive Oils of Catalonia, which confirms the status of the oils.

A further 24 samples were obtained by randomly mixing 5 original oils. The 84 unadulterated Arbequina samples were randomly divided into training (54 samples) and test set (30 samples). To obtain the adulterated samples, the 30 test set samples were adulterated with sunflower oil, each at 13 different levels (between 0.5% and 35%).

Fluorescence spectra were obtained using a Shimadzu RF-5301PC ((Shimadzu Corporation, Kyoto, Japan). The slit width was 5 nm for emission spectra, which were collected between 360 and 800 nm using an excitation wavelength of 350 nm. The integration time was 0.1s, and the increasing wavelength while scanning the spectrum was 10 nm. No sample pretreatment was applied.

The recorded data was treated by using MATLAB software, version 8.0.0.783-R2012b (Natick, MA, USA) and PLS Toolbox 7.0.2 (Eigenvector Research Inc., Wenatchee, WA, USA).

## 3. Theoretical background

*Principal component analysis* (PCA) is one of the techniques most commonly used to compress, describe, and interpret large sets of multidimensional data. It should always be used for a preliminary exploratory analysis of every dataset, even when the final aim is to perform a supervised classification for predictive purposes.

*Partial Least Square Discriminant Analysis* (PLS-DA) is a PLS regression technique adapted to a classification technique. It requires two matrices, one with independent variables (matrix **X**), which in our case are the fluorescence spectra, and the other with dependent variables (matrix **Y**), which in our case is a binary code (0 and 1) where 1 indicates sample membership and 0 does not.

There is an extensive bibliography that describes the theoretical and practical aspects of both PCA and PLS-DA. Without being exhaustive in the references, two recent reviews can be consulted, which in turn provide multiple references [18,24].

### *3.1. Performance parameters*

To validate the classification models, the main performance parameters are typically sensitivity, specificity, and efficiency, terms which have been extensively defined. Other parameters that are directly related to those mentioned (such as Youden's index, and likelihood ratio …) are also reported. For further details on figures of merit and validation strategies, the interested reader is referred to [3,6,21,25].

Three additional performance parameters for semi-quantifiable qualitative methods have also been defined: decision limit, detection capability, and unreliability region [3,12,13]. The three parameters can be obtained from

the performance characteristic curve (PCC). The PCC is a plot of the probability of having a positive classification output, P(X), versus the corresponding concentration of the analyte [23,26-28]. Note that in the literature, PCC curves have also been referred to by other names.

The probability of positive classification (P(X)) is the probability of the classification PLS-DA method giving a positive result. This probability is dependent upon the amount of analyte present in the sample. In qualitative methods such as in this case, the probability of a positive classification should be zero or close to zero when the analyte is not present and would increase to the final value of 100% as the adulterant concentration or mass increases. In this study, the probability of positive classification was experimentally obtained for the different adulterant levels (13 levels, between 0.5% and 35%). The PCC curve is obtained by fitting the P(X) values to a sigmoid function by minimizing the root mean square of the residuals (RMSE) following Eq. (1) [3,28].

$$P(X) = 1 - e^{-(x/a)^{\wedge}b} \qquad \text{Equation 1}$$

Where $x$ is the adulterant concentration, and $a$ and $b$ are the regression coefficients fitted to minimize the RMSE.

The goodness parameters of the PCC non-linear fit are evaluated by means of the root mean square of errors (RMSE) (Eq. (2)), and the adjusted coefficient of determination ($R^2_{adj}$) which measures how accurately the calculated curve fits the original data (Eq. (3)) [12]:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-m}} \qquad \text{Equation 2}$$

$$R^2_{adj} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \cdot (n-1)}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \cdot (n-m)} \qquad \text{Equation 3}$$

Where $n$, is the number of adulteration levels (in that work n=14, from 0% to 35%); $i$ refers to a specific level; $y_i$, is the value of P(X) for a given level; $\bar{y}_i$, is the P(X) predicted value for a given level; $\bar{y}$, the mean of values

of the fourteen P(X) values and *m*, the number of equation parameters (in that case *m*=2).

From the PCC curve, the three semi-quantitative performance parameters are calculated [3,12,13]:

*Decision limit (CCα)* is the minimum concentration of a compound that will give a positive output when indeed it is positive with a particular probability (usually P(X)=5%). Below this limit, there is a 95% probability or higher of obtaining a negative output N(X)=100-P(X) (that is to say, that the sample is not adulterated or adulterated at lower levels. *CCα* is obtained from the intersection of the PCC curve with the horizontal black dashed lines placed at α=5%.

*Detection capability (CCβ)* is the concentration of a compound in a sample that can be reliably detected and/or identified with statistical certainty (usually (P(X)=95%). Above this limit, there is a 95% probability or higher of obtaining a positive output for a sample adulterated at any level above it. *CCβ* is obtained from the intersection of the PCC curve with the horizontal black dashed lines placed at 1-β=95%.

*Unreliability region (UR)* is the range of concentration between the two limits where there is certain probability (P(X) between 5% and 95%) of false negative errors. It means that there is a certain probability that a sample is not adulterated when indeed it is.

## 4. Results and discussion

The flowchart of the screening strategy proposed is described in Fig. 1. Dealing with an adulteration problem, after selecting a multivariate (rather than a univariate) classification approach, the first step is to look at the spectra and the data structure (Fig. 1, preliminary studies).

Analytica Chimica Acta 1026 (2022) 339785

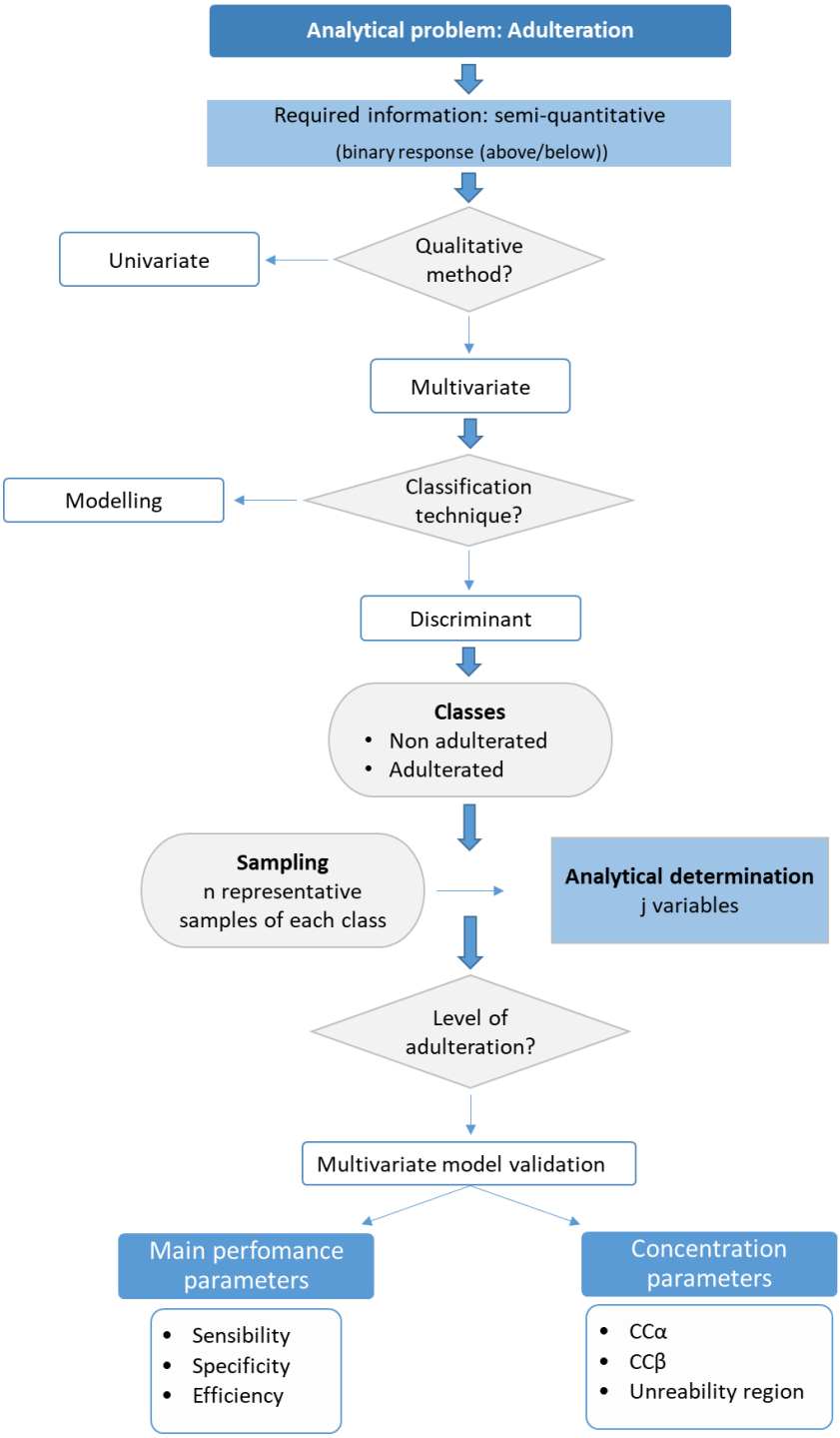



**Fig 1.** Flow chart showing the steps to develop a qualitative multivariate method with semi-quantitative purpose.

Fig. 2 shows the mean spectrum of the EVOOs non-adulterated and adulterated at the different sunflower oil percentages. It can be seen that the fluorescence increases as the percentage of adulterant does, mainly in the bands from 360 to 480 nm. Additionally, shifts are observed in the maximum at wavelengths around 520 nm (towards shorter wavelengths) and around 380, 440, and 475 nm (towards longer wavelengths). The appearance of a band can be perceived at 410 nm, which becomes more evident as the concentration of adulterant increases.

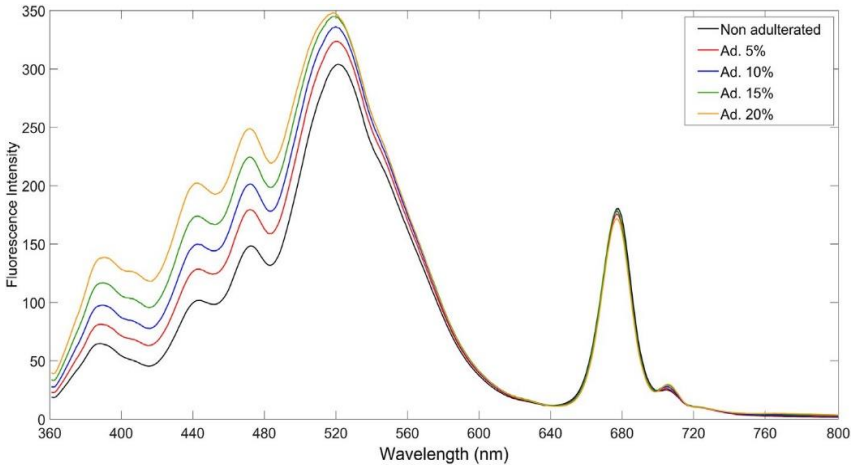



**Fig 2.** Mean spectra of non-adulterated olive oil samples and samples adulterated at several percentage of sunflower oil.

PCA was used to analyze the data structure and the presence of outliers. Fig. 3 shows the scores for the first two PCs, for all the samples analyzed no matter the % of adulteration. With a variance of 97%, it can be seen that there are no clear groups, but samples show a clear trend: the greater the percentage of sunflower oil, the higher the score is on PC1.

The next step (flowchart Fig. 1) is to decide on the type of classification technique to use, modelling or discriminant. In an adulteration problem in which the possible adulterant is known and will be studied (sunflower oil in this case study), a discriminant technique is the most appropriate since the class membership is recognized 30 as non-adulterated assigned unambiguously.

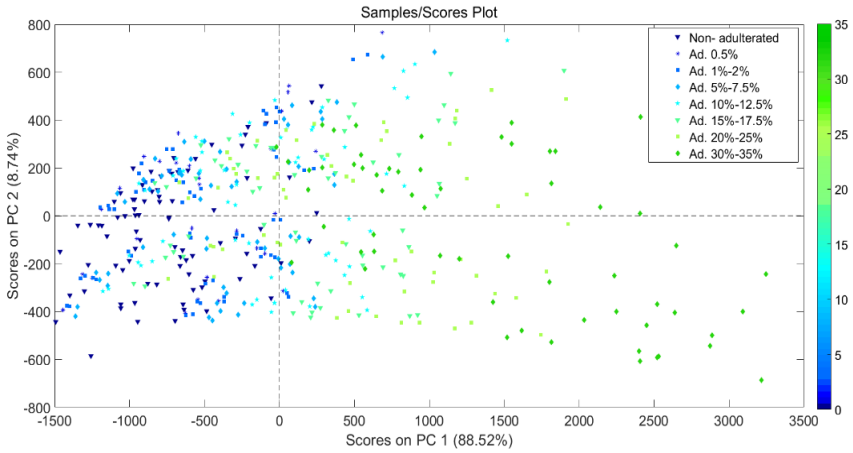



**Fig 3.** PCA score plot for the non-adulterated olive oil samples and samples adulterated at several percentages of sunflower oil. Color bar representing the adulteration levels.

In this paper, a PLS-DA method was implemented, and two classes were defined: class 1 for non-adulterated samples and class 2 for samples adulterated with sunflower oil. To establish the two-class model, the level of adulteration for class 2 (adulterated class) has to be defined. This level is called the cut-off level so the samples in class 2 have a percentage of adulterant equal to the cut-off value, but the non-adulterated class (class 1) is not changed.

The cut-off value that fits for the purpose of the present problem is unknown. Therefore, to define class 2 (adulterated class), various percentages of adulterant are studied. In this study, four cut-off values were considered: 5, 10, 15, and 20% of adulterant. So, four PLS-DA two-class models were established. In all four models, class 1 is the same and was built with 54 non-adulterated training samples.

The PLS-DA models differ in class 2, Model 1 was built with 30 samples adulterated at 5% sunflower. Likewise, Class 2 in Model 2, Model 3, and Model 4 were built with 30 samples adulterated at 10%, 15%, and 20% of sunflower, respectively.

Before the model development, the spectra were mean-centered. A total of 4 LVs were kept in all four models, with the explained variance being around 98% in all four models. Then, the main quality parameters were calculated, all of which were based on the well-known TP, TN, TP, and TN rates calculated from the model output. Table 1 shows the main performance parameters of the adulterated class (class 2) obtained with the four PLS-DA models. Since it is a two-class model, the main quality performance values for the non-adulterated class (class 1) are the same but swapping the value for sensitivity instead of specificity (and vice versa). As expected, all performance parameters increase as the cut-off level increases since it discriminates between non-adulterated samples (class 1) and samples adulterated at higher percentages (class 2). This increase is less and less important and when the level of adulteration increases from 15% (Model 3) to 20% (Model 4) the value of the performance parameters was the same.

**Table 1.** Main performance parameters of the adulterated class (class 2), for different adulterant levels, Model 1, Model 2, Model 3, and Model 4, class 2 built with samples adulterated at 5, 10, 15, and 20% of sunflower oil, respectively.

| Parameter (%) | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| **Sensitivity** | 70.0 | 76.7 | 93.3 | 93.3 |
| **Specificity** | 86.7 | 93.3 | 96.7 | 96.7 |
| **Efficiency** | 78.3 | 85.0 | 95.0 | 95.0 |

A closer look at the main quality performance values of one of the models (for instance, Model 1, cut-off set at 5%, Table 1) indicates that of every 100 non-adulterated samples, the model would properly recognize 87 as non-adulterated and wrongly recognize 13 as adulterated (specificity=87%).

Similarly, of 100 samples adulterated at or higher than 5% of sunflower oil, the model would properly recognize 70 as adulterated and wrongly recognize 30 as non-adulterated. Similar conclusions can be drawn from the values obtained with the other three models.

From the main performance parameters (Table 1), useful information can be obtained for a fixed adulterant percentage (the cut-off values). To obtain information on levels of adulteration above and below the cut-off value, the performance characteristic curves (PCC) were adjusted. To build it, the developed two-class PLS-DA models (Models 1 to 4) were used to predict all the samples no matter the real level of adulterant they contained. Specifically, Fig. 4 shows the PCC curves for the four PLS-DA models. Each PCC curve was obtained from the PLS-DA model predicted values, expressed as P(X), of the whole data set. For instance, the first PCC curve (Fig. 4a) was obtained from the P(X) values obtained from Model 1 predictions of the 30 non-adulterated test samples (0% of adulterant) and all adulterated samples (30 samples at each of the 13 adulteration levels, from 0.5% to 35%), including the 30 samples of class 2. Likewise, PCC curves for Models 2, 3, and 4, from their model predictions of the whole data set (Fig. 4b, c, d).

**Table 2.** Fit parameters of the PCC curves and semi-quantitative performance parameters, for the four PLS-DA models. Model 1, Model 2, Model 3, and Model 4, class 1 built with 54 non-adulterated samples, and class 2 built with samples adulterated at 5, 10, 15, and 20% of sunflower oil, respectively.

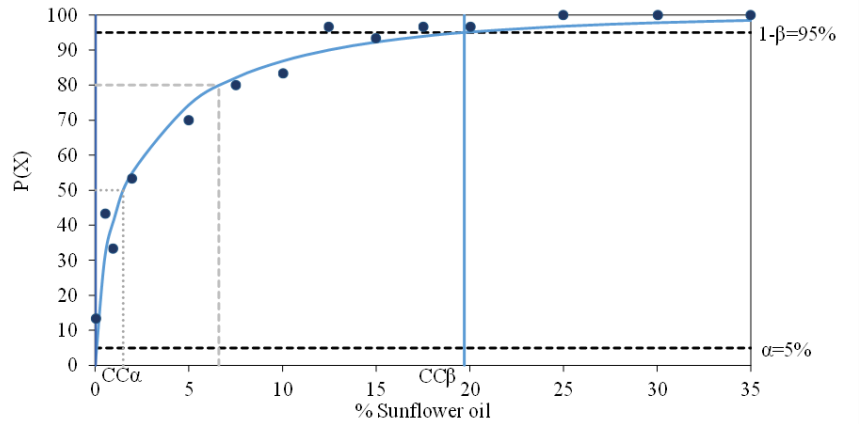
| Parameter (%) | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| $\mathbf{R^2_{adj}}$ | 0.9468 | 0.9635 | 93.3 | 93.3 |
| **RMSE** | 0.0663 | 0.0669 | 96.7 | 96.7 |
| **Equation** | $1-e^{-\left(\frac{x}{2.93}\right)^{0.57}}$ | $1-e^{-\left(\frac{x}{6.27}\right)^{0.93}}$ | $1-e^{-\left(\frac{x}{8.23}\right)^{1.34}}$ | $1-e^{-\left(\frac{x}{10.72}\right)^{1.57}}$ |
| $\boldsymbol{CC\alpha}$ | 0.0 | 0.3 | 0.9 | 1.6 |
| $\boldsymbol{CC\beta}$ | 17.7 | 20.3 | 18.7 | 21.6 |
| **UR** | 0.0-17.7 | 0.3-20.3 | 0.9-18.7 | 1.6-21-6 |

Table 2 contains the corresponding fit parameters as well as the performance parameters for each PCC. In general, no big differences were found in the *CCα*, *CCβ* (intersection of the horizontal black dashed lines with the PCC

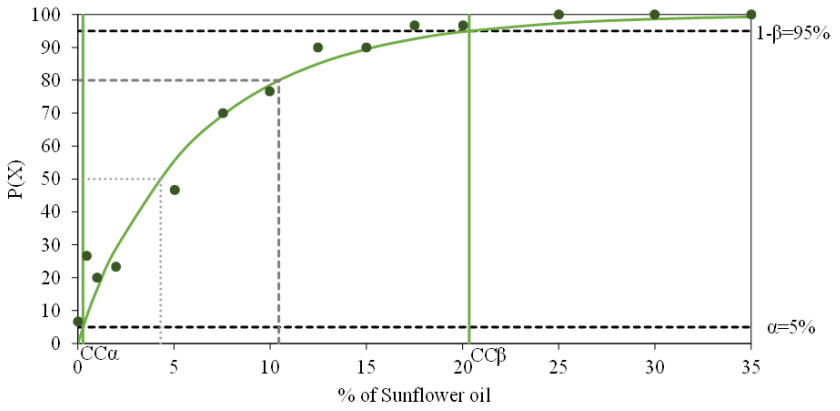
curve, Fig. 4a, b, c, d), and unreliability regions obtained for the four PLS-DA models.

Nevertheless, the *CCα* and *CCβ* values increased slightly with the cut-off value and the values were largest for the model built with an adulterant level of 20%.

**a) Model 1**




**b) Model 2**



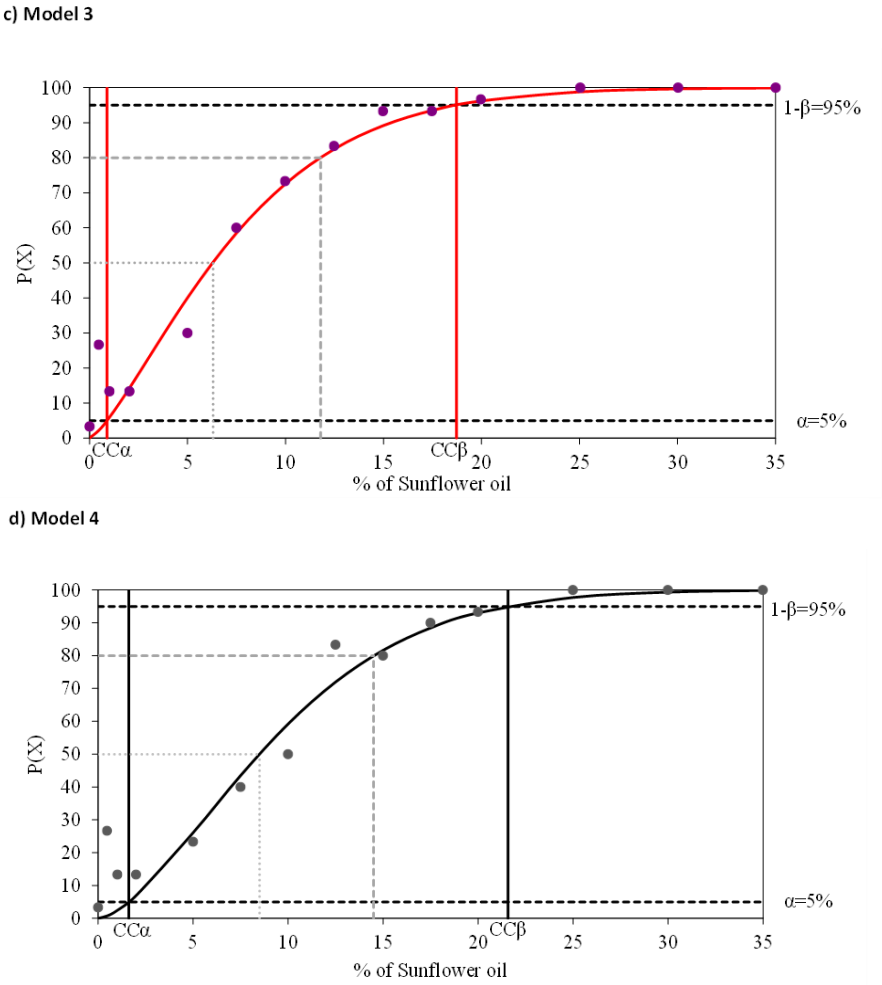



**Fig 4.** Performance characteristic curves (PCC) and semi-quantitative parameters for obtained from the four PLS-DA Models: a) PCC for model 1 (cut-off value at 5%), b) Model 2 (cut-off value at 10%), c) Model 3 (cut-off value at 15%) and d) Model 4 (cut-off value at 20%). CCα and CCβ values calculated from the intersection of the horizontal black dashed lines with the PCC curves, vertical grey lines indicate the adulteration level from which a sample will be detected as adulterated with 50% or higher probability (dotted line) and 80% or higher probability (dashed line).

It can be observed that as the cut-off value increases, the shape of the curve changes from exponential to sigmoidal (Fig. 4). Therefore, the slope of the PCC curve decreases as the cut-off value increases, the slope being highest at a cut-off value of 5%. This has implications for the performance parameters calculated, but more significantly for the probabilities of properly assigning an adulterated sample. When the shape of the PCC is exponential, the adulterant level corresponding to the cut-off value is placed in the first part of the curve within the unreliability region.

However, when the shape is sigmoidal, the cut-off value is placed close to the $CC\beta$ value. This has a considerable impact on the probability of properly classifying a sample whose level of adulteration is unknown. For instance, considering the fit equations of the PCC curves (Table 2) obtained with a cut-off value of 5% (Fig. 4a), an adulterated sample will have a probability higher than 50% (Fig. 4a, grey dotted line) of being properly classified as adulterated (class 2 assignation) when it is indeed adulterated with 1.6% of sunflower oil or higher (indicated with a dotted line). Similar conclusions can be drawn from the models obtained with a cut-off value of 10% (Fig 4b), 15% (Fig 4c), and 20% (Fig. 4d) when predicting samples adulterated at or higher than 4.6%, 6.3% and 8.5%, respectively. As expected, for a fixed probability of error, a low cut-off value enables adulterated samples to be discriminated from non-adulterated at lower adulteration levels.

The slope and shape of the PCC curves show that as the level of adulteration increases the percentage of correct classification also increases. As an example, for the model built at a cut-off value of 5% (Fig. 4a), the percentage of correct recognition of an adulterated sample is higher than 70% for an adulteration level above 4.1%; higher than 80% (Fig. 4a, grey dashed line) for an adulteration level above 6.8% and higher than 90% for an adulteration level above 12.5%. As the cut-off value increases, the adulterant level that will more probably be correctly recognized also increases.

For instance, for the model built at a cut-off value of 20% (Fig. 4d), the percentage of correct recognition of an adulterated sample is higher than 70% for an adulteration level above 12%; higher than 80% (grey dashed line) for an adulteration level above 15% and higher than 90% for an adulteration level above 19%. The PCC curve provides information on the behavior of the adulterated samples as a function of their level of adulteration. The behavior of the non-adulterated samples is indicated by the specificity of the class 2 model and is also PCC=100-P(X) for an adulterant level equal to zero.

Thus, to build a classification model with the appropriate cut-off value, a compromise must be reached between the probability of recognizing a non-adulterated sample as well as an adulterated sample. The first probability can be set by the main performance parameters of the two-class PLS-DA model. However, the second probability can be more accurately extracted from the PCC curves rather than from the sensitivity of the adulterated class. This approach involves checking different cut-off values with extra experimental data.

The proposed strategy makes it possible to adapt the screening methodology to the laboratory's requirements. To implement it, two types of adulterants should be considered: adulterants with an impact on health and adulterants for economic reasons. It is to be expected that the majority of cases would be of the second type.

In the case of adulteration for economic reasons, the choice of the cut-off point is determined by the value at which the adulteration is profitable This value and some close to it would be suitable cut-off values. What's more, several situations can be differentiated, for example, if adulteration is not expected, a laboratory might be interested in a screening approach that ensures that the nonadulterated samples are properly recognized. If that is the case, a high cut-off value is the best option.

Another situation could be when the laboratory submits all samples classified as adulterated to a confirmatory analysis. In this case, it is more important that the adulterated samples be recognized as such, than wrongly classifying a non-adulterated sample as adulterated. Therefore, the best option will be to set a low-cut-off value.

In the case of a prohibited contaminant with an impact on health, the best option will be to set the cut-off value defining the adulterated class at low adulteration levels. As a result, the PCC will have an exponential shape and several non-adulterated samples will be submitted to a confirmatory analysis but only a small number of samples with an adulteration level below $CC\alpha$ will be erroneously assigned as non-adulterated.

If the case study aims to discriminate samples that contain a compound below a specific threshold, therefore, the cut-off level should be placed below the threshold. For instance, if we want to discriminate samples adulterated below 20% of sunflower oil, the best option is to set the cut-off value at 10% (Fig. 4b).

The errors that laboratories accept are different in every case. An error in non-adulterated samples means unnecessary confirmatory analysis and an error in adulterated samples means that there will be fraud with economic benefits or a health impact. In any case, the proposed strategy makes it possible to understand the risks and consequences that are being assumed and adapt them to the case under study by selecting the cut-off value.

Once the methodology has been established based on the above considerations, its validation can be carried out, first, from samples of known composition. Subsequently, with data obtained throughout its application, it can be readjusted through updating strategies.

## 5. Conclusions

A strategy using fluorescence measurements and a two-class (adulterated and non-adulterated) PLS-DA classification model has been developed to

determine the adulteration of olive oil with sunflower oil. Dealing with a two-class approach, it is a key point to have representative samples of the non-adulterated class and the adulterated class. Additionally, the adulterant concentration in the samples of the adulterated class (cut-off) should be properly defined.

Even more, this work provides evidence that the choice of the level of adulteration is a relevant factor to consider in the design of the adulterated class. Four PLS-DA models have been established at four cut-off levels. The performance of each model was evaluated by calculating the main quality parameters (sensitivity, specificity, and efficiency) and the three additional semi-quantitative parameters (decision limit, detection capability, and unreliability region) from the performance characteristic curve.

All main performance parameters increase as the cut-off level increases since it discriminates between non-adulterated (class 1) and adulterated samples at higher percentages (class 2), although above a certain percentage, the increase is irrelevant. This trend is also observed in the semi-quantitative parameters.

This paper contributes to the implementation of the PCC curves as a tool to calculate figures of merit of qualitative methods. PCC are extensively used in univariate qualitative methods, but not in multivariate as in this case. Performance characteristic curves (PCC) have been shown as a successful tool to provide information on the probability of properly assigning adulterated samples. Even more, it has been demonstrated that it allows adapting the screening method to the laboratory requirements. For instance, setting a high cut-off value if adulteration is not expected since one is more interested in recognizing the non-adulterated samples as such. Or setting a low cut-off value when it is more important to recognize adulterated samples as such. For example, when all samples classified as adulterated will be submitted to a confirmatory analysis.

## CRediT authorship contribution statement

**Itziar Ruisánchez:** Writing - original draft, In charge of the whole process that has given rise to this work, Design of the experimental part, Chemometric treatment, Discussion of results, and Drafting of the manuscript. **Glòria Rovira:** Participated in the selection of samples, Sample measures, and the chemometric treatment of the data. **M. Pilar Callao:** Writing - original draft, In charge of the whole process that has given rise to this work, Design of the experimental part, Chemometric treatment, Discussion of results, and Drafting of the manuscript.

## Declaration of competing interest

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

## Acknowledgments

This study was supported by the research program “Program of Research Activity (2020PMF-PIPF) at the Rovira i Virgili University, Tarragona, Spain. Special mention to M. Angels Calvo, head of the Official Tasting Panel of Virgin Olive Oils of the Catalonia Government.

## References

[1] O.Y. Rodionova, A.L. Pomerantsev, Chemometric tools for food fraud detection: the role of target class in non-targeted analysis, Food Chem. 317 (2020), 126448, https://doi.org/10.1016/j.foodchem.2020.126448.

[2] E.J. Rifna, R. Pandiselvam, A. Kothakota, K.V.S. Rao, M. Dwivedi, M. Kumar, R. Thirumdas, S.V. Ramesh, Advanced process analytical tools for identification of adulterants in edible oils-A review, Food Chem. 369 (2022), 130898, https://doi.org/10.1016/j.foodchem.2021.130898.

[3] M.I. López, M.P. Callao, I. Ruisánchez, A tutorial on the validation of qualitative methods: from the univariate to the multivariate approach, Anal. Chim. Acta 891 (2015) 62-72, https://doi.org/10.1016/j.aca.2015.06.032.

[4] A.L. Pomerantsev, O.Y. Rodionova, New trends in qualitative analysis: performance, optimization, and validation of multi-class and soft models, TrAC Trends Anal. Chem. 143 (2021), 116372, https://doi.org/10.1016/j.trac.2021.116372.

[5] D. Ballabio, F. Grisoni, R. Todeschini, Multivariate comparison of classification performance measures, Chemometr. Intell. Lab. Syst. 174 (2018) 33-44, https://doi.org/10.1016/j.chemolab.2017.12.004.

[6] L. Cuadros-Rodríguez, E. Pérez-Castaño, C. Ruiz-Samblá, Quality performance metrics in multivariate classification methods for qualitative analysis, TrAC Trends Anal. Chem. 80 (2016) 612-624, https://doi.org/10.1016/j.foodcont.2018.04.057.

[7] A.M. Jiménez-Carvelo, L. Cuadros-Rodríguez, The occurrence: a meaningful parameter to be considered in the validation of multivariate classification- based screening methods e application for authenticating virgin olive oil, Talanta 208 (2020), 120467, https://doi.org/10.1016/j.talanta.2019.120467.

[8] D.N. Vera, I. Ruisánchez, M.P. Callao, Establishing time stability for multivariate qualitative methods. Case study: Sudan I and IV adulteration in food spices, Food Control 92 (2018) 341-347, https://doi.org/10.1016/j.foodcont.2018.04.057.

[9] F.C. Lemyre, B. Desharnais, J. Laquerre, M.A. Morel, C. Côté, P. Mireault, C.D. Skinner, Qualitative threshold method validation and uncertainty evaluation: a theoretical framework and application to a 40 analytes LC-MS/MS method, Drug Test. Anal. 12 (2020) 1287-1297, https://doi.org/10.1002/dta.2867.

[10] C. de S. Gondim, O.A.M. Coelho, R.L. Alvarenga, R.G. Junqueira, S.V.C. de Souza, An appropriate and systematized procedure for validating qualitative methods: its application in the detection of sulfonamide residues in raw milk, Anal. Chim. Acta 830 (2014) 11-22, https://doi.org/10.1016/j.aca.2014.04.050.

[11] A.I. Corps, N. Rodriguez, F.J. Guzman, R.C. Rodriguez, A. Rios, Screening-confirmation strategy for nanomaterials involving spectroscopic analytical techniques and its application to the control of silver nanoparticles in pastry samples, Spectrochim. Acta Mol. Biomol. Spectrosc. 246 (2021), 119015, https://doi.org/10.1016/j.saa.2020.119015.

[12] C. de S. Gondim, R.G. Junqueira, S.V.C. de Souza, M.P. Callao, I. Ruisánchez, Determining performance parameters in qualitative multivariate methods using probability of detection (POD) curves. Case study: two common milk adulterants, Talanta 168 (2017) 23-30, https://doi.org/10.1016/j.talanta.2016.12.065.

[13] E. Trullols, I. Ruisánchez, F.X. Rius, J. Huguet, Validation of qualitative methods of analysis that use control samples, TrAC Trends Anal. Chem. 24 (2005) 516-524, https://doi.org/10.1016/j.trac.2005.04.001.

[14] S.L.R. Ellison, T. Fearn, Characterising the performance of qualitative analytical methods: statistics and terminology, TrAC Trends Anal. Chem. 24 (2005) 468-476, https://doi.org/10.1016/j.trac.2005.03.007.

[15] C. de S. Gondim, R.G. Junqueira, S.V.C. de Souza, I. Ruisánchez, M.P. Callao, Detection of several common adulterants in raw milk by MID-infrared spectroscopy and one-class and multi-class multivariate strategies, Food Chem. 230 (2017) 68-75, https://doi.org/10.1016/j.foodchem.2017.03.022.

[16] C.S.W. Miaw, M.M. Sena, S.V.C. de Souza, I. Ruisánchez, M.P. Callao, Variable selection for multivariate classification aiming to detect individual adulterants and their blends in grape nectars, Talanta 190 (2018) 55-61, https://doi.org/10.1016/j.talanta.2018.07.078.

[17] C.S.W. Miaw, M.M. Sena, S.V.C. de Souza, M.P. Callao, I. Ruisánchez, Detection of adulterants in grape nectars by attenuated total reflectance Fourier-transform mid-infrared spectroscopy and multivariate classification, Food Chem. 266 (2018) 254-261, https://doi.org/10.1016/j.foodchem.2018.06.006.

[18] P. Oliveri, C. Malegori, E. Mustorgi, M. Casale, Qualitative pattern recognition in chemistry: theoretical background and practical guidelines, Microchem. J. 162 (2021), 105725, https://doi.org/10.1016/j.microc.2020.105725.

[19] O.Y. Rodionova, P. Oliveri, A.L. Pomerantsev, Rigorous and compliant approaches to one-class classification, Chemometr. Intell. Lab. Syst. 159 (2016) 89-96, https://doi.org/10.1016/j.chemolab.2016.10.002.

[20] P. Oliveri, Class-modelling in food analytical chemistry: development, sampling, optimisation, and validation issues-A tutorial, Anal. Chim. Acta 982 (2017) 9-19, https://doi.org/10.1016/j.aca.2017.05.013.

[21] M.P. Callao, I. Ruisánchez, An overview of multivariate qualitative methods for food fraud detection, Food Control 86 (2018) 283-293, https://doi.org/10.1016/j.foodcont.2017.11.034.

[22] A.I.C. Ricardo, S.A. García, F.J.G. Bernardo, A. Ríos, R.C.R. Martín-Doimeadios, Rapid assessment of silver nanoparticle migration from food containers into food simulants using a qualitative method, Food Chem. 361 (2021), 130091, https://doi.org/10.1016/j.foodchem.2021.130091.

[23] R. Macarthur, C. von Holst, A protocol for the validation of qualitative methods of detection, Anal. Methods 4 (2012) 2744-2754, https://doi.org/10.1039/C2AY05719K.

[24] A. Biancolillo, F. Marini, C. Ruckebusch, R. Vitale, Chemometric strategies for spectroscopy-based food authentication, Appl. Sci. 10 (2020) 6544, https://doi.org/10.3390/app10186544.

[25] P. Oliveri, C. Malegori, M. Casale, Chemometrics and statistics: multivariate classification techniques, Encycl. Anal. Sci. (2019) 481-486, https://doi.org/10.1016/B978-0-12-409547-2.14239-8, 3d Edition.

[26] A. Ríos, D. Barcelo, L. Buydens, S. Cardenas, K. Heydorn, B. Karlberg, et al., Quality assurance of qualitative analysis in the framework of the European project 'MEQUALAN', Accred Qual. Assur. 8 (2003) 68-77, https://doi.org/10.1007/s00769-002-0556-x.

[27] R. Song, P.C. Schlecht, K. Ashley, Field screening test methods: performance criteria and performance characteristics, J. Hazard Mater. 83 (2001) 29-39, https://doi.org/10.1016/S0304-3894(00)00325-3.

[28] P. Wehling, R.A. LaBudde, S.L. Brunelle, M.T. Nelson, Probability of detection (POD) as a statistical model for the validation of qualitative methods, J. AOAC Int. 94 (2011) 335-347, https://doi.org/10.1093/jaoac/94.1.335.

## Paper 2

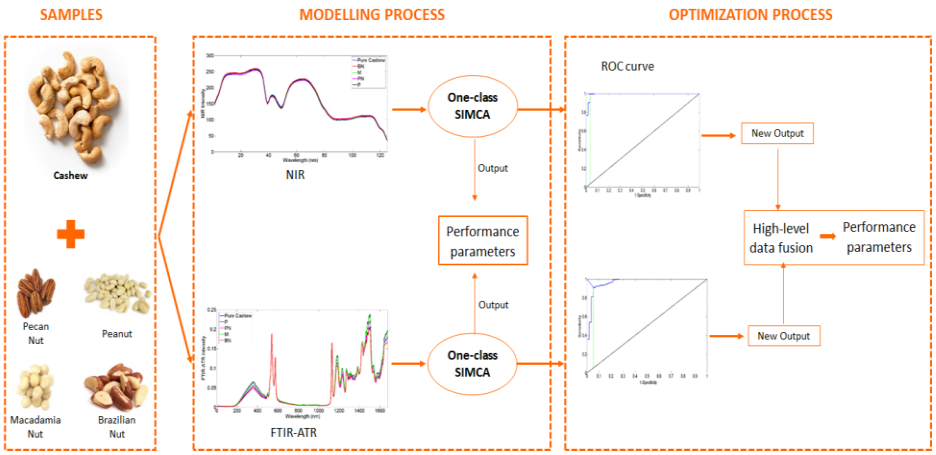
**In depth chemometric strategy to detect up to four adulterants in cashew nuts by IR spectroscopic techniques**

Glòria Rovira, Carolina Sheng Whei Miaw, Mário Lúcio Campos Martins, Marcelo Martins Sena, Scheilla Vitorino Carvalho de Souza, Itziar Ruisánchez, M. Pilar Callao

*Microchemical Journal,* **2022**, 181, 107816

https://doi.org/10.1016/j.microc.2022.107816

### Graphical Abstract




**Keywords:** Untargeted chemometrics, one-class SIMCA, NIR, ATR-FTIR, ROC curve, High-level data fusion.

# In-depth chemometric strategy to detect up to four adulterants in cashew nuts by IR spectroscopic techniques

Glòria Rovira[a], Carolina Sheng Whei Miaw[b], Mário Lúcio Campos Martins[b], Marcelo Martins Sena[c,d], Scheilla Vitorino Carvalho de Souza[b], Itziar Ruisánchez[a,*], M. Pilar Callao[a]

[a] Chemometrics, Qualimetric, and Nanosensors Group, Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo S/n, 43007, Tarragona, Spain.

[b] Department of Food Science, Faculty of Pharmacy (FAFAR), Federal University of Minas Gerais (UFMG), Av. Antônio Carlos, 6627, Campus da UFMG, Pampulha, 31270-010, Belo Horizonte, MG, Brazil.

[c] Chemistry Department, Institute of Exact Sciences (ICEX), Federal University of Minas Gerais (UFMG), Av. Antônio Carlos, 6627, Campus da UFMG, Pampulha, 31270-010, Belo Horizonte, MG, Brazil.

[d] Instituto Nacional de Ciência e Tecnologia em Bioanalítica (INCT-Bio), Campinas, SP 13083-970, Brazil.

Microchemical Journal 181 (2022) 107816

**Abstract**

An untargeted strategy was developed to determine cashew nuts adulteration with Brazilian nuts, pecan nuts, macadamia nuts, and peanuts. A one-class SIMCA model was developed for the cashew non-adulterated samples by means of two spectroscopic techniques: Near-Infrared (NIR) and Attenuated Total Reflectance-Fourier Transform Infrared (ATR-FTIR). Receiver operating characteristic (ROC) curves have been proven to be useful to optimize class limits, both for the NIR and ATR-FTIR models, allowing to balancing of the values of the performance parameters. An increase in the sensitivity of the training and test set has been obtained from 79% with NIR and 85% with ATR-FTIR to 93% in both cases. As a result, the specificity has slightly decreased from 100% with NIR and a range of 90–98% with ATR-FTIR to a range of 82–98% and 84–96%, respectively. The implementation of high-level data fusion to the classification results obtained from NIR and ATR-FTIR, considering the limit value optimized by ROC curves, allowed the improvement of the performance parameters of the untargeted strategy. Obtaining sensitivity values for the training and test set of 100% and 93%, respectively. Specificity values of 100% were obtained for the detection of Brazilian nuts, macadamia nuts, and peanuts, while for pecans it was 98%.

## 1. Introduction

The consumption of nuts and peanuts is widespread throughout the world not only due to their high organoleptic value but also due to their beneficial effects on human health [1]. This food class includes a wide range of products such as almonds, Brazil nuts, cashew nuts, hazelnuts, macadamia nuts, peanuts, and pecans, among others. Some of them present medium or high risk of food fraud due to adulteration, usually by adding cheaper and lower quality products [2]. Particularly, Brazil is among the major world producers of cashew nuts, which are also one of the preferred nuts of Brazilian consumers due to their pleasant taste. Thus, they are potential targets of fraud. The market prices of these food products vary substantially. Currently, in Brazil, cashew nuts are sold for R$51–65, while pecan, Brazil nuts, and peanuts are sold for R$25–50, R $37–45, and R$4–10; macadamia nuts, which are primarily imported, present prices comparable to cashew nuts [3].

Detecting food adulteration is important for economic reasons but it is especially important when the non-declared substance involves a health risk. Food adulteration is always a concern due to the high complexity of food and the difficulty in detecting the presence of an adulterant in an easy and rapid way.

In such a scenario, it is of great interest the development of screening methods since it generally implies low time analysis, permitting a high throughput of samples at low cost, thus making them suitable for routine analysis. Today, the application of multivariate instrumental techniques together with chemometrics is a consolidated strategy in the field of food fraud detection. Multivariate supervised classification models might provide a binary response of the type there is/there is no adulteration. Some examples of these strategies for detecting nut fraud due to the addition of adulterants have been recently referenced [4], such as adulteration of almonds [5–8], pistachio [9], and hazelnut [10–12].

There is a wide range of both instrumental techniques and classification models that can be implemented to solve a given adulteration problem. Methods developed for nuts adulteration have included a study on hazelnut paste adulterated with almonds and chickpea four, which applied near-infrared spectroscopy (NIRS) combined with target and non-target modelling [10]; another article has detected hazelnut paste adulterated with almonds by applying data fusion of NIR and Raman spectra combined with class modelling [12]. A very recent article has applied and compared several multivariate classification models to short-wave infrared hyperspectral images aiming to detect contaminants in edible pistachio nuts, such as inedible pistachio nuts, pistachio shells, husks, twigs, and stones [13]. Among the analytical techniques most used to detect nut adulteration, it can be cited inductively coupled plasma optical emission spectroscopy (ICP-OES) [5], gas chromatography with flame ionization detector (GC-FID) [6], high performance liquid chromatography (HPLC) with fluorescence and UV detection [7,8], FT-Raman hyperspectral imaging [9], NIR [10,12], mid-infrared [11] and FT-Raman [12] spectroscopies. Regarding chemometric models, principal component analysis (PCA) has been used mainly as a data exploration tool, while the main supervised classification methods have been soft independent modelling of class analogies (SIMCA) [10–12], partial least squares discriminant analysis (PLS-DA) [7,8], support vector machine (SVM) [5,6] and linear discriminant analysis (LDA) [6]. Among chemometricians, the most used are SIMCA as a class modelling model and PLS-DA as a discriminant model.

A key step in the development of a screening method based on multivariate supervised classification is the choice of a suitable chemometric strategy. This implies a proper data pre-processing, model parameter optimization, and the possible exploitation of the synergies between different data sources using data fusion, if more than one instrumental technique was used. Signal pre-processing is applied to correct/remove the contribution of undesired

phenomena ranging from stochastic measurement noise to various sources of systematic errors. Different possibilities have been critically discussed [14,15].

Regarding the classification technique, a choice must be performed between discriminating and class modelling. While discriminant methods establish a delimiter between two (or more) classes and split the hyperspace in a number of regions corresponding to the number of classes, class modelling methods model each class individually, irrespectively of the others. In view of this, some authors [16,17] have recently criticized the predominant use of discriminant methods in the literature, arguing that class modelling methods are more robust than discriminant ones considering the practical limitations in acquiring a sample set representative of all possible types of adulterations [18]. Class modelling methods offer the possibility of building just one class (untargeted modelling or one-class approach), being this strategy considered one of the most appropriate for food fraud detection [10,16,19]. If that the case, the untargeted/one-class approach builds a class from the non-adulterated samples and detects which new samples resemble them no matter which adulterant under study is present. In fact, this approach is not a novelty, but its application has recently been incremented. The model can be optimized by selecting suitable class limits that define whether a sample is considered unadulterated or not (fits the model). Studies of this type have been recently published [11,16,18–21].

In the case of the present study, it is important to note that the use of one-class modelling ensures the representativeness of the model, avoiding the need to expand the comprehensiveness of the samples by incorporating other types of nuts or adulterants as non-authentic classes. This is in contrast with the use of discriminant models, such as PLS-DA. When using one-class modelling, the decision regarding compliance is not at all influenced by out-of-class samples. Therefore, this avoids any need to collect a representative

data set that includes all the possible sources of out-of-specification variations [10,22].

In this paper, an untargeted strategy was developed to determine the possible adulteration of cashew nuts with other types of nuts (Brazilian nuts, pecan nuts, macadamia nuts, and peanuts). A one-class SIMCA model was developed for cashew non-adulterated/authentic samples by means of two spectroscopic techniques, near-infrared (NIR) and attenuated total reflection-Fourier transform infrared (ATR-FTIR). Aiming to optimize the performance parameters of the method, different tools have been sequentially applied. First, different signal processing methods were studied. Second, model optimal limits have been determined by developing receiver operating characteristic (ROC) curves. Finally, high-level data fusion has been applied and compared with the result obtained from the models established with the two individual techniques (NIR and ATR-FTIR). In spite of the recent publication of many articles developing multivariate qualitative methods to detect adulterations, particularly in food analysis, very few of them have exhaustively explored all the chemometric possibilities available today.

## 2. Materials and methods

### *2.1. Samples*

Commercial batches of each nut (Cashew nut, Brazilian nut, Macadamia nut, Peanut, and Pecan) were acquired from certified producers. They were crushed in a sample processor (Arno Magiclean WWBC Blender), homogenized, sieved to size 40 mesh using calibrated tamis, packed in polyethylene packaging, sealed, and kept at room temperature (25 ± 3ºC) until preparation of the formulated batches. This processing aimed to simulate ground nuts products. Unadulterated/authentic samples of cashew nuts were composed of seven formulated batches prepared in eight variations, giving a total of 56 samples. Adulterated samples were prepared from batches of the corresponding adulterant nut (Brazilian nut, Macadamia

nut, Peanut, and Pecan) plus different amounts of the seven formulated batches of the unadulterated samples (8 levels of adulteration: 10.0; 5.0; 2.5; 1.3; 0.6; 0.3; 0.2 and 0.1% w/ w).

The total of adulterated samples was 224 (4 × 56 samples). The experimental design used to formulate the samples were previously described [26]. The 56 samples of non-adulterated cashew nuts were systematically divided into training (42 samples) and test (14 samples) sets by employing the Kennard-Stone algorithm [27].

*2.2. NIR spectroscopy*

NIR analysis was conducted using a portable MicroNIR® 1700 equipment from Viavi Solution (San Jose, CA, USA) in the diffuse reflectance mode. Its dispersive element is a linear variable filter (LVF), and its detector is a 128-pixel InGaAs photodiode array. This detector is a variable-band semiconductor with excellent optical properties. Spectralon was used as a reflectance standard reference. A sample portion, previously homogenized, was placed on a Petri dish (3.5 cm in diameter × 1.2 cm in height) until complete covered. Then, the plate was placed on the MicroNIR®, a reading was performed with 20 scans at a resolution of 6.25 nm and spectra were recorded in the wavelength range from 908 to 1676 nm (11013–5967 $cm^{-1}$). Readings were performed randomly, under repeatability conditions. The reflectance values were converted into pseudo-absorbance, log(1/R) prior to data processing.

*2.3. ATR-FTIR spectroscopy*

ATR-FTIR analysis was carried out using a Perkin Elmer Frontier spectrophotometer (Waltham, MA, USA) equipped with a deuterated triglycine sulphate detector and a single-reflection diamond crystal ATR accessory.

A sample portion, previously homogenized, was placed on the ATR crystal. Sample was pressed at a constant pressure level with a metallic tip

accessory. The spectrum of each sample was recorded with 16 scans at a resolution of 4 $cm^{-1}$, from 4000 to 650 $cm^{-1}$. Readings were performed randomly, under repeatability conditions. The reflectance values were converted into pseudo-absorbance, log(1/R), prior to data processing.

### *2.4. Software*

Recorded data were processed and models were built by using MATLAB software, version 8.0.0.783 – R2012b (Natick, MA, USA) and PLS_Toolbox version 7.0.2 (Eigenvector Research Inc., Wenatchee, WA, USA).

### *2.5. Simca*

*SIMCA* is a multivariate supervised class-modelling technique that models each class independently from all the others [28]. Assignation of unknown samples has evolved from the first criteria proposed by Wold et al. in 1976, but whatever the criteria, it is always related to calculating a distance to the model [28]. One of the SIMCA modifications implies defining the limits of the two scalar statistics, Hotelling $T^2$ and Q residues (Hotelling $T^2_{lim}$ and $Q_{lim}$) at a specific significance level (α), normally set at 0.05. Once defined, there are several criteria to assign or classify a sample in a certain class. One criterion is that a sample should have values of both statistic parameters lower than the two statistic limits to be considered as belonging to the class model.

Another criterion of sample assignation is based on calculating the distance of a sample from the class. The distance of a sample *i* from the class *j* ($d_{ij}$) is a combination of its reduced statistic parameters expressed as in the following equation (Eq. (1)).

$$d_{ij} = \sqrt{\left(Q_{r,i}\right)^2 + \left(T^2_{r,i}\right)^2} \qquad \text{Equation 1}$$

where "*r*" stands for the ratio between the statistics of sample "*i*" ($T_i^2$ and $Q_i$) and the corresponding class frontiers ($T^2_{lim}$ and $Q_{lim}$).

Once the distance value is calculated, a sample could be assigned to a class model when its distance value is lower than 1 [29,30], $\sqrt{2}$ [28,31], or an optimized distance class limit calculated by means of the ROC curves [19].

In our study, the first criterion, a distance limit of 1.0 for the target class, was initially adopted, and then the model's limit was optimized using more robust criteria based on ROC curves.

*2.6. Receiver operating curves (ROC)*

ROC curve represents sensitivity versus 1-specificity for a considered parameter (score) used as a criterion to classify. Therefore, it allows visualizing, organizing, and selecting classifiers (scores) on the basis of their performance [17,32,33]. As has been previously stated, ROC curves can be implemented to optimize class limit values of a class.

*2.7. High-level data fusion*

High level data fusion combines the assignation results obtained from the classification models of each individual data source. In this work, the fusion has been performed following the fuzzy set theory by choosing as operators minimum, maximum, average, and product values. The final decision (*ensemble decision*) is obtained by the majority vote provided by all the fuzzy operators [12,34,35].

## 3. Results and discussion

Fig. 1a and Fig. 2a show the average original spectra for each class, the non-adulterated cashew nut samples and the adulterated ones with Brazilian nuts (BN), pecan nuts (PN), macadamia nuts (M), and peanuts (P), recorded with the two instrumental techniques employed, NIR and ATR-FTIR, respectively.

By observing Fig. 1a, the largest NIR bands are present approximately between 8550 and 7690 $cm^{-1}$ and 7140–6670 $cm^{-1}$. The first band can be assigned to the second overtone of the C–H stretching, while the second band to the first overtone of the O–H stretching.

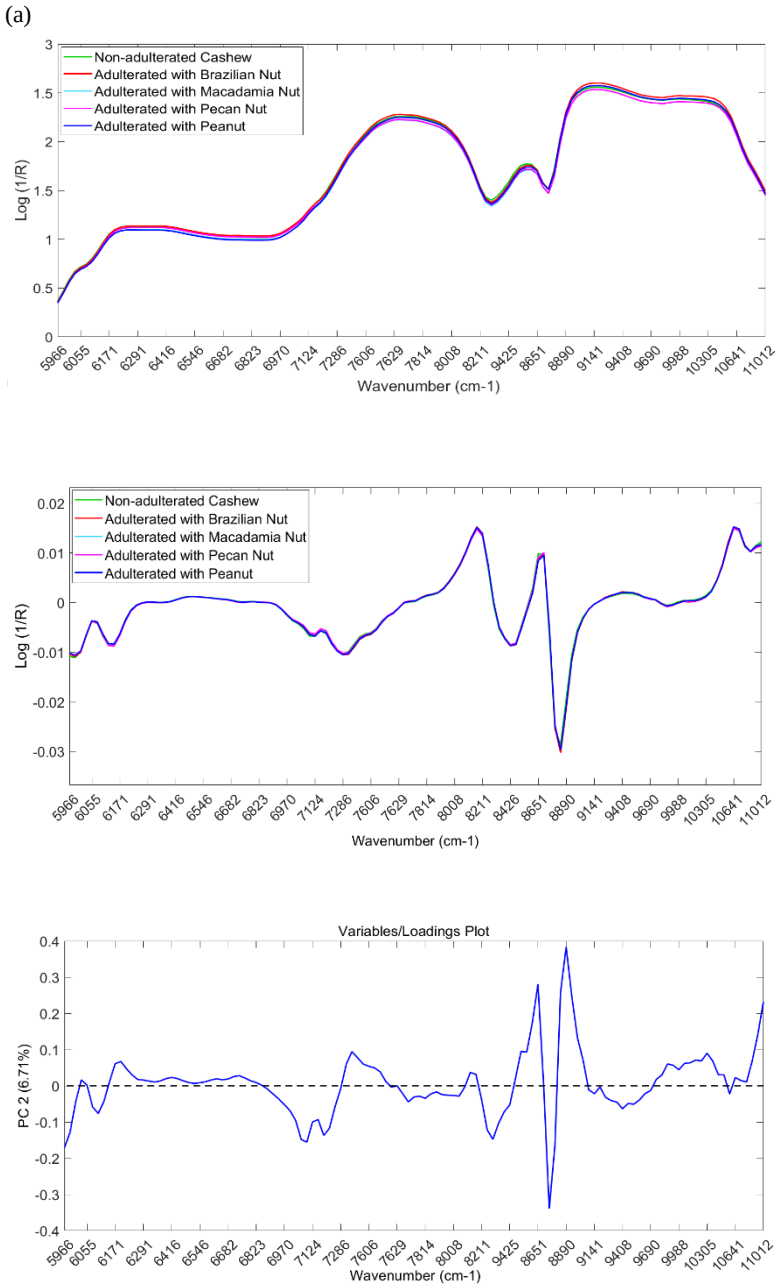



**Fig. 1.** NIR spectra of non-adulterated and adulterated samples. (a) the average original spectra, (b) the average pretreated spectra and (c) PCA loadings values of the second PC obtained by NIR data. Color code: green for non-adulterated cashew nuts, red for Brazilian nuts, pink for pecan nuts, light blue for macadamia nuts and dark blue for peanuts.

Particularly, the spectral band centered around 8330 $cm^{-1}$ has been reported as a discriminant of nuts in relation to other food materials (wheat, milk, and cocoa) [36]. The smaller band between 7245 and 7090 $cm^{-1}$ can be assigned to the combination band of C–H vibrations [37].

By observing Fig. 2a, the most important absorption regions were observed in the fingerprint region (1750–1050 $cm^{-1}$) assigned to carbonyl ester stretching, -C–N amide II and III stretching, and -C–H symmetric stretching vibration modes; around 2850 $cm^{-1}$, related to C–H asymmetric and symmetric stretching vibrations of long-chain fatty acids; and 3500–3000 $cm^{-1}$, related to axial bending of OH and NH bonds. The peak around 1750 $cm^{-1}$ is assigned to the carbonyl group of fatty acid esters in fats, while absorption around 1560 $cm^{-1}$ is associated with C–N stretching and N–H bending modes in proteins. The region between 1300 and 1100 $cm^{-1}$ presents C–H stretching vibrations of carbohydrates [25]. Thus, the discrimination between different nuts provided by NIR and FTIR spectra in combination with chemometrics might be related to their different contents of components, such as proteins, lipids, and carbohydrates.

Before chemometric modelling, some pre-processing was necessary aiming to eliminate non-linear baseline deviations caused by multiplicative scatter and to improve the signal-to-noise ratio. After some trials (supervised models), the first derivative followed by mean centering was applied to NIR spectra. For ATR-FTIR spectra, smoothing with the Savitsky-Golay algorithm with a window width of 5, followed by multiplicative scatter correction (MSC) and generalized least squares weighting (GLSW) with an alpha of 0.01 were applied [12,24,38,39].

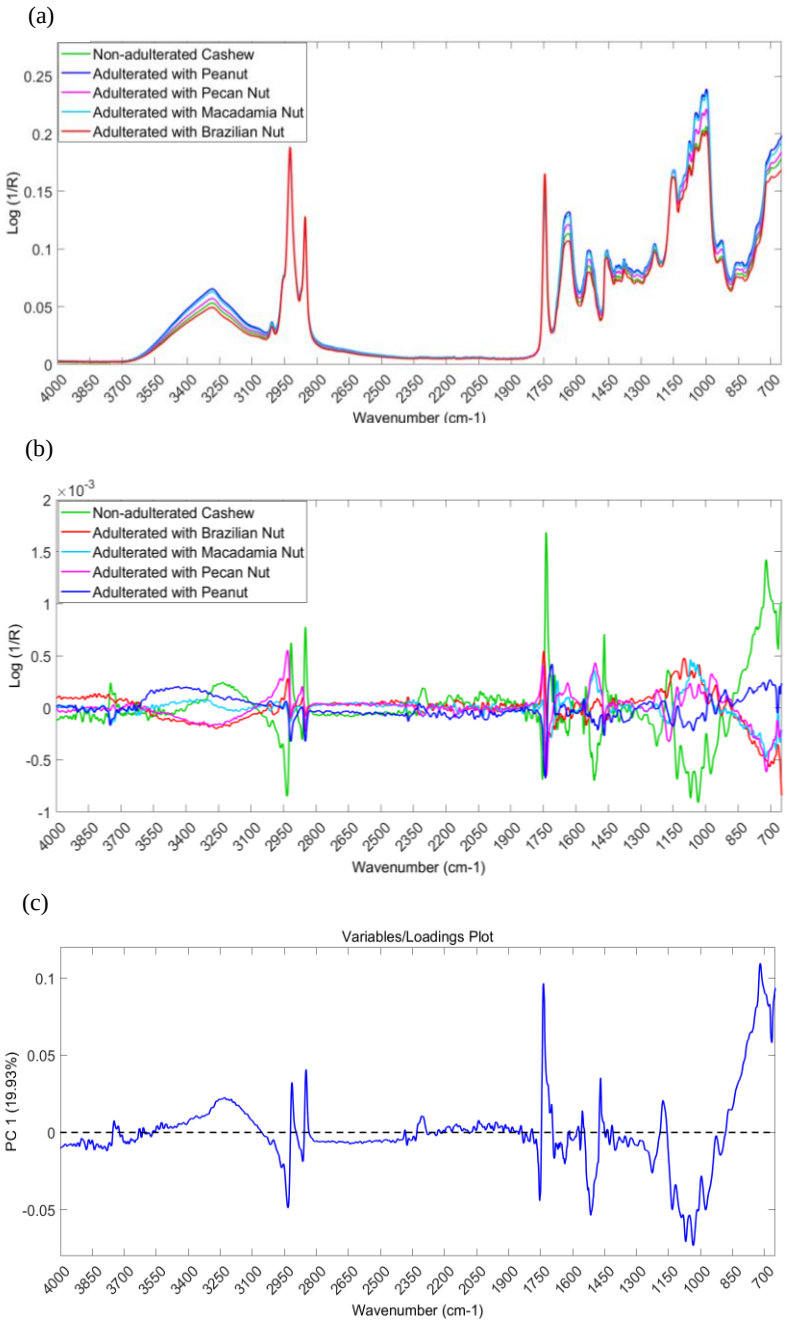



**Fig. 2.** ATR-FTIR spectra of non-adulterated samples. (a) the average original spectra, (b) the average pre-treated spectra and (c) PCA loadings values of the first PC obtained by ATR-FTIR data. Color code: green for non-adulterated cashew nuts, red for Brazilian nuts, pink for pecan nuts, light blue for macadamia nuts and dark blue for peanuts.

It is always advisable to apply PCA previously to developing the supervised classification. Thus, PCA was applied to each pre-treated spectroscopic dataset, aiming at observing any possible discriminating trend between cashew nuts authentication and adulteration. Fig. 3 shows PC1 versus PC2 score plots for each dataset. For NIR spectra (Fig. 3a), the first two PCs accounted for 88.1% of the total variance. PC2 (6.7%) clearly showed a discriminating trend with most of the non-adulterated cashew nuts presenting positive scores, in contrast with adulterated samples, whose scores were predominantly negative. As expected, the comparison between the average pre-treated spectra (Fig. 1b) and the PCA loadings values of the second PC (Fig. 1c) shows shape similarities. More specifically around 6100 $cm^{-1}$, 7200 $cm^{-1}$, and 8600 $cm^{-1}$, which in a certain way indicate their relevance in the discrimination of non-adulterant related to adulterate samples.

For ATR-FTIR spectra (Fig. 3b), the variance was more partitioned between several PCs, and the first two accounted for only 27.8% of the total variance. In this case, PC1 (19.9%) clearly discriminated non-adulterated samples in its positive part. In contrast, almost all the samples adulterated with Brazilian, pecan, and macadamia nuts showed negative scores on PC1, while samples adulterated with peanuts presented scores in an intermediate region. This intermediate behavior of samples adulterated with peanuts was already observed along PC2 in the NIR PCA model (Fig. 3a). When performing the ATR-FTIR spectra pre-treatment (Fig. 2b), the differences between the average signal of the non-adulterated samples with respect to the average signals of the adulterated samples were magnified. Likewise, it can be seen that the loadings of the first PC (Fig. 2c) present shape similarities with the pre-treated average spectrum of the non-adulterated samples (green line, Fig. 2b).

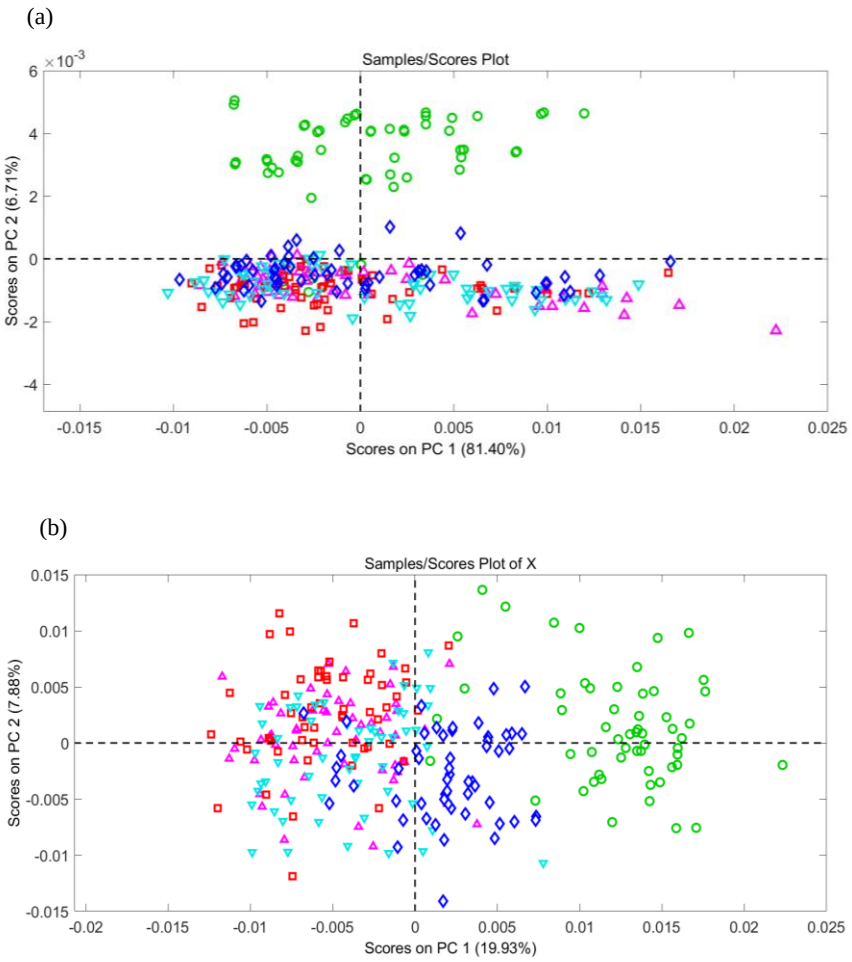



**Fig. 3.** Score plot of PC1 vs PC2 for (a) NIR data and (b) ATR-FTIR data. Color code: green circles for non-adulterated cashew nuts, red squares for Brazilian nuts, pink triangles for pecan nuts, light blue triangles for macadamia nuts and dark blue diamonds for peanuts.

For both instrumental techniques, no differences were observed in PCA score plots related to the adulteration levels in any of the adulterants. Thus, these two PCA score plots suggested that it is possible to split variances related to cashew nuts authentication and adulteration by utilizing one-class modelling.

In the sequence, SIMCA models were built individually for each technique. These models were obtained using 42 training samples from non-adulterated cashew nuts.

For the validation/test set, 14 non-adulterated samples were used together with all samples containing the four adulterants (BN, PN, M, and P). Thus, two one-class classification models were constructed, one with NIR spectra and another with ATR-FTIR spectra. Leave-one-out cross-validation on the training samples has been used to decide the number of retained PCs for each model, based on the lowest cross-validation classification error (CVCE). For NIR SIMCA model, 5 PC and were chosen accounting for a 98.3% of the spectral variance. For ATR-FTIR SIMCA model, 10 PCs were selected representing a cumulative spectral variance of 91.8%.

Sample assignation to authentic class was performed using the distance value according to Eq. (1). Initially, the criterion adopted to assign samples to the target class was a distance value lower than 1.0. Performance parameters for these models are shown in Table 1, which include efficiency as a global parameter calculated as the ratio between the number of true assignments (TP + TN) and the total number of samples. One-class SIMCA results obtained for the NIR model indicated a training sensitivity (rate of true positives) of 95%, which means that the model properly assigned the own samples used to build it. While the test set sensitivity was lower, around 79%. A specificity (rate of true negatives) of 100% was achieved for all adulterants, meaning that the model properly recognized them as adulterated. The results obtained for ATR-FTIR model showed lower performance than NIR model, with the sensitivity of the training and test sets of 83 and 86%, respectively. The specificity values for this model were slightly lower than 100% (91–98%).

Both individual models showed an ability of detecting adulterated samples higher than the rate of recognition of non-adulterated samples. This behavior suggests that a change in the class limit could lead to higher sensitivity values or a more suitable balance between both parameters (sensitivity and specificity).

**Table 1.** Performance parameters of one-class SIMCA models for NIR and ATR-FTIR data. NA: non-adulterated cashew nuts; BN: Brazilian nuts; PN: Pecan nuts; M: Macadamia nuts; P: Peanuts and BN/PN/M/P: all adulterants.

| | | NA | BN | PN | M | P | BN/PN/M/P |
|---|---|---|---|---|---|---|---|
| **NIR** | **Sensitivity training** | 95.24% | | | | | |
| | **Sensitivity test** | 78.57% | | | | | |
| | **Specificity** | | 100% | 100% | 100% | 100% | 100% |
| | **Efficiency** | | 95.71% | 95.71% | 95.71% | 95.71% | 98.74% |
| **ATR-FTIR** | **Sensitivity training** | 83.33% | | | | | |
| | **Sensitivity test** | 85.71% | | | | | |
| | **Specificity** | | 98.21% | 98.21% | 91.07% | 92.86% | 95.09% |
| | **Efficiency** | | 95.71% | 95.71% | 90.00% | 91.43% | 94.54% |

In order to optimize distance class limits for each SIMCA model, ROC curves were constructed. Fig. 4 shows ROC curves obtained for one-class classification models built with NIR (Fig. 4a) and ATR-FTIR (Fig. 4b) spectra. ROC curves were estimated using the distance defined in Eq. (1) as the basis (score) for calculating the performance parameters.

Specifically, the distances of the test samples (14 non-adulterated samples, and 56 samples for the presence of each adulterant) were used. The optimal distance is the one closest to the point (0, 1), which corresponds to both sensitivity and specificity equal to 100%. In Fig. 4, optimal distances are marked with blue points, corresponding to 1.44 and 1.11, for the NIR and ATR-FTIR models, respectively. These distances were therefore the class limits, which means that samples with distances lower or equal than 1.44 in the NIR model and 1.11 in the ATR-FTIR model will be considered as belonging to the target class.

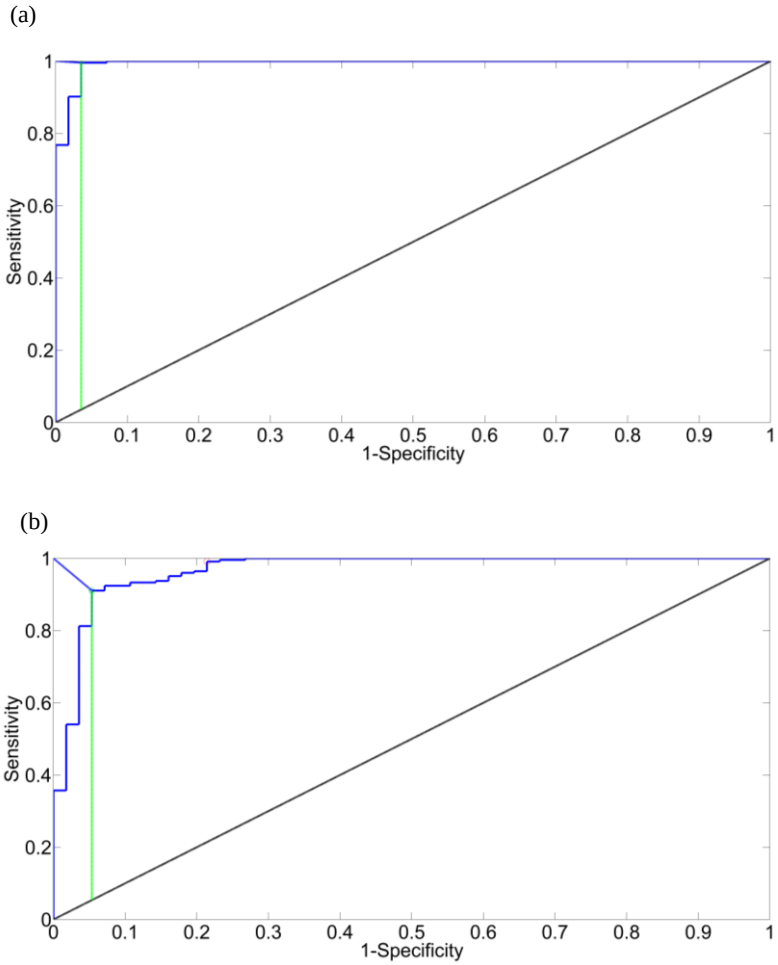



**Fig. 4.** Receiver operating characteristic (ROC) curves estimated for one-class SIMCA models built with (a) NIR data set and (b) ATR-FTIR data set.

Reciprocally, samples with larger distances will be considered as not belonging to the target class. Figures of merit obtained by applying the optimal distances estimated based on ROC curves are shown in Table 2. For the NIR model, sensitivities of both training and test sets were improved. The most remarkable improvement was observed for the test set, from 79% to 93%. Specificity and efficiency reasonably decreased, regardless of the adulterant, presenting values between 82% and 98%.

**Table 2.** Performance parameters of one-class SIMCA models optimized with optimal distances based on ROC curves for NIR and ATR-FTIR data. NA: non-adulterated cashew nuts; BN: Brazilian nuts; PN: Pecan nuts; M: Macadamia nuts: P: Peanuts and BN/PN/M/P: all adulterants.

| | | NA | BN | PN | M | P | BN/PN/M/P |
|---|---|---|---|---|---|---|---|
| **NIR** | **Sensitivity training** | 100% | | | | | |
| | **Sensitivity test** | 92.86% | | | | | |
| | **Specificity** | | 85.71% | 82.14% | 94.64% | 98.21% | 90.18% |
| | **Efficiency** | | 87.14% | 84.29% | 94.29% | 97.14% | 90.34% |
| **ATR-FTIR** | **Sensitivity training** | 92.86% | | | | | |
| | **Sensitivity test** | 92.86% | | | | | |
| | **Specificity** | | 96.43% | 96.43% | 91.07% | 83.93% | 91.96% |
| | **Efficiency** | | 95.71% | 95.71% | 91.43% | 85.71% | 92.02% |

For the ATR-FTIR model, sensitivities of the training and the test were also improved, increasing from 83 to 86% to 93%. Following a trend similar to the NIR model, specificity slightly decreased for all adulterants except for peanuts, which showed a greater decrease, from 93% to 84%. Similarly, the efficiency did not change or slightly increase for all adulterants except for Peanut.

It was observed that implementing class limits based on ROC curves balanced sensitivity and specificity values of SIMCA models. In this study, choosing minimum distances using ROC curves helped to maximize the sensitivity, which is equivalent to increasing the ability of the model to recognize target samples (non-adulterated/authentic). In contrast, specificity decreased, while efficiency remained the same. This statement can be clearly realized by evaluating the total specificity value, that is, the specificity regardless of the type of adulterant used (from 100% to 90% and from 95% to 92%, for NIR and ATR-FTIR models, respectively).

**Table 3**. Performance parameters of one-class SIMCA models optimized with optimal distances based on ROC curves for NIR and ATR-FTIR data. NA: non-adulterated cashew nuts; BN: Brazilian nuts; PN: Pecan nuts; M: Macadamia nuts: P: Peanuts and BN/PN/M/P: all adulterants.

| | | NA | BN | PN | M | P | BN/PN/M/P |
|---|---|---|---|---|---|---|---|
| Optimized | **Sensitivity training** | 100% | | | | | |
| | **Sensitivity test** | 92.86% | | | | | |
| | **Specificity** | | 100% | 98.21% | 100% | 100% | 99.55% |
| | **Efficiency** | | 98.57% | 97.14% | 98.57% | 98.57% | 99.16% |

In the sequence, high-level data fusion was implemented aiming to improve the results obtained with SIMCA models built with individual techniques. In order to build a high-level fusion model, samples misclassified by the NIR model but correctly classified by the ATR-FTIR model, or vice-versa, were selected. In total, 45 samples were chosen as susceptible to apply the fuzzy operators. Before applying them, distance values obtained from one-class SIMCA models of each instrumental technique were normalized to 1.0. In such a way, the contributions of both models were balanced.

Figures of merit obtained for the high-level data fusion model (Table 3) clearly demonstrated the improvement over NIR and ATR-FTIR individual models. Training sensitivity was 100%, the same value as for the NIR model based on a ROC curve (Table 2) and higher than for the ATR-FTIR model (93%). Sensitivity for the test set obtained with high-level data fusion was the same as in Table 2. However, data fusion provided simultaneously specificities close to 100% (above 97% for all adulterants), much better than SIMCA models built with individual vibrational techniques (Tables 1 and 2). The efficiency was also improved (between 97% and 99%) as compared to both situations, using the class distances obtained by the ROC curve (between 85% and 97%) or equal to 1.0 (between 91% and 96%).

Thus, the use of a high-level data fusion strategy was justified considering the improvement of the classification results in comparison with individual spectral data.

## 4. Conclusions

An untargeted strategy for the rapid detection of adulteration in cashew nuts using portable NIR and ATR-FTIR spectroscopies jointly with one-class SIMCA models was developed. The specificity against four types of nuts (Brazilian nuts, pecan nuts, macadamia nuts, and peanuts) was established. The implementation of the receiver operating characteristic (ROC) curves allows optimizing target class limits, that is, the limit of authentic cashew nuts class below which a sample will be considered as non-adulterated. As a result, it has been proved that the proposed strategy allowed to balance the values of the performance parameters (sensitivity and specificity), providing information on the probability of success in the assignments of non-adulterated and adulterated samples.

A high-level data fusion model based on Fuzzy operators was constructed resulting in an improvement of the performance parameters as compared to models based on individual techniques. It should be emphasized that the development of a high-level data fusion implied the need of sample measurements by at least two instrumental techniques. However, once the classification model was built and optimized, no additional effort is necessary as in the case of applying low- or mid-level data fusion.

The proposed strategy of one-class modelling is preferred when the target class to be modelled is the non-adulterated/authentic, regardless of the possible adulterants. Given the need for constant improvements in food fraud detection, this study represents a contribution to the food scientific community that can easily be extended to other types of nut fraud or involving other products/matrices.

The developed analytical methodology is simple, rapid, green (does not consume reagents or solvents nor generates chemical waste), and non-destructive, thus being considered suitable for screening analysis.

**CRediT authorship contribution statement**

**Glòria Rovira:** Formal analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. **Carolina Sheng Whei Miaw:** Formal analysis, Investigation, Methodology, Writing – review & editing. **Mário Lúcio Campos Martins:** Formal analysis, Methodology. **Marcelo Martins Sena:** Resources, Supervision, Writing – review & editing, Funding acquisition. **Scheilla Vitorino Carvalho de Souza:** Conceptualization, Resources, Supervision, Writing – review & editing, Funding acquisition. **Itziar Ruisánchez:** Conceptualization, Investigation, Methodology, Supervision, Validation, Writing – review & editing, Funding acquisition. **M. Pilar Callao:** Conceptualization, Investigation, Methodology, Supervision, Validation, Writing – review & editing, Funding acquisition.

**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**

This study was supported by the research program “Program of Research Activity (2020PMF-PIPF) at the Rovira i Virgili University, Tarragona, Spain. The acquisition of a portable NIR spectrometer (Viavi MicroNIR® 1700) was supported by the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) through project APQ03457-16.

## References

[1] R.G.M. de Souza, R.M. Shincaglia, G.D. Pimentel, J.F. Mota, Nuts and human health outcomes: A systematic review, Nutrients 9 (2017) 1311, https://doi.org/10.3390/nu9121311.

[2] P. Visciano, M. Schirone, Food frauds: Global incidents and misleading situations, Trends Food Sci. Technol. 114 (2021) 424–442, https://doi.org/10.1016/j.tifs.2021.06.010.

[3] M.F. Rural, Castanhas à venda com preço, Mercado Físico Rural, Marília, Brazil, 2022 https://www.mfrural.com.br/produtos/2-681/alimentos-castanhas, (accessed in May 2022).

[4] A. Valdés, A. Beltrán, C. Mellinas, A. Jiménez, M.C. Garrigós, Analytical methods combined with multivariate analysis for authentication of animal and vegetable food products with high fat content, Trends Food Sci. Technol. 77 (2018) 120–130, https://doi.org/10.1016/j.tifs.2018.05.014.

[5] M. Esteki, Y.V. Heyden, B. Farajmand, Y. Kolahderazi, Qualitative and quantitative analysis of peanut adulteration in almond powder samples using multi-elemental fingerprinting combined with multivariate data analysis methods, Food Contr. 82 (2017) 31–41, https://doi.org/10.1016/j.foodcont.2017.06.014.

[6] M. Esteki, Y.V. Heyden, B. Farajmand, Y. Kolahderazi, Chromatographic fingerprinting with multivariate data analysis for detection and quantification of apricot kernel in almond powder, Food Anal. Methods 10 (2017) 3312–3320, https://doi.org/10.1007/s12161-017-0903-5.

[7] G. Campmajó, G.J. Navarro, N. Nuñez, L. Puignou, J. Saurina, O. Nuñez, Non-Targeted HPLC-UV Fingerprinting as chemical descriptors for the classification and authentication of nuts by multivariate chemometric methods, Sensors 19 (2019) 1388, https://doi.org/10.3390/s19061388.

[8] G. Campmajó, R. Saez-Vigo, J. Saurina, O. Nuñez, High-performance liquid chromatography with fluorescence detection fingerprinting combined with chemometrics for nut classification and the detection and quantitation of almond-based product adulterations, Food Contr. 114 (2020), 107265, https://doi.org/10.1016/j.foodcont.2020.107265.

[9] H. Eksi-Kocak, O. Mentes-Yilmaz, I.H. Boyaci, Detection of green pea adulteration in pistachio nut granules by using Raman hyperspectral imaging, Eur. Food Res. Technol. 242 (2016) 271–277, https://doi.org/10.1007/s00217-015-2538-3.

[10] M.I. López, E. Trullols, M.P. Callao, I. Ruisánchez, Multivariate screening in food adulteration: Untargeted versus targeted modelling, Food Chem. 147 (2014) 177–181, https://doi.org/10.1016/j.foodchem.2013.09.139.

[11] M.I. López, N. Colomer, I. Ruisánchez, M.P. Callao, Validation of multivariate screening methodology. Case study: Detection of food fraud, Anal. Chim. Acta 827 (2014) 28–33, https://doi.org/10.1016/j.aca.2014.04.019.

[12] C. Márquez, M.I. López, I. Ruisánchez, M.P. Callao, FT-Raman and NIR spectroscopy data fusion strategy for multivariate qualitative analysis of food fraud, Talanta 161 (2016) 80–86, https://doi.org/10.1016/j.talanta.2016.08.003.

[13] G. Bonifazi, G. Capobianco, R. Gasbarrone, S. Serranti, Contaminant detection in pistachio nuts by different classification methods applied to short-wave infrared

hyperspectral images, Food Contr. 130 (2021), 108202, https://doi.org/10.1016/j.foodcont.2021.108202.

[14] J.M. Roger, J.C. Boulet, M. Zeaiter, F. Marini, Pre-processing Methods, in: S. D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive Chemometrics, Elsevier, Chemical and Biochemical Data Analysis, 2020, pp. 1–75.

[15] P. Mishra, A. Biancolillo, J.M. Roger, F. Marini, D.N. Rutledge, New data preprocessing trends based on ensemble of multiple preprocessing techniques, TrAC Trends Anal. Chem. 132 (2020), 116045, https://doi.org/10.1016/j.trac.2020.116045.

[16] O. Rodionova, P. Oliveri, A.L. Pomerantsev, Rigorous and compliant approaches to one-class classification, Chemometr. Intell. Lab. Syst. 159 (2016) 89–96, https://doi.org/10.1016/j.chemolab.2016.10.002.

[17] P. Oliveri, Class-modelling in food analytical chemistry: Development, sampling, optimisation and validation issues - A tutorial, Anal. Chim. Acta 982 (2017) 9–19, https://doi.org/10.1016/j.aca.2017.05.013.

[18] M.P. Callao, I. Ruisánchez, An overview of multivariate qualitative methods for food fraud detection, Food Contr. 86 (2018) 283–293, https://doi.org/10.1016/j.foodcont.2017.11.034.

[19] I. Ruisánchez, A.M. Jiménez-Carvelo, M.P. Callao, ROC curves for the optimization of one-class model parameters A case study: Authenticating extra virgin olive oil from Catalan protected designation of origin, Talanta 222 (2021), 121564, https://doi.org/10.1016/j.talanta.2020.121564.

[20] R. Vitale, F. Marini, C. Ruckebusch, SIMCA modeling for overlapping classes: fixed or optimized decision limit? Anal. Chem. 90 (2018) 10738–10747, https://doi.org/10.1021/acs.analchem.8b01270.

[21] B. Quintanilla-Casas, J. Bustamante, F. Guardiola, D.L. García-González, S. Barbieri, A. Bendini, T.G. Toschi, S. Vichi, A. Tres, Virgin olive oil volatile fingerprint and chemometrics: Towards an instrumental screening tool to grade the sensor quality, LWT – Food Sci Technol. 121 (2020), 108936, https://doi.org/10.1016/j.lwt.2019.108936.

[22] P. Oliveri, G. Downey, Multivariate class modeling for the verification of food-authenticity claims, TrAC Trends Anal. Chem. 35 (2012) 74–86, https://doi.org/10.1016/j.trac.2012.02.005.

[23] E. Borrás, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment- A review, Anal. Chim. Acta 891 (2015) 1–14, https://doi.org/10.1016/j.aca.2015.04.042.

[24] C. Alamprese, M. Casale, N. Sinelli, S. Lanteri, E. Casiraghi, Detection of minced beef adulteration with turkey meat by UV-vis, NIR and MIR spectroscopy, LWT – Food Sci, Technol. 53 (2013) 225–232, https://doi.org/10.1016/j.lwt.2013.01.027.

[25] D.P. Aykas, A. Menevseoglu, A rapid method to detect green pea and peanut adulteration in pistachio by using portable FT-MIR and FT-NIR spectroscopy combined with chemometrics, Food Contr. 121 (2021), 107670, https://doi.org/10.1016/j.foodcont.2020.107670.

[26] C.S.M. Miaw, M.L.C. Martins, M.M. Sena, S.V.C. de Souza, Screening method for the detection of other allergenic nuts in cashew nuts using chemometrics and portable near-infrared spectrophotometer, Food Anal, Methods 15 (2022) 1074–1084, https://doi.org/10.1007/s12161-021-02184-0.

[27] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148, https://doi.org/10.1080/00401706.1969.10490666.

[28] M. Bevilaqua, R. Bucci, A.D. Magrì, R. Nescatelli, F. Marini, Classification and class-modelling in: F. Marini (Editor), Data handling in science and Technology, Elsevier 28 (2013) 171-233. https://doi.org/10.1016/B978-0-444-59528-7.00005-3.

[29] C.S.M. Miaw, M.M. Sena, S.V.C. de Souza, M.P. Callao, I. Ruisánchez, Detection of adulterants in grape nectars by attenuated total reflectance Fourier-transform mid-infrared spectroscopy and multivariate classification, Food Chem. 266 (2018) 254–261, https://doi.org/10.1016/j.foodchem.2018.06.006.

[30] C.S. Gondim, R.G. Junqueira, S.V.C. de Souza, I. Ruisánchez, M.P. Callao, Detection of several common adulterants in raw milk by MID-infrared spectroscopy and one-class and multi-class multivariate strategies, Food Chem. 230 (2017) 68–75, https://doi.org/10.1016/j.foodchem.2017.03.022.

[31] C. Durante, R. Bro, M. Cocchi, A classification tool for N-way array based on SIMCA methodology, Chemometr. Intell. Lab. Syst. 106 (2011) 73–85, https://doi.org/10.1016/j.chemolab.2010.09.004.

[32] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (2006) 35–46, https://doi.org/10.1016/j.patrec.2005.10.010.

[33] M. De Figueiredo, C.B.Y. Cordella, D.J.R. Bouveresse, X. Archer, J.M. Bégué, D. N. Rutledge, A variable selection method for multiclass classification problems using two-class ROC analysis, Chemometr. Intell. Lab. Syst. 177 (2018) 35–46, https://doi.org/10.1016/j.chemolab.2018.04.005.

[34] Y. Li, Y. Huang, J. Xia, Y. Xiong, S. Min, Quantitative analysis of honey adulteration by spectrum analysis combined with several high-level data fusion strategies, Vib. Spectrosc. 108 (2020), 103060, https://doi.org/10.1016/j.vibspec.2020.103060.

[35] C.V. di Anibal, M.P. Callao, I. Ruisánchez, $^{1}$H NMR and UV-visible data fusion for determining Sudan dyes in culinary spices, Talanta 84 (2011) 829–833, https://doi.org/10.1016/j.talanta.2011.02.014.

[36] S. Ghosh, P. Mishra, S.N.H. Mohamad, R.M. de Santos, B.D. Iglesias, P.B. Elorza, Discrimination of peanuts from bulk cereals and nuts by near infrared reflectance spectroscopy, Biosyst. Eng. 151 (2016) 178–186, https://doi.org/10.1016/j.biosystemseng.2016.09.008.

[37] H.E. Genis, S. Durna, I.H. Boyaci, Determination of green pea and spinach adulteration in pistachio nuts using NIR spectroscopy, LWT – Food Sci, Technol. 136 (2021), 110008, https://doi.org/10.1016/j.lwt.2020.110008.

[38] M.G. Nespeca, W.D. Pavini, J.E. Oliveira, Multivariate filters combined with interval partial least square method: A strategy for optimizing PLS models developed with near infrared data of multicomponent solutions, Vib. Spectrosc. 102 (2019) 97–102, https://doi.org/10.1016/j.vibspec.2019.05.001.

[39] O. Anjos, A.J.A. Santos, L.M. Estevinho, I. Caldeira, FTIR-ATR spectroscopy applied to quality control of grape-derived spirits, Food Chem. 205 (2016) 28–35, https://doi.org/10.1016/j.foodchem.2016.02.128.

## Paper 3

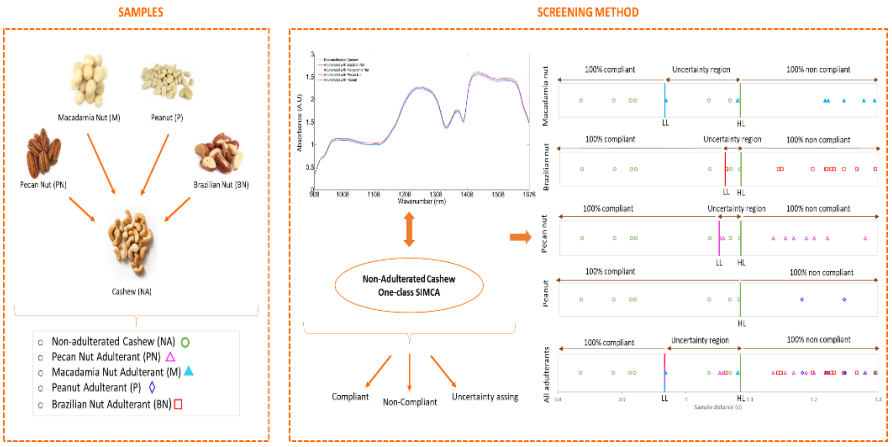
**One-class with two decision thresholds for the rapid detection of cashew nuts adulteration by other nuts**

Glòria Rovira, Carolina Sheng whei Miaw, Mário Lúcio Campos Martins, Marcelo Martins Sena, Scheilla Vitorino Carvalho de Souza, M.Pilar Callao, Itziar Ruisámchez

*Talanta,* **2023**, 253, 123916

https://doi.org/10.1016/j.talanta.2022.123916

### Graphical Abstract




**Keywords:** Nut adulteration, Multivariate screening, Soft independent modelling of class analogy, Portability, Uncertainty intervals, Decision Thresholds.

# One-class model with two decision thresholds for the rapid detection of cashew nuts adulteration by other nuts

Glòria Rovira[a], Carolina Sheng Whei Miaw[b], Mário Lúcio Campos Martins[b], Marcelo Martins Sena[c,d], Scheilla Vitorino Carvalho de Souza[b], M. Pilar Callao[a,*], Itziar Ruisánchez[a]

[a] Chemometrics, Qualimetric, and Nanosensors Group, Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo S/n, 43007, Tarragona, Spain.

[b] Department of Food Science, Faculty of Pharmacy (FAFAR), Federal University of Minas Gerais (UFMG), Av. Antônio Carlos, 6627, Campus da UFMG, Pampulha, 31270-010, Belo Horizonte, MG, Brazil.

[c] Chemistry Department, Institute of Exact Sciences (ICEX), Federal University of Minas Gerais (UFMG), Av. Antônio Carlos, 6627, Campus da UFMG, Pampulha, 31270-010, Belo Horizonte, MG, Brazil.

[d] Instituto Nacional de Ciência e Tecnologia em Bioanalítica (INCT-Bio), Campinas, SP 13083-970, Brazil.

Talanta 253 (2023) 123916

**Abstract**

A green screening method to determine cashew nut adulteration with Brazilian nut, pecan nut, macadamia nut, and peanut was proposed. The method was based on the development of a one-class soft independent modelling of class analogy (SIMCA) model for non-adulterated cashew nuts using near-infrared (NIR) spectra obtained with portable equipment. Once the model is established, the assignment of unknown samples depends on the threshold established for the authentic class, which is a key aspect in any screening approach. The authors propose innovatively to define two thresholds: lower model distance limit and upper model distance limit. Samples with distances below the lower threshold are assigned as non-adulterated with a 100% probability; samples with distance values greater than the upper threshold are assigned as adulterated with a 100% probability; and samples with distances within these two thresholds will be considered uncertain and should be submitted to confirmatory analysis. Thus, the possibility of error in the sample assignment significantly decreases. In the present study, when just one threshold was defined, values greater than 95% for the optimized threshold were obtained for both selectivity and specificity. When two class thresholds were defined, the percentage of samples with uncertain assignment changes according to the adulterant considered, highlighting the case of peanuts, in which 0% of uncertain samples were obtained. Considering all adulterants, the number of samples that were submitted to a confirmatory analysis was quite low, with 5 of 224 adulterated samples and 3 of 56 non-adulterated samples.

## 1. Introduction

Concerns about food safety from most stakeholders, such as consumers, producers, and regulators grow every year. It is well known to these stakeholders that food scandals continue to occur despite national and international regulations [1]. Therefore, there is an increasing demand for developing and improving innovative analytical methods.

The vast potential for food fraud hinders its detection. Often, food fraud detection requires the use of expensive and sophisticated equipment. Currently, alternative and green screening methods, which use less expensive instrumentation, do not consume reagents, and minimize the number of steps and sample manipulation, are gaining relevance [2–9]. Screening methods are very convenient for deciding whether a sample is adulterated or not adulterated (yes/no) and if it is necessary to submit the suspicious samples to a confirmatory analysis. Due to the high complexity of food, it is difficult to develop screening methodologies based on a sample single property or signal. Therefore, the application of spectroscopic techniques together with multivariate classification models is an alternative that gains importance.

Some examples for detecting nut fraud due to the addition of adulterants such as almond [10,11], pistachio [12], and hazelnut [9,13,14] have been referenced. For screening purposes, obtaining multivariate signals should be a process that requires minimal sample treatment prior to measurement. Methods based on vibrational techniques offer these advantages. In particular, handheld near-infrared (NIR) spectrophotometers present the advantage of portability, allowing in situ measurements, time savings for obtaining analytical results and increasing the analytical frequency of the method. On the other hand, portable NIR equipment presents some disadvantages in comparison to benchtop instruments, such as lower resolution and signal-to-noise ratio, and reduced wavelength range. In

general, the optical performance of this type of spectrophotometer has still not reached the level of mainstream commercial instruments [15].

Recently, portable NIR spectroscopy has been successfully applied to food analysis with different aims, such as the varietal discrimination of walnuts [16], the determination of total antioxidant capacity in gluten-free grains [17], the prediction of stable isotopes and fatty acids in subcutaneous fat of Iberian pigs [18], and the on-line monitoring of quality parameters in intact olives for determining optimal harvesting times [19].

Regarding chemometric classification tools, one-class modelling has been considered a better option than the most commonly employed discriminant models for food authenticity problems. One-class modelling methods build a class that is focused on authentic/non-fraudulent samples. This is a proper approach since it aims to detect whether new samples belong to the authentic class regardless of the type of fraud being investigated. This strategy has several benefits in relation to the multi-class approach, in which, in addition to the non-adulterated class, one or more adulterant classes also have to be modelled, since it is impossible in practice to cover all possible adulterants in a representative way [20].

As with any analytical method, multivariate screening methods should be validated prior to their implementation in the routine of quality control laboratories. This process involves establishing performance parameters; however, the validation of multivariate screening methods is still not fully established. Efforts to develop a harmonized validation procedure have been performed in recent years [21–24]. The main figures of merit, sensitivity, and specificity, evaluated from true and false assignments of the samples that belong to the modelled class and those that do not belong to the modelled class, respectively, are currently accepted by the scientific community. Many studies have been carried out aiming to optimize the models and to obtain better performance parameters. Variable selection [25,26] and data fusion [27–29] are notable research topics in this area.

Whether a sample belongs to the modelled class (fits the model) depends on the threshold (limit) established for the class, which in the one-class modelling approach is a distance. As a result, samples having a sample distance lower than the class threshold are the samples that fit the model. Therefore, establishing an optimum class threshold is a key aspect in any screening model. Recently, published articles have discussed alternatives to the establishment of an optimal class threshold [13,14,30,31].

For an analytical method to be considered validated, high sensitivity and specificity values must be obtained, but unless 100% sensitivity and specificity are obtained, there are certain probabilities of error, respectively, for samples belonging to the model class and not belonging to the model class, in the final assignment.

The objective of this article is to propose a new alternative to one-class modelling that estimates two class thresholds (lower- and upper-class model distances) instead of only one class threshold, thus providing three distance intervals:

1) Samples with distance values below the lower-class threshold will fit the model, so they will be unambiguously assigned as authentic/non-fraudulent; 2) Samples with distance values greater than the upper-class threshold will not fit the model, so they will be assigned as fraudulent/adulterated; and 3) samples having distance values between the lower- and the upper-class threshold will be considered inconclusive and, if necessary, will be submitted to confirmatory analysis.

Attempts have also been made to define an uncertainty region in multivariate qualitative analysis [21,32–34], but these approaches have focused on experimenting at concentration levels above and below the concentration of the cut-off value [34], which is the concentration limit established or specified by the end user or legislation. Nevertheless, even when defining the uncertainty region, most of the published examples do

not assure both a sensitivity of 100% and a specificity of 100%. As a case study, the adulteration of cashew nut samples has been considered, and Brazilian nut, macadamia nut, pecan nut, and peanut have been studied as possible adulterants with concentrations within the interval between 0.1 and 10.0%. Adulterated and non-adulterated samples were measured by portable NIR spectroscopy (NIRS) and a one class soft independent modelling of class analogy (SIMCA) model of authentic/non-adulterated samples was established.

## 2. Materials and methods

### *2.1. Samples*

Commercial batches of cashew nut, Brazilian nut (BN), macadamia nut (M), pecan nut (PN), and peanut (P) were obtained from different certified producers. All batches of each nut, individually, were crushed in a sample processor (Arno Magiclean WWBC Blender), homogenized, sieved using a calibrated sieve to size 40 mesh, packed in polyethylene packaging, sealed, and kept at room temperature (25 ± 3 ºC) until preparation of the non-adulterated and adulterated samples.

Seven formulated batches in eight variations were prepared to form the non-adulterated samples of cashew nuts, resulting in a total of 56 samples. Adulterated samples were prepared from each batch of adulterants (Brazilian nut, macadamia nut, pecan nut, and peanut). These samples were added in different quantities to the seven formulated batches of non-adulterated samples, to obtain 8 levels of adulteration (10.0; 5.0; 2.5; 1.3; 0.6; 0.3; 0.2 and 0.1%). The total number of adulterated samples was 224 (56 for each adulterant). The non-adulterated samples were divided into 42 samples of training and 14 samples of test set using the Kennard-Stone algorithm [35]. The objective of this algorithm was to select the most diverse samples for the training set. This selection should be representative (samples homogeneously distributed throughout the whole composition range) and reproducible, based on a systematic criterion.

*2.2. Instrumental measurements*

NIR analysis was performed using portable MicroNIR® 1700 equipment from Viavi Solutions (San Jose, CA, USA).

A homogenized sample portion was placed on a Petri dish (diameter of 3.5 cm x height of 1.2 cm) until the dish was covered up to its maximum height of 1.2 cm. The plate was placed on the MicroNIR®, and for each sample, a reading was carried out with 20 scans at a resolution of 6.25 nm, providing stable and smooth spectra. NIR spectra were recorded in the wavelength range from 908 to 1676 nm. All 280 sample spectra were recorded in random order on the same day.

NIR analysis was performed using portable MicroNIR® 1700 equipment from Viavi Solutions (San Jose, CA, USA). A homogenized sample portion was placed on a Petri dish (diameter of 3.5 cm x height of 1.2 cm) until the dish was covered up to its maximum height of 1.2 cm. The plate was placed on the MicroNIR®, and for each sample, once reading was carried out with 20 scans at a resolution of 6.25 nm, providing stable and smooth spectra. NIR spectra were recorded in the wavelength range from 908 to 1676 nm. All 280 sample spectra were randomly recorded in random order on the same day.

*2.3. Software*

The recorded data were processed, and models were built by using MATLAB software, version 8.0.0.783 – R2012b (Natick, MA, USA) and PLS Toolbox 7.0.2 (Eigenvector Research Inc., Wenatchee, WA, USA).

*2.4. Data processing*

Principal component analysis (PCA) should be performed for a preliminary exploratory analysis of any dataset, even when the final aim is to build a supervised classification model using a class modelling method, such as soft independent modelling of class analogy (SIMCA), for predictive purposes. There is an extensive bibliography that describes the theoretical and

practical aspects of both PCA and SIMCA. Without being exhaustive in the references, a recent review can be consulted, which provides multiple references [20].

SIMCA is a class modelling method that assumes the main systematic variability characterizing the samples of a category, as retained by a principal component model of opportune dimensionality, and builds the model with training samples of that class.

SIMCA assignments are obtained considering the model distance value from a sample "*i"*, Eq. (1).

$$d_{r,i} = \sqrt{\left(Q_{r,i}\right)^2 + \left(T^2_{r,i}\right)^2} \qquad \text{Equation 1}$$

where $T^2_{r,i}$ and $Q_{r,i}$ are the reduced statistics of *Hotelling's* $T^2$ and $Q$, respectively, of a sample; "*i"* and "*r*" denote for reduced values, which comprise the ratio between the statistics of sample "*i*" and the corresponding statistical class limit ($T^2_{lim}$ and $Q_{lim}$) at a significance level of significance [36].

Up to now, just one class threshold is set to decide whether a sample fits the model and several criteria have been applied to determining the threshold: 1, $\sqrt{2}$ and a value obtained through experimentation applying receiver operating characteristic (ROC) curves [31].

To evaluate the quality of the classification models the main performance parameters, such as sensitivity, specificity, and efficiency, were considered [21,23, 24]. These parameters are calculated from the four well-known possibilities of sample model assignment: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Since TP, FP, TN, and FN values depend on the value considered as the class threshold distance, differences in the quality performance parameters can be obtained.

In this work, it is proposed to set two class thresholds, the upper threshold ($d_{upper\_th}$) and the lower threshold ($d_{low\_th}$). $d_{upper\_th}$ corresponds to the

maximum sample distance value ($d_{r,i}$, Eq. (1)) obtained in the prediction of the non-adulterated samples. Similarly, $d_{low\_th}$ corresponds to the minimum $d_{r,i}$ (Eq. (1)) obtained in the prediction of the adulterated samples. As a result, for a given class model, three types of sample assignations can occur: 1) If $d_{r,i}$,<$d_{low\_th}$, the sample will be assigned as non-adulterated (compliant sample) with a 100% probability. 2) If $d_{r,i}$>$d_{upper\_th}$, it will be assigned as adulterated (non-compliant sample) with a 100% probability. 3) If $d_{r,i}$ falls between the two thresholds, the sample falls into the uncertainty region and should undergo a confirmatory analysis.

## 3. Results and discussion

Fig. 1 shows the average raw NIR spectra for the non-adulterated cashew nut samples and adulterated samples with Brazilian nuts (BN), pecan nuts (PN), macadamia nuts (M), and peanuts (P). The largest NIR bands appear between approximately 1170-1300 nm and 1400–1500 nm. The first band can be assigned to the second overtone of the C–H stretching, while the second band can be assigned to the first overtone of the O–H stretching. In particular, the spectral band placed around 1200 nm has been reported as a discriminant of nuts in relation to other food materials (wheat, milk, and cocoa) [37]. The smaller band between 1380 and 1410 nm can be assigned to the combination band of C–H vibrations [38].

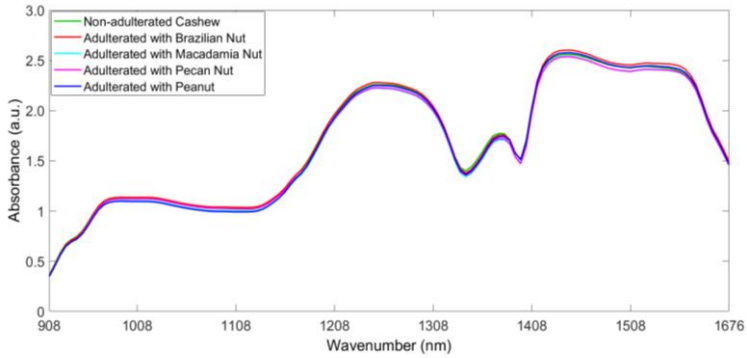



**Fig 1**. Average raw NIR spectra. Color code: green for non-adulterated cashew nuts, red for Brazilian nuts, pink for pecan nuts, light blue for macadamia nuts and dark blue for peanuts.

Before establishing the classification model, the first derivative followed by mean centering pre-processing was applied to the spectra to eliminate nonlinear baseline deviations caused by multiplicative scatter and to improve the signal-to-noise ratio.

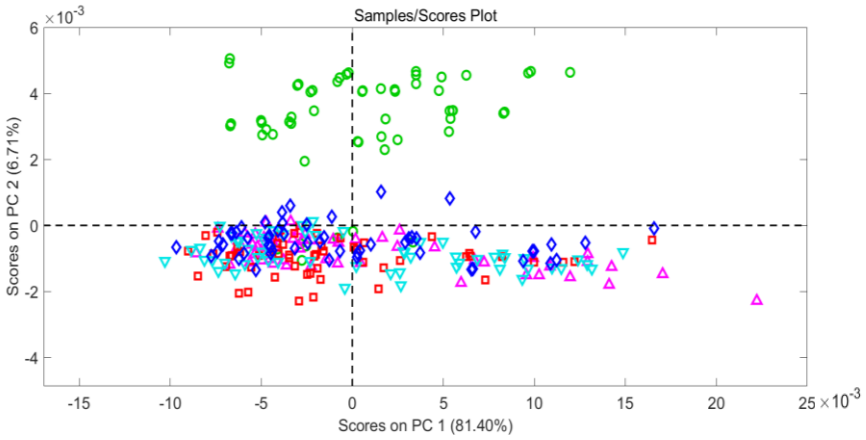



**Fig 2.** Score plot of PC1 vs PC2 for NIR data. Color and symbol code: green circle for non-adulterated cashew nuts, red squares for Brazilian nuts, pink triangles for pecan nuts, filled light blue triangles for macadamia nuts and blue diamonds for peanuts.

Unsupervised PCA was applied before building the supervised classification model, aiming to perform an exploratory analysis with all the samples (non-adulterated and adulterated).

The score plot of the first two PCs for all samples is shown in Fig. 2 (88% of the total explained variance). Regarding the sample distribution along PC 1 (81.4%), no clear grouping was observed in relation to their classes (non-adulterated and adulterated samples with each of the adulterants). This situation is expected because the major chemical components of the samples are identical and, therefore, the most relevant information of the spectra is common to all of them. Regarding PC2 (6.7%) score values, a distinct separation was observed between all non-adulterated samples with positive values and the adulterated samples with mostly negative values. Thus, this lower part of the total variance is responsible for discriminating adulteration. Within the adulterated sample grouping (negative values on PC2), there is

no distinct tendency to discriminate among the studied adulterants or among the percentages of adulteration.

In the sequence, a one-class SIMCA model was built for the authentic/non-adulterated cashew nut samples with the 42 training samples selected by the Kennard-Stone algorithm. Five PCs were employed based on the leave-out-one cross-validation classification error (CVCE). For the test set, 14 non-adulterated samples were combined with all adulterated samples containing each of the four studied adulterants (BN, PN, M, and P).

It is known that the developed model and therefore its quality parameters, highly depends on the criteria used to obtain the training and test sub-sets. Defining both data sets has been always and continues to be an issue when establishing a multivariate methodology and different algorithms have been proposed together with just do it randomly [39]. Kennard-Stone is a possible algorithm that selects the training samples in such a way that they cover the maximum of the multivariate space defined by the available samples. As a result, K–S emphasizes the training samples and thus could lead to slightly optimistic test set quality parameter results. But it is a well-known and defined algorithm with reliable and reproducible results.

If random selection is used, by its very nature, the values of the quality parameters can present more differences, since the possible random sets increase exponentially with the number of samples and may not be very reproducible [40].

In order to evaluate the main quality parameters for the developed model, three criteria to set just one distance threshold ($d_{class\_th}$) have been considered: two criteria fix $d_{class\_th}$ at 1.00 and $\sqrt{2}$ and the third criterion fix $d_{class\_th}$ at 1.14, calculated by means of a ROC curve. If a sample has a distance value ($d_{r,i}$ Eq. 1) lower than $d_{class\_th}$, it is considered to belong to the non-adulterated class (compliant samples); the opposite holds for $d_{r,i}$ greater than $d_{class\_th}$. Fig. 3 shows the reduced distance values calculated from Eq.

(1) ($d_{r,i}$) for all analyzed samples (non-adulterated and adulterated), and those for the three considered $d_{class\_th}$ values (vertical lines). From these results, the main performance parameters were obtained (Table 1).

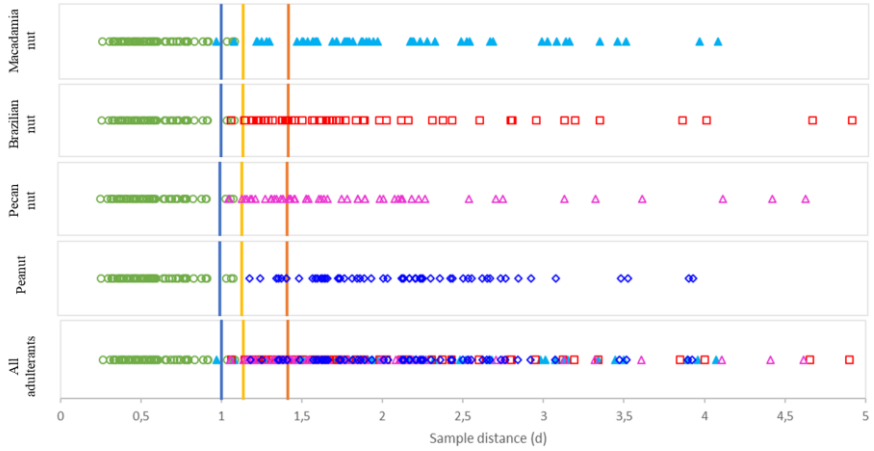



**Fig 3.** Distances for all the analyzed samples to the one-class SIMCA model. Color and symbol codes are the same from Fig.2. Class limit: blue d=1; yellow d=$\sqrt{2}$; orange d=1.08, optimized by ROC curves.

As expected, the figure of merit values varied according to the considered threshold. The threshold value obtained through the ROC curve ($d_{class\_th}$ = 1.14) is the one that better balances both sensitivity and specificity.

When comparing these parameters with each other, as the threshold increases, sensitivity improves, and specificity worsens. Therefore, the choice of the threshold should be based on the practical interest in minimizing the percentage of error associated with the assignment of adulterated or authentic/non-adulterated samples.

Even considering that quite good classification results have been initially obtained, there is still placed to improve the model, from the point of view of its application as a screening method. The goal is to identify with certainty whether a sample is compliant or non-compliant, and in the case of a non-conclusive prediction, submit it to a confirmatory analysis. With this idea,

the strategy proposed in this article implies defining two thresholds (low and upper).

**Table 1.** Figures of merit for one-class SIMCA models considering different class limits. NA: non-adulterated cashew nuts; BN: Brazilian nuts; PN: Pecan nuts; M: Macadamia nuts; P: Peanuts; and BN/PN/M/P: all adulterants.

| | | | NA (%) | BN (%) | PN (%) | M (%) | P (%) | BN/PN /M/P (%) |
|---|---|---|---|---|---|---|---|---|
| **Distance class threshold ($d_{class\_th}$)** | ***d*<1** | **Sensitivity training** | 95.24 | | | | | |
| | | **Sensitivity test** | 92.86 | | | | | |
| | | **Specificity** | | 100 | 100 | 98.21 | 100 | 99.55 |
| | | **Efficiency** | | 98.57 | 98.57 | 97.14 | 98.57 | 99.55 |
| | $d<\sqrt{2}$ | **Sensitivity training** | 100 | | | | | |
| | | **Sensitivity test** | 100 | | | | | |
| | | **Specificity** | | 69.64 | 71.43 | 85.71 | 87.5 | 78.57 |
| | | **Efficiency** | | 74.29 | 75.71 | 87.14 | 88.57 | 79.41 |
| | ***d(ROC)*<1.1359** | **Sensitivity training** | 100 | | | | | |
| | | **Sensitivity test** | 100 | | | | | |
| | | **Specificity** | | 98.21 | 94.64 | 96.43 | 100 | 97.32 |
| | | **Efficiency** | | 98.57 | 95.71 | 97.14 | 100 | 97.49 |

Fig. 4 shows the results obtained with two thresholds considering the adulterants both individually and concurrently. In the problem under study, the upper threshold ($d_{upper\_th}$) value was equal to 1.08 (green vertical lines, Fig. 4). Since the upper threshold was set from the maximum $d_{r,i}$ of the non-adulterated samples, it is the same regardless of the adulterant that is considered. The lower threshold ($d_{low\_th}$) can be calculated independently for each adulterant considered, resulting in four $d_{low\_th}$ values at 0.97, 1.05, 1.06,

and 1.18 for macadamia nuts, pecan nuts, Brazilian nuts, and peanuts, respectively.

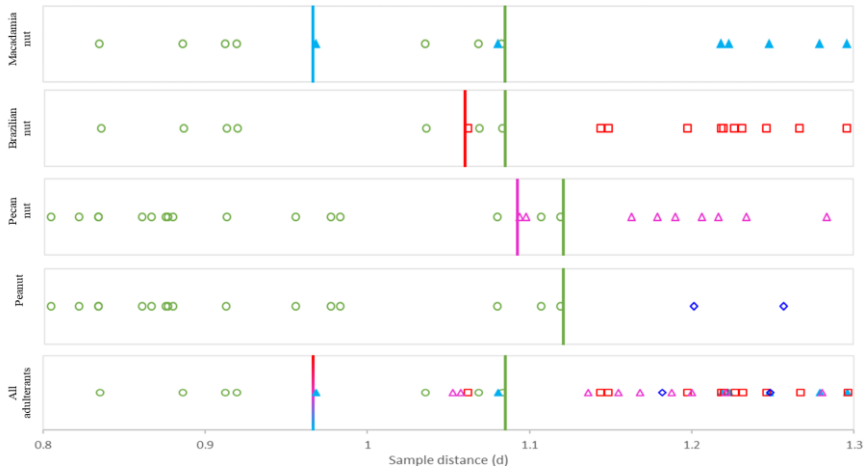



**Fig 4.** One-class SIMCA model based on two thresholds, configuring the distance that define uncertainty intervals. Color and symbol codes are the same from Fig. 2. The abscissa axis is shown between 0.75 and 1.25 sample distance aiming to highlight the samples in between and around the uncertainty region.

That value changed for each adulterant (one color line for each adulterant, Fig. 4), providing a different uncertainty range for each adulterant (region between both lines). When the differentiation between the four adulterants was not considered, a single $d_{low\_th}$ equal to 0.97 was obtained. Note that the lower threshold considering all adulterants coincides with the lowest value individually considering each adulterant and, in that case, this was the threshold for macadamia nut adulteration.

Below the low threshold, there are only non-adulterated samples, and above the upper threshold, there are only adulterated samples. Therefore, there is no ambiguity in the assignments. Between the thresholds, there are both adulterated samples and non-adulterated samples, whose assignments were inconclusive.

**Table 2.** Uncertainty intervals and percentage of samples of uncertain assignment. NA: non-adulterated cashew nuts; BN: Brazilian nuts; PN: Pecan nuts; M: Macadamia nuts; P: Peanuts and BN/PN/M/P: all adulterants.

| Uncertain interval (d) | | Uncertain assignment (%) | |
|---|---|---|---|
| | | Adulterated | Non-adulterated |
| **PN** | 1.05-1.08 | 3.6 | 3.6 |
| **BN** | 1.06-1.08 | 1.8 | 3.6 |
| **M** | 0.97-1.08 | 3.6 | 5.4 |
| **P** | - | 0.0 | 0.0 |
| **BN/PN/M/P** | 0.97-1.08 | 2.2 | 5.4 |

Table 2 shows the distance values that define the uncertainty region and the percentage of samples that fall into it for each adulterant. The number of samples from which the uncertainty assignment percentages were calculated were 56 non-adulterated samples, 56 samples for each individual adulterant (PN, BN, M, and P), and 224 when considering all adulterated samples, regardless of the adulterant (PN + BN + M + P).

The percentage of samples that should be submitted to a confirmatory analysis is quite low (Table 2): 2.2%, corresponding to 5 out of 224 adulterated samples, and 5.4%, corresponding to 3 out of 56 non-adulterated samples. A comprehensive analysis of these 5 adulterated samples indicates that 3 of them (1 sample from Brazilian nut and 2 samples from pecan nuts) correspond to adulterated samples at very low levels (0.15% and 0.6%), while the two remaining samples have been adulterated with macadamia nuts at levels higher than 1.0 (1.3% and 2.5%).

Particular attention should be given to adulteration with peanuts since the minimum predicted distance of the adulterated samples ($d_{r,i}$ = 1.18) was higher than the upper threshold (1.08). In these cases, there is no uncertainty interval. Moreover, two thresholds are not necessary since with just one

threshold, both the sensitivity and specificity were 100%, that is, only the case in which the class distance threshold was set by means of the ROC curve ($d_{class\_th}$, Table 1).

## 4. Conclusions

A multivariate screening method that jointly uses portable NIR spectroscopy with one-class SIMCA models was developed to determine cashew nut adulteration with Brazilian nut, pecan nut, macadamia nut, and peanut. When just one class threshold was estimated for the one-class SIMCA model, values greater than 95% for the optimized threshold were obtained for both selectivity and specificity. The establishment of two threshold values (low and upper limits) generates an uncertainty region. Outside this interval, the samples can be assigned as authentic/non-adulterated (distances less than the lower threshold) or non-authentic/adulterated (distances greater than the upper threshold) with a 100% probability of success. Samples predicted within the interval between these two thresholds are assigned to the uncertainty region and should undergo further confirmatory analysis.

When two class thresholds were defined, it was possible to detect with certainty if a sample was compliant or non-compliant (100% for both sensitivity and specificity) by defining an uncertainty region.

In this region, the percentage of samples within the uncertain assignment changed according to the adulterant that was considered. In all cases, the percentage of samples that should be submitted to a confirmatory analysis was quite low, including both non-adulterated samples and adulterated samples, even when they were simultaneously considered regardless of the adulterant (PN + BN + M + P).

The developed analytical methodology is simple, rapid, green (neither consumes reagents or solvents nor generates chemical waste) non-destructive, and thus is considered suitable for screening analysis. Given the need for constant improvements in food fraud detection, this study

represents a contribution to food and analytical scientific communities that can easily be extended to other types of food fraud, or even fraud involving other types of products/matrices.

**Author Statement**

Glòria Rovira: Formal analysis, Chemometric analysis, Investigation, Methodology, writing, Carolina Sheng Whei Miaw: Methodology, Investigation, Formal analysis, Writing – review & editing, Mário Lúcio Campos Martins: Methodology, Formal analysis, Marcelo Martins Sena: Resources, Supervision, Writing – review & editing, Funding acquisition, Scheilla Vitorino Carvalho de Souza: Conceptualization, Resources, Supervision, Writing – review & editing, Funding acquisition. Itziar Ruisánchez: Conceptualization, Investigation, Methodology, Supervision, Validation, writing, reviewing and Editing, Funding acquisition, M. Pilar Callao: Conceptualization, Investigation, Methodology, Supervision, Validation, writing, reviewing and Editing, Funding acquisition.

**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**

This study was supported by the research program “Program of Research Activity (2020PMF-PIPF)” at the Rovira i Virgili University, Tarragona, Spain. The acquisition of a portable NIR spectrometer (Viavi MicroNIR® 1700) was supported by the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) through project APQ03457-16.

## References

[1] A.S. Tsagkaris, J.L.D. Nelis, G.M.S. Ross, S. Jafari, J. Guercetti, K. Kopper, Y. Zhao, K. Rafferty, J.P. Salvador, D. Migliorelli, G.I.J. Salentijn, K. Campbell, M.P. Marco, C.T. Elliot, M.W.F. Nielen, J. Pulkrabova, J. Hajslova, Critical assessment of recent trends related to screening and confirmatory analytical methods for selected food contaminants and allergens, TrAC, Trends Anal. Chem. 121 (2019), 115688, https://doi.org/10.1016/j.trac.2019.115688.

[2] C.S.M. Miaw, M.L.C. Martins, M.M. Sena, S.V. C de Souza, Screening method for the detection of other allergenic nuts in cashew nuts using chemometrics and a portable near-infrared spectrophotometer, Food Anal. Methods 15 (2022) 1074–1084, https://doi.org/10.1007/s12161-021-02184-0.

[3] A. de Girolamo, M.C. Arroyo, V. Lippolis, S. Cervellieri, M. Cortese, M. Pascale, A. F. Logrieco, C. von Holst, A simple design for the validation of a FT-NIR screening method: application to the detection of durum wheat pasta adulteration, Food Chem. 333 (2020), 127449, https://doi.org/10.1016/j.foodchem.2020.127449.

[4] M. Spiteri, E. Jamin, F. Thomas, A. Rebours, M. Lees, K.M. Rogers, D.N. Rutledge, Fast and global authenticity screening of honey using 1H-NMR profiling, Food Chem. 189 (2015) 60–66, https://doi.org/10.1016/j.foodchem.2014.11.099.

[5] B. Quintanilla-Casas, J. Bustamante, F. Guardiola, D.L. García-González, S. Barbieri, A. Bendini, T.G. Toschi, S. Vichi, A. Tres, Virgin olive oil volatile fingerprint and chemometrics: towards an instrumental screening tool to grade the sensory quality, LWT–Food Sci. Technol. 121 (2020), 108936, https://doi.org/10.1016/j.lwt.2019.108936.

[6] C.S. Gondim, R.G. Junqueira, S.V. C de Souza, I. Ruisánchez, M.P. Callao, Detection of several common adulterants in raw milk by MID-infrared spectroscopy and one-class and multi-class multivariate strategies, Food Chem. 230 (2017) 68–75, https://doi.org/10.1016/j.foodchem.2017.03.022.

[7] L.S.A. Pereira, F.L.C. Lisboa, J.C. Neto, F.N. Valladão, M.M. Sena, Screening method for rapid classification of psychoactive substances in illicit tablets using mid-infrared spectroscopy and PLS-DA, Forensic Sci. Int. 288 (2018) 227–235, https://doi.org/10.1016/j.forsciint.2018.05.001.

[8] B. Quintanilla-Casas, G. Strocchi, J. Bustamante, B. Torres-Cobos, F. Guardiola, W. Moreda, J.M. Martínez-Rivas, E. Valli, A. Bendini, T.G. Toschi, A. Tres, S. Vichi, Large-scale evaluation of shotgun triacylglycerol profiling for the fast detection of olive oil adulteration, Food Control 123 (2021), 107851, https://doi.org/10.1016/j.foodcont.2020.107851.

[9] M.I. López, N. Colomer, I. Ruisánchez, M.P. Callao, Validation of multivariate screening methodology. Case study: detection of food fraud, Anal. Chim. Acta 827 (2014) 28–33, https://doi.org/10.1016/j.aca.2014.04.019.

[10] M. Esteki, Y. Heyden, B. Farajmand, Y. Kolahderazi, Qualitative and quantitative analysis of peanut adulteration in almond powder samples using multi-elemental fingerprinting combined with multivariate data analysis methods, Food Control 82 (2017) 31–41, https://doi.org/10.1016/j.foodcont.2017.06.014.

[11] G. Campmajó, R. Saez-Vigo, J. Saurina, O. Núñez, High-performance liquid chromatography with fluorescence detection fingerprinting combined with chemometrics for

nut classification and the detection and quantitation of almond-based product adulterations, Food Control 114 (2020), 107265, https://doi.org/10.1016/j.foodcont.2020.107265.

[12] H. Eksi-Kocak, O. Mentes-Yilmaz, I.H. Boyaci, Detection of green pea adulteration in pistachio nut granules by using Raman hyperspectral imaging, Eur. Food Res. Technol. 242 (2016) 271–277, https://doi.org/10.1007/s00217-015-2538-3.

[13] M.I. López, E. Trullols, M.P. Callao, I. Ruisánchez, Multivariate screening in food adulteration: untargeted versus targeted modelling, Food Chem. 147 (2014) 177–181, https://doi.org/10.1016/j.foodchem.2013.09.139.

[14] C. Márquez, M.I. López, I. Ruisánchez, M.P. Callao, FT-Raman and NIR spectroscopy data fusion strategy for multivariate qualitative analysis of food fraud, Talanta 161 (2016) 80–86, https://doi.org/10.1016/j.talanta.2016.08.003.

[15] C. Zhu, X. Fu, J. Zhang, K. Qin, C. Wu, Review of portable near infrared spectrometers: current status and new techniques, J. Near Infrared Spectrosc. 30 (2022) 51–66, https://doi.org/10.1177/09670335211030617.

[16] 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.

[17] V. Wiedemair, C.W. Huck, Evaluation of the performance of three hand-held near-infrared spectrometers through investigation of total antioxidant capacity in gluten-free grains, Talanta 189 (2018) 233–240, https://doi.org/10.1016/j.talanta.2018.06.056.

[18] M.I. González-Martín, O. Escuredo, M. Hernández-Jiménez, I. Revilla, A.M. A. Vivar-Quintana, I. Martínez-Martín, P. Hernández-Ramos, Prediction of stable isotopes and fatty acids in subcutaneous fat of Iberian pigs by means of NIR: a comparison between benchtop and portable systems, Talanta 224 (2021), 121817, https://doi.org/10.1016/j.talanta.2020.121817.

[19] A.J. Fernández-Espinosa, Combining PLS regression with portable NIR spectroscopy to on-line monitor quality parameters in intact olives for determining optimal harvesting time, Talanta 148 (2016) 216–228, https://doi.org/10.1016/j.talanta.2015.10.084.

[20] P. Oliveri, C. Malegori, E. Mustorgi, M. Casale, Qualitative pattern recognition in chemistry: theoretical background and practical guidelines, Microchem. J. 162 (2021), 105725, https://doi.org/10.1016/j.microc.2020.105725.

[21] M.I. López, M.P. Callao, I. Ruisánchez, A tutorial on the validation of qualitative methods: from the univariate to the multivariate approach, Anal. Chim. Acta 891 (2015) 62–72, https://doi.org/10.1016/j.aca.2015.06.032.

[22] A.L. Pomerantsev, O.Y. Rodionova, New trends in qualitative analysis: performance, optimization, and validation of multi-class and soft models, TrAC, Trends Anal. Chem. 143 (2021), 116372, https://doi.org/10.1016/j.trac.2021.116372.

[23] D. Ballabio, F. Grisoni, R. Todeschini, Multivariate comparison of classification performance measures, Chemometr. Intell. Lab. Syst. 174 (2018) 33–44, https://doi.org/10.1016/j.chemolab.2017.12.004.

[24] L. Cuadros-Rodríguez, E. Pérez-Castaño, C. Ruiz-Samblás, Quality performance metrics in multivariate classification methods for qualitative analysis, TrAC, Trends Anal. Chem. 80 (2016) 612–624, https://doi.org/10.1016/j.trac.2016.04.021.

[25] Y.H. Yun, H.D. Li, B.C. Deng, D.S. Cao, An overview of variable selection methods in multivariate analysis of near-infrared spectra, TrAC, Trends Anal. Chem. 113 (2019) 102–115, https://doi.org/10.1016/j.trac.2019.01.018.

[26] A.A. Gomes, S.M. Azcarate, P.H.G.D. Diniz, D.D.S. Fernandes, G. Veras, Variable selection in the chemometric treatment of food data: a tutorial review, Food Chem. 370 (2022), 131072, https://doi.org/10.1016/j.foodchem.2021.131072.

[27] A. Mir-Cerdà, B. Granell, A. Izquierdo-Llopart, A. Sahuquillo, J.F. López-Sánchez, J. Saurina, S. Sentellas, Data fusion approaches for the characterization of musts and wines based on biogenic amine and elemental composition, Sens 22 (2022) 2132, https://doi.org/10.3390/s22062132.

[28] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment –A review, Anal. Chim. Acta 891 (2015) 1–14, https://doi.org/10.1016/j.aca.2015.04.042.

[29] M.P. Callao, I. Ruisánchez, An overview of multivariate qualitative methods for food fraud detection, Food Control 86 (2018) 283–293, https://doi.org/10.1016/j.foodcont.2017.11.034.

[30] R. Vitale, F. Marini, C. Ruckebusch, SIMCA modeling for overlapping classes: fixed or optimized decision threshold? Anal. Chem. 90 (2018) 10738–10747, https://doi.org/10.1021/acs.analchem.8b01270.

[31] I. Ruisánchez, A.M. Jiménez-Carvelo, M.P. Callao, ROC curves for the optimization of one-class model parameters. A case study: authenticating extra virgin olive oil from a Catalan protected designation of origin, Talanta 222 (2021), 121564, https://doi.org/10.1016/j.talanta.2020.121564.

[32] F.C. Lemyre, B. Desharnais, J. Laquerre, M.A. Morel, C. Côté, P. Mireault, C. D. Skinner, Qualitative threshold method validation and uncertainty evaluation: a theoretical framework and application to a 40 analytes LC-MS/MS method, Drug Test. Anal. 12 (2020) 1287–1297, https://doi.org/10.1002/dta.2867.

[33] C.S. Gondim, R.G. Junqueira, S.V.C. de Souza, M.P. Callao, I. Ruisánchez, Determining performance parameters in qualitative multivariate methods using probability of detection (POD) curves. Case study: two common milk adulterants, Talanta 168 (2017) 23–30, https://doi.org/10.1016/j.talanta.2016.12.065.

[34] I. Ruisánchez, G. Rovira, M.P. Callao, Multivariate qualitative methodology for semi-quantitative information. A case study: adulteration of olive oil with sunflower oil, Anal. Chim. Acta 1206 (2022), 339785, https://doi.org/10.1016/j.aca.2022.339785.

[35] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148, https://doi.org/10.1080/00401706.1969.10490666.

[36] A. Rius, M.P. Callao, F.X. Rius, Multivariate statistical process control applied to sulfate determination by sequential injection analysis, Analyst 122 (1997) 737–741, https://doi.org/10.1039/A607954G.

[37] S. Ghosh, P. Mishra, S.N.H. Mohamad, R.M. de Santos, B.D. Iglesias, P.B. Elorza, Discrimination of peanuts from bulk cereals and nuts by near infrared reflectance spectroscopy, Biosyst. Eng. 151 (2016) 178–186, https://doi.org/10.1016/j.biosystemseng.2016.09.008.

[38] H.E. Genis, S. Durna, I.H. Boyaci, Determination of green pea and spinach adulteration in pistachio nuts using NIR spectroscopy, LWT – Food Sci, Technol. 136 (2021), 110008, https://doi.org/10.1016/j.lwt.2020.110008.

[39] N. Shetty, Å. Rinnan, R. Gislum, Selection of representative calibration sample sets for near-infrared reflectance spectroscopy to predict nitrogen concentration grasses, Chemometr. Intell. Lab. Syst. 111 (2012) 59–65, https://doi.org/10.1016/j.chemolab.2011.11.013.

[40] A. Fort, I. Ruisánchez, M.P. Callao, Chemometric strategies for authenticating extra virgin olive oils from two geographically adjacent Catalan protected designations of origin, Microchem. J. 169 (2021), 106611, https://doi.org/10.1016/j.microc.2021.106611.

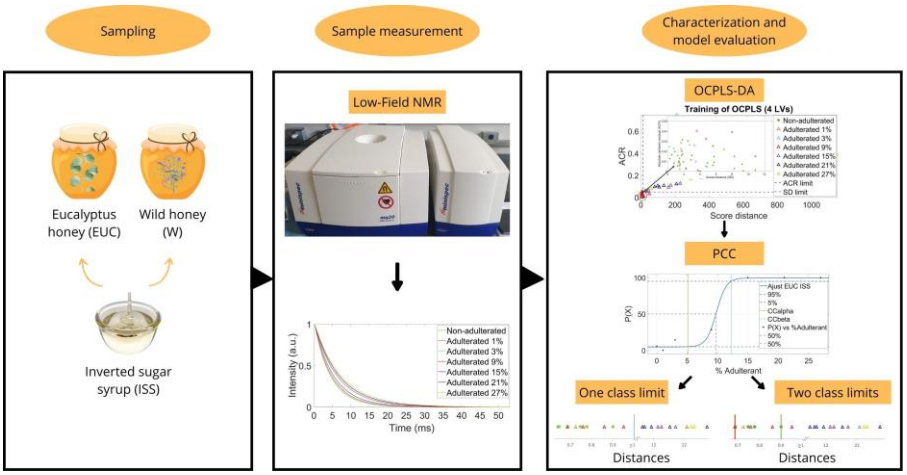
## Paper 4

**A semi-quantitative one-class modelling method for detecting honey adulteration using two-class limits**

Glòria Rovira, Carolina Sheng Whei Miaw, Laura Lima de Oliveira, Marcus Vinicius de Oliveira Andrade, Poliana Macedo Santos, Marcelo Martins Sena, Scheilla Vitorino Carvalho de Souza, Itziar Ruisánchez, M.Pilar Callao

*Submitted*

### Graphical Abstract




**Keywords:** Multivariate classification, Low-Field NMR, Honey adulteration, Semi-quantitative performance parameters, Uncertainty interval, Decision limit.

# Semi-quantitative one-class modelling method for detecting honey adulteration using two-class limits

Glòria Rovira[a], Carolina Sheng Whei Miaw[b], Laura Lima de Oliveira[c], Marcus Vinicius de Oliveira Andrade[d], Poliana Macedo Santos[e], Marcelo Martins Sena[f,g], Scheilla Vitorino Carvalho de Souza[g], M. Pilar Callao[a,*], Itziar Ruisánchez[a]

[a] Chemometrics and Sensorics for Analytical Solutions (CHEMOSENS) Group, Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo s/n, 43007, Tarragona, Spain.

[b] Food Science Graduate Program (PPGCA), Faculty of Pharmacy (FAFAR), Federal University of Minas Gerais (UFMG), Av. Antônio Carlos, 6627, Campus da UFMG, Pampulha, 31270-010, Belo Horizonte, MG, Brazil.

[c] Faculty of Pharmacy (FAFAR), Federal University of Minas Gerais (UFMG), Av. Antônio Carlos, 6627, Campus da UFMG, Pampulha, 31270-010, Belo Horizonte, MG, Brazil.

[d] Brazilian Federal Police, Regional Superintendence in Minas Gerais, 30441-170, Belo Horizonte, MG, Brazil.

[e] Chemistry and Biology Department (DAQBI), Federal Technological University of Paraná (UTFPR), Rua Dep. Heitor Alencar Furtado, 5000, Cidade Industrial, 81280-340, Curitiba, PR, Brazil.

[f] Chemistry Department, Institute of Exact Sciences (ICEx), Federal University of Minas Gerais (UFMG), Av. Antônio Carlos, 6627, Campus da UFMG, Pampulha, 31270-010, Belo Horizonte, MG, Brazil. Instituto Nacional de Ciência e Tecnologia em Bioanalítica (INCT-Bio), Campinas, SP, 13083-970, Brazil.

[g] Food Science Graduate Program (PPGCA) and Department of Food Science (ALM), Faculty of Pharmacy (FAFAR), Federal University of Minas Gerais (UFMG), Av. Antônio Carlos, 6627, Campus da UFMG, Pampulha, 31270-010, Belo Horizonte, MG, Brazil.

**Abstract**

A screening method to determine honey adulteration with inverted sugar syrup was proposed. The method was based on the development of a one-class Partial Least Squares Discriminant Analysis (OCPLS) model for non-adulterated honey using Low-Field Nuclear Magnetic Resonance (LF-NMR) spectroscopy. A total of 77 honey samples were analyzed (35 authentic and 42 adulterated). Once the model was established, it was characterized by semi-quantitative performance parameters obtained from the performance characteristic curves (PCC). The decision limit ($CC\alpha$) was estimated at approximately 5% (w/w) for eucalyptus (EUC) and wild (W) honey. Consequently, samples adulterated close to this level will not be distinguishable from the non-adulterated ones. For the unknown samples' prediction stage, we propose defining two class limits that generate an uncertainty region (UR). The samples assigned as inconclusive (within the UR) should be submitted for further confirmatory analysis. The main differences concerning adopting only one limit were observed at the medium levels of adulteration (around 9%), for which the error rate was reduced from 70% (EUC) and 30% (W) to 14% for both types of honey. Additionally, 11% (EUC) and 37% (W) of the adulterated samples, respectively, were predicted as inconclusive, thus demanding to be submitted for confirmatory analysis.

## 1. Introduction

According to the Standard for Honey CXS 12-1981 from the Codex Alimentarius, honey is defined as the natural sweet substance produced by honeybees from the nectar of plants or secretions of living parts of plants [1]. Honey is widely recognized for its good and attractive organoleptic properties and high health benefits [2,3]. The price of high-quality honey has gone up due to the rising demand and limited availability. Consequently, honey is susceptible to adulteration [4]. This type of fraud violates the standards established by Codex Alimentarius [1] and regulatory bodies such as the European Commission [5,6] and the Brazilian Ministry of Agriculture and Livestock [7], due to the possible adverse health impact of the consumption of adulterated honey [8].

There are different types of fraud in honey but the most common is the direct addition of sugar or syrups. The most used syrups are high-fructose corn syrup [9,10], rice syrup [11,12], inverted sugar [4], cane sugar [5,13], and other sugar syrups [3,14,15].

In this work, a strategy was established to detect the adulteration of honey by the addition of inverted sugar syrup, one of the most common adulterants reported in fraud [4]. The official reference method for the identification of syrup-adulterated honey is based on stable carbon isotope ratio analysis [16]. This method is time-consuming, sample-destructive, and requires expensive instruments. In recent years, an effort has been made to implement faster, environmentally friendly, and non-destructive techniques, such as molecular spectroscopy [3]. Many articles have described the use of spectroscopic techniques, such as near-infrared (NIR) [3,9,17], UV-Visible [5], Fourier-transform infrared-attenuated total reflectance (FTIR-ATR) [4,18,19], spectrofluorimetry [20], Raman [21], and nuclear magnetic resonance (NMR) [22,23], coupled with chemometrics for the quantitative and qualitative determination of honey adulteration. Qualitative methods relying on multivariate classification techniques, such as partial least

squares discriminant analysis (PLS-DA) [20,24], soft independent modelling of class analogy (SIMCA) [5], and one-class partial least squares discriminant analysis (OCPLS) [5], have increasingly been employed for detecting food fraud.

The validation of these classification methods typically involves assessing main performance parameters, such as sensitivity, specificity, accuracy, precision, and occurrence [25,26,27]. These parameters are calculated from the four binary possible responses: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) indicating whether a sample belongs to or does not belong to the modelled class. To determine if a sample belongs to the modelled class, it is necessary to set the limit for this class, which is usually defined as a model distance. Samples with a distance value lower than the class limit are considered to fit within the class model [27,28,29].

The performance parameters mentioned above refer to a binary response, e.g., the unknown sample being compliant or not. When dealing with adulteration problems, it is important to consider that the model's ability will depend on the level of sample adulteration. Therefore, in addition to strictly qualitative performance parameters, concentration-related performance (semi-quantitative) parameters should be established. Recently, the detection limit ($CC\alpha$), the detection capability ($CC\beta$), and the unreliability region (UR) have been referenced [31] with their estimates derived from performance characteristic curves (PCC) [31,32,33].

The strategy proposed in this article aims to establish an OCPLS classification model and characterize it using semi-quantitative performance parameters ($CC\alpha$, $CC\beta$, and UR) estimated from PCC. These parameters will be used to define two class limits (lower and upper) as model distances, resulting in three distance regions: i) non-adulterated region with distance values below the lower limit, ii) adulterated region with distance values greater than the upper limit, and iii) uncertainty region with distance values

falling between both limits [28]. The assignment of unknown samples will be based on the region where their calculated distance value falls. If a sample falls in the uncertainty region, it will be considered inconclusive and should be submitted for further confirmatory analysis.

Numerous studies in food analysis have utilized multivariate qualitative methods to detect potential adulteration. However, only a few of them have determined parameters related to the limits of concentration detectable for the adulterant, thus establishing semi-quantitative models [31,34,35,36,37]. This paper makes a noteworthy contribution in this sense, taking advantage of PCC to determine the decision limit ($CC\alpha$), which will help us to define two class limits when predicting unknown samples. Furthermore, the paper highlights the benefits of evaluating the model before prediction using semi-quantitative parameters.

## 2. Experimental part

### *2.1. Samples*

Two batches of eucalyptus (EUC) and wild (W) types of honey were obtained directly from traceable producers in Minas Gerais State, Brazil. Seven proportions of those batches were prepared in five replicates to form 35 unadulterated samples for each type. Inverted sugar syrup (ISS) was added in different quantities to the seven formulated batches of unadulterated samples to obtain six levels of adulteration (1.0, 3.0, 9.0, 15.0, 21.0, and 27.0% w/w), resulting in 42 adulterated samples.

All samples were weighed using a Shimadzu AUX 220 analytical balance with a calibrated scale in 300 g flasks and in 15 mL or 50 mL Falcon tubes, followed by manual homogenization and storage at room temperature (19ºC to 25ºC) until the moment of the analysis. In the case of a honey sample crystallized, it was placed in a water bath at 39-40ºC for 5 minutes.

### *2.2. Instrumental measurements*

Low-field NMR analysis was performed using the Minispec ND mq-20 from Buker Biospin GmbH (Rheinstetten, Germany), operating at a $^{1}H$ frequency of 20 MHz (0.47 T) with a 10 mm probe head. The Transverse relaxation decays ($T^2$) were measured by a Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence with a 90° and 180° pulse length of 2.76 and 5.42, respectively, an echo time of 500 μs, a recycle delay of 2 s, and 54 dB of gain. A homogenized sample portion (approximately 5 mL) was placed on a crystal NMR tube and heated in a thermostatic bath at 40ºC. The tube was introduced into the Minispec and analyzed with three readings of 16 scans each, generating a relaxation curve of 500 data points for each sample.

### *2.3. Software*

The data obtained were processed using MATLAB, version 8.0.0.783-R2012b (Natick, MA, USA), and PLS Toolbox 7.0.2 (Eigenvector Research Inc., Wenatchee, WA, USA). The Matlab routine for OCPLS was provided by Lu Xu [38]. For the Inverse Laplace Transform the webapp rmn-ilt.streamlit.app was used, developed by Tiago B. Moraes [39].

## 3. Theoretical background

### *3.1. One-class partial least squares classifier (OCPLS)*

OCPLS classifier is based on partial least squares (PLS). For each sample, two statistical parameters emerge once the number of latent variables (LV) is fixed. The first, Hotelling's $T^2$, relies on the score distance (*SD*), quantifying the distance from an object to the center of the class. The second is the absolute centered residual (*ACR*), which is a measure of dispersion or residuals in the projection onto the vector of the OCPLS regression coefficients. A confidence limit is established for both statistics, $SD_{lim}$ and $ACR_{lim}$, that corresponds to the limit of the class at a specified confidence level (typically 95%) [40,41]. A sample must exhibit values for both statistical parameters below these limits to be classified as belonging to a target class. Another criterion for sample assignment involves calculating

the reduced distance of a sample from its class ($d_{i,r}$) which is calculated according to the following equation, Eq. (1). If that is the case, for sample prediction its distance value ($d_{i,r}$) should be compared to a reduce distance class limit. Different distance values can be used as a limit to assign a sample to a specific class, for instance, 1 [42,43], √2 [44,45], or an optimized distance obtained by applying ROC curves [29]. Usually, it is used the distance value established at 1, so $d_{i,r} \leq 1$.

$$d_{i,r} = \sqrt{\left(\frac{SD_i}{SD_{lim}}\right)^2 + \left(\frac{ACR_i}{ACR_{lim}}\right)^2} \quad \text{Eq. (1)}$$

where $SD_i$ and $ACR_i$ are the statistical parameters of a sample "*i*" and the $SD_{lim}$ and $ACR_{lim}$ are the corresponding statistical class limits at a determinate level of significance.

Following the scheme in Fig. 3, if the model is considered appropriate for the case under study, the next step is the prediction of unknown samples. Instead of just using a distance class limit (usually $d_{i,r} \leq 1$), it is suggested to define two distance class limits, the upper ($d_{upper\_lim}$) and the lower ($d_{lower\_lim}$) class limits. Recent articles have also proposed the use of two decision limits [28,46]. In the sample prediction step, the sample is assigned according to $d_{i,r} \leq d_{lower\_lim}$ or $d_{i,r} \geq d_{upper\_lim}$.

*3.2. Semi-quantitative figures of merit*

Sensitivity, specificity, and accuracy [25,26,47] are the main performance parameters for qualitative method validation. When detecting food adulteration, these parameters can be related to the percentage of adulterant present in the sample. This dependency can be observed by establishing PCC. PCCs have been referred to by different names, such as the probability of detection [48], performance curves [49], and probability of identification [50], among others. PCC are represented through a plot of the probability of the method providing a positive response (P(X)) as a function of each concentration/percentage of the adulterant [48,51]. A positive response indicates the presence of an adulterant or, in other words, the sample does

not belong to the model established for unadulterated/authentic samples. So, when the adulterant is absent, the probability of estimating a positive result should be zero or close to zero. But when the concentration of the adulterant increases, the probability should increase until it reaches 100%. The P(X) values versus % adulterant are fitted to a sigmoid function by minimizing the root mean square of the residuals (RMSE) to obtain PCC according to Eq. 2 [25].

$$P(X) = \frac{a}{1-e^{(b+(c*x))}} + d \qquad \text{Eq. (2)}$$

From PCC, three semi-quantitative additional parameters can be obtained [25,52,53]:

- The decision limit (*CCα*) is the minimum concentration of the analyte (adulterant) that can be reliably detected or identified in a sample with a low statistical certainty, usually 5%. This value is obtained from the intersection of the PCC with a horizontal line at a certain value of P(X), usually 5%.
- The detection capability (*CCβ)* is the concentration of the analyte (adulterant), from which its presence can be reliably detected or identified in a sample with a statistical certainty of 95%. The *CCβ* is obtained from the intersection of PCC with an upper horizontal line at P(X)=95%.
- The uncertainty region (UR) is the range between *CCα* and *CCβ*, if a sample falls into UR, it will be assigned as inconclusive and should undergo a confirmatory analysis.

## 4. Results and discussion

Fig. 1 shows the relaxation curves of the non-adulterated EUC honey samples and adulterated EUC honey with inverted sugar at different percentages (1-27%). The intensity of the signals of mean curves of maximum normalized $T^2$ relaxation profiles obtained with the CPMG pulse sequence increases with the level of adulteration (Fig. 1a). Similarly, Fig.

1b shows the inverse Laplace transform (ILT) applied to these CPMG decays, where can be observed the distributed $T^2$ relaxation time in the logarithmic scale for non-adulterated and adulterated samples. Non-adulterated (dark green line) and adulterated samples at 1% (orange line) and 3% (blue line) have relaxation populations between 4.76-4.85 ms, from the start of their bands (Fig 1b), as can be seen in Table 1.

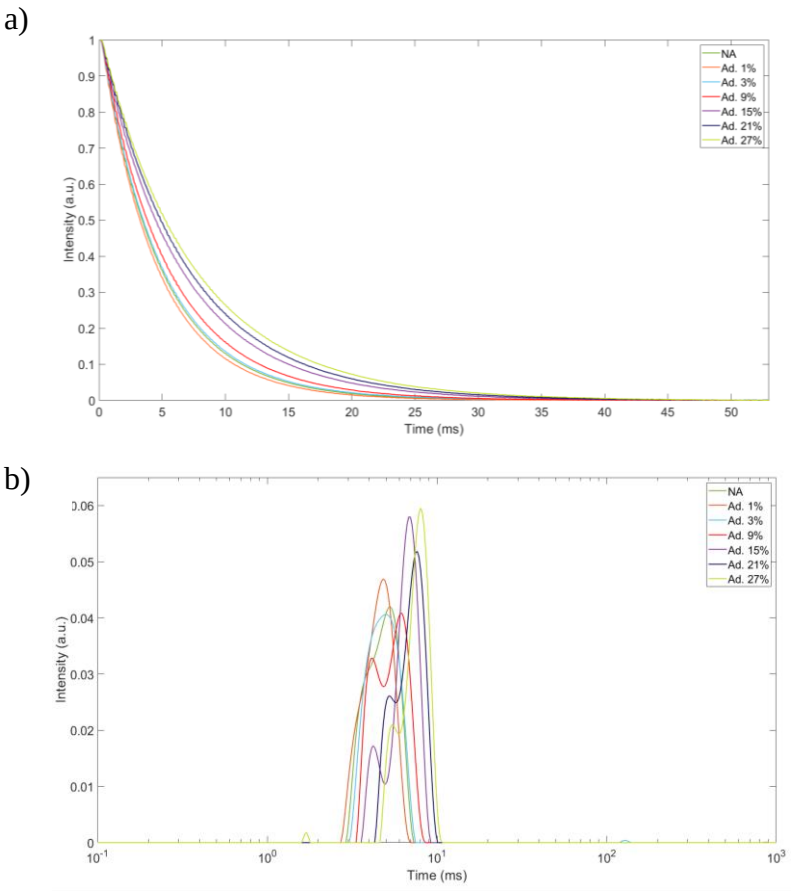



**Fig 1. (a)** Maximum normalized $T^2$ mean relaxation curves obtained with CPMG pulse sequence for unadulterated/authentic and adulterated (1-27% w/w) eucalyptus honey samples. **(b)** $T^2$ relaxation spectra of the respective honey samples obtained via ILT. Color codes: dark green for non-adulterated honey, orange for samples adulterated at 1%, blue for adulterated at 3%, red for adulterated at 9%, purple for adulterated at 15%, dark blue for adulterated at 21% and light green for adulterated at 27%.

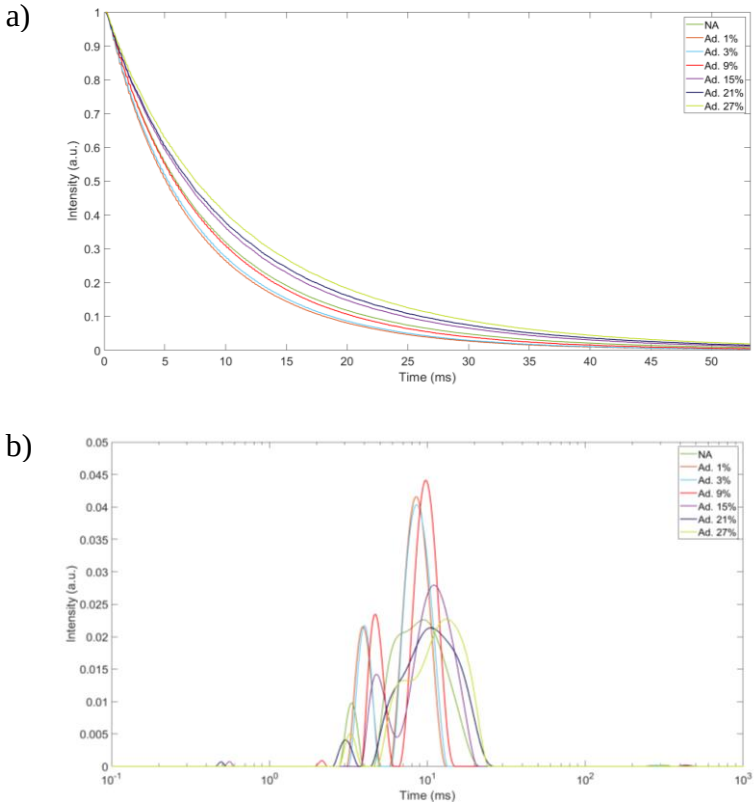



**Fig 2. (a)** Maximum normalized $T^2$ mean relaxation curves obtained with CPMG pulse sequence for unadulterated/authentic and adulterated (1-27% w/w) wild honey samples. **(b)** $T^2$ relaxation spectra of the respective honey samples obtained via ILT. Color codes: dark green for non-adulterated honey, orange for samples adulterated at 1%, blue for adulterated at 3%, red for adulterated at 9%, purple for adulterated at 15%, dark blue for adulterated at 21% and light green for adulterated at 27%.

As the level of adulteration increases, these bands are shifted reaching relaxation populations between 6.85 – 7.37 ms (Table 1) for adulteration at 21% (black line) and 27% (light green line). Therefore, the relaxation time depends on the percentage of adulteration, indicating that water mobility was lower for non-adulterated than for adulterated honey. Similar plots (Figure 2 and Table 2) are obtained for W samples.

**Table 1.** $T^2$ parameters obtained in EUC honey according to the level of adulteration with ISS.

| Adulteration level | $T^2$ (ms) |
|---|---|
| **0%** | 4.76 |
| **1%** | 4.50 |
| **3%** | 4.85 |
| **9%** | 5.34 |
| **15%** | 6.34 |
| **21%** | 6.85 |
| **27%** | 7.37 |

**Table 2.** $T^2$ parameters obtained in W honey according to the level of adulteration with ISS.

| Adulteration level | $T^2$ (ms) |
|---|---|
| **0%** | 9.15 |
| **1%** | 7.64 |
| **3%** | 7.89 |
| **9%** | 8.70 |
| **15%** | 10.41 |
| **21%** | 11.01 |
| **27%** | 11.88 |

The nature of these differences in Low-field NMR (LF-NMR) signals allowed the problem to be addressed through a semi-quantitative classification approach. Fig. 3 schematically shows a flowchart of the steps involved in the development and validation of a multivariate classification model. The first step was the selection of the proper classification method. As the aim was to detect whether a sample is adulterated or not, a one-class classification model was chosen, namely OCPLS [38,39].

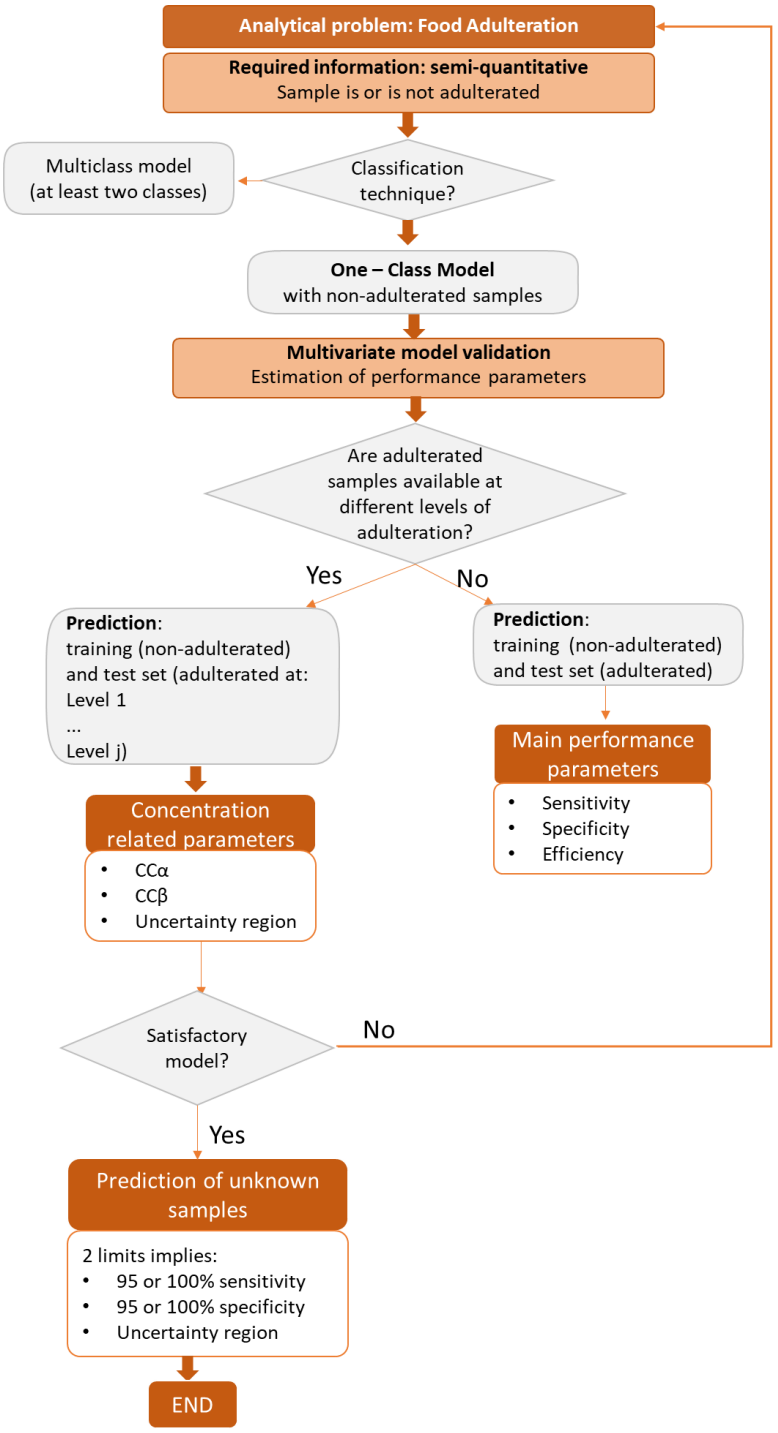



**Fig 3.** Schematic flowchart describing the semi-quantitative chemometric approach developed in this work.

Two independent one-class models were built for each of the two origins of honey samples, EUC and W. Due to the reduced number of authentic samples, models were built with all 35 non-adulterated samples. Four and three LVs were selected for EUC and W models respectively, based on the minimum cross-validation classification error.

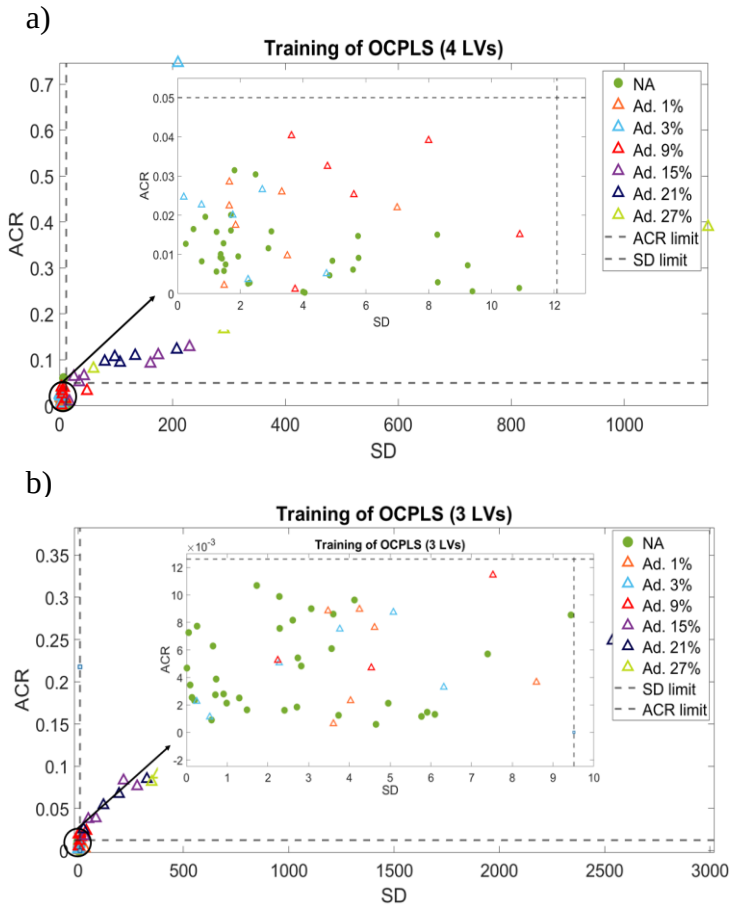



**Fig 4.** Prediction results obtained by OCPLS models for detecting adulteration with inverted sugar syrup in (a) eucalyptus and (b) wild honey. Dashed lines indicate significance levels of 0.05 for both score distances (SD) and centered model residuals (ACR). Boxes inside the plots represent a zoomed view of the acceptance region (SD and ACR values below the limits). Color and symbol codes: Full green circles for non-adulterated honey, empty orange triangles for samples adulterated at 1%, empty light blue triangles for adulterated at 3%, empty red triangles for adulterated at 9%, empty purple triangles for adulterated at 15%, empty dark blue triangles for adulterated at 21% and empty light green triangles for adulterated at 27%.

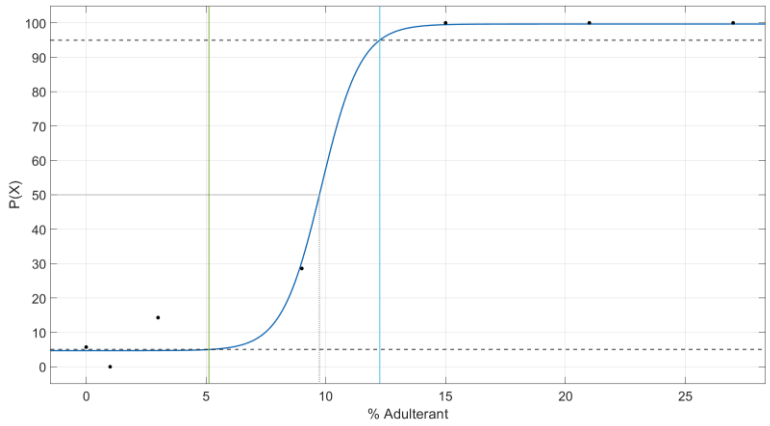
Figs. 4a and 4b show output plots, *SD* versus *ACR*, obtained in the prediction step for all adulterated and non-adulterated EUC and W samples, respectively. Almost all non-adulterated samples and most samples adulterated at low levels lay within the model limits of *SD* and *ACR*. As could be expected, for samples adulterated at different concentration levels, a gradual increase in the sample values prediction ($SD_i$ and $ACR_i$ or $d_{i,r}$, according to Eq. (1)) was observed. The higher the level of adulteration the higher the distance values (both *SD* and *ACR*).

Following the flowchart (Fig. 3), once the model is built, the next step is to validate it by estimating its performance parameters. Addressing an adulteration issue becomes challenging when there is a lack of adulterated samples at various levels. In such cases, the computation of main performance parameters, including sensitivity, specificity, and efficiency, becomes the primary approach. In case there are samples available at different adulteration levels, specificity provides overall information about the prediction of the adulterated samples (TP values). However, specificity does not provide information about the model behaviour for each level of adulteration. Useful information can be obtained related to the level of adulteration from PCC.

Fig. 5 shows PCC estimated from OCPLS models for EUC (Fig. 5a) and W (Fig. 5b) honey samples. In addition to authentic samples, 42 samples adulterated with inverted sugar syrup (7 samples for each level of adulteration, between 1% and 27%) were used for testing each model. Values were expressed as probability P(X) of detection of the whole data set. Table 3 presents the equations and parameters related to fitted PCC. From Fig. 5, $CC\alpha$ was estimated as equal to 5% for EUC honey and 2% for W honey. Below these values, samples are considered with a 95% and 90% certainty as non-adulterated samples, respectively. Samples adulterated at percentage levels close to $CC\alpha$ cannot be differentiated from the non-adulterated. $CC\beta$ was estimated at 12% for both models. So, above 12% of

adulterant content, all samples can be assigned as adulterated with 95% certainty. Above 15% of adulterant content, they can be detected with 100% certainty. Samples adulterated at concentration levels between *CCα* and *CCβ* (5-12% for EUC honey and 2-12% for W honey) fell in the uncertainty region and thus will be assigned as inconclusive and further submitted for confirmatory analysis.

a)




b)

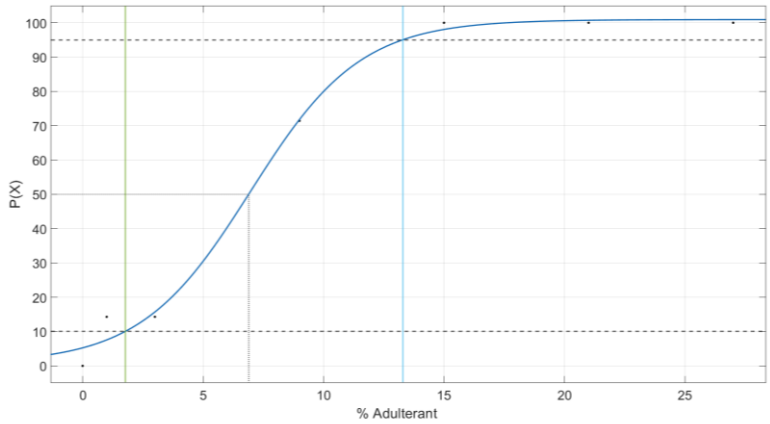



**Fig 5.** Performance characteristic curves (PCC) constructed for estimating CCα and CCβ for OCPLS models built with adulterated (a) eucalyptus and (b) wild honey samples.

**Table 3.** Fit parameters of performance characteristic curves (PCC) and semi-quantitative parameters for both OCPLS models (EUC and W).

| Parameter | EUC Model | W Model |
|---|---|---|
| **$R^2$** | 0.9948 | 0.9937 |
| **RMSE (%)** | 3.68 | 5.17 |
| **Adjusted R-square** | 0.9934 | 0.9873 |
| **Equation** | $\frac{94.98}{(1+e^{(11.88+(-1.211\cdot x))})}+4.67$ | $\frac{-100.2}{(1+e^{(-3.05+(0.44\cdot x))})}+100.9$ |
| ***CCα* (%)** | 5 | 2 |
| ***CCβ*(%)** | 12 | 12 |
| **UR (%)** | 5-12 | 2-12 |

It is important to emphasize that although $\alpha$ and $\beta$ values are usually set at 0.05 (5%), the user can define other values, which will change *CCα*, *CCβ*, and consequently the uncertainty interval. In addition to *CCα*, and *CCβ*, other probability values could be worth exploring. For instance, the adulteration level at which there are 50% of properly obtaining a positive assignation (sample is adulterated). From Fig. 5 (grey dotted lines), these values correspond to a percentage of adulteration of 10 % and 7% for EUC and W, respectively.

Following the scheme in Fig. 3, if the model is considered appropriate for the case under study, the next step is the prediction of unknown samples. Instead of just using the class limit, usually established at $d_{i,r} \leq 1$, it is suggested to define two class limits, the upper limit ($d_{upper\_lim}$) and the lower limit ($d_{lower\_lim}$). Both limits were established with 95% confidence. The $d_{upper\_lim}$ was obtained from the maximum sample distance value ($d_{i,r}$, Eq. 1), considering non-adulterated sample predictions. Similarly, $d_{lower\_lim}$ corresponds to the minimum sample distance value ($d_{i,r}$, Eq. 1) from the adulterated samples. For the EUC model, these parameters were defined

without considering samples adulterated at 1% and 3% because, during the characterization of the model with PCC, it was not possible to differentiate these levels of adulteration. Applying the same criteria, the W model was defined without considering samples adulterated at 1%.

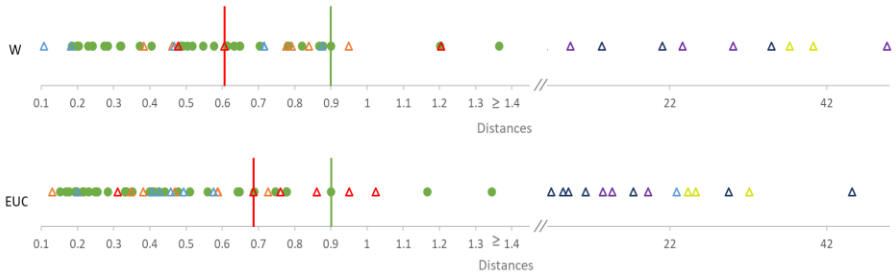



**Fig 6.** Distances of all analyzed samples to the OCPLS model. Vertical red and green solid lines represent the lower and upper limits estimated for the uncertainty region, respectively. Color and symbol codes: Full green circles for non-adulterated honey, empty orange triangles for samples adulterated at 1%, empty light blue triangles for adulterated at 3%, empty red triangles for adulterated at 9%, empty purple triangles for adulterated at 15%, empty dark blue triangles for adulterated at 21% and empty light green triangles for adulterated at 27%.

In Table 4a and Fig. 6, the two-class limits ($d_{upper\_lim}$ and $d_{lower\_lim}$) for each type of honey are shown. From those limits, the percentages of samples that were predicted as adulterated and uncertain were calculated. In both models (EUC and W), 6% of the non-adulterated samples were wrongly predicted as adulterated. Additionally, 11% of EUC samples and 37% of W samples were predicted as inconclusive, demanding further confirmatory analysis. For samples adulterated at the lowest levels (1 and 3%), predictions varied depending on the type of honey considered. For EUC honey, the model predominantly predicted them as non-adulterated. In the case of W honey, a small percentage were recognized as adulterated, and approximately half of them were classified as inconclusive. When observing the prediction of the adulterated samples at 9% in both models, it can be concluded that beyond this limit, adulterated samples were predicted with minimal error rate.

They were either identified as adulterated or categorized as inconclusive, thus requiring confirmatory analysis. For both types of honey, all adulterated samples at an adulteration level of 15% or above were correctly predicted, without error or inconclusive assignments.

To highlight the advantages of establishing two class limits, Table 4b shows the percentage of samples assigned as adulterated when applying the most common approach, with only one class limit ($d_{i,r} \leq 1$). The main differences can be seen at the medium levels of adulteration (around 9%). In the case of using two limits, the error rate of classifying adulterated samples as non-adulterated was around 14% for both types of honey. When employing a single limit ($d_{i,r} \leq 1$), this error rate was approximately 70% for EUC and 30% for W honey samples. Higher levels of adulteration (15%, 21%, and 27%) exhibited consistent behaviour for both approaches, with one and two limits, correctly assigning all these samples as adulterated. At lower levels of adulteration (1% and 3%), prediction errors decreased when using two limits, particularly for W honey.

Regarding the non-adulterated samples assigned as adulterated, the same percentage of error was observed employing one or two limits (Table 4a-b), since all class limit distances have been fixed at 5%. If there is an interest in minimizing errors in the prediction of authentic samples, $d_{lower\_lim}$ can be reduced or even set at 0%, which indicates no error in classifying a non-adulterated sample as adulterated. However, this decision will widen the uncertainty region with the adoption of a larger $d_{upper\_lim}$, leading to the drawback of requiring more samples to be submitted for confirmatory analysis.

**Table 4.** a) Uncertainty intervals and percentage of samples assigned as adulterated and as uncertain considering two class limits ($d_{lower\_lim}$ and $d_{upper\_lim}$). b) Percentage of samples assigned as adulterated considering one class limit ($d_{i,}\leq 1$) for eucalyptus (EUC) and wild (W) honey.

a)

| Model | *Uncertainty interval (d)* | | 0% (NA) | 1% | 3% | 9% | 15% | 21% | 27% |
|---|---|---|---|---|---|---|---|---|---|
| *EUC* | 0.69-0.90 | % Adulterant assignment | 6 | 0 | 14 | 43 | 100 | 100 | 100 |
| | | % Uncertain assignment | 11 | 14 | 0 | 43 | 0 | 0 | 0 |
| *W* | 0.60-0.90 | % Adulterant assignment | 6 | 29 | 14 | 71 | 100 | 100 | 100 |
| | | % Uncertain assignment | 37 | 43 | 43 | 14 | 0 | 0 | 0 |

b)

| Model | *Uncertainty interval (d)* | | 0% (NA) | 1% | 3% | 9% | 15% | 21% | 27% |
|---|---|---|---|---|---|---|---|---|---|
| *EUC* | $d\leq 1$ | % Adulterant assignment | 6 | 0 | 0 | 29 | 100 | 100 | 100 |
| *W* | | | 6 | 14 | 14 | 71 | 100 | 100 | 100 |

From the practical point of view, end-users have the flexibility to define the UR limits based on the specific nature of the problem. For instance, in the current study, if the end-user is aware that honey samples are not expected to be adulterated below 10%, adjusting the $d_{lower\_lim}$ to slightly higher values can effectively reduce the UR. Defining the two limits involves striking a balance between the percentage of samples requiring confirmatory analysis and the acceptable error percentage for the end user.

## 5. Conclusions

Two one-class models using OCPLS were developed to detect honey adulteration with inverted sugar syrup, with two different types of honey, eucalyptus and wild. These models were validated by establishing semi-quantitative parameters (*CCα*, *CCβ*, and UR) obtained from performance characteristic curves (PCC). The developed strategy allowed the examination of the model's behaviour across different adulteration levels before the prediction step. The estimate of *CCα* provided insights into which/how many samples at the lowest levels of adulteration cannot be discriminated from non-adulterated/authentic samples. The estimate of *CCβ* indicated the adulteration level from which samples will be identified as adulterated with certainty.

The proposed strategy of defining two class limits allowed the establishment of an uncertainty region, which implies determining with 100% certainty whether a sample is compliant or non-compliant. Consequently, the error rate in the prediction of unknown samples during the prediction step is reduced by an approach using only one estimated limit. For both types of honey, all samples adulterated at the highest levels (15%, 21%, and 27% w/w) were correctly assigned as adulterated, while all samples at 9% w/w, and some samples at 1-3% w/w of adulteration, were classified in the uncertainty region, thus demanding to be subjected to confirmatory analysis with an official method (e.g. IRMS, chromatography) which are a more laborious, time-consuming and high-cost technique. The approach

developed in this article can be easily adapted to the detection of other types of adulterants in honey.

**Acknowledgments**

The authors would like to thank the Brazilian government agencies CNPq, FAPEMIG, and CAPES for financial support. FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais) is particularly acknowledged for the financial support of the project "Rede Mineira de Ciências Forenses" (RED00042-16) and the project "Programa de Cooperação Acadêmica em Segurança Pública e Ciências Forenses" (Edital CAPES 16/2020). The authors also would like to thank Professor Ricardo Geraldo de Sousa, coordinator of the Polymer Science and Technology Laboratory (LCTP) at UFMG, for the lend of low-field nuclear magnetic resonance equipment.

This research study is part of a project supported through joint funding under the collaborative agreement between the Diputació de Tarragona and the Universitat Rovira i Virgili, covering the period from 2020 to 2023. This project was carried out in the year 2023, with the reference 2023PIN-DIPTA-URV01, and focuses on the "Training of pre-doctoral research staff".

**References**

[1] Codex Alimentarius Comission (2001). Revised codex standard for honey. Alinorm, 19–26, Food and Agriculture Organization of the United Nations, Rome.

[2] K.W. Se, R.A. Wahab, S.N.S. Yaacob, S.K. Ghoshal, Detection techniques for adulterants in honey: Challenges and recent trends, J. Food Compos. Anal. 80 (2019) 16-32. https://doi.org/10.1016/j.jfca.2019.04.001.

[3] F. Huang, H. Song, L. Guo, P. Guang, X. Yang, L. Li, H. Zhao, M. Yang, Detection of adulteration in Chinese honey using NIR and ATR-FTIR spectral data fusion, Spectrochim. Acta A 235 (2020) 118297. https://doi.org/10.1016/j.saa.2020.118297.

[4] P. Ciursă, D. Pauliuc, F. Dranca, S. Ropciuc, M. Oroian, Detection of honey adulterated with agave, corn, inverted sugar maple, and rice syrups using FTIR analysis, Food Control 130 (2021) 108266. https://doi.org/10.1016/j.foodcont.2021.108266.

[5] R.R. De Souza, D.D.d.S. Fernandes, P.H.G.D. Diniz, Honey authentication in terms of its adulteration with sugar syrups using UV-Vis spectroscopy and one-class classifiers, Food Chem. 365 (2021) 130467. https://doi.org/10.1016/j.foodchem.2021.130467.

[6] European Commission. (2018). Meeting Report of Technical Round Table on Honey Authentication, JRC - Geel, Belgium. https://ec.europa.eu/jrc/sites/jrcsh/files/ares1815690741_technical_round_table_on_honey_adulteration_report.pdf. (accessed 10 November 2023)

[7] Regulamento técnico de identidade e qualidade do mel (2000). Instrução Normativa Nº 11, de 20 de outubro de 2000. Ministério da Agricultura, Pecuária e Abastecimento, Brazil. Available at: https://www.gov.br/agricultura/pt-br/assuntos/defesa-agropecuaria/suasa/regulamentos-tecnicos-de-identidade-e-qualidade-de-produtos-de-origem-animal-1/IN11de2000.pdf (accessed 14 November 2023).

[8] R. Fakhlaei, J. Selamat, A. Khatib, A.F.A. Razis, R. Sukor, S. Ahmad, A.A. Babadi, The toxic impact of honey adulteration: A review, Foods, 9 (2020) 1538. https://doi.org/10.3390/foods9111538.

[9] B. Başar, D. Özdemir, Determination of honey adulteration with beet sugar and corn syrup using infrared spectroscopy and genetic-algorithm-based multivariate calibration, J. Sci. Food Agric. 98 (2018) 5616-5624. https://doi.org/10.1002/jsfa.9105.

[10] S. Li, X. Zhang, Y. Shan, D. Su, Q. Ma, R. Wen, J. Li, Qualitative and quantitative detection of honey adulterated with high-fructose corn syrup and maltose syrup by using near-infrared spectroscopy, Food Chem. 2018 (2017) 231-236. https://doi.org/10.1016/j.foodchem.2016.08.105.

[11] Q. Li, J. Zeng, L. Lin, J. Zhang, J. Zhu, L. Yao, Z. Wu, Low risk of category misdiagnosis of rice syrup adulteration in three botanical origin honey by ATR-FTIR and general model, Food Chem. 332 (2020) 127356. https://doi.org/10.1016/j.foodchem.2020.127356.

[12] W. Limm, S.R. Karunathilaka, M.M. Mossoba, Fourier transform infrared spectroscopy and chemometrics for the ràpid screening of economically motivated adulteration of honey spiked with corn or rice syrup, J. Food Prot. 86 (2023) 100054. https://doi.org/10.1016/j.jfp.2023.100054.

[13] L.G. Dias, A.R.S. Bruni, C.P. Anizelli, M. Zangirolami, P.C. Lima, L.M. Estevinho, E. Bona, Semi-quantitative discrimination of honey adulterated with cane sugar solution by an ETongue, Chem. Biodivers. 19 (2022) e2022006. https://doi.org/10.1002/cbdv.202200698.

[14] D.S. Brar, K. Pant, R. Krishnan, S. Kaur, P. Rasane, V. Nanda, S. Saxena, S. Gautam, A comprehensive review on unethical honey: Validation by emerging techniques, Food Control 145 (2023) 109482. https://doi.org/10.1016/j.foodcont.2022.109482.

[15] C. Egido, J. Saurina, S. Sentellas, O. Núñez, Honey fraud detection based on sugar syrup adulterations by HPLC-UV fingerprinting and chemometrics, Food Chem. 436 (2024) 137758. https://doi.org/10.1016/j.foodchem.2023.137758.

[16] AOAC, C-4 plant sugars in honey, Internal Standard Stable Carbon Isotope Ratio Method. Association of Official Analytical Chemists 1998 p.4.

[17] A. Guelpa, F. Marini, A. Plessis, R. Slabbert, M. Manley, Verification of authenticity and fraud detection in South African honey using NIR spectroscopy, Food Control 73 (2017) 1388-1396. https://doi.org/10.1016/j.foodcont.2016.11.002.

[18] J. Cárdenas-Escudero, D. Galán-Madruga, J.O. Cáceres, Rapid, reliable and easy-to-perform *chemometric-less* method for rice syrup adulterated honey detection using FITR-ATR, Talanta 253 (2023) 123961. https://doi.org/10.1016/j.talanta.2022.123961.

[19] M. Sahland, S. Karwita, M. Gozan, H. Hermansyah, M. Yohda, Y.J. Yoo, D.K. Pratami, Identification and classification of hone's authenticity by attenuated total reflectance Fourier-transform infrared spectroscopy and chemometric method, Vet. World 12 (2019) 1304-1310. https://doi.org/10.14202/vetworld.2019.1304-1310.

[20] D.C. Antônio, D.C.S. de Assis, B.G. Botelho, M.M. Sena, Detection of adulterations in a valuable Brazilian honey by using spectrofluorimetry and multiway classification, Food Chem. 370 (2022) 131064. https://doi.org/10.1016/j.foodchem.2021.131064.

[21] M. Orioian, S. Ropciuc, S. Paduret, Honey adulteration detection using Raman spectroscopy, Food Anal. Methods 11 (2018) 959-968. https://doi.org/10.1007/s12161-017-1072-2.

[22] K. Rachineni, V.M.R. Kakita, N.P. Awasthi, V.S. Shirke, R.V. Hosur, S.C. Shukla, Identifying type of sugar adulterants in honey: Combined application of NMR spectroscopy and supervised machine learning classification, Curr. Res. Food Sci. 5 (2022) 272-277. https://doi.org/10.1016/j.crfs.2022.01.008.

[23] R.O.R. Ribeiro, E.T. Mársico, C.S. Carneiro, M.L.G. Monteiro, C.C. Júnior, E.F.O. Jesus, Detection of honey adulteration of high fructose corn syrup by Low Field Nuclear Magnetic Resonance (LF $^{1}$H NMR), J. Food Eng. 135 (2014) 39-43. https://doi.org/10.1016/j.jfoodeng.2014.03.009.

[24] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis: taking the magic away, J. Chemom. 28 (2014) 213-225. https://doi.org/10.1002/cem.2609.

[25] M.I. López, M.P. Callao, I. Ruisánchez, A tutorial on the validation of qualitative methods: from the univariate to the multivariate approach, Anal. Chim. Acta 891 (2015) 62-72. https://doi.org/10.1016/j.aca.2015.06.032.

[26] L. Cuadros-Rodríguez, E. Pérez-Castaño, C. Ruiz-Samblá, Quality performance metrics in multivariate classification methods for qualitative analysis, TrAC Trends Anal. Chem. 80 (2016) 612-624. https://doi.org/10.1016/j.trac.2016.04.021.

[27] A.M. Jiménez-Carvelo, L. Cuadros-Rodríguez, The occurrence: a meaningful parameter to be considered in the validation of multivariate classification-based screening methods-application for authenticating virgin olive oil, Talanta 208 (2020) 120467. https://doi.org/10.1016/j.talanta.2019.120467.

[28] G. Rovira, C.S.W. Miaw, M.L.C. Martins, M.M. Sena, S.V.C. de Souza, M.P. Callao, I. Ruisánchez, One-class model with two decision thresholds for the rapid detection of cashew nuts adulteration by other nuts, Talanta 253 (2023) 123916. https://doi.org/10.1016/j.talanta.2022.123916.

[29] I. Ruisánchez, A.M. Jiménez-Carvelo, M.P. Callao, ROC curves for the optimization of one-class model parameters. A case study: authenticating extra virgin olive oil from a Catalan

protected designation of origin, Talanta 222 (2021) 121564. https://doi.org/10.1016/j.talanta.2020.121564.

[30] R. Vitale, F. Marini, C. Ruckebusch, SIMCA modeling for overlapping classes: fixed or optimized decision threshold? Anal. Chem. 90 (2018) 10738-10747. https://doi.org/10.1021/acs.analchem.8b01270.

[31] I. Ruisánchez, G. Rovira, M.P. Callao, Multivariate qualitative methodology for semi-quantitative information. A case study: adulteration of olive oil with sunflower oil, Anal. Chem. Acta 1206 (2022) 339785. https://doi.org/10.1016/j.aca.2022.339785.

[32] A.I.C. Ricardo, S.A. García, F.J.G. Bernardo, A. Ríos, R.C.R. Martín-Doimeadios, Rapid assessment of silver nanoparticle migration from food containers into food simulants using a qualitative method, Food Chem. 361 (2021) 130091. https://doi.org/10.1016/j.foodchem.2021.130091.

[33] A.I. Corps, N. Rodriguez, F.J. Guzman, R.C. Rodriguez, A. Rios, Screening-confirmation strategy for nanomaterials involving spectroscopic analytical techniques and its application to the control of silver nanoparticles in pastry samples, Spectrochim. Acta A 246 (2021) 119015. https://doi.org/10.1016/j.saa.2020.119015.

[34] Q. Yang, H. Lin, J. Ma, N. Chen, C. Zhao, D. Guo, B. Niu, Z. Zhao, X. Deng, Q. Chen, An improved POD model for fast semi-quantitative analysis of carbendazim in fruit by surface enhanced raman spectroscopy, Mol. 27 (2022) 4230. https://doi.org/10.3390/molecules27134230.

[35] M.I. López, N. Colomer, I. Ruisánchez, M.P. Callao, Validation of multivariate screening methodology. Case study: Detection of food fraud, Anal. Chim. Acta 827 (2014) 28-33. https://doi.org/10.1016/j.aca.2014.04.019.

[36] A.C.C. Fulgêncio, G.A.P. Resende, M.C.F. Teixeira, B.G. Botelho, M.M. Sena, Screening method for the rapid detection of diethylene glycol in beer based on chemometrics and portable near-infrared spectroscopy, Food Chem. 391 (2022) 133258. https://doi.org/10.1016/j.foodchem.2022.133258.

[37] A.L. Pomerantsev, D.N. Vtyurina, O.Y. Rodionova, Limit of detection in qualitative analysis: Classification Analytical Signal approach, Microchem. J. 195 (2023) 109490. https://doi.org/10.1016/j.microc.2023.109490

[38] L. Xu, M. Goodarzi, W. Shi, C.B. Cai, J.H. Jiang, A MATLAB toolbox for class modeling using one-class partial least squares (OCPLS) classifiers, Chemom. Intell. Lab. Syst. 139 (2014) 58-63. https://doi.org/10.1016/j.chemolab.2014.09.005.

[39] T.B. Moraes, L.P. Mazzero, W.S. Mendes, Inverse Laplace Transform WebApp. Available at: https://rmn-ilt.streamlit.app/ (visited last time: 16 Nov. 23)

[40] L. Xu, S.M. Yan, C.B. Cai, X.P. Yu, One-class partial least squares (OCPLS) classifier, Chemom. Intell. Lab. Syst. 126 (2013) 1-5. https://www.doi.org/10.1016/j.chemolab.2013.04.008.

[41] M.A. Fageerzada, S. Lohumi, R. Joshi, M.S. Kim, I. Baek, B-K. Cho, Non-targeted detection of adulterants in almond powder using spectroscopic techniques combined with chemometrics, Foods 7 (2020) 876, https://doi.org/10.3390/foods9070876.

[42] C.S.M. Miaw, M.M. Sena, S.V.C. de Souza, M.P. Callao, I. Ruisánchez, Detection of adulterants in grape nectars by attenuated total reflectance Fourier-transform mid-infrared spectroscopy and multivariate classification, Food Chem. 266 (2018) 254–261. https://doi.org/10.1016/j.foodchem.2018.06.006.

[43] C.S. Gondim, R.G. Junqueira, S.V.C. de Souza, I. Ruisánchez, M.P. Callao, Detection of several common adulterants in raw milk by MID-infrared spectroscopy and one-class and multi-class multivariate strategies, Food Chem. 230 (2017) 68–75. https://doi.org/10.1016/j.foodchem.2017.03.022.

[44] M. Bevilaqua, R. Bucci, A.D. Magrì, R. Nescatelli, F. Marini, Classification and class-modeling in: F. Marini (Editor), Data handling in science and Technology, Elsevier 28 (2013) 171-233. https://doi.org/10.1016/B978-0-444-59528-7.00005-3.

[45] C. Durante, R. Bro, M. Cocchi, A classification tool for N-way array based on SIMCA methodology, Chemom. Intell. Lab. Syst. 106 (2011) 73–85. https://doi.org/10.1016/j.chemolab.2010.09.004.

[46] B. Quintanilla-Casas, J. Bustamante, F. Guardiola, D.L. García-González, S. Barbieri, A. Bendini, T.G. Toschi, S. Vichi, A. Tres, Virgin olive oil volatile fingerprint and chemometrics: towards an instrumental screening tool to grade the sensory quality, LWT Food Sci. Technol. 121 (2020) 108936. https://doi.org/10.1016/j.lwt.2019.108936.

[47] D. Ballabio, F. Grisoni, R. Todeschini, Multivariate comparison of classification performance measures, Chemom. Intell. Lab. Syst. 174 (2018) 33–44. https://doi.org/10.1016/j.chemolab.2017.12.004.

[48] P. Wehling, R.A. LaBudde, S.L. Brunelle, M.T. Nelson, Probability of detection (POD) as a statistical model for the validation of qualitative methods, J. AOAC Int. 94 (2011) 335-347. https://doi.org/10.1093/jaoac/94.1.335.

[49] R. Song, P.C. Schlecht, K. Ashley, Field screening test methods: performance criteria and performance characteristics, J. Hazard. Mater. 83 (2001) 29-39. https://doi.org/10.1016/S0304-3894(00)00325-3.

[50] R.A. LaBudde, J.M. Harnly, Probability of identification: a statistical model for the validation of qualitative botanical identification methods, J. AOAC Int. 95 (2012) 273-285. https://doi.org/10.5740/jaoacint.11-266.

[51] R. Macarthur, C. von Holst, A protocol for the validation of qualitative methods of detection, Anal. Methods 4 (2012) 2744-2754. https://doi.org/10.1039/C2AY05719K.

[52] C. de S. Gondim, R.G. Junqueira, S.V.C. de Souza, M.P. Callao, I. Ruisánchez, Determining performance parameters in qualitative multivariate methods using probability of detection (POD) curves. Case study: two common milk adulterants, Talanta 168 (2017) 23-30. https://doi.org/10.1016/j.talanta.2016.12.065.

[53] E. Trullols, I. Ruisánchez, F.X. Rius, J. Huguet, Validation of qualitative methods of analysis that use control samples, TrAC Trends Anal. Chem. 24 (2005) 516-524. https://doi.org/10.1016/j.trac.2005.04.001.

# Section 3.2. Food Authentication

## Section 3.2.

The main goal of this thesis is to develop multivariate models capable of accurately predicting future samples. Consequently, our scope extends beyond model creation to encompass validation procedures that verify the model's predictive ability over time. Within this framework, it becomes crucial to continuously assess the validity of the model, particularly in the face of potential disruptions such as changes in experimental conditions or the emergence of new sources of variability in future samples. By emphasizing the importance of ongoing validation, we aim to ensure the reliability and efficacy of the predictive model in real-world applications.

When changes are detected that make the model no longer valid, they can be corrected by applying tools known as standardization techniques. These techniques involve relating data from samples measured before and after the change. By applying a mathematical function derived from this comparison to the samples in the new situation, they are transformed as if they had been measured in the original situation.

In this section, the main purpose is to apply a standardization approach to increase the usefulness of the models in a food authentication problem. The utilization of multivariate transfer techniques within the research group has served as a valuable tool over the years. Table 3.1. summarizes the scientific publications of the research group through the years.

**Table 8.** Scientific publications of the research group applying standardization techniques.

| Sample | Standardization techniques | Study | Analysis | Reference |
|---|---|---|---|---|
| Water | PDS, SWS, and CC | $Ca^{2+}$ and $Mg^{2+}$ | Quantitative | [1] |
| Water | SBC, SWS, and PDS | $Ca^{2+}$, $Mg^{2+}$, $K^{+}$ and $Na^{+}$ | Quantitative | [2] |
| Tannin sewage | SBC and PDS | Chromium | Quantitative | [3] |
| Sudan | PDS | Spices | Qualitative | [4] |

As can be observed in Table 3.1, different standardization techniques have been applied: single wavelength standardization (SWS), centering correction (CC), slope/bias correction (SBC), and piecewise direct standardization (PDS). The last one is the most used technique and has been applied in all scientific publications. Only one of the scientific publications applies multivariate transfer techniques in the field of food fraud, specifically in food adulteration.

In this section, a standardization methodology is proposed for food authentication to study and eliminate the variability of the harvest. To do it, extra virgin olive oil samples from two Catalan Protected Denominations of Origin (PDO) have been collected over different years.

## References

[1] Sales, F.; Callao, M.P.; Rius, F.X. Multivariate standardization techniques using UV-Vis data. *Chemom. Intell. Lab. Syst.* **1997**, 38, 63-73, doi: 10.1016/S0169-7439(97)00051-8.

[2] Sales, F.; Callao, M.P.; Rius, F.X. Multivariate standardization techniques on ion-selective arrays. *Anal.* **1999**, 124, 1045-1051, doi: 10.1039/A902585E.

[3] Sales, F.; Rius, A.; Callao, M.P.; Rius, F.X. Standardization of a multivariate calibration model applied to the determination of chromium in tanning sewage. *Talanta* **2000**, 52, 329-336, doi: 10.1016/S0039-9140(00)00366-0.

[4] Di Anibal, C.V.; Ruisánchez, I.; Fernández, M.; Forteza, R.; Cerdà, V.; Callao, M.P. Standardization of UV-visible data in food adulteration classification problem. *Food Chem.* **2012**, 134, 2326-2331, doi: 10.1016/j.foodchem.2012.03.100.

## Paper 5

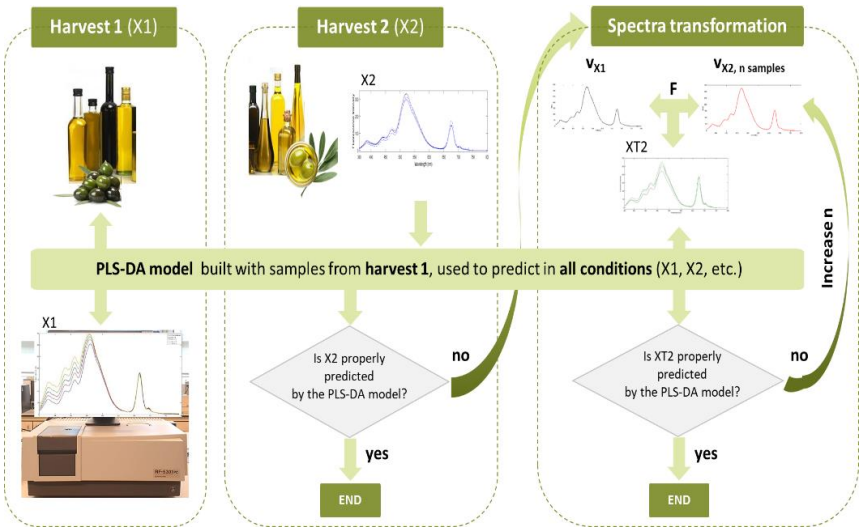
**Data standardization strategy to correct the effect of seasonality in the authentication of virgin olive oil**

Glòria Rovira, Itziar Ruisánchez, M. Pilar Callao

*Microchemical Journal,* **2023**, 195, 109520

https://doi.org/10.1016/j.microc.2023.109520

### Graphical Abstract




**Keywords:** Multivariate standardization, PLS-DA, Olive oil authentication, Multivariate screening, Food authentication, Extra-virgin olive oil.



# Data standardization strategy to correct the effect of seasonality in the authentication of virgin olive oil

Glòria Rovira, Itziar Ruisánchez*, M. Pilar Callao

Chemometrics, Qualimetric and Nanosensors Group, Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo S/n, 43007, Tarragona, Spain.

Microchemical Journal 195 (2023) 109520

**Abstract**

This study proposes a standardization strategy for dealing with seasonal variability in the authentication of extra virgin oils from the PDOs Les Garrigues and Siurana. Samples from four harvests were measured by fluorescence spectroscopy. A PLS-DA two-class model was developed and validated from samples from one of the harvests. When samples from three other harvests were predicted with the model developed, it was observed that the sensitivity and specificity were lower than when the model was validated. For the standardization process, we adapted the PDS technique to obtain the transfer function. The results obtained from the transformed spectra show that standardization is a good strategy for extending the usefulness of the models if the samples to be predicted are subject to seasonal variability.

## 1. Introduction

Olive oil is of great interest because of its nutritional value and major health benefits. This has generated considerable research interest in this field. A recent bibliometric review describes the global situation and evolution of olive oil research, considering various aspects (trends in the number of publications and distribution of them by countries and institutions, journals and research areas, analysis of authors, citations and co-citations and analysis of most relevant keywords) [1].

Compared to other types of oil, olive oil is more expensive, so it is susceptible to fraud. Recently, this issue has been the subject of an interesting paper [2] that considers various types of fraud, of which two are particularly important: adulteration with other cheaper oils and lack of authenticity due to fraudulent labeling [3].

Since authentication requires a qualitative response (yes or no), the most appropriate methodologies are those that use instrumentation to obtain the multivariate signals and multivariate classification techniques to treat them. A recent review summarizes the use of various analytical techniques coupled with multivariate data analysis to trace the geographical origin of edible oils [4]. In general terms, the main problem involved in establishing and validating a multivariate classification model is finding a large number of representative samples whose class membership is unambiguously known (training/test sets). From them, the main performance parameters, sensitivity, and specificity are estimated [5], which give information about the model's ability to predict future samples. If the future samples have the same variability as those used to build the model, the model's performance parameters are maintained. However, when the future samples contain new sources of variability, the model may not be suitable for predicting them since they do not fit it, so the model's performance parameters are no longer kept.

This might be the case in authenticating seasonal products (or their derivatives) because the samples may contain sources of variability not considered when the model was developed. Therefore, making the model robust to seasonal variability, which involves maintaining the performance parameters, is still an important challenge in the field of authentication nowadays. To deal with new sources of variability not considered in the training data set, three strategies can be followed:

1) Develop a new model for each season [6]. The main limitation of this strategy is that it cannot take advantage of historical data. In addition, the collection of sufficient training samples for each season can be time-consuming and challenging.

2) Update the model by adding samples of the new conditions, so the variability due to the season is considered. In principle, this alternative seems the most promising, since enough representative samples of the possible variations would eventually be obtained, although it may not be very effective in the first years. Additionally, increasing the sources of variability usually decreases the model's performance parameters.

3) Correct the spectra obtained in the new conditions so that they resemble the spectra of the training set used to build the multivariate model by calculating and applying a transform function. As a result, the model can be used to predict samples measured in the new conditions. It has the advantage of taking historical data into account, the multivariate model need not be built or updated, and few samples are required, the authenticity of which is unambiguously known, for standardization. This approach is implemented by chemometric tools known as standardization or transfer techniques.

The standardization strategy was initially designed to extend the applicability of multivariate calibration NIR models developed with a

master instrument to be used by secondary instruments (hence, this type of technique is also known by the alternative name of transfer methods).

The first standardization methods introduced by Wang [7] were direct standardization (DS) and piecewise direct standardization (PDS). The latter is still the most widely used. It determines a transformation matrix, F, which relates the spectra in the two conditions. Once established, the spectrum of a sample measured under the second conditions is transformed by applying matrix F so that it resembles the spectra obtained under the first conditions. This transformation allows it to be properly predicted by the model developed in the first conditions.

Most applications of transfer techniques have been developed for IR signals, although they have also been used with other types of multivariate signals, such as UV–visible data [8–10], sensor arrays [11–13], polarography [14], mass spectrometry [15], fluorescence [16–19] and HPLC [20]. They have been used above all in multivariate calibrations [10,21–23], but also to a lesser extent in multivariate classification problems and 2nd order calibrations [16,19].

Likewise, they have been used to correct for effects other than the change of instrument, for example, the effect of time [10] or temperature (Chen, Morris & Martin, 2005) [24]. Generally, to establish the transformation function F, the same sample has to be measured under both conditions, but this is not always the case and a recent review has focused on studies that determine the transfer function without the need to use the same samples in the two situations considered [25].

The goal of this paper is to propose a standardization strategy for dealing with seasonal variability in the authentication of extra virgin oils from two denominations of origin (Les Garrigues and Siurana). We worked with four data sets from four harvests that were measured using fluorescence spectroscopy.

Microchemical Journal 195 (2023) 109520

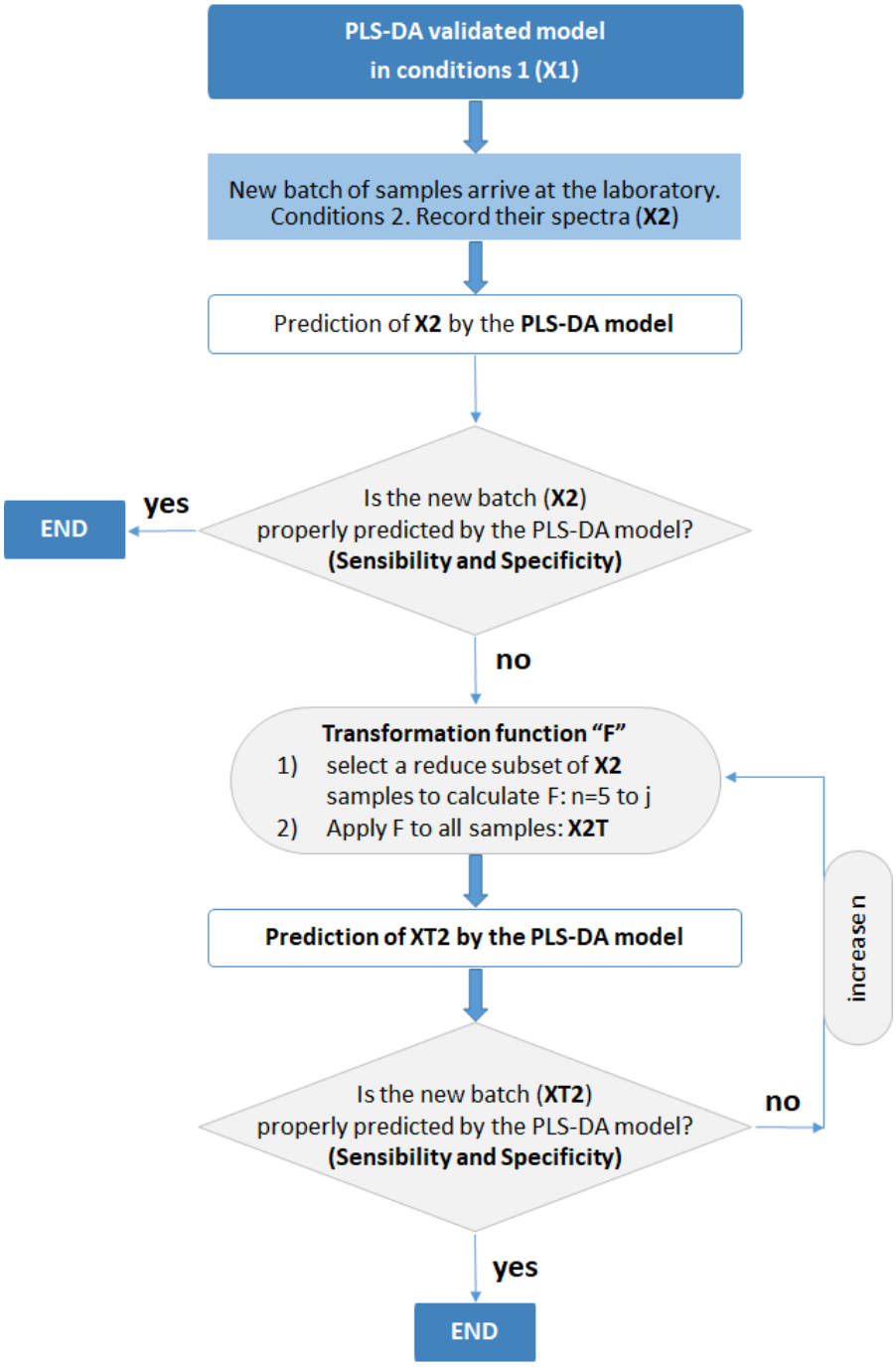



**Fig 1.** Flow chart of the standardization strategy proposed in this paper.

Therefore, in some cases the environmental conditions in each harvest were different. One of these data sets was used to develop a Partial Least Squares Discriminant Analysis (PLS-DA) two-class model for the first conditions. Since it is impossible for samples from different harvests to be the same, we propose to obtain the transfer function from the average of the spectra of both conditions. This strategy is described in Fig. 1, which schematically shows the steps involved.

Our aim is to show the potential of standardization techniques in classification from olive oil spectroscopic data problems. The strategy is simple and easy to implement in a routine laboratory and it has practical advantages such as maintaining the quality performance of the multivariate model developed.

## 2. Theoretical background

PLS-DA is a PLS regression technique adapted to a classification technique. It requires two matrices, one with independent variables (matrix **X**), which in our case are the fluorescence spectra, and the other with dependent variables (matrix **Y**), which in our case is a binary code (0 and 1) where 1 indicates sample membership and 0 does not. There is an extensive bibliography on the theoretical and practical aspects of PLS-DA. We have included references to just two recent reviews here, which is by no means exhaustive, but these two references provide many more [26,27].

To validate the classification model, the main performance parameters – sensitivity (SEN) and specificity (SPC) – [5,28] are calculated from the probabilities of the four well-known binary responses [29]: true positive (TP) and true Negative (TN), when the qualitative method rightly considers a sample to be positive or negative; and false positive (FP) and false negative (FN) when the qualitative method wrongly considers a sample to be positive or negative.

The theoretical foundations of various standardization processes can be found in the references [7,8,25]. To carry out the standardization, we propose adapting the PDS technique.

The PDS technique considers that when the experimental conditions change, the spectra may drift in any direction, both vertically (changes in sensitivity) or horizontally, which means a slight shift in the real wavelengths at which the sample absorbs. To apply PDS, a subset of samples measured in first and second conditions is available.

From these two submatrices, a multivariate regression (PCR or PLS) is established between the absorbance in first conditions at a wavelength i and a vector of absorbances around wavelength i, in second conditions.

$$a_{1i} = x_i^T f_i + f_{0i} \quad (1)$$

where $x_i^T = [a_{2(i-j)}, \ldots, a_{2(i)}, \ldots, a_{2(i+j)}]$, and the length of the vector (2j + 1) is known as window size.

The process is repeated for all the wavelengths and n vectors with f coefficients, and n values of f0 are obtained, which are grouped in the corresponding matrix **F**.

$$F = diag(\boldsymbol{f}) \quad (2)$$

In this study, the **f** is established by a linear relation between **v1** and **v2**, as shown in Equation (3):

$$v_1 = v_2 \cdot \boldsymbol{f} \quad (3)$$

being, **v1** the average of the **X1** matrix containing the spectra of the samples with which the model was developed (harvest in conditions 1). Similarly, **v2** is the average of the **X2** matrix containing the spectra of a subseries of samples from the harvests to be predicted (harvest in conditions 2). Where **f** is the transformation or transfer vector which contains as many values as the

spectra dimension (*f1, f2, f3,…fn*), with *n* being the number of wavelengths considered. From **f**, the transformation matrix F is obtained (Eq. (2)).

Eq. (4) is applied so that the spectra of the new conditions (harvest in conditions (2) resemble the spectra of the first conditions:

$$X_{T2} = X_2 \cdot F \quad (4)$$

where $\mathbf{X_{T2}}$ is the transform matrix of spectra in conditions (2) as if they had been obtained in conditions (1). Therefore, $\mathbf{X_{T2}}$ will be used in the prediction step.

## 3. Samples, instrumentation, and software

The data set consisted of 330 samples from four different harvests. Table 1 shows a summary of the number of samples from each harvest and each category, Les Garrigues (LG) and Siurana (S). All samples were supplied by the Catalan Government's Official Tasting Panel of Virgin Olive Oils of Catalonia, which confirms the authentication of the oils.

**Table 1.** Number of samples available in each harvest (conditions) and in the two DOs studied.

| Conditions | Nº of samples | Les Garrigues | Siurana |
|---|---|---|---|
| **1** | 156 | 96 | 60 |
| **2A** | 72 | 36 | 36 |
| **2B** | 58 | 19 | 39 |
| **2C** | 44 | 19 | 25 |
| **Total** | 330 | 170 | 160 |

The two-class PLS-DA model was built with samples from the harvest in conditions (1), the **X1** matrix. The 156 samples available were divided into two subsets: the training and test set. The samples were assigned to the two subsets using the Kennard-Stone algorithm [30]. The criteria determining the number of samples from each class in the training set (50 for Les Garrigues and 44 for Siurana) were that it should be similar for both classes and be around 75% of the samples available.

Microchemical Journal 195 (2023) 109520

Fluorescence analysis was carried out using a Shimadzu RF-5301PC (Shimadzu Corporation, Kyoto, Japan). The emission spectra were collected between 360 and 800 nm using an excitation wavelength of 350 nm and a slit width of 5 nm. The integration time was 0.1 nm and the wavelength was increased every 10 nm during the scanning of the spectra.

The spectra obtained were processed and the models were established using MATLAB software, version 8.0.0.783 – R2012b (Natick, MA, USA) and PLS Toolbox 7.0.2 (Eigenvector Research Inc., Wenatchee, WA, USA).

## 4. Results and discussion

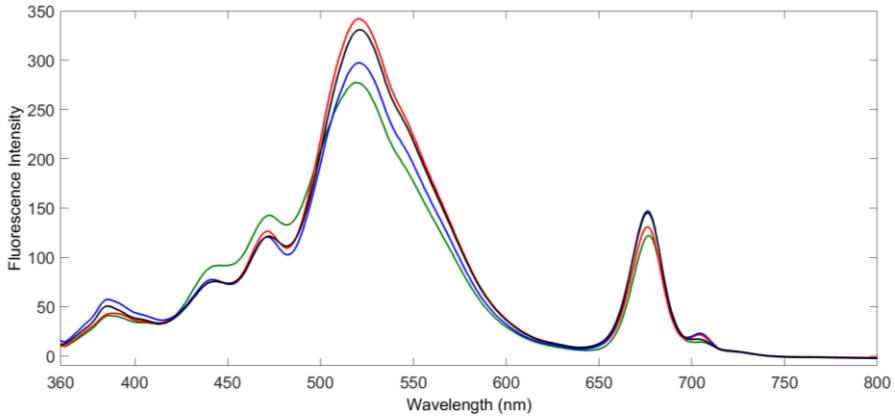
Fig. 2 shows the average spectrum of the samples for the four harvests and for Les Garrigues (Fig. 2a), and Siurana (Fig. 2b).

The major differences with the Siurana spectrum recorded in conditions (1) (black line) were observed in the spectra recorded in conditions 2B and 2C. And for the Les Garrigues spectrum, the differences were observed, above all, in conditions 2A and 2C.

Although the mean spectra in the different conditions are expected to present a certain variability, the graphs suggest that the harvest year is a factor that can present new sources of variability.

The proposed standardization strategy is summarized in Fig. 1. The first step is to establish and validate the PLS-DA model with samples from the harvest in conditions (1). Before modelling, all samples were pre-treated with baseline correction and centering. The two-class PLS-DA classification model was built for the two DOs (Les Garrigues and Siurana). The model required 5 LVs with an explained variance of around 99%. The second step is to predict the samples from other harvests (in our case, three conditions: 2A, 2B, and 2C) to check the model's ability.

(a)




(b)

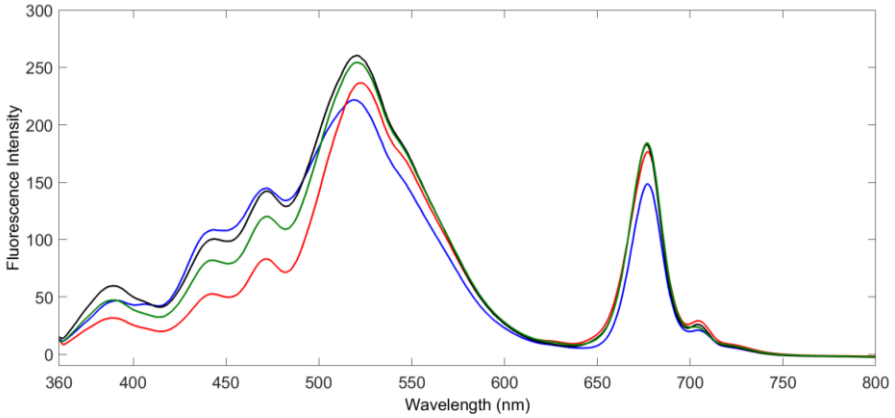



**Fig 2.** Average spectrum of the samples for the four harvests studied for a) Les Garrigues and b) Siurana. Color code: black for condition (1); blue, red and green for conditions 2A, 2B and 2C, respectively.

Table 2 shows the confusion matrix. From it, the different performance parameters (Accuracy, False and True Positive rate, False and True Negative rate, Precision, and F-Score) are obtained [5].

**Table 2.** Confusion Matrix for Les Garrigues (LG) and Siurana (S).

| Conditions | TP | FP | FN | TN |
|---|---|---|---|---|
| **1 training (LG)** | 45 | | 5 | |
| **1 training (S)** | | 5 | | 39 |
| **1 test (LG)** | 37 | | 6 | |
| **1 test (S)** | | 0 | | 14 |
| **2A (LG)** | 27 | | 9 | |
| **2A (S)** | | 0 | | 36 |
| **2B (LG)** | 17 | | 2 | |
| **2B (S)** | | 13 | | 26 |
| **2C (LG)** | 18 | | 1 | |
| **2C (S)** | | 6 | | 19 |

Table 3 shows the main performance parameters for each harvest and one class (Les Garrigues).

**Table 3.** Quality parameters obtained for the Les Garrigues class when the PLS-DA model built with the original data (conditions (1) was used to predict all data sets conditions (1) and (2).

| Conditions | Sensitivity (%) | Specificity (%) |
|---|---|---|
| **1 training** | 90 | 89 |
| **1 test** | 86 | 100 |
| **2A** | 75 | 100 |
| **2B** | 90 | 67 |
| **2C** | 95 | 76 |

Since it is a discriminant two-class model, the values for class 1 (Les Garrigues) are the same as for class 2 (Siurana) but with the sensitivity and specificity values inverted, that is to say, the sensitivity of the Siurana class is the same as the specificity of the Les Garrigues class. The opposite occurs with the results of the prediction of the Les Garrigues samples.

It can be seen that the sensitivity of the samples in conditions 2B and 2C is of the order of that of the samples in conditions (1), in both the training and the test set. But the sensitivity in conditions 2A is worse (value in bold). In

other words, the samples from the Les Garrigues vintages 2B and 2C are properly recognized by the model as authentic Les Garrigues just as the Les Garrigues samples are recognized by the model in conditions (1). But this is not the case in conditions 2A. The specificity value is notably worse than the values obtained in conditions (1) for the samples from harvests 2B and 2C (values in bold). Thus, the ability to recognize that the Siurana samples are not from Les Garrigues in these conditions was lower than in conditions (1).

On the basis of these results and following the diagram in Fig. 1, the samples that have to be transformed are vintages 2B and 2C from Siurana and vintage 2A from Les Garrigues since these new batches (X2) were not properly predicted by the PLS-DA model. The next step (Fig. 1) is to establish the transformation function "F" (Eq. (2)).

The transfer vector "**f**" is calculated by selecting a smaller subset of **X2** samples to find the average spectrum in the second condition.

To start with, a small number of samples should be selected and then, if necessary, gradually increased. For greater representativeness in the results, the selection has been carried out in triplicate, so the samples have been selected randomly and not following a pre-established criterion such as Kennard-Stone, for example. So, as a result, three independent vectors were obtained for each class and for each one of the second conditions that are to be corrected ($\mathbf{v2A_{LG,3s}}$; $\mathbf{v2B_{S,3s}}$; and $\mathbf{v2C_{S,3s}}$). By increasing the number of selected samples, new vectors were obtained under conditions (2) such as $\mathbf{v2A_{LG,\,nS,}}$ etc.

Similarly, the vector for the average values of the spectra of harvest conditions (1) was obtained for each category ($\mathbf{v1_{LG}}$, and $\mathbf{v1_{S}}$). The transfer vector "**f**" was obtained by relating **v1** with the corresponding vectors under conditions (2), for instance, $\mathbf{v1_{LG}}$ with $\mathbf{v2A_{LG,3S}}$ (equation (1). Then, the corresponding transfer matrices "**F**" were obtained by applying equation (2).

Microchemical Journal 195 (2023) 109520

Finally, equation (3) was applied to obtain the transformed data matrices $\mathbf{XT}_{2A,\,LG}$, $\mathbf{XT}_{2B,\,S,}$ and $\mathbf{XT}_{2C,\,S}$ in triplicate, one for each of the three random replicates used to select the number of samples for the standardization.

Table 4. shows the confusion matrix obtained from the standardized spectra. Table 5 shows the performance parameters obtained when predicting the transformed matrices with the PLS-DA model built in conditions (1). We will regard the results as satisfactory if the values of these parameters are similar to those obtained under conditions (1). A comparison of the results of transforming the data from the average vector with three samples with the results of no transformation (Tables 2 and 3) shows that the sensitivity of transformed condition 2A is higher than that obtained with the untransformed data but not as high as that obtained under conditions (1) in two of the three random standardizations.

**Table 4.** Confusion Matrix after standardization for Les Garrigues (LG) and Siurana (S).

| Conditions | Nº of samples | Random selection | TP | FP | FN | TN |
|---|---|---|---|---|---|---|
| **2A** | 3 | 1 | 28 | | 5 | |
| | | 2 | 33 | | 0 | |
| | | 3 | 28 | | 5 | |
| | 5 | 1 | 26 | | 5 | |
| | | 2 | 30 | | 1 | |
| | | 3 | 25 | | 6 | |
| **2B** | 3 | 1 | | 2 | | 34 |
| | | 2 | | 2 | | 34 |
| | | 3 | | 2 | | 34 |
| | 5 | 1 | | 1 | | 33 |
| | | 2 | | 1 | | 33 |
| | | 3 | | 2 | | 32 |
| **2C** | 3 | 1 | | 0 | | 22 |
| | | 2 | | 1 | | 21 |
| | | 3 | | 7 | | 15 |
| | 5 | 1 | | 0 | | 20 |
| | | 2 | | 3 | | 17 |
| | | 3 | | 1 | | 18 |

Following the procedure proposed in the scheme of Fig. 1, the same process is carried out but with five samples. The results are shown below in the same

table and the values are satisfactory. For the Siurana conditions 2B and 2C, with the standardization starting from three samples, results were satisfactory. Although the scheme presented in Fig. 1 suggests that three samples are sufficient to obtain the average vector, we also tested with five samples to observe the effect of the number of samples. It can be seen that the more samples there are, the greater the quality of the parameters obtained. But it should be pointed out that this improvement was not significant, since the results cannot be expected to be better than the results obtained in the original conditions in which the model was established (conditions (1).

**Table 5.** Prediction of the results of the transformed spectra in three different conditions (2A, 2B, and 2C) with the model established with spectra in conditions (1).

| Conditions | Nº of samples | Random selection | Sensitivity (%) | Specificity (%) |
|---|---|---|---|---|
| **2A** | 3 | 1 | 81 | |
| | | 2 | 100 | |
| | | 3 | 81 | |
| | 5 | 1 | 90 | |
| | | 2 | 94 | |
| | | 3 | 97 | |
| **2B** | 3 | 1 | | 94 |
| | | 2 | | 94 |
| | | 3 | | 94 |
| | 5 | 1 | | 97 |
| | | 2 | | 97 |
| | | 3 | | 94 |
| **2C** | 3 | 1 | | 100 |
| | | 2 | | 91 |
| | | 3 | | 97 |
| | 5 | 1 | | 100 |
| | | 2 | | 85 |
| | | 3 | | 90 |

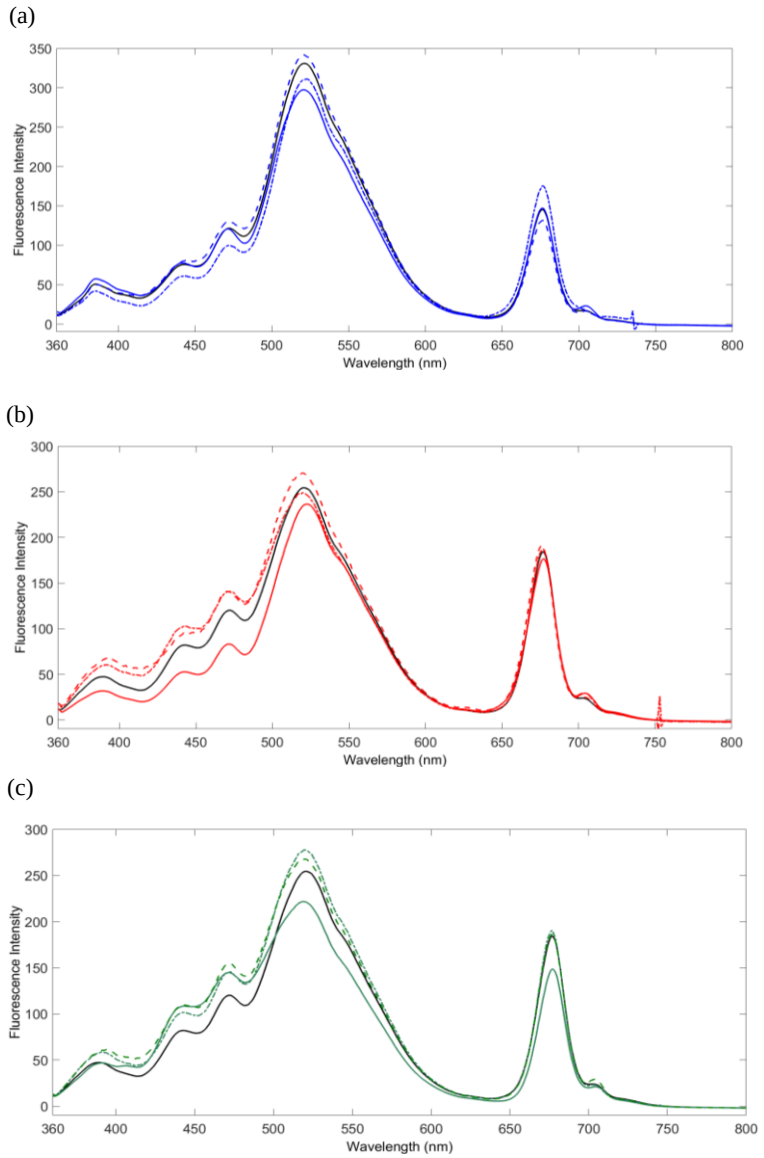



**Fig 3.** Average spectrum of samples measured in original conditions and after the standardization using different numbers of randomly selected samples: a) Les Garrigues, conditions (1) (black solid), conditions 2A (blue solid) and conditions 2A after standardization with three samples (blue dashes) and with five samples (blue dots); b) Siurana, conditions (1) (black solid), conditions 2B (red solid), and conditions 2B after standardization with three samples (red dashes) and with five samples (red dots); c) Siurana, conditions (1) (black solid), conditions 2C (green solid), and conditions 2C after standardization with three samples (green dashes) and with five samples (green dots).

Fig. 3 shows the mean spectra under conditions (1), under conditions (2), and under conditions (2) after standardization (**XT2**) with different numbers of randomly selected samples. Fig. 3a shows the spectra from Les Garrigues under conditions 1 and 2A, Fig. 3b the spectra from Siurana under conditions 1 and 2B, and Fig. 3b the spectra from Siurana under conditions 1 and 2C. In all cases, as might be expected, the transformed spectra under conditions (2) were similar to, but not exactly the same as, the spectra under conditions (1). We should also point out that there were no patterns in terms of the number of samples used. Therefore, if a small number of standardization samples provides satisfactory results in terms of the quality parameters, the process can be ended.

## 5. Conclusions

A two-class PLS-DA classification model was developed to classify virgin olive oil from two Catalan PDOs (Les Garrigues and Siurana). The model was developed with samples from one harvest using fluorescence measurements.

When samples were predicted from three other harvests, the quality parameters (sensitivity and specificity) were lower since the seasonal variability was not the same as the one used in the model. Therefore, the PDS standardization technique was adapted to extend the model's usefulness. As a general trend, the results of the standardization process are comparable to those obtained from the initial PLS-DA model.

A great advantage of the proposed standardization strategy is that can be implemented with a small number of samples. From a practical point of view, it is useful when it is difficult to obtain samples that are known to belong to a particular class.

Standardization methods have proved to be valuable tools for preserving the performance of a multivariate classification model when the samples to be

predicted are subject to new sources of variability, particularly the effect of seasonality on agricultural products and/or their derivatives.

## 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

The research work presented in this paper is the outcome of a project funded by both institutions under the collaboration framework agreement between the Diputació de Tarragona and the Universitat Rovira i Virgili for the period 2020–2023, year 2023, with the reference number 2023PIN-DIPTA-URV01: “Training of pre-doctoral research staff”. Special mention should be made of M. Angels Calvo, head of the Official Tasting Panel of Virgin Olive Oils of the Catalan Government.

## References

[1] J. Shao, X. Huang, J. Liu, D. Di, Characteristics and trends in global olive oil research: A bibliometric analysis, Int. J. Food Sci. Technol. 57 (2022) 3311–3325, https://doi.org/10.1111/ijfs.15659.

[2] J. Yan, S.W. Erasmus, M.A. Toro, H. Huang, S.M. van Ruth, Food fraud: assessing fraud vulnerability in the extra virgin olive oil supply chain, Food Control 111 (2020), 107081, https://doi.org/10.1016/j.foodcont.2019.107081.

[3] M.P. Callao, I. Ruisánchez, An overview of multivariate qualitative methods for food fraud detection, Food Control 86 (2018) 283–293, https://doi.org/10.1016/j.foodcont.2017.11.034.

[4] H.E. Tahir, M. Arslan, G.K. Mahunu, A.A. Mariod, S.B.H. Hashim, Z. Xiaobo, S. Jiyong, H.R. El-Seedi, T.H. Musa, The use of analytical techniques coupled with chemometrics for tracing the geographical origin of oils: a systematic review (2013–2020), Food Chem. 366 (2022), 130633, https://doi.org/10.1016/j.foodchem.2021.130633.

[5] M.I. López, M.P. Callao, I. Ruisánchez, A tutorial on the validation of qualitative methods: from the univariate to the multivariate approach, Anal. Chim. Acta 891 (2015) 62–72, https://doi.org/10.1016/j.aca.2015.06.032.

[6] P. Rungpichayapichet, B. Mahayothee, M. Nagle, P. Khuwijitjaru, J. Müller, Robust NIRS models for non-destructive prediction of postharvest fruit ripeness and quality in mango, Postharvest Biol. Technol. 111 (2016) 31–40, https://doi.org/10.1016/j.postharvbio.2015.07.006.

[7] Y. Wang, D.J. Veltkamp, B.R. Kowalski, Multivariate instrument standardization, Anal. Chem. 63 (1991) 2750–2756, https://doi.org/10.1021/ac00023a016.

[8] F. Sales, M.P. Callao, F.X. Rius, Multivariate standardization techniques using UV-Vis data, Chemom. Intel. Lab. Syst. 38 (1997) 63–73, https://doi.org/10.1016/S0169-7439(97)00051-8.

[9] F. Sales, M.P. Callao, F.X. Rius, Standardization of a multivariate calibration model applied to the determination of chromium in tanning sewage, Talanta 52 (2000) 329–336, https://doi.org/10.1016/S0039-9140(00)00366-0.

[10] C.V. Di Anibal, I. Ruisánchez, M. Fernández, R. Forteza, V. Cerdà, M.P. Callao, Standardization of UV–visible data in a food adulteration classification problem, Food Chem. 134 (2012) 2326–2331, https://doi.org/10.1016/j.foodchem.2012.03.100.

[11] F. Sales, M.P. Callao, F.X. Rius, Multivariate standardization techniques on ion-selective sensor arrays, Analyst 124 (1999) 1045–1051, https://doi.org/10.1039/A902585E.

[12] D. Galvan, E. Bona, D. Borsato, E. Danieli, M.H.M. Killner, Calibration transfer of partial least squares regression models between desktop nuclear magnetic resonance spectrometers, Anal. Chem. 92 (2020) 12809–12816, https://doi.org/10.1021/acs.analchem.0c00902.

[13] S. Lindner, R. Burguer, D.N. Rutledge, X.T. Do, J. Rumpf, B.W.K. Diehl, M. Schulze, Y.B. Monakhova, Is the calibration transfer of multivariate calibration models between high- and low-field NMR instruments possible? A case study of lignin molecular weight, Anal. Chem. 94 (2022) 3997–4004, https://doi.org/10.1021/acs.analchem.1c05125.

[14] A. Herrero, M.C. Ortiz, Multivariate calibration transfer applied to the routine polarographic determination of copper, lead, cadmium and zinc, Anal. Chim. Acta 348 (1997) 51–59, https://doi.org/10.1016/S0003-2670(97)00154-2.

[15] M.D. Coleman, P.J. Brewer, J.M. Smith, P.M. Harris, M.G. Clift, M.J.T. Milton, Calibration transfer strategy to compensate for instrumental drift in portable quadrupole mass spectrometers, Anal. Chim. Acta 601 (2007) 189–195, https://doi.org/10.1016/j.aca.2007.08.031.

[16] X.-D. Sun, H.-L. Wu, Y. Chen, J.-C. Chen, R.-Q. Yu, Chemometrics-assisted calibration transfer strategy for determination of three agrochemicals in environmental samples: Solving signal variation and maintaining second-order advantage, Chemom. Intel. Lab. Syst. 194 (2019), 103869, https://doi.org/10.1016/j.chemolab.2019.103869.

[17] J.C.L. Alves, R.J. Poppi, Pharmaceutical analysis in solids using front face fluorescence spectroscopy and multivariate calibration with matrix correction by piecewise direct standardization, Spectrochim. Acta A Mol. Biomol. Spectrosc. 103 (2013) 311–318, https://doi.org/10.1016/j.saa.2012.10.074.

[18] L. Nørgaard, Direct standardisation in multi wavelength fluorescence spectroscopy, Chemom. Intel. Lab. Syst. 29 (1995) 283–293, https://doi.org/10.1016/0169-7439(95)80103-G.

[19] K.D.T.M. Milanez, T.C.A. Nóbrega, D.S. Nascimento, M. Insausti, M.J.C. Pontes, Transfer of multivariate classification models applied to digital images and fluorescence spectroscopy data, Microchem. J. 133 (2017) 669–675, https://doi.org/10.1016/j.microc.2017.03.004.

[20] Y. Chen, H.-L. Wu, T. Wang, B.-B. Liu, Y.-J. Ding, R.-Q. Yu, Piecewise direct standardization assisted with second-order calibration methods to solve signal instability in high-performance liquid chromatography-diode array detection systems, J. Chromatogr. A 1667 (2022), 462851, https://doi.org/10.1016/j.chroma.2022.462851.

[21] A.J. Myles, T.A. Zimmerman, S.D. Brown, Transfer of multivariate classification models between laboratory and process near-infrared spectrometers for the discrimination of green Arabica and Robusta coffee beans, Appl. Spectrosc. 60 (2006) 1198–1203, https://doi.org/10.1366/000370206778664581.

[22] N.C. Silva, M.F. Pimentel, R.S. Honorato, M. Talhavini, A.O. Maldaner, F. A. Honorato, Classification of Brazilian and foreign gasolines adulterated with alcohol using infrared spectroscopy, Forensic Sci. Int. 253 (2015) 33–42, https://doi.org/10.1016/j.forsciint.2015.05.011.

[23] K.D.T.M. Milanez, A.C. Silva, J.E.M. Paz, E.P. Medeiros, M.J.C. Pontes, Standardization of NIR data to identify adulteration in ethanol fuel, Microchem. J. 124 (2016) 121–126, https://doi.org/10.1016/j.microc.2015.08.013.

[24] Z.-P. Chen, J. Morris, E. Martin, Correction of temperature-induced spectral variations by loading space standardization, Anal. Chem. 77 (2005) 1376–1384, https://doi.org/10.1021/ac040119g.

[25] P. Mishra, R. Nikzad-Langerodi, F. Marini, J.M. Roger, A. Biancolillo, D. N. Rutledge, S. Lohumi, Are standard sample measurements still needed to transfer multivariate calibration models between near-infrared spectrometers? The answer is not always, TrAC Trends Anal. Chem. 143 (2021), 116331, https://doi.org/10.1016/j.trac.2021.116331.

[26] A. Biancolillo, F. Marini, C. Ruckebusch, R. Vitale, Chemometric strategies for spectroscopy-based food authentication, Appl. Sci. 10 (2020) 6544, https://doi.org/10.3390/app10186544.

[27] P. Oliveri, C. Malegori, E. Mustorgi, M. Casale, Qualitative pattern recognition in chemistry: theoretical background and practical guidelines, Microchem. J. 162 (2021), 105725, https://doi.org/10.1016/j.microc.2020.105725.

[28] L. Cuadros-Rodríguez, E. Pérez-Castaño, C. Ruiz-Samblás, Quality performance metrics in multivariate classification methods for qualitative analysis, TrAC Trends Anal. Chem. 80 (2016) 612–624, https://doi.org/10.1016/j.trac.2016.04.021.

[29] S.L.R. Ellison, T. Fearn, Characterising the performance of qualitative analytical methods: statistics and terminology, TrAC Trends Anal. Chem. 24 (2005) 468–476, https://doi.org/10.1016/j.trac.2005.03.007.

[30] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148, https://doi.org/10.1080/00401706.1969.10490666.

# Chapter 4. General conclusions

# Chapter 4

This chapter provides a summary of the overall conclusions derived from the work developed in this thesis, given that the conclusions for each study were previously discussed at the end of each paper in Chapter 3.

The following conclusions can be made concerning the objectives defined in Chapter 1:

*Develop multivariate qualitative methodologies based on molecular spectroscopy measurement and classification techniques, for different types of foods in authentication and adulteration problems.*

- Four different methodologies were developed to detect food adulteration or authentication problems in different types of food:
    - Olive oil adulteration with sunflower oil using fluorescence as a spectroscopic technique coupled with PLS-DA as a classification technique.
    - Cashew nuts adulteration with Brazilian nut, pecan nut, macadamia nut, and peanut using ATR-FTIR and NIR coupled with SIMCA.
    - Honey adulteration with inverted sugar syrup using LF-NMR as analytical techniques coupled with OCPLS.
    - Geographical origin authentication of olive oil from two Catalonia PDOs using fluorescence and PLS-DA.

*In the case of adulteration, propose tools associated with semi-quantitative information, that is, analysis with a binary response (yes/no it is adulterated) for different levels of adulteration.*

- PCC curves have been applied which makes it possible to characterize the model with additional semi-quantitative performance parameters.
- The semi-quantitative parameters as CCα and CCβ, allow defining an uncertainty region (range of concentrations) where samples will

be assigned by the model as inconclusive and therefore should be submitted to a further confirmatory analysis.

- In a two-class model, a cut-off value has been set to specify the adulteration level that defines the adulterated class.

*Propose strategies to optimize the performance parameters of the developed methods.*

- Data fusion strategy has allowed an enhancement of the model performance parameters compared to those obtained from models developed using individual instrumental techniques.
- In a one-limit class strategy, ROC curves enable the determination of an optimal class limit by jointly considering the sensitivity and specificity model parameters.
- The establishment of two-class limits implies the definition of an uncertainty region. Consequently, the error rate is reduced, by sending all inconclusive samples to a confirmatory analysis.

*Propose strategies such as multivariate transfer techniques, to increase the usefulness of models developed under certain conditions* .

- Standardization techniques are effective tools for preserving the ability of the classification models when the samples to be predicted are subject to new sources of variability, such as the seasonality effects.

As a summary, we can conclude that a notable contribution of this thesis lies in its contribution to the standardization of multivariate qualitative parameters and the innovative introduction of semiquantitative performance metrics.

The methodologies developed in this thesis offer efficient strategies for controls that require rapid results. These methods can be generalized to solve other adulteration/authentication problems including different analytical approaches such as instrumental configurations and classification strategies.

Of great importance is the application of these methodologies to real-world food fraud scenarios, including nut and honey adulteration and olive oil authentication.

UNIVERSITAT
ROVIRA i VIRGILI