UNIVERSIDAD
DE GRANADA

Evaluación de la autenticidad de productos alimenticios mediante el empleo de técnicas analíticas rápidas y poco invasivas – Hacia el desarrollo de métodos analíticos 'verdes'

Tesis Doctoral

Programa de Doctorado en Química

Alejandra Arroyo Cerezo

2024

Directores:

Dr. Luis Cuadros Rodríguez

Dra. Ana M. Jiménez Carvelo

Dpto. de Química Analítica
Prof. Fermín Capitán García

Editor: Universidad de Granada. Tesis Doctorales
Autor: Alejandra Arroyo Cerezo
ISBN: 978-84-1195-667-3
URI: https://hdl.handle.net/10481/102159

Presentación

Los estudios de doctorado constituyen el tercer ciclo de la educación universitaria oficial en España, están orientados a la adquisición de las competencias y habilidades relacionadas con la investigación de calidad y su desarrollo. Éstos se encuentran regidos por el Real Decreto 99/2011, de 28 de enero, del Ministerio de Educación por el que se regulan las enseñanzas oficiales de doctorado en España. Desde su entrada en vigor, este Real Decreto ha sido parcialmente modificado en cuatro ocasiones, en julio 2013 (Real Decreto 534-2013), febrero 2015 (Real Decreto 43/2015), junio 2016 (Real Decreto 195/2016) y por último en julio 2023 (Real Decreto 576/2023). Ninguna de estas reformas modificó radicalmente las enseñanzas de doctorado, pero sí se introdujeron innovaciones parciales vinculadas sobre todo con la calidad y la internacionalización del doctorado.

En su redacción original, el Real Decreto 99/2011 ya define que:

> "La tesis doctoral consistirá en un trabajo original de investigación elaborado por el candidato en cualquier ámbito de estudio. La tesis debe capacitar al doctorando para el trabajo autónomo en el ámbito de la I+D+i."

Igualmente, en el mismo RD 99/2011 se establecen las competencias básicas y generales, capacidades y destrezas personales que deben ser adquiridas durante el periodo de ejecución de la tesis doctoral. Dichas competencias han sido recogidas en el Programa de Doctorado en Química de la Universidad de Granada (UGR), donde la presente tesis doctoral ha sido desarrollada.

Competencias básicas y generales:

CB11 - Comprensión sistemática de un campo de estudio y dominio de las habilidades y métodos de investigación relacionados con dicho campo.

CB12 - Capacidad de concebir, diseñar o crear, poner en práctica y adoptar un proceso sustancial de investigación o creación.

CB13 - Capacidad para contribuir a la ampliación de las fronteras del conocimiento a través de una investigación original.

CB14 - Capacidad de realizar un análisis crítico y de evaluación y síntesis de ideas nuevas y complejas.

CB15 - Capacidad de comunicación con la comunidad académica y científica y con la sociedad en general acerca de sus ámbitos de conocimiento en los modos e idiomas de uso habitual en su comunidad científica internacional.

CB16 - Capacidad de fomentar, en contextos académicos y profesionales, el avance científico, tecnológico, social, artístico o cultural dentro de una sociedad basada en el conocimiento.

<u>Capacidades y destrezas personales</u>:

CA01 - Desenvolverse en contextos en los que hay poca información específica.

CA02 - Encontrar las preguntas claves que hay que responder para resolver un problema complejo.

CA03 - Diseñar, crear, desarrollar y emprender proyectos novedosos e innovadores en su ámbito de conocimiento.

CA04 - Trabajar tanto en equipo como de manera autónoma en un contexto internacional o multidisciplinar.

CA05 - Integrar conocimientos, enfrentarse a la complejidad y formular juicios con información limitada.

CA06 - La crítica y defensa intelectual de soluciones.

Estas competencias han sido abordadas y trabajadas a lo largo de la ejecución de la presente tesis doctoral.

Dentro del Programa de Doctorado en Química, que pertenece a la Escuela de Doctorado en Ciencias, Tecnologías e Ingenierías (EDCTI) de la UGR, la tesis se enmarca en la línea de investigación "Metodologías de obtención de información analítica en sistemas reales".

Realización de la tesis

La presente tesis se ha llevado a cabo en el Grupo de Investigación "Análisis en Alimentación y Medio Ambiente (AnAMA)" (Código PAIDI: FQM-232), perteneciente al departamento de Química Analítica de la Facultad de Ciencias de la Universidad de Granada, bajo la supervisión del Prof. Dr. Luis Cuadros Rodríguez y la Dra. Ana M. Jiménez Carvelo.

No obstante, durante el periodo predoctoral se han llevado a cabo dos estancias en centros de investigación internacionales:

- Estancia en el "Dipartimento di Medicina Animale, Produzioni e Salute" de la "Università degli Studi di Padova", en Padua (Italia), de una duración de tres meses entre abril y julio de 2022, bajo la supervisión del Prof. Paolo Berzaghi. Los resultados derivados de la estancia se recogen en el Capítulo 3. La realización de esta estancia fue posible gracias a la financiación otorgada por la COST Action (CA19145) SensorFINT para la realización de una *Short-Term Scientific Mission,* cuyos fondos provienen de la Unión Europea.

- Estancia en el “Department of Food Science” de la “Københavns Universitet”, en Copenhague (Dinamarca), de una duración de tres meses entre septiembre y diciembre de 2023, bajo la supervisión del Prof. Dr. Rasmus Bro. La estancia fue posible gracias a la ayuda complementaria de movilidad destinada a beneficiarios FPU (Ref.: EST23/00572) concedida por el Ministerio de Ciencia, Innovación y Universidades, Gobierno de España.

La realización de la tesis doctoral ha sido posible gracias a los recursos del grupo de investigación de pertenencia. Además, ha contado con una Ayuda para la Formación del Profesorado Universitario (FPU) financiada por el Ministerio de Ciencia, Innovación y Universidades, Gobierno de España (Ref.: FPU20/04711). Asimismo, parte de los recursos, resultados y fase experimental derivan del Proyecto de Colaboración Público-Privada (Ref.: CPP2021-008672), en el marco del Plan de Recuperación, Transformación y Resiliencia, titulado “Implantación de la resonancia magnética nuclear de baja frecuenta de campo (LF-NMR) en laboratorios de control para estudios cuantitativos y de clasificación de productos alimenticios y de otros sectores industriales (NMR-CONTROL)”, del Ministerio de Ciencia e Innovación y Universidades, Gobierno de España, financiado por la Unión Europea.

Estructura de la tesis

La presente tesis doctoral comienza con una introducción recogida en el Capítulo 1 tras exponer la justificación e hipótesis planteadas. El objetivo es contextualizar los temas abordados en los siguientes capítulos, así como describir el estado del arte y las metodologías empleadas para la consecución de los objetivos perseguidos. En este capítulo de la tesis se recoge además un capítulo de libro recientemente enviado para su publicación que describe el estado del arte de la evaluación de la calidad alimentaria desde una perspectiva sostenible. Es de destacar que tanto los directores de esta tesis, como la propia doctoranda son los coordinadores de la obra aludida.

Seguidamente, cada capítulo comienza con una breve presentación del tema principal abordado y las metodologías aplicadas en detalle. Continúa con los estudios que dieron lugar a publicaciones científicas y cuya temática se enmarca en el capítulo acometido. Por último, se enumeran las comunicaciones presentadas en congresos nacionales y/o internacionales fruto de los estudios expuestos en el capítulo correspondiente.

El Capítulo 2 aborda el uso de una técnica espectrométrica rápida y no invasiva y, por tanto, no destructiva, y su aplicación a derivados lácteos, con el objeto de evaluar su potencialidad como alternativa para la autentificación de productos alimenticios, junto con el uso de la inteligencia artificial para el tratamiento y análisis de los datos adquiridos. Los experimentos que se describen en dicho capítulo se realizaron durante el primer año de tesis doctoral, dando lugar a dos artículos publicados en revistas científicas de alto impacto.

A continuación, el Capítulo 3 versa sobre la exploración del uso de técnicas espectrométricas para el desarrollo de métodos de cribado enfocados a la evaluación rápida de la calidad de aceites de oliva virgen. Da comienzo con una publicación que describe una nueva metodología propuesta para la evaluación y comparación de métodos analíticos no destructivos, y su aplicación práctica en métodos analíticos multivariable para el análisis de la calidad de aceite de oliva. A continuación, se describen dos estudios que coinciden con las dos técnicas empleadas para el fin descrito: espectrometría infrarroja y de resonancia magnética nuclear de baja frecuencia de campo. El primero se encuentra enmarcado en la estancia internacional realizada durante el primer año de tesis en Italia, mientras que el segundo es parte del proyecto de Colaboración Público-Privada mencionado (CPP2021-008672), el cual dio comienzo de forma paralela a la tesis doctoral durante el segundo año de la misma.

En el Capítulo 4 se detallan otros estudios que han formado parte de la tesis doctoral, con un enfoque destinado a mejorar el tratamiento de los datos espectrales obtenidos durante la etapa experimental, tratando de resolver problemas de la química analítica alimentaria actual mediante la ciencia conocida como quimiometría. Comienza con un capítulo de libro recientemente enviado para su publicación prevista en 2025. En él se presenta por primera vez la propuesta de inclusión del análisis de similitud como parte de la quimiometría. A continuación, se recoge un estudio que se encuentra relacionado con el Capítulo 2, dado que surge como consecuencia de la exploración de la técnica analítica sobre la que versa el segundo capítulo. El segundo estudio surge con el comienzo del proyecto mencionado (CPP2021-008672), y describe la optimización de las condiciones instrumentales para la adquisición de señales espectrales de la máxima calidad informativa, usando para ello la metodología de diseño de experimentos.

Un último capítulo pone el broche final a la tesis doctoral mediante la discusión integrada de los anteriores. Antes de finalizar con las conclusiones y perspectivas futuras planteadas, se enumeran otros estudios complementarios y tareas llevados a cabo de forma paralela durante el periodo predoctoral, que no forman parte íntegra de la tesis doctoral.

RESUMEN

Esta Tesis Doctoral desarrolla, aplica y explora el potencial del tándem espectrometría-quimiometría en el control de la calidad y autenticidad, rápido y poco invasivo, de productos alimenticios.

El objetivo principal es el de desarrollar métodos analíticos sostenibles para el análisis de alimentos, explorando para tal fin diversas técnicas espectrométricas capaces de proporcionar señales analíticas que encierren la información química necesaria. Considerando dichas señales como una huella instrumental no específica, se han explorado diversas herramientas quimiométricas para el desarrollo de métodos analíticos multivariable.

Una visión general de este enfoque ha sido revisado y puesto en valor en el Capítulo 1 de la tesis, y materializado en forma de capítulo de libro.

Se han aplicado diversas técnicas analíticas para la consecución del objetivo general, como son la espectrometría Raman en su modalidad no invasiva SORS, la espectrometría de infrarrojo cercano (NIR) y la espectrometría de resonancia magnética nuclear de baja frecuencia de campo (LF-NMR). Esto condujo a la adquisición de huellas instrumentales características de varios alimentos, concretamente: productos lácteos (margarinas y quesos en lonchas) y aceites vegetales.

El tratamiento y análisis de los datos adquiridos se llevó a cabo mediante la aplicación de diversas herramientas quimiométricas, que podrían agruparse en: (i) métodos no supervisados para el análisis exploratorio y de similitud, (ii) métodos supervisados para el desarrollo de aplicaciones cualitativas y cuantitativas, y (iii) métodos para la optimización de procesos analíticos de adquisición de señales instrumentales. Esto permitió ejecutar los estudios presentados a lo largo de la presente tesis, demostrando la relevancia de la quimiometría como parte indispensable para la resolución de los problemas propios de la química analítica.

Además, se han realizado propuestas innovadoras, dos de ellas dirigidas a la evaluación *ex-ante*, por un lado, (i) de la 'blancura' de un nuevo método analítico basado en técnicas no destructivas / no invasivas para el análisis de alimentos, y (ii) de la calidad informativa de una señal analítica antes del desarrollo del método. Asimismo, se recoge un capítulo de libro que pone de manifiesto la relevancia del análisis de similitud de señales analíticas en las diversas etapas del desarrollo de una metodología o procedimiento analítico.

Aparte de los dos capítulos de libro ya mencionados, se han publicado 6 artículos científicos en revistas internacionales de alto impacto, que recogen la mayoría de los resultados obtenidos en los estudios experimentales.

SUMMARY

This Doctoral Thesis develops, applies and explores the potential of the spectrometry-chemometrics tandem for the rapid and non-invasive control of the quality and authenticity of food products.

The main objective is to develop sustainable analytical methods for food analysis by exploring different spectrometric techniques providing analytical signals that contain the required chemical information. Considering these signals as a non-specific instrumental fingerprint, various chemometric tools have been explored for the development of multivariate analytical methods.

An overview of this approach has been reviewed and highlighted in Chapter 1 of the thesis and presented in the form of a book chapter.

Several analytical techniques have been applied to achieve the general aim, such as Raman spectrometry under the non-invasive SORS mode, near infrared spectrometry (NIR) and low-field nuclear magnetic resonance spectrometry (LF-NMR). This led to the acquisition of instrumental fingerprints characteristic of several foods, namely: dairy products (margarines and sliced cheeses) and vegetable oils.

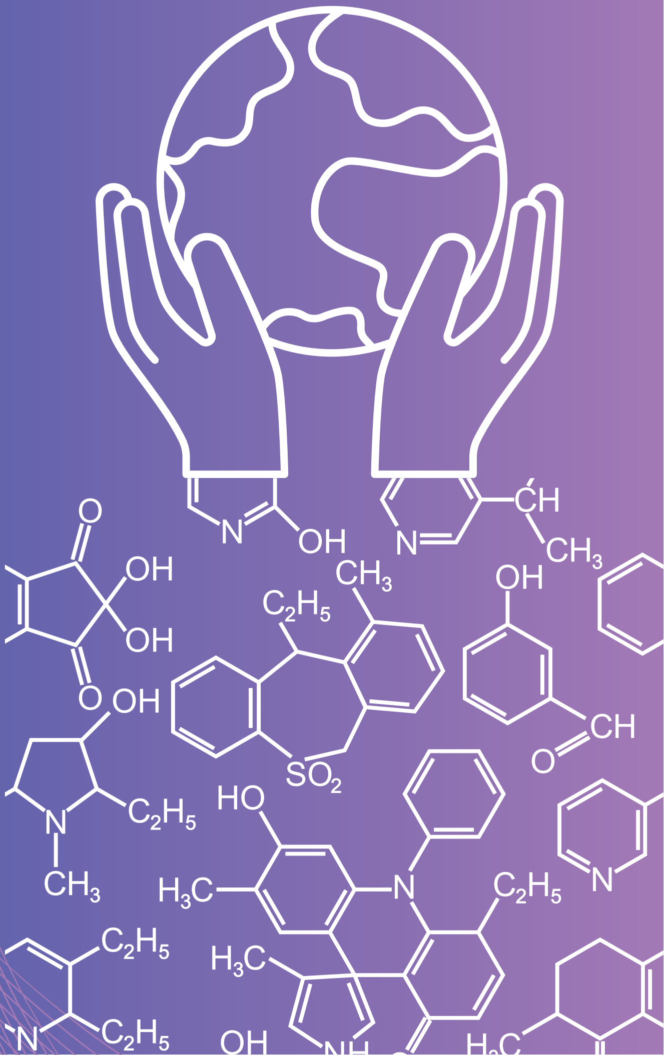
The handling and analysis of the acquired data was performed by applying various chemometric tools, which could be grouped into: (i) unsupervised methods for exploratory and similarity analysis, (ii) supervised methods for the development of qualitative and quantitative applications, and (iii) methods for process optimization related to instrumental signal acquisition. This allowed the studies presented throughout this thesis to be carried out, demonstrating the relevance of chemometrics as an essential part for solving analytical chemistry problems.

Furthermore, innovative proposals have been presented, two of them aimed at the *ex-ante* evaluation of: (i) the ‘whiteness’ of a new analytical method based on non-destructive / non-invasive techniques for food analysis, and (ii) the informative quality of an analytical signal prior to method development. In addition, a book chapter highlights the relevance of analytical signal similarity analysis throughout the various stages of the development of an analytical methodology or procedure.

In addition to the two book chapters, six scientific articles have been published in high impact international journals, which include most of the results obtained in the experimental studies.

CAPÍTULO 1

Introducción



1.1. Justificación, hipótesis y objetivos

Justificación e hipótesis

La creciente demanda de más información y que ésta sea de mayor calidad por parte de uno de los tres actores principales de la seguridad y calidad alimentaria: los consumidores, que asegure su autenticidad, la ausencia de fraudes y el cumplimiento de la legislación y normativas de calidad, confirma el interés existente por el desarrollo de métodos analíticos rápidos y de aplicación general para dicho control.

Sin embargo, actualmente la autentificación de los productos alimenticios se limita a conclusiones basadas en análisis de fracciones específicas de los alimentos mediante el uso de métodos analíticos clásicos.

Por ello, es necesario modificar los esquemas convencionales de los métodos analíticos actuales que proporcionan datos univariable, identificando y cuantificando compuestos/analitos o familias de ellos de forma individual, y sustituirlos por nuevos métodos analíticos, fundamentados en la aplicación de un enfoque multivariable no dirigido, empleando la metodología de huellas instrumentales. Dado que hasta el alimento más sencillo es una compleja matriz multicomposicional, el enfoque para investigar su composición o estructura química no debería ser otro que el multivariable.

El tándem espectrometría-quimiometría ha demostrado tener una enorme potencialidad para el desarrollo de estos nuevos métodos analíticos. La posibilidad que ofrecen las técnicas espectrométricas de llevar a cabo análisis rápidos, poco o nada invasivos, de forma directa, y algunos incluso *in situ*, sumado a la gran cantidad de información química que proporcionan, hace de este tipo de técnicas analíticas unas perfectas candidatas para ser exploradas en este sentido, manteniendo el compromiso y tendencia actual de reducir el impacto medioambiental.

Por todo ello, la presente tesis tiene como hipótesis de partida:

> Los alimentos presentan una huella instrumental característica que depende de su composición y estructura química particular, y que encierra la información necesaria para evaluar su calidad y/o autenticidad, aunque ésta no se presente de forma explícita y evidente.

Objetivos

Fundamentado en esta hipótesis, el objetivo general de la tesis es el de:

> Desarrollar un método analítico global para cada uno de los productos alimenticios analizados, que permita verificar su autenticidad y/o calidad en un único análisis rápido y poco invasivo, utilizando la huella instrumental adquirida mediante técnicas espectrométricas, y posteriormente analizada mediante herramientas quimiométricas y minería de datos.

De mismo, se derivan los siguientes objetivos específicos:

I. Obtener bancos de muestras de diferentes categorías alimentarias, representativos de la variedad encontrada en el mercado.

II. Evaluar la aplicabilidad de diversas técnicas analíticas espectrométricas avanzadas que permitan llevar a cabo análisis rápidos y poco invasivos de dichos alimentos, proporcionando su huella instrumental característica.

III. Desarrollar y validar modelos de aprendizaje automático para la clasificación de alimentos de una misma gama y/o cuantificación de parámetros característicos mediante el uso de herramientas de minería de datos.

IV. Establecer protocolos de transferencia para el control de calidad y/o autenticidad de los alimentos analizados que incluyan el uso del tándem explorado espectrometría-quimiometría.

1.2. Calidad y seguridad en la industria alimentaria

En un mundo cada vez más consciente de la importancia de la alimentación para la salud y el bienestar, asegurar la calidad y seguridad de los alimentos es una cuestión fundamental, y un desafío cada vez más complejo para esta industria, debido en gran parte a la globalización del mercado. Por un lado, la **seguridad alimentaria** (*food security*) es *la situación que existe cuando todas las personas, en todo momento, tienen acceso físico, social y económico a alimentos suficientes, inocuos y nutritivos que satisfagan sus necesidades dietéticas y sus preferencias alimentarias para llevar una vida activa y saludable* [1]*.* Tener acceso a alimentos seguros es un derecho fundamental de todas las personas. La Organización de las Naciones Unidas para la Agricultura y la Alimentación (conocida como FAO, del inglés *Food and Agriculture Organization of the United Nations*), en conjunto con otras organizaciones internacionales, entre ellas la Organización Mundial de la Salud (OMS, o WHO, del inglés *World Health Organization*), se encargan de dirigir y establecer las bases sobre cómo lograr el alcance de este derecho a nivel mundial, y así queda reflejado en el informe publicado de forma anual que recoge las actualizaciones sobre los progresos mundiales en línea con las metas 2.1 y 2.2 de los objetivos de desarrollo sostenible (ODS) [2]. La definición dada para seguridad alimentaria hace referencia al término inocuidad alimentaria (*food safety*), entendida como *la garantía de que los alimentos no causarán daño al consumidor cuando se preparen y/o consuman de acuerdo con el uso al que se destinan*. La existencia de alimentos inocuos es vital para lograr la seguridad alimentaria a nivel mundial. Existen diferentes organismos encargados de velar por la inocuidad de los alimentos, a los que se hará referencia más adelante, siempre respaldados al mismo tiempo por la FAO como recoge el documento de prioridades dentro del marco estratégico 2022-2031 [3].

Por otro lado, la **calidad** no solo se basa en la inocuidad del alimento, sino que va más allá. A nivel general, la noción de calidad puede ser definida como *el grado en el que un conjunto de características inherentes cumple con los requisitos establecidos* [4]. En otras palabras, un producto de calidad sería aquel que cumple con determinados criterios, definidos tanto en el ámbito obligatorio como en el ámbito voluntario, y que habitualmente conectan con las necesidades

[1] FAO. The State of Food Insecurity in the World 2001. Rome, Italy, 2001.

[2] FAO, IFAD, UNICEF, WFP and WHO. The State of Food Security and Nutrition in the World 2023. Rome, Italy, 2023. DOI: 10.4060/cc3017en.

[3] FAO. FAO Strategic Priorities for Food Safety within the FAO Strategic Framework 2022-2031, Rome, Italy, 2023.

[4] ISO 9000:2015. Quality management systems – Fundamentals and vocabulary; International Organization for Standardization, Geneva, Switzerland, 2015.

y expectativas del consumidor. En el campo específico de la alimentación, no es fácil encontrar una única y mundialmente aceptada definición. La calidad de los alimentos es, por tanto, un término muy amplio, que puede ser abordado desde diferentes enfoques, y que es una conjunción de diversos conceptos que deben tenerse en cuenta.

El concepto de calidad alimentaria ha sufrido cambios a lo largo de los años. Inicialmente, se entendía como la simple ausencia de defecto y fraude. Más adelante, se asimiló al concepto de calidad industrial originalmente acuñado a finales del siglo XX por la Organización Internacional de Normalización (ISO, del inglés *International Organization for Standardization*) como el *conjunto de características de una entidad que le confieren su aptitud para satisfacer las necesidades expresadas y las implícitas* [**5**], dándole al concepto un enfoque difícil de medir de manera objetiva, dado que implica satisfacer las expectativas del consumidor a diversos niveles, como el organoléptico (sensorial) y el nutricional en el caso concreto de los alimentos, que varían ampliamente de persona a persona. A partir del año 2000, y con el advenimiento de la nueva serie de normas ISO 9000 que describen los sistemas de gestión de la calidad, el término "necesidad" fue sustituido por "requisito" para dotarlo de objetividad [**6**], y dando lugar por tanto a la definición actual de calidad que se mantiene vigente en la actual versión ISO 9000:2015 [**4**].

No cabe duda de que estos dos conceptos, seguridad y calidad, no solo implican a las industrias alimentarias como actor principal que busca ofrecer un producto inocuo y de calidad a la población, sino que consumidores y autoridades también están intrínsecamente involucrados en el proceso de la búsqueda de calidad y seguridad de los alimentos [**7**]. Precisamente, la comisión del *Codex Alimentarius* se funda en 1963 de la mano de la FAO y la OMS, con el objetivo fundamental de contribuir a la seguridad, calidad y equidad de alimentos a nivel internacional, mediante el desarrollo de normas, directrices y códigos de buenas prácticas. Sin embargo, el *Codex Alimentarius* debe entenderse como una guía de referencia, y carece de carácter autoritario. Son los gobiernos y diferentes agencias reguladoras a lo largo de todo el mundo los responsables de elaborar leyes y reglamentos que aseguren la existencia de alimentos seguros y de calidad en el mercado. Cabe destacar que todo producto alimenticio que no cumpla con los requisitos de seguridad, automáticamente no lo hace con los de calidad. Sin embargo, existe la posibilidad de que un producto seguro no

[5] ISO 8402:1994. Quality management and quality assurance. Vocabulary; International Organization for Standardization, Geneva, Switzerland, 1995.

[6] ISO 9000:2000. Quality management systems. Fundamentals and vocabulary; International Organization for Standardization, Geneva, Switzerland, 2000.

[7] Alli, I. *Food Quality Assurance: Principles and Practices*; CRC Press: Florida, U.S., 2004.

cumpla con los requisitos de calidad. Así pues, la seguridad alimentaria ha estado siempre integrada como parte de la calidad alimentaria [7].

Actualmente, el concepto de calidad alimentaria abarca además otros factores que sin duda influyen y justifican el valor añadido, es decir, aquello que lo hace excelente de entre sus similares. Estos factores son la forma de producción, zona de producción, y tradiciones asociadas, entre otros. Sin embargo, las definiciones dadas de calidad no deben considerarse versiones diferentes, sino que se integran entre ellas para formar lo que podría llamarse como los tres niveles de la calidad de los alimentos: (i) ausencia de defectos, (ii) propiedades esperadas y (iii) valor añadido. En realidad, los dos primeros niveles podrían aunarse en un solo término: "calidad genérica", mientras que el valor añadido podría denominarse "calidad específica" o "calidad diferenciada" [8].

La calidad genérica conforma los requisitos mínimos exigibles a un alimento para que pueda ser comercializado, de forma que es responsabilidad de las autoridades garantizar que esto ocurra así. Llegado este punto, conviene recordar la definición de calidad alimentaria establecida específicamente en la legislación española [9], que se refiere al *conjunto de propiedades y características de un producto alimenticio o alimento relativas a las materias primas o ingredientes utilizados en su elaboración, a su naturaleza, composición, pureza, identificación, origen, y trazabilidad, así como a los procesos de elaboración, almacenamiento, envasado y comercialización utilizados y a la presentación del producto final, incluyendo su contenido efectivo y la información al consumidor final especialmente el etiquetado*.

Mientras que la calidad genérica conforma los requisitos mínimos para que un alimento pueda ser comercializado, las distinciones de la calidad específica no son de obligado cumplimiento. Ésta constituye un tipo de garantías con un valor añadido que diferencia al alimento del resto de productos similares [8]. Es la industria alimentaria la que tiene la potestad de elegir si adherirse de forma voluntaria a este tipo de calidad, y cumplir entonces con la normativa correspondiente. Actualmente, a nivel europeo se pueden agrupar en tres regímenes de calidad, que son: (i) indicadores geográficos, (ii) especialidades tradicionales garantizadas y (iii) términos de calidad facultativos [10], que se

[8] FAO. Food Safety and Quality in Europe: Aspects Concerning Quality, Nutritional Balance, Importance of Agricultural Land and Cultural Heritage (“Terroirs”). ERC/04/4, 24th FAO Regional Conference for Europe, Montpellier, France, 2004.

[9] Ley 28/2015, de 30 de julio, para la defensa de la calidad alimentaria. Boletín Oficial del Estado 182, 2015.

[10] Regulation (EU) No 2024/1143 of the European Parliament and of the Council of 11 April 2024 geographical indications for wine, spirit drinks and agricultural products, as well as traditional specialities guaranteed and optional quality terms for agricultural products. Official Journal of the European Union, 2024.

encuentran resumidos en la Tabla 1.1. A éstos se le suma la distinción de producción ecológica, recogida en el reglamento 2018/848 [11] que deroga y sustituye al anterior del año 2007.

Tabla 1.1. Regímenes de calidad diferenciada reconocidos en la Unión Europea.

Régimen de calidad	Características	Sellos
Sello geográfico	Cualidades vinculadas a la zona de producción: denominación de origen e indicación geográfica	
Especialidad tradicional	Aspectos tradicionales de elaboración del producto sin vinculación geográfica	
Términos facultativos	Productos elaborados en zonas naturales difíciles: producto de montaña y agricultura insular	
Producción ecológica	Alimentos producidos mediante buenas prácticas en materia de medio ambiente y clima	

Para poder acogerse a uno de estos regímenes de calidad diferenciada, la empresa debe seguir un pliego de condiciones que recoge los requisitos que debe cumplir el producto alimenticio en cuestión. Cabe destacar que estos son los sellos de calidad específica reconocidos a nivel europeo en la actualidad, pero existen otros a nivel nacional o autonómico, e incluso otros sellos de calidad privados [12]. Prueba de ello es el sello de producción integrada que se encuentra regulado a nivel autonómico por la Junta de Andalucía y que, a diferencia de la producción ecológica, permite el uso mínimo de ciertos abonos o pesticidas [13].

A nivel estatal, en España actualmente son cuatro los Ministerios responsables que forman parte del Plan Nacional de Control Oficial de la Cadena Alimentaria (PNCOCA) 2021-2025, que describe los sistemas de control en materia de

[11] Regulation (EU) 2018/848 of the European Parliament and of the Council of 30 May 2018 on organic production and labelling of organic products and repealing Council Regulation (EC) No 834/2007. Official Journal of the European Union L 150/1, 2018.

[12] Decreto 245/2003 de Consejería de Agricultura y Pesca, de 2 septiembre. Regula la producción integrada y su indicación en productos agrarios y sus transformados. Boletín Oficial de la Junta de Andalucía 91 de 15/5/2021 (consolidated version).

[13] Agriculture and Rural Development (European Commission). Geographical indications and quality schemes. https://agriculture.ec.europa.eu/farming_en

seguridad alimentaria, cuya existencia en cada país miembro de la Unión Europea es obligatoria [14]. Estos son el Ministerio de Agricultura, Pesca y Alimentación, Ministerio de Derechos Sociales, Consumo y Agenda 2030, Ministerio de Sanidad y Ministerio de Economía, Comercio y Empresa. Además, adscrita al primero se encuentra la Agencia Española de Seguridad Alimentaria y Nutrición (AESAN) [15], fundada en 2001, y cuyo estatuto ha sido recientemente actualizado [16]. Coordina las actuaciones de las diferentes administraciones públicas a nivel regional del país, y es el nexo comunicativo con la Autoridad Europea de Seguridad Alimentaria (EFSA, del inglés *European Food Safety Authority*), el órgano de referencia a nivel europeo, fundada en 2002 [17]. Mientras que, a nivel internacional, la Red Internacional de Autoridades de Inocuidad de los Alimentos (INFOSAN), coordinada por la FAO y la OMS, se crea para cuestiones más comunicativas en materia de seguridad alimentaria [18]. La Figura 1.1 ilustra una visión general de la organización en materia de seguridad e inocuidad alimentaria a los diferentes niveles.

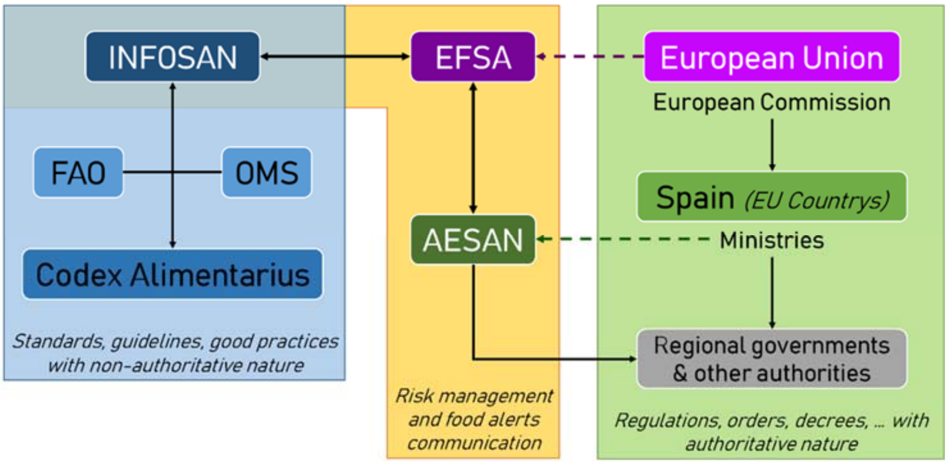

INFOSAN
EFSA
European Union
European Commission
FAO
OMS
Spain (EU Countrys)
AESAN
Ministries
Codex Alimentarius
Standards, guidelines, good practices with non-authoritative nature
Regional governments & other authorities
Risk management and food alerts communication
Regulations, orders, decrees, ... with authoritative nature

Figura 1.1. Visión general de la organización en materia de seguridad e inocuidad alimentaria a nivel internacional, europeo y nacional (España).

[14] Regulation (EU) 2017/625 of the European Parliament and of the Council of 15 March 2017 on official controls and other activities performed to ensure the application of food and feed law (...). Official Journal of the European Union L 95/1, 2017.

[15] Agencia Española de Seguridad Alimentaria y Nutrición (AESAN). https://www.aesan.gob.es

[16] Real Decreto 697/2022, de 23 de agosto, por el que se aprueba el Estatuto del Organismo Autónomo Agencia Española de Seguridad Alimentaria y Nutrición. Boletín Oficial del Estado 203, 2022.

[17] Regulation (EC) No 178/2002 of the European Parliament and of the Council of 28 January 2002 laying down the general principles and requirements of food law, establishing the European Food Safety Authority and laying down procedures in matters of food safety. Official Journal of the European Union L 31/1, 2002 (consolidated version on 2022).

[18] International Food Safety Authorities Network (INFOSAN). https://www.fao.org/food-safety/emergencies/infosan/en/

En 2020, el consejo de la Unión Europea presenta la estrategia "De la granja a la mesa" [19] como parte del Pacto Verde Europeo, con la intención de elaborar un plan de contingencia que garantice la seguridad de los alimentos, y que llegaría en 2021 [20]. Del eslogan "De la granja a la mesa" es posible inferir el hecho de que todos los aspectos relacionados con los alimentos, incluido la calidad y seguridad, abarcan todas las fases por las que atraviesa un producto alimenticio durante su desarrollo: desde las materias primas hasta su consumo, y por lo tanto, calidad y seguridad deben ser consideradas en todas sus etapas (concepción, diseño y desarrollo) [7], siguiendo el principio rector ya enunciado en el Libro Blanco sobre Seguridad Alimentaria, publicado en el año 2000 [21].

Una vez establecida la participación indispensable de los tres actores principales (administración, industria y consumidores) en lo que respecta a seguridad y calidad alimentaria, es de suponer la importancia de la existencia de una interacción comunicativa entre las partes. En este contexto, los ya mencionados reglamentos, decretos, órdenes y/o directrices, son el instrumento principal de comunicación desde las autoridades hacia las industrias. A ellos se les suma la ejecución de auditorías e inspecciones, que podría considerarse un instrumento de comunicación bidireccional entre autoridades e industrias, dado que es la forma de comunicar por parte de la empresa que se está siguiendo la normativa mediante la presentación de informes de cumplimiento.

Si bien la herramienta comunicativa que conecta a los tres pilares es el **etiquetado**, dado que su presencia en los alimentos se hace obligatoria mediante normativas (autoridades), para que el consumidor tenga a su disposición toda información detallada del producto por parte de la industria productora. Se entiende como etiquetado *las menciones, indicaciones, marcas de fábrica o comerciales, dibujos o signos relacionados con un alimento y que figuren en cualquier envase, documento, rótulo, etiqueta, faja o collarín, que acompañen o se refieran a dicho alimento* [22]. Todo etiquetado de producto alimenticio debe especificar como mínimo: el nombre, la presencia de alérgenos si los hubiese, cantidad neta, y fecha de consumo preferente o caducidad, según corresponda. Además, por medio de este instrumento es cómo la industria comunica al

[19] European Commission, A Farm to Fork Strategy for a fair, healthy and environmentally-friendly food system, COM/2020/381 final, Document 52020DC0381, Brussels, 2020.

[20] Contingency plan for ensuring food supply and food security in times of crisis, COM/2021/689 final. European Commission, Brussels, 2021.

[21] White paper on food safety of 12 January 2000, COM/99/0719 final. Commission of the European Communities, Brussels, 2000.

[22] Regulation (EU) No 1169/2011 of the European Parliament and of the Council of 25 October 2011 on the provision of food information to consumers. Official Journal of the European Union L 304/18, 2011 (consolidated version on 2018).

consumidor la información que éste necesita para escoger un producto y cumplir con sus requisitos de calidad, como es: el listado de ingredientes, la información nutricional, o los sellos de calidad diferenciada. El objetivo principal de las auditorías e inspecciones es corroborar que la industria está siendo fiel a lo que declara en el etiquetado, o en otras palabras, autentificar los alimentos. La autentificación de un alimento es el proceso de verificación que confirma que el producto es genuino y sus características reales coinciden con las alegaciones asociadas a su etiqueta en cuanto a calidad, origen, producción y técnicas de procesado utilizadas [23,24].

Una herramienta indispensable para la autentificación de alimentos, y cuya existencia es además obligatoria por normativa, es la trazabilidad, que queda definida como *la posibilidad de encontrar y seguir el rastro, a través de todas las etapas de producción, transformación y distribución, de un alimento, un pienso, un animal destinado a la producción de alimentos o una sustancia destinados a ser incorporados en alimentos o piensos o con probabilidad de serlo* [17]. La trazabilidad alimentaria está destinada a proteger por una parte a los consumidores del fraude alimentario, y por otra a las empresas del sector de una posible competencia desleal por parte de otras [23]. Asegurar un correcto sistema de trazabilidad beneficia a los tres actores principales de la calidad y seguridad alimentaria, ya que permite a los consumidores tener información sobre el producto, a las empresas una mayor facilidad de registrar y corregir posibles fallos y a las autoridades una mayor eficacia a la hora de gestionar incidencias o alertas alimentarias.

Cuando un alimento no cumple con los criterios de autenticidad, se considera un caso de **fraude alimentario**. El fraude alimentario podría definirse como *el engaño intencionado por parte de la industria hacia el consumidor en lo que a calidad y/o contenido del alimento se refiere, normalmente con un interés económico de la empresa* [25,26]. Los tipos de fraude son adulteración, sustitución, dilución, manipulación, simulación, falsificación y tergiversación, aunque pueden existir más clases de prácticas fraudulentas. Algunos tipos de fraude pueden incluso poner en riesgo la salud de los consumidores, especialmente aquellos casos en los que se adultera el alimento con la adición

[23] El Sheikha, A.F. Food authentication: Introduction, techniques, and prospects. In *Food Authentication and Traceability*, Galanakis, C.M., Ed.; Elsevier: London, U.K., 2021; pp 1-34.

[24] Morin, J.F.; Lees, M. Definition of food fraud and food authenticity. In *FoodIntegrity Handbook: A guide to food authenticity issues and analytical solutions*; Morin, J.F.; Lees, M., Eds.; Eurofins Analytics France: Nantes, France, 2018.

[25] Discussion paper on food integrity and food authenticity, CX/FICS 18/24/7. Codex Alimentarius Commission, Brisbane, Australia, 2018.

[26] FAO, Food fraud - Intention, detection and management. In *Food safety technical toolkit for Asia and the Pacific No 5*. Bangkok, Thailand, 2021.

de sustancias no aptas. De hecho, se han llegado a producir a lo largo de la historia intoxicaciones masivas, con la consecuente muerte de personas, debido a la adulteración de alimentos. Es el ejemplo de aquel fatídico episodio vivido en España en 1981, cuando alrededor de 4000 muertes que se sucedieron con el paso de los años, fueron presuntamente causadas por la venta de aceite para freír que contenía aceite de colza desnaturalizado para usos industriales tratado con anilina, una sustancia tóxica que además provocó más de 25000 enfermos, y consecuencias en éstos que aún sufren a día de hoy. Años después, el Tribunal Supremo se declararía responsable civil de dicha tragedia, precisamente por no haber llevado a cabo protocolos de actuación que lo evitasen [27]. Este ha sido el mayor caso de intoxicación alimentaria vivido en España, aunque no el único. Afortunadamente, en los últimos años ha aumentado la capacidad de detección, y la comunicación es cada vez más fluida, de forma que se agiliza el proceso de evitar que estos productos lleguen al consumidor y produzca daños en la salud de los consumidores. No obstante, la prevención debería ser el foco de atención [28].

Si el fraude se basa en un etiquetado incorrecto, uno de los casos más comunes, y este implica la no declaración de ingredientes potencialmente alérgenos o perjudiciales en algunas enfermedades, estaría poniéndose en riesgo también la salud de los consumidores, como la de personas alérgicas a los frutos secos, o bien diabéticos cuya dosis de insulina depende de la cantidad de azúcar consumida, y no se declara correctamente la presencia de azúcares o almidones [29]. Otros casos de fraude no relacionados con la salud, se vinculan con la calidad del alimento. Por ejemplo, al hacer uso de una denominación asociada a sellos de calidad diferenciada en un alimento que realmente no está siguiendo la normativa, y por tanto no podría denominarse como tal, con el objetivo de venderlo a un precio superior. Lo mismo ocurre con el etiquetado que induce a confusión en el consumidor por la terminología o incluso imágenes utilizadas en el diseño del envase para captar su atención. Un ejemplo claro de esta perversión fue el uso extensivo del término "bio" en marcas de alimentos, hasta que fue regulado por la Unión Europea en 2018 [11]. En definitiva, es inmensa la tipología de fraude alimentario que se puede dar hoy día en los alimentos.

[27] Aceite de colza, el origen del mayor caso de intoxicación alimentaria en la historia de España. ABC Historia, September 6, 2021, updated October 19, 2021.

[28] Spink, J.W. *Food Fraud Prevention. Introduction, Implementation, and Management,* Doyle, M.P., Ed.; Food Microbiology and Food Safety (FMFS); Springer: New York, U.S., 2019.

[29] Sammut, J.; Gopi, K; Saintilan, N.; Mazumder, D. Facing the challenges of food fraud in the global food system. In *Food Authentication and Traceability*, Galanakis, C.M., Ed.; Elsevier: London, U.K., 2021; pp 35-64.

Mensualmente la Comunidad Europea publica un informe de los casos de fraude detectados a nivel mundial, que puede consultarse en la página web del Centro de Conocimiento sobre el Fraude y la Calidad de los Alimentos (KC-FFQ, del inglés *Knowledge Center for Food Fraud and Quality*) [**30**].

Otra gran consecuencia del fraude es que merma la confianza del consumidor en la información que se le proporciona, lo que deriva en una demanda de más información, y que ésta sea de calidad. Por tanto, la detección del fraude es un reto cada vez mayor, dado que las prácticas fraudulentas son progresivamente más sofisticadas [**23**]. No son pocos los esfuerzos para evitarlo a nivel mundial, mediante normativas, políticas, etc. Pero desgraciadamente, la sofisticación de las técnicas empleadas para cometer fraude alimentario parece avanzar a mayor velocidad que aquellas desarrolladas para detectarlo a tiempo. Por ello, en literatura científica en la última década ha habido una creciente tendencia de publicaciones relacionadas con la calidad, seguridad y autentificación de alimentos [**31**].

1.3. Química analítica alimentaria: pasado, presente y futuro

La disciplina que provee las herramientas necesarias para ser aplicadas en la detección del fraude, autentificación alimentaria y/o control de calidad alimentaria, es la **química analítica**. La IUPAC (del inglés *International Union of Pure and Applied Chemistry*) la define como *la disciplina científica que desarrolla y aplica estrategias, instrumentos y procedimientos para obtener información sobre la composición y naturaleza de la materia en el espacio y el tiempo* [**32**]. Aunque asegurar una adecuada calidad y seguridad de los alimentos es una tarea sin duda multidisciplinar, es responsabilidad del químico analítico desarrollar nuevas metodologías que persigan los objetivos citados en el campo del análisis de alimentos.

La química analítica es una de las seis grandes disciplinas en las que se divide la química: química inorgánica, química orgánica, bioquímica, químico-física, ingeniería química y química analítica. Esta disciplina también es conocida como aquella que estudia la composición química de la materia de forma cualitativa y cuantitativa. Para obtener la información buscada se mide una propiedad del

[**30**] Knowledge Centre for Food Fraud and Quality (KC-FFQ), European Commission. https://knowledge4policy.ec.europa.eu/food-fraud-quality_en.

[**31**] Kyrgiakos, L.S.; et al. The food fraud landscape: A brief review of food safety and authenticity. *Proceedings* **2024**, *94*, 6. 10.3390/proceedings2024094006.

[**32**] Hibbert, D.B.; Korte, E.H.; Örnemark, U. Metrological and quality concepts in analytical chemistry (IUPAC Recommendations 2021). *Pure & Appl. Chem*. **2021**, *93*, 997-1048. DOI: 10.1515/pac-2019-0819.

material en estudio que esté directa o indirectamente relacionada. Sin embargo, esta última definición podría acuñarse más bien al término "análisis químico", es decir, es la aplicación del conocimiento a lo que sería el análisis rutinario. Pero la química analítica no es la simple aplicación de lo que ya existe, sino que trata de mejorar, desarrollar y ampliar los métodos analíticos existentes con el objetivo de extraer la información química que caracteriza la materia bajo estudio [**33**].

El primer paso imprescindible en cualquier análisis químico es la selección de la metodología de análisis que va a ser aplicada, cuya decisión depende de un conjunto de factores, entre los que se encuentran: la naturaleza del material a medir, el resultado perseguido, el nivel de exactitud que se pretende alcanzar, o el coste económico, entre otros. Los siguientes pasos que forman parte del proceso analítico estarán definidos por esta elección. La Figura 1.2 muestra un diagrama de flujo general de los pasos a seguir que forman parte de cualquier proceso analítico, tras la elección de la metodología de análisis [**34**]. Cabe destacar que no todo proceso analítico está compuesto por todas las etapas, sino que, en función de las características de la metodología, es posible que no sea necesario llevar a cabo alguna/s de ellas.

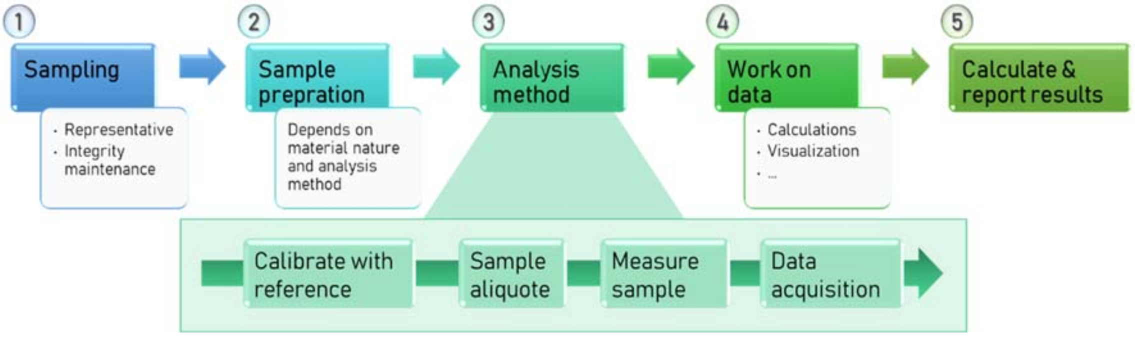


Figura 1.2. Esquema general del proceso analítico. *Adaptada de [**35**].*

Comenzando por la primera, esta es una etapa crítica, pues de ella dependerán la calidad y veracidad de los resultados. La medida de una propiedad química de un material habitualmente se realiza en una pequeña porción de éste, conocida como muestra, la cual debe ser representativa y reunir todas las características,

33. Harvey, D.; Introduction to Analytical Chemistry. In *Analytical Chemistry 2.1*; LibreTexts, 2021, pp 1-12.

34. Nielsen, S.S. Introduction to Food Analysis. In *Food Analysis,* 5th ed.; Ismail, B.P.; Nielsen, S.S., Eds.; FSTS; Springer: Cham, Switzerland, 2017, pp 3-16.

35. Kenkel, J. Introduction to Analytical Science. In *Analytical Chemistry for Technicians*, 4th ed.; CRC Press: Florida, U.S., 2014, pp 1-18.

para poder finalmente deducir unos resultados que sean aplicables al material completo bajo estudio [36]. En el caso de la química analítica aplicada al análisis de alimentos, la materia bajo estudio es un multicomposicional y, por tanto, compleja en comparación con otros campos de aplicación como el farmacéutico o el petrolero. Debe elaborarse y seguirse un plan de muestreo adaptado a las características específicas del producto alimenticio y la fase en la que se encuentre. Dicho plan dependerá también del objetivo perseguido con el análisis [37]. Además, también es importante mantener la integridad de la muestra, evitando cualquier tipo de contaminación o condiciones que puedan poner en riesgo su composición original. A continuación, la muestra representativa debe ser acondicionada de acuerdo con el método que va a utilizarse para el análisis. Esto implica a menudo la ejecución de una serie de pasos que se agrupan en la etapa denominada preparación de muestra. Generalmente, estos pasos a seguir pueden involucrar procesos de digestión, homogeneización, disolución, extracción, filtración, separación, …, entre otros [34,37].

Una vez la muestra se encuentra en el estado óptimo para medir la magnitud analítica previamente seleccionada, llega el momento de aplicar el método de análisis seleccionado. Ésta es una compleja etapa compuesta por diferentes subetapas dentro de la misma, que de forma general se pueden ver resumidas en la figura anterior, pero cuya manera de llevarse a cabo es muy diversa según el método de análisis escogido.

Una vez adquiridos los datos, éstos deben ser tratados matemáticamente, ya que son pocos los casos en los que en química analítica se consiga llegar a un resultado interpretable sobre el material bajo estudio simplemente midiendo de forma directa la propiedad química de interés y sin aplicar ningún calculo. En otras palabras, la información química de interés se adquiere de forma indirecta, midiendo una propiedad química que esté relacionada con la información que da respuesta al problema analítico. Por lo tanto, para poder llegar a la última etapa y dictaminar los resultados, es necesario estudiar, tratar y analizar los datos y transformarlos en la información que se busca [38]. Además, para llegar a unas conclusiones válidas, es necesario realizar el análisis de la misma muestra un número determinado de veces, que a menudo suele ser tres o cinco repeticiones. Por lo que el resultado final será, normalmente, el promedio de las

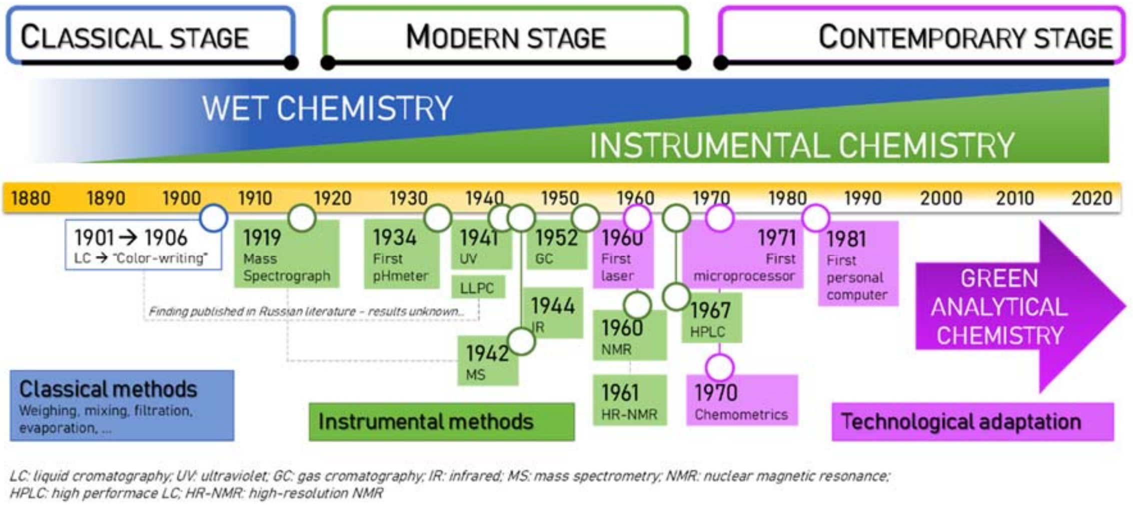
[36]. Kenkel, J. Sampling and Sample Preparation. In *Analytical Chemistry for Technicians*, 4th ed.; CRC Press: Florida, U.S., 2014, pp 19-48.

[37]. Morawicki, R.O. Sampling and Sample Preparation. In *Food Analysis,* 5th ed.; Ismail, B.P.; Nielsen, S.S., Eds.; FSTS; Springer: Cham, Switzerland, 2017, pp 61-75.

[38]. Cuadros-Rodríguez, L.; Jiménez-Carvelo, A.M.; Andrade, J.M. A multivariate approach to Analytical Chemistry. In *Problem-Oriented Analytical Chemistry Driven by Chemometrics,* Cuadros-Rodríguez, L.; Jiménez-Carvelo, A.M.; Andrade, J.M., Eds.; Elsevier: Amsterdam, Netherlands, 2025 [*submitted, in process*].

repeticiones realizadas, tras llevar a cabo un debido análisis estadístico de cómo se comportan los datos, y si es necesario descartar el resultado de alguna de las repeticiones, con el objetivo de eliminar posibles errores que se hayan producido durante el proceso analítico [39].

El trabajo del químico analítico en el análisis de alimentos ha sido durante años, es, y seguirá siendo, integral. Está presente en las diversas etapas por las que atraviesa un alimento durante su desarrollo, "de la granja a la mesa", abarcando todos los niveles: aseguramiento de la calidad, seguridad, estabilidad, propiedades organolépticas y nutricionales, etc., tal y como recoge la revisión titulada "Cien años de avances en el análisis de alimentos" de McGorrin [40]. Este artículo hace un repaso de la historia de la química analítica de los alimentos, dividiéndola en diferentes etapas marcadas por diversos hitos. La química alimentaria se ha beneficiado históricamente de los avances tecnológicos que se han desarrollado para otras industrias, adaptándolas a los problemas analíticos de la misma. La Figura 1.3 recoge una visión a lo largo del tiempo de la química analítica alimentaria.



LC: liquid cromatography; UV: ultraviolet; GC: gas cromatography; IR: infrared; MS: mass spectrometry; NMR: nuclear magnetic resonance; HPLC: high performace LC; HR-NMR: high-resolution NMR

Figura 1.3. Cronología de la química analítica de los alimentos: las etapas y los hitos más relevantes [40].

No es fácil establecer la fecha exacta en la que se comienza a aplicar la química analítica al análisis de alimentos, aunque muchos autores coinciden en que no fue hasta finales del siglo XIX. En sus inicios, los métodos de análisis empleados forman parte de lo que se conoce como *wet chemistry* (química analítica

[39] Harvey, D.; Evaluating Analytical Data. In *Analytical Chemistry 2.1*, LibreTexts, 2023, pp 63-152.

[40] McGorrin, R.J. One Hundred Years of Progress in Food Analysis. *J. Agric. Food Chem.* **2009**, *57*, 8076-8088. DOI: 10.1021/jf900189s.

húmeda, o química analítica en disolución). Se refiere a lo que aparece en la Figura 1.3 como "métodos clásicos", aquellos que hacen uso de una combinación de diferentes procedimientos analíticos básicos como mezclado, filtración, evaporación, destilación, etc., llevados a cabo de forma manual y sin el uso de instrumentación [35]. El objetivo es obtener la información química de interés mediante la determinación de volumen o valoración, mientras que el resultado se deduce "a ojo" del químico analítico. Un ejemplo de método clásico basado en la química húmeda es el Kjeldahl, que se convirtió en método estándar reconocido para la determinación de proteínas mediante análisis volumétrico, y que a día de hoy aún se sigue usando en algunos laboratorios. En general, son métodos que proporcionan resultados con alto grado de exactitud y precisión, a veces sujetos a la subjetividad del analista, pero requieren largos periodos de tiempo para su ejecución [40].

Este análisis clásico pronto comenzó a ser sustituido a mediados del siglo XX por el análisis instrumental, gracias al desarrollo de equipos e instrumentos y el avance tecnológico, consecuencia del surgimiento de nuevas técnicas analíticas. La química analítica instrumental incluye el uso de instrumentación de alta tecnología, proporciona análisis más rápidos y prácticos, ya que permite analizar un número mayor de muestras. Un hito importante en esta nueva etapa fue la comercialización del primer pHmetro Beckman modelo G en 1937, diseñado por Arnold Beckman en 1934, que por primera vez eliminaba la subjetividad a la hora de realizar un análisis químico [41]. Otro hito importante en la química analítica alimentaria fue el primer espectrómetro ultravioleta (UV) modelo Beckman DU en 1941 [42], seguido poco después por el primer espectrómetro de infrarrojo (IR).

Aunque sin duda, la técnica instrumental más significativa para el análisis de alimentos ha sido la cromatografía, tanto de líquidos (LC) como de gases (GC), y los diferentes hitos en torno a ella, como el desarrollo de la cromatografía de líquidos de altas prestaciones (HPLC) [43]. Numerosos métodos analíticos que existen hoy en día en la química analítica alimentaria se basan en la aplicación de técnicas cromatográficas.

Otras técnicas instrumentales relevantes en la química analítica alimentaria que han ido mejorando con el paso de los años, también fueron introducidas en el siglo XX, como la espectrometría de masas (MS) o la resonancia magnética

[41] Moore, C.E.; Jaselskis, B. The pH meter, a product of technological crossovers. *Bull. Hist. Chem.* **1998**, *21*, 32-37.

[42] Schmidt, W. Introduction to Optical Spectroscopy. In *Optical Spectroscopy in Chemistry and Life Science*; Wiley-VCH: Weinheim, Germany, 2005, pp 1-11.

[43] Cifuentes, A. Food Analysis: Present, Future, and Foodomics. *ISRN Anal. Chem.* **2012**, 201607. DOI: 10.5402/2012/801607.

nuclear (NMR). En definitiva, el desarrollo de estas técnicas ha hecho que el análisis químico por vía húmeda sea sustituido de forma progresiva a lo largo de los años por el análisis instrumental, aunque en la actualidad algunos métodos de análisis aún se basan en el primero [40].

Tras la aparición de nuevos instrumentos durante el siglo XX, fue posible desarrollar nuevos métodos y procedimientos analíticos aplicados al análisis de alimentos. Posteriormente, la adaptación tecnológica dio paso a la etapa contemporánea. Según McGorrin, tres avances tecnológicos fueron significativos en el terreno de la química analítica alimentaria [40]. Estos fueron el desarrollo del primer microprocesador, del primer láser, y del primer ordenador portátil. A su vez, fue desembocando en una mejora de los equipos existentes, mediante el empleo de técnicas de forma secuencial (p.ej., GC-MS) o en tándem (LC-MS/MS). Estos avances permitieron mejorar la sensibilidad y los límites de detección de los métodos de análisis, incluso posibilitar la detección de ciertos analitos que antes no era posible, con su consecuente contribución a la mejora de la seguridad alimentaria, ya que se pudieron incorporar límites máximos para la presencia de compuestos cancerígenos o tóxicos en los alimentos, los cuales anteriormente eran indetectables.

Los avances tecnológicos también han permitido automatizar alguno de los pasos que forman parte de los métodos analíticos clásicos, ejemplo de ello son la extracción con líquidos presurizados, con fluidos supercríticos, o la extracción en fase sólida, como parte de la preparación de muestra (etapa 2 en la Figura 1.2). Todo ello derivó en una mejora de la productividad del laboratorio, y reducción de la duración del proceso analítico [40]. En cuanto al método de análisis (etapa 3), la tendencia en el siglo XXI persigue el objetivo de desarrollar alternativas más rápidas, automatizadas, reduciendo la necesidad de preparación de muestra y disponer de instrumentación de tamaño reducido. Esta búsqueda persigue a su vez una mejora del rendimiento, y todo ello sin comprometer la fiabilidad y calidad de los resultados en términos de sensibilidad, precisión, o límite de detección.

En relación con estos objetivos descritos, en la década de los 90 aparecen los primeros "métodos analíticos limpios" o "métodos analíticos verdes", abriendo la puerta a la **química analítica verde** (GAC, del inglés *green analytical chemistry*) [44]. El enfoque verde de la química analítica surge como consecuencia no solo del impacto medioambiental que supone el uso de reactivos y solventes tóxicos y los desechos químicos generados, sino también de los riesgos a los que se expone el analista en según qué tipo de análisis. La

[44] de la Guardia, M.; Garrigues, S. The concept of Green Analytical Chemistry. In *Handbook of Green Analytical Chemistry*.; de la Guardia, M.; Garrigues, S., Eds.; John Wiley & Sons: Chichester, U.K., 2012, pp 3-16.

GAC nace con el objetivo principal de buscar alternativas más respetuosas con el medio ambiente, minimizando el uso de disolventes tóxicos y los residuos generados [45].

La GAC se puede considerar una parte de la química verde [46], o bien la aplicación de un enfoque analítico a la química verde [44]. Para hacer más fácil su entendimiento, se formularon los 12 principios de la química verde, que más tarde se han adaptado y reformulado para la GAC en la propuesta publicada por el grupo de Namieśnik en Polonia [46], que se enumeran brevemente a continuación:

1. Técnica analítica directa
2. Tamaño de muestra mínimo
3. Medida *in situ*
4. Reducir el uso de disolventes
5. Métodos automatizados y miniaturizados
6. Evitar la formación de derivados
7. Evitar residuos analíticos
8. Métodos multiparamétricos
9. Minimizar el uso de energía
10. Reactivos de fuentes renovables
11. Eliminar agentes tóxicos
12. Aumentar seguridad del analista

Estos principios se basan en la prevención de efectos contaminantes, mediante la reducción parcial o total del uso de disolventes y reactivos, especialmente si son agentes tóxicos, evitando la generación de residuos, y priorizando el ahorro energético [44]. Por otro lado, algunos de los principios hacen referencia a la técnica analítica empleada para el análisis. Los principios 1, 3, 5 y 8 se relacionan con la tendencia perseguida en las últimas décadas en el análisis instrumental.

Las técnicas analíticas que cumplen los principios de este enfoque verde son mayoritariamente las técnicas espectrométricas. El hecho de que sea posible analizar la muestra de forma directa hace que reúna prácticamente todos los principios de la GAC: no requiere preparación de muestra, lo que implica un ahorro de tiempo, y a su vez del uso de disolventes, con el consecuente aumento de la seguridad del analista [47]. A pesar de que las técnicas cromatográficas están a la orden del día en el análisis de alimentos, por las innumerables ventajas en cuanto a los resultados que proporciona, cada vez son más los estudios que se publican en literatura científica enfocados a desarrollar

[45] Armenta, S.; Garrigues, S.; de la Guardia, M. Green Analytical Chemistry. *Trends Anal. Chem.* **2008**, *27*, 497-511. DOI: 10.1016/j.trac.2008.05.003.

[46] Gałuszka, A., Migaszewski, Z., & Namieśnik, J. The 12 principles of green analytical chemistry and the SIGNIFICANCE mnemonic of green analytical practices. *Trends Anal. Chem.* **2013**, *50*, 78-84. DOI: 10.1016/j.trac.2013.04.010.

[47] Armenta, S.; de la Guardia, M. Direct analysis of samples. In *Handbook of Green Analytical Chemistry.*; de la Guardia, M.; Garrigues, S., Eds.; John Wiley & Sons: Chichester, U.K., 2012, pp 85-102.

métodos analíticos basados en técnicas que sigan los principios de la GAC [**48**,**49**,**50**].

Algunas de las técnicas espectrométricas también permiten realizar medidas *in situ* del material en estudio. Además, los avances tecnológicos han posibilitado desarrollar instrumentación de menor tamaño (miniaturización). Si bien es cierto que la cualidad más destacada de estas técnicas es el hecho de ser no destructivas o poco invasivas, lo cual permite que la muestra pueda ser medida de nuevo, ya que su composición no se ha visto alterada. Gracias a la cantidad de información química que proporcionan, es posible desarrollar métodos multiparamétricos, cumpliendo así también con el principio número 8 [**51**], que será abordado con más detalle en el siguiente apartado.

Existe una ampla gama de técnicas espectrométricas que difieren en su fundamento, por lo que la selección debe adecuarse al objetivo perseguido (información buscada, tipo y estado de la muestra, calidad del resultado esperada, etc.). La espectroscopia estudia la interacción entre la radiación electromagnética y la materia. Según las recomendaciones de la IUPAC [**52**], los términos *espectroscopia*, *espectrometría*, *espectrofotometría* o *espectrografía* se usan para describir la medida de la intensidad irradiada por el material como función de la frecuencia o longitud de onda. Sin embargo, existe una diferencia destacable entre los términos **espectroscopia** y **espectrometría**. Mientras que la primera hace referencia al estudio de la interacción entre la radiación electromagnética (REM) y la materia con fines de caracterización o estudio del comportamiento de la materia desde un punto de vista teórico, la espectrometría es la aplicación práctica de la medida de la intensidad de la REM para generar resultados relacionados con propiedades de los sistemas materiales. Dada la aplicación que se le da en el ámbito analítico, y en concreto en la presente tesis, se utiliza el término espectrometría para referirse al uso aplicado de técnicas espectroscópicas.

48. Ballesteros-Vivas, D.; et al. Green food analysis: Current trends and perspectives. *Curr. Opin. Green Sustain. Chem.* **2021**, 31, 100522. DOI: 10.1016/j.cogsc.2021.100522.

49. Hassoun, A.; et al. Food quality 4.0: From traditional approaches to digitalized automated analysis. *J. Food Eng.* **2023**, 337, 111216. DOI: 10.1016/j.jfoodeng.2022.111216.

50. Aslam, R.; et al. A systematic account of food adulteration and recent trends in the non-destructive analysis of food fraud detection. *J. Food Meas. Char.* **2023**, 17, 3094-3114. DOI: 10.1007/s11694-023-01846-3.

51. Jiménez-Carvelo, A.M.; Arroyo-Cerezo, A.; Cuadros-Rodríguez, L. Evaluating the whiteness of spectroscopy-based non-destructive analytical methods – Application to food analytical control. *Trends Anal. Chem.* **2024**, *170*, 117463. DOI: 10.1016/j.trac.2023.117463.

52. Infante, H.G.; et al. Glossary of methods and terms used in analytical spectroscopy (IUPAC Recommendations 2019). *Pure Appl. Chem.* **2021**, *93*, 647–776. DOI: 10.1515/pac-2019-0203.

De forma general, estas técnicas se basan en irradiar el material bajo estudio para provocar una perturbación o alteración de su estado energético natural (a nivel molecular o atómico), que dependerá de la composición química del mismo. Esta perturbación da lugar a una respuesta por parte del material, que el equipo recoge en forma de **señal**, lo que se conoce como **espectro**. De forma que la señal contiene información química sobre la composición del material, que puede ser utilizada de forma cualitativa y cuantitativa para conocer su composición química. Mención aparte merece la espectrometría de masas (MS), ya que su fundamento es diferente y no se basa en la REM, aunque la señal analítica resultante se conoce también como espectro. La MS proporciona información sobre la masa de los iones en los que se ioniza el material bajo estudio [53].

La presente tesis se ha centrado en la aplicación de técnicas espectrométricas ópticas, por lo que éstas serán las protagonistas de los siguientes párrafos. Las técnicas espectrométricas ópticas pueden clasificarse atendiendo (i) al tipo de radiación empleada, (ii) a la transición energética provocada o (iii) a la interacción detectada [54]. En función de la fuente utilizada (i) para emitir la radiación que provocará la perturbación del estado natural del material, se pueden clasificar según la región del espectro electromagnético en la que trabaje la fuente de luz incidente. Cada tipo de REM provoca una perturbación diferente en el estado natural del material bajo estudio, que viene definida por el nivel energético al que se dé la transición de energía. La Figura 1.4 muestra las regiones del espectro electromagnético y el tipo de transición que genera (ii) [55].

La energía de la REM incidente es absorbida, lo que provoca un salto entre algunos de los diferentes niveles de energía en los átomos y/o moléculas que componen el material desde su estado fundamental (el estado de menor energía) a un nivel superior o estado excitado. A nivel molecular, el salto de energía puede provenir de una transición electrónica, vibracional o rotacional, mientras que a nivel atómico solo se describen transiciones de tipo electrónico [56]. En la química analítica alimentaria requiere mucha más atención la espectrometría a nivel molecular. Sin embargo, la espectrometría de NMR

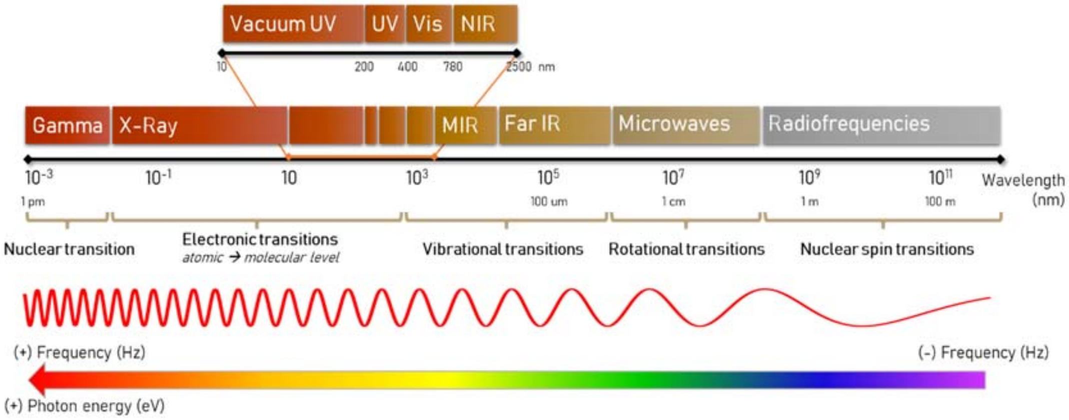
[53] Kenkel, J. Mass Spectrometry. In *Analytical Chemistry for Technicians*, 4th ed.; CRC Press: Florida, U.S., 2014, pp 371-381.

[54] Penner, M.H. Basic Principles of Spectroscopy. In *Food Analysis,* 5th ed.; Ismail, B.P.; Nielsen, S.S., Eds.; FSTS; Springer: Cham, Switzerland, 2017, pp 79-88.

[55] Pavia, D.L.; Lampman, G.M.; Kriz, G.S.; Vyvyan, J.R. Infrared Spectroscopy. In *Introduction to Spectroscopy*, 4th ed.; Brooks/Cole: Cham, U.S., 2009, pp 15-104.

[56] Skoog, D.A.; Holler, F.J.; Crouch, S.R. An Introduction to Spectrometric Methods. In *Principles of Instrumental Analysis*; 7th ed.; Cengage Learning: Boston, U.S., 2017, pp 119-147.

estudia otro tipo de niveles de energía distinto a los descritos, que son observables solo bajo la aplicación de un campo magnético externo, se trata de transiciones de espín nuclear. El fundamento de esta técnica será descrito con más detalle en el Capítulo 3.



UV = ultraviolet; Vis = visible; IR = infrared; NIR = near IR; MIR = Mid IR.

Figura 1.4. Regiones del espectro electromagnético y tipo de transición generada.

Por último, cuando el material bajo estudio recibe la REM, parte de la intensidad incidente (I_0) se absorbe (I_A), y como consecuencia se produce la transición a un estado excitado. Además, pueden darse distintos fenómenos, como que el material refleje (I_R), disperse (I_S) transmita (I_T) o emita (I_E) parte o toda la intensidad recibida [57]. La Figura 1.5 muestra un esquema de los fenómenos que pueden tener lugar durante la interacción entre la REM y la materia. Las técnicas espectrométricas pueden clasificarse en función del tipo de fenómeno que detecte (iii), lo cual caracteriza el tipo de fuente emisiva y el detector del equipo empleado.

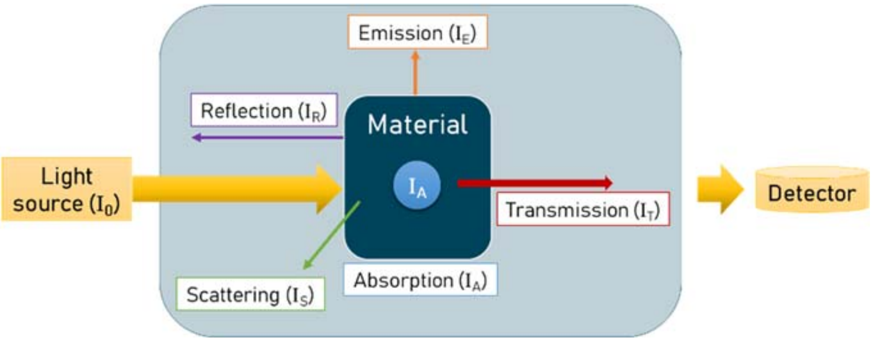


Figura 1.5. Fenómenos ópticos causados por la interacción entre radiación y materia.

[57] Robinson, J.W.; Frame, E.M.S.; Frame II, G.M. Introduction to Spectroscopy. In *Instrumental Analytical Chemistry. An Introduction*; CRC Press: Florida, U.S., 2021, pp 61-100.

Atendiendo a esta última clasificación, se pueden diferenciar:

- Espectrometría de absorción
- Espectrometría de transmisión
- Espectrometría de reflexión
- Espectrometría de dispersión
- Espectrometría de emisión [**58**]

El fenómeno de absorción puede ser medido en todas las regiones del espectro electromagnético (véase Figura 1.4), tanto a nivel atómico como molecular, y es el fundamento de toda técnica espectrométrica óptica [**56**]. Su uso permite calcular la absorbancia del material a una longitud de onda dada. El nombre de la técnica viene dado por la región del espectro de la REM incidente [**54**]. La información proporcionada por los fenómenos de transmisión y reflexión (transmitancia y reflectancia, respectivamente), dependiendo de que el material sea o no transparente a la REM incidente, está ligada a la de absorción. La primera es la cantidad de radiación que atraviesa el material y que éste no ha absorbido, y la reflexión es la cantidad de luz que refleja la superficie del material tras ser irradiado. Por lo tanto, son dos modos diferentes, transmisión y reflexión, en los que puede registrarse el espectro de absorción. Ambos modos son comunes especialmente en la región del infrarrojo [**58**]. Por otro lado, la dispersión tiene lugar cuando la radiación incidente en la región del IR produce una emisión de la misma frecuencia o a una frecuencia inferior o superior por parte del material. En el primer caso, se conoce como dispersión de Rayleigh, mientras que el segundo se conoce como dispersión inelástica o dispersión Raman [**59**].

Por último, la espectrometría de emisión hace referencia a fenómenos de emisión o de fluorescencia atómica, o bien de luminiscencia (fluorescencia o fosforescencia) a nivel molecular. El que ocurra una u otra emisión luminiscente dependerá de la naturaleza química del material excitado. Tiene lugar en moléculas que tras la excitación en la región UV-Vis emiten parte de la energía cuando vuelven a su estado fundamental a una longitud de onda superior a la de absorción [**54**].

La espectrometría NMR requiere de nuevo una mención especial en este contexto. Algunos autores la clasifican como una espectrometría de emisión, ya

58. Chu, X.; Huang, Y.; Yun, Y.H.; Bian, X. Modern Spectral Analysis Techniques. In *Chemometrics Methods in Analytical Spectroscopy Technology*, Springer: Singapore, 2022, pp 31-88.

59. Rodriguez-Saona, L.; Ayvaz, J.; Wehling, R.L. Infrared and Raman Spectroscopy. In *Food Analysis,* 5th ed.; Ismail, B.P.; Nielsen, S.S., Eds.; FSTS; Springer: Cham, Switzerland, 2017, pp 107-128.

que mide la frecuencia que emite el núcleo atómico tras irradiar el material en la región de las radiofrecuencias (véase Figura 1.4). Sin embargo, el dato adquirido realmente se utiliza para describir la frecuencia del campo magnético que ha sido absorbida por el núcleo atómico [**60**], por lo que debería considerarse dentro de la espectrometría de absorción. En la práctica, lo cierto es que a la hora de clasificar las técnicas espectrométricas ópticas, se suele enumerar la NMR de forma independiente, al igual que la espectrometría Raman [**58**].

De todas las citadas, las técnicas más exploradas en el campo del análisis de alimentos son las espectrometrías vibracionales (espectrometría de IR, tanto en la región del IR cercano (NIR) como en la región del IR medio (MIR o FT-IR), y la espectrometría Raman), la espectrometría de fluorescencia molecular y la espectrometría de NMR [**54, 61**]. Como ya se ha adelantado, todas ellas comparten la característica de que la señal detectada se recoge en forma de espectro, que se representa gráficamente como la intensidad del fenómeno óptico medido frente al valor de un parámetro espectral, p.ej., la longitud de onda, el número de ondas o frecuencia a la que ha tenido lugar.

En la práctica, un espectro es un conjunto de parejas de números que describen la posición espectral y la intensidad observada (la intensidad de la señal analítica correspondiente a cada valor de posición), tal y como muestra la Figura 1.6 a modo de ejemplo. Si bien es cierto que recientemente está ganando cada vez más atención otro tipo de técnicas más sofisticadas como las basadas en imágenes. Dentro de éstas, se encuentran aquellas que como señal resultante proporcionan desde una fotografía convencional donde cada píxel es un conjunto de tres números (coordenadas RGB), hasta imágenes multi o hiperespectrales, donde cada píxel de la imagen encierra 10 o más valores de intensidad, respectivamente, que recogen la información espectral que caracteriza el material bajo estudio [**62**]. Por lo tanto, en estos casos los datos adquiridos se conforman en una estructura más compleja. También alguna de las espectrometrías mencionadas puede dar lugar a datos más complejos como el

60. Spyros, A.; Dais, P. Theoretical aspects. In *NMR Spectroscopy in Food Analysis*; RSC: Cambridge, U.K., 2013, pp 5-66.

61. Kharbach, M.; Mansouri, M.A.; Taabouz, M.; Yu, H. Current Application of Advancing Spectroscopy Techniques in Food Analysis: Data Handling with Chemometric Approaches. *Foods* **2023**, *12*, 2753. DOI: 10.3390/foods12142753.

62. Tian, S.; Xu, H. Nondestructive methods for the quality assessment of fruits and vegetables considering their physical and biological variability. *Food Eng. Rev.* **2022**, *14*, 380-407. DOI: 10.1007/s12393-021-09300-0.

espectro de excitación/emisión de fluorescencia total (TFS) o experimentos de dos o más dimensiones en NMR [63].

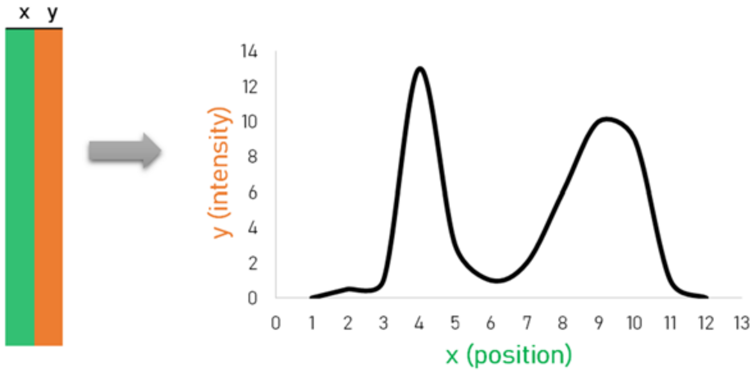


Figura 1.6. Ejemplo de representación gráfica de una señal bidimensional (espectro).

Tras adquirir la señal analítica, independientemente de la complejidad de su estructura, la siguiente etapa consiste en el tratamiento y análisis de los datos para finalmente obtener la información buscada que da respuesta al problema analítico (etapas 4 y 5 en la Figura 1.3). Para trabajar con unos datos de tal complejidad es necesario aplicar la disciplina conocida como **quimiometría**, ya que visualmente no es factible deducir conclusiones de estos datos. Su aparición en la década de los 70 ha permitido que el tratamiento y análisis de las señales adquiridas mediante técnicas espectrométricas se realice desde un enfoque holístico, enfrentando problemas relacionados con la calidad de alimentos desde sus inicios [64].

1.4. Huellas instrumentales y minería de datos

La quimiometría queda definida por la IUPAC como la *ciencia que relaciona las medidas realizadas en un sistema o proceso químico con el estado del sistema mediante la aplicación de métodos matemáticos o estadísticos* [65]. Su objetivo es extraer la información química útil mediante el uso de la **minería de datos** (*data mining*), que emplea métodos de inteligencia artificial y aprendizaje automático (*machine learning*). Desde su nacimiento y hasta la fecha, el uso de

[63] Boqué Martí, R.; Ferré Baldrich, J. Fundamentals of PARAFAC. In *Data Handling in Science and Technology*, de la Peña, A. et al., Eds.; Elsevier: Amsterdam, Netherlands, 2015, pp 7-35.

[64] Marini, F. Introduction. In *Chemometrics in Food Chemistry*, Marini, F., Ed.; Elsevier: Amsterdam, Netherlands, 2013, pp 1-5.

[65] Hibbert, D.B. Vocabulary of concepts and terms in chemometrics (IUPAC Recommendations 2016). *Pure Appl. Chem.* **2016**, *88*, 407-443. DOI: 10.1515/pac-2015-0605.

esta disciplina científica no ha hecho más que crecer, convirtiéndose a día de hoy en una parte imprescindible de la química analítica moderna [66].

Aunque la aplicación de quimiometría en la presente tesis se refiere en especial a las etapas finales del proceso analítico tras la adquisición del dato, ésta también se utiliza en diseño de experimentos (DoE) y la optimización de procesos analíticos, contribuyendo a los principios de la GAC [66,67].

Las herramientas quimiométricas se aplican fundamentalmente a señales cromatográficas y espectrométricas. Siguiendo la línea del apartado anterior, este se centrará en las segundas. La señal adquirida en forma de espectro, que se definió como un conjunto de parejas de números, se conoce como una señal de dimensión 2 (señal 2D). Sin embargo, en quimiometría cuando se trabaja con un conjunto de señales de diversas muestras que han sido adquiridas con el mismo equipo, se prescinde de los números que indican la posición (es decir, el valor del parámetro espectral al que tuvo lugar el fenómeno medido), y de esta forma se reduce una dimensión de los datos, simplificando su tratamiento y análisis. Por lo tanto, en lenguaje matemático, un espectro pasa a ser un tensor de datos de orden 1, también conocido como vector, que encierra cientos o miles de variables. Mientras que las señales más complejas como las adquiridas mediante TFS o 2D-NMR se organizan en un tensor de datos de orden 2, o matriz de datos, tras eliminar los vectores de datos relativos a la posición de la señal. Señales aún más complejas pueden estar contenidas en un tensor de datos de orden 3 o un cubo de datos, o bien órdenes superiores [68]. Por el contrario, se habla de tensor de orden 0 o escalar cuando la señal adquirida es un único número, como puede ser el valor de absorbancia a una determinada y única longitud de onda [69]. La Figura 1.7 representa los diferentes niveles de complejidad de datos que se pueden adquirir en espectrometría para una única muestra del material en estudio.

Sin importar la complejidad de la estructura de los datos, en todos los casos estas técnicas proporcionan un conjunto de números que encierra la información química que caracteriza el material bajo estudio, la cual está

[66] Brereton, R.G.; et al. Chemometrics in analytical chemistry – Part I: history, experimental design and data analysis tools. *Anal. Bioanal. Chem*. **2017**, *409*, 5891-5899. DOI: 10.1007/s00216-017-0517-1.

[67] Kalinowska, K.; Bystrzanowska, M.; Tobiszewski, M. Chemometrics approaches to green analytical chemistry procedure development. *Curr. Opin. Green. Sustain. Chem*. **2021**, *30*, 100498. DOI: 10.1016/j.cogsc.2021.100498.

[68] Cuadros-Rodríguez, L.; et al. Chromatographic fingerprinting and food identity/quality: potentials and challenges. *J Agric. Food Chem*. **2021**, *69*, 14428-14434. DOI: 10.1021/acs.jafc.1c05584.

[69] Brereton, R.G.; et al. Chemometrics in analytical chemistry – Part II: modeling, validation, and applications. *Anal. Bioanal. Chem*. **2018**, *409*, 6691-6704. DOI: 10.1007/s00216-018-1283-4.

presente de manera implícita pero no evidente, y por ello debe ser extraída mediante el uso de la minería de datos.

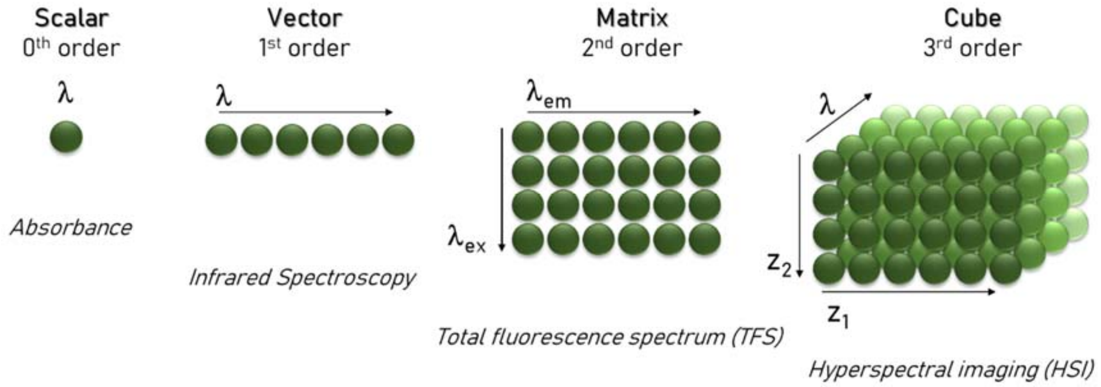


Figura 1.7. Niveles de estructura de datos químicos adquiridos por diferentes técnicas analíticas. *Adaptada de [69].*

Dicha información puede ser tratada y analizada desde dos tipos de enfoque:

- ❖ Enfoque dirigido (*targeted approach*): busca y utiliza una parte o varias partes de la señal adquirida para establecer una relación con objetivos analíticos predefinidos, o marcadores químicos. También forma parte de este enfoque la elaboración de un perfil de componentes (*profiling*) construido a partir de marcadores químicos seleccionados [**70**].
- ❖ Enfoque no dirigido (*non-targeted approach*): trabaja con la señal al completo para detectar numerosos objetivos analíticos no predefinidos. La metodología de **huellas instrumentales** (*fingerprinting methodology*) se enmarca como parte de este enfoque [**71**].

En el campo de la química analítica de los alimentos, trabajar e investigar siguiendo el enfoque no dirigido es cada vez más común [**70**]. Y no es de extrañar, ya que cada uno de los ingredientes, compuestos o analitos que constituyen un alimento, conviven y dependen unos de otros. Están integrados y, por lo tanto, la información debe ser tenida en cuenta en todo su conjunto, y no analizando los diversos analitos por separado. La integración de métodos analíticos no dirigidos en el análisis de alimentos rutinario se hace cada vez más inminente. Haciendo un símil, la metodología de huellas instrumentales aplicada a señales

70. Ballin, N.Z.; Laursen, K.H. To target or not to target? Definitions and nomenclature for targeted versus non-targeted analytical food authentication. *Trends Food Sci. Technol.* **2019**, *86*, 537-543. DOI: 10.1016/j.tifs.2018.09.025.

71. Jiménez-Carvelo, A.M.; Martín-Torres, S.; Cuadros-Rodríguez, L.; González-Casado, A. Nontargeted fingerprinting approaches. In *Food Authentication and Traceability*, Galanakis, C.M., Ed.; Elsevier: London, U.K., 2021, pp 163-194.

espectrométricas es como el músico que golpea un diapasón para afinar su instrumento. La radiación de luz incidente hacia el material en estudio es el mazo con el que el músico golpea su diapasón. La frecuencia a la que el material transmita, refleje, disperse o emita la luz es específica de cada molécula, al igual que la onda de sonido generada por el diapasón es específica de una nota musical en particular [**72**].

La metodología de huellas instrumentales aplicada al análisis de alimentos trata de buscar un patrón o pauta que se repite (*pattern recognition*) en distintos productos alimenticios de características similares [**73**]. No hay duda de que su nombre proviene del fundamento que sigue la identificación de personas. Un método analítico convencional desarrollado para medir un analito específico sería como tratar de identificar a una persona mediante sus características físicas o antropomórficas de forma individual: el color del pelo, la altura, color de la retina, la composición de grasa corporal, parámetros sanguíneos, etc. Mientras que un método analítico basado en huellas instrumentales es análogo a la identificación de una persona mediante su huella dactilar, que es única de cada persona. Cada alimento también puede tener su huella instrumental única y personal, que lo identifica y caracteriza, aunque no necesariamente lo describa.

La primera revisión bibliográfica publicada acerca del uso de huellas instrumentales para el análisis de alimentos es relativamente reciente, donde ya se pone de manifiesto el gran potencial del tándem espectrometría-quimiometría para el análisis de alimentos [**74**]. Desde entonces, el estudio de este enfoque ha ido en aumento [**49,73,75,76,77**]. Sin embargo, a pesar de los grandes beneficios demostrados que aporta la aplicación de la minería de datos desde un enfoque no dirigido en el campo del análisis de alimentos, éste aún

72. Ansede, M. "Cada cáncer tiene una huella de luz infrarroja diferente". El País, November 14, 2023, pp 27.

73. Cuadros-Rodríguez, L.; et al. Chromatographic fingerprinting: An innovative approach for food 'identitation' and food authentication – A tutorial. *Anal. Chim. Acta.* **2016**, *909*, 9-23. DOI: 10.1016/j.aca.2015.12.042.

74. Ellis, D.I.; et al. Fingerprinting food: current technologies for the detection of food adulteration and contamination. *Chem. Soc. Rev.* **2012**, *41*, 5706-5727. DOI: 10.1039/c2cs35138b.

75. Cubero-León, E.; Peñalver, R.; Maquet, A. Review on metabolomics for food authentication. *Food Res. Int.* **2014**, *60*, 95-107. DOI: 10.1016/j.foodres.2013.11.041.

76. Medina, S.; et al. Food fingerprints – A valuable tool to monitor food authenticity and safety. *Food Chem.* **2019**, *278*, 144-162. DOI: 10.1016/j.foodchem.2018.11.046.

77. Mialon, N.; Roing, B.; Capodanno, E.; Cadiere, A. Untargeted metabolomic approaches in food authenticity: A review that showcases biomarkers. *Food Chem.* **2023**, *398*, 133856. DOI: 10.1016/j.foodchem.2022.133856.

queda lejos de ser parte de métodos oficiales y estandarizados [**77,78**]. No obstante, en los últimos años se observa cierto progreso con publicaciones como la norma ASTM E2617-17 que recoge recomendaciones prácticas para la etapa de validación de modelos quimiométricos, referidos como "calibraciones multivariable" o la guía USP (del inglés *United States Pharmacopeia*) para el desarrollo y validación de métodos no dirigidos para la detección de adulteraciones [**79**]. También existen otros documentos más específicos, como las normas ASTM E1790-04 y ASTM E1655-17, dirigidas al uso concreto de espectrometría IR para el análisis multivariable cualitativo y cuantitativo, o la norma ISO 12099:2010 para la aplicación de la misma técnica acoplada a herramientas de minería de datos para la determinación de constituyentes en alimentos para animales. A ellas se le suman documentos publicados por un grupo de trabajo de la AOAC International (del inglés *Association of Official Agricultural Chemists*) que recogen los requisitos mínimos para realizar análisis no dirigidos de ingredientes para la evaluación del fraude y autenticidad del aceite de oliva virgen extra y de la miel: AOAC SMPR® 2020.006 y 2020.007. Siguiendo esta línea, destaca la iniciativa impulsada en 2022 por el Servicio de Acreditación del Reino Unido (UKAS) instando a los laboratorios a solicitar la acreditación de análisis no dirigidos para la autentificación de alimentos o piensos bajo la acreditación ISO/IEC 17025:2017 [**80**]. Aún más recientemente, en otros campos de análisis distintos al alimentario, se han publicado otras normas ASTM específicas del ámbito farmacéutico y petrolífero, ASTM E2891-20 y ASTM D8321-22, respectivamente. Por todo ello, aunque aún queda camino por recorrer, la implementación de la metodología de huellas instrumentales y la minería de datos en el análisis (rutinario y oficial) de los alimentos está cada vez más cerca.

El desarrollo de modelos de aprendizaje automático requiere disponer previamente de un gran conjunto de datos, provenientes de diferentes muestras que compartan una o más características. Haciéndolo extensivo al análisis alimentario, la aplicación de minería de datos precisa tener un amplio conjunto de productos alimenticios de una misma categoría [**68**]. El número mínimo exacto no es un dato conocido, ya que depende de muchos factores, y es difícil encontrar un consenso entre todos los quimiómetras.

---

78. Squeo, G.; et al. Considerations about the gap between research in Near Infrared spectroscopy and official methods and recommendations of analysis in foods. *Curr. Opin. Food Sci.* **2024**, 101203. DOI: 10.1016/j.cofs.2024.101203.

79. Food Chemicals Codex, Appendix XVIII: Guidance on Developing and Validating Non-targeted Methods for Adulteration Detection. US Pharmacopoeia Convention: Rockville, U.S., 2019.

80. Expression of Interest – Food non-targeted authenticity testing pilot. UKAS, August 5, 2022. https://www.ukas.com/resources/latest-news/expression-of-interest-food-non-targeted-authenticity-testing/.

"PUT 1000 CHEMOMETRICIANS IN A ROOM AND THERE WILL BE 1001 OPINIONS AS TO ITS FUTURE DIRECTION" [69].

Aunque parece haber cierto consenso en que para fines de autentificación de alimentos es necesario que el conjunto de muestras abarque todas las fuentes de variabilidad vinculadas a la propiedad diana perseguida [81].

A nivel matemático en minería de datos, cuando se trabaja con un conjunto de materiales o muestras, es necesario construir un único conjunto de datos concatenando las señales adquiridas de todas ellas, y por lo tanto añadiendo una dimensión más a la estructura de datos. Según la dimensionalidad de la señal adquirida, esto desencadena en matrices de datos de dos, tres, cuatro o más vías respectivamente para tensores de orden 1, 2, 3 o superior (véase Figura 1.7) [82].

Una vez construido el conjunto de datos y antes de comenzar su análisis, a veces es necesario llevar a cabo una etapa de preprocesado de los datos. Según la técnica analítica empleada para la adquisición de señales, y el método quimiométrico a aplicar posteriormente, existen métodos y/o herramientas de preprocesado más o menos adecuados. Esta etapa se lleva a cabo con diferentes fines, siendo los métodos y herramientas de preprocesado más comunes para ello los siguientes:

(i) Eliminar información irrelevante presente en la señal que pueda interferir en los resultados: mejora de la relación señal/ruido mediante suavizado (*smoothing*), filtrado o corrección de la línea base.

(ii) Preparar y depurar los datos: transformación de dominio mediante transformada de Fourier (FT) o transformación *wavelet*; reducción de variables; normalización de las intensidades; alineamiento de las señales o normalización de la posición.

(iii) Aumentar las diferencias entre las señales de unas muestras y otras: corrección de las diferencias entre las variables mediante centrado en la media o autoescalado [83].

81. Kemsley, E.K.; Defemez, M.; Marini, F. Multivariate statistics: Considerations and confidences in food authenticity problems. *Food Control* **2019**, *105*, 102-112. DOI: 10.1016/j.foodcont.2019.05.021.

82. Escandar, G.M.; Goicoechea, H.C.; Muñoz de la Peña, A.; Olivieri, A.C. Second- and higher-order data generation and calibration: A tutorial. *Anal. Chim. Acta* **2014**, *806*, 8-26. DOI: B.V. 10.1016/j.aca.2013.11.009.

83. Chu, X.; Huang, Y.; Yun, Y.H.; Bian, X. Spectral preprocessing methods. In *Chemometrics Methods in Analytical Spectroscopy Technology*, Springer: Singapore, 2022, pp 111-168.

Concluida la etapa de preprocesado, llega el momento de aplicar el/los métodos quimiométricos. Existe una gran variedad, y pueden clasificarse atendiendo al objetivo perseguido en: identificación, resolución, comparación, clasificación y cuantificación, que en lenguaje quimiométrico podría traducirse como: análisis exploratorio, resolución, análisis de la similitud, clasificación y cuantificación multivariable respectivamente. Los tres primeros engloban los conocidos como métodos no supervisados, mientras que los otros dos comprenden el modelado supervisado, o métodos supervisados [**84**]. La Figura 1.8 ilustra esta clasificación, que a continuación se desarrolla brevemente.

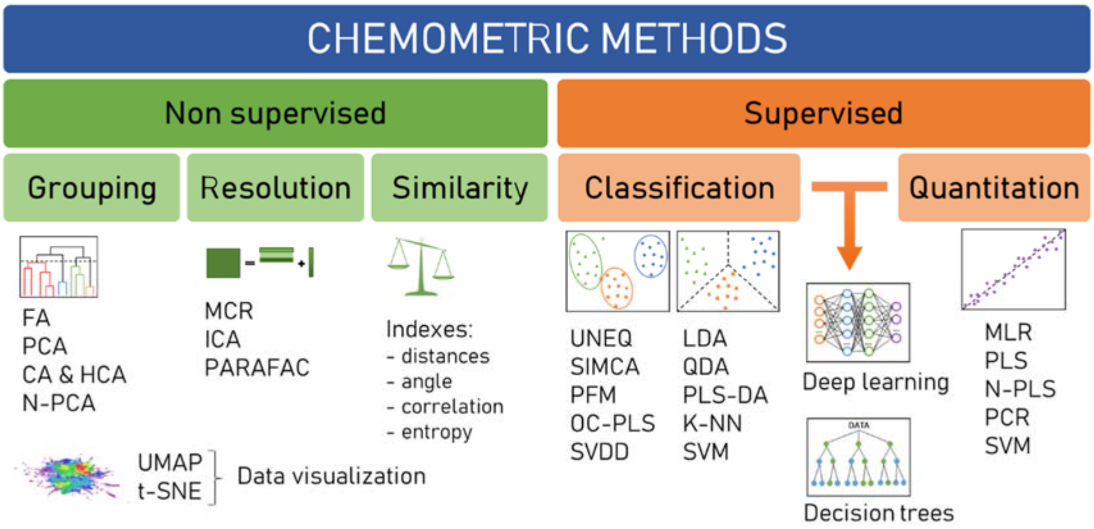


Figura 1.8. Clasificación de métodos quimiométricos.

FA: análisis de factores; PCA: análisis de componentes principales; CA: análisis de conglomerados; HCA: análisis jerárquico de conglomerados; MCR: resolución de curvas multivariable; ICA: análisis de componentes independientes; PARAFAC: análisis paralelo de factores; UMAP: proyección de colectores uniformes; t-SNE: inclusión estocástica de vecinos t-distribuida; UNEQ: modelos de clases desiguales; SIMCA: modelado flexible e independiente por analogía de clases; PFM: métodos de función potencial; PLS: regresión parcial de mínimos cuadrados; OC-PLS: PLS de una clase; SVDD: descripción del dominio mediante vectores de soporte; LDA: análisis discriminante lineal; QDA: análisis discriminante cuadrático; PLS-DA: análisis discriminante mediante PLS; k-NN: clasificación basada en los k vecinos más cercanos; SVM: sistemas de aprendizaje automático mediante vectores de soporte; MLR: regresión lineal multivariable; PCR: regresión por componentes principales.

Primeramente, cabe destacar que trabajar con inteligencia artificial y aprendizaje automático, además de requerir un amplio número de muestras, también requiere disponer de una caracterización previa e, incluso, de

84. Biancolillo, A.; Marini, F.; Ruckebush, C.; Vitale, R. Chemometric Strategies for Spectroscopy-Based Food Authentication. *Appl. Sci.* **2020**, *10,* 6544. DOI: :10.3390/app10186544.

información adicional (metadatos). En el análisis de alimentos acoplado a la minería de datos, esto se traduce en tener a disposición información sobre los alimentos bajo estudio: su origen geográfico o botánico, composición nutricional, o cualquier otro dato relevante que lo caracteriza, y que es el objetivo para el que se pretende desarrollar un modelo matemático capaz de predecirlo en el futuro aplicándolo a nuevas muestras. Esta información "extra" en quimiometría define la propiedad diana, y se considera la variable Y, mientras que el conjunto de datos formado por las señales analíticas adquiridas constituye la variable X [**65**,**85**].

➡ ANÁLISIS EXPLORATORIO (MÉTODOS NO SUPERVISADOS)

Se trata de un conjunto de métodos destinados a detectar patrones comunes entre las variables observadas (X). Se denomina análisis no supervisado porque estudia el comportamiento de las muestras en el espacio de las variables X, es decir, considerando únicamente los datos que conforman la X, sin tener en cuenta la información contenida en Y.

Dentro de este grupo de métodos no supervisados, se pueden diferenciar tres subgrupos, atendiendo al objetivo perseguido con la aplicación del método y son: (i) agrupamiento (*grouping*), (ii) resolución (*resolution*) y (iii) análisis de la similitud (*similarity*).

❖ Agrupamiento: Se estudia el agrupamiento natural de las muestras, a veces como paso previo para llevar a cabo un modelado supervisado y formular la/las hipótesis de partida. Los métodos más comunes y empleados en el análisis de alimentos son el análisis de factores (FA), análisis de componentes principales (PCA) y el análisis de conglomerados o agrupamientos (*clustering*) (CA) y las diferentes estrategias para aplicar este último: jerárquico (HCA), K-medias, … [**86**,**87**].

❖ Resolución: Comúnmente se aplican métodos de resolución de señales para extraer las señales y proporciones de componentes puros que conforman una señal de una matriz compleja como puede ser un alimento. No requieren tener a disposición información complementaria para desarrollarlo, y por ello se consideran métodos no supervisados. Pueden ser utilizados con fines de identificación. Entre ellos se encuentran el

[85] Buvé, C.; et al. Application of multivariate data analysis for food quality investigations: An example-based review. *Food Res. Int.* **2022**, *151*, 110878. DOI: 10.1016/j.foodres.2021.110878.

[86] Li Vigni, M.; Durante, C.; Cocchi, M. Exploratory Data Analysis. In *Chemometrics in Food Chemistry*, Marini, F., Ed.; Elsevier: Amsterdam, Netherlands, 2013, pp 55-126.

[87] Chu, X.; Huang, Y.; Yun, Y.H.; Bian, X. Pattern Recognition Methods. In *Chemometric Methods in Analytical Spectroscopy Technology*, Springer: Singapore, 2022, pp 329-380.

método de resolución de curvas multivariable (MCR) y el análisis de componentes independientes (ICA) [84].

- ❖ Análisis de la similitud: El análisis de la similitud entre señales analíticas tiene por objetivo la comparación entre las mismas y la medida de cuánto se parecen o difieren entre ellas, a través del cálculo de índices de similitud (distancias, correlación, ángulo, …), bien para la obtención de un escalar que proporcione la información buscada (similitud entre las señales), o bien para la obtención de un perfil completo de similitud variable-a-variable [88]. A pesar de existir una gran cantidad de aplicaciones en literatura científica, pocas veces se hace referencia a éste como una herramienta más de la quimiometría, que bien podría encajarse dentro del análisis exploratorio. Esta propuesta forma parte de un libro, que será publicado próximamente, recogido en forma de capítulo y que se presenta en el Capítulo 4 de la presente tesis doctoral.

Otras utilidades de los métodos no supervisados son la descomposición y visualización de datos, ya que su complejidad a veces lo impide o consume mucho tiempo; o la detección de valores atípicos (*outliers*), con la consecuente toma de decisiones (eliminación de dichos valores para los siguientes pasos). Existen herramientas desarrolladas recientemente para la visualización de datos como la aproximación y proyección de colectores uniformes (UMAP) o la inclusión estocástica de vecinos t-distribuida (t-SNE) [89,90].

Por último, el análisis paralelo de factores (PARAFAC) es un método de descomposición de datos multi-vía que podría entenderse como la extensión del método MCR, por lo que también se incluye como parte de los métodos no supervisados en algunas referencias, al igual que el PCA N-vías, que transforma un conjunto de datos de tres vías en uno de dos para poder aplicar el método PCA [91].

➡ MÉTODOS SUPERVISADOS

Por el contrario, los métodos que forman parte del reconocimiento de pautas supervisado sí tienen en cuenta la información contenida en Y. Estos métodos

[88] Zeng, R.; et al. How similar is "similar", or what is the best measure of soil spectral and physiochemical similarity? *PLoS ONE* **2021**, *16*, e0247028. DOI: 10.1371/journal.pone.0247028.

[89] Becht, E.; et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **2019**, *37*, 38-44. DOI: 10.1038/nbt.4314.

[90] Soni, J.; Prabakar, N.; Upadhyay, H. Visualizing High-Dimensional Data Using t-Distributed Stochastic Neighbor Embedding Algorithm. In *Principles of Data Science*, Arabnia, H.R. et al., Eds.; Springer: Cham, Switzerland, 2020; pp 189-206.

[91] Otto, M. Pattern Recognition and Classification. In *Chemometrics. Statistics and Computer Application in Analytical Chemistry*, 3rd ed.; Wiley-VCH: Weinheim, Germany, 2017, pp 135-212.

conllevan el desarrollo de un modelo de aprendizaje automático que identifica la información relevante en X para predecir Y [70].

El conjunto de datos se divide en dos subconjuntos: entrenamiento y validación, en una proporción 70-80% y 30-20% respectivamente. Existen métodos para realizar esta selección como el algoritmo CADEX (del inglés, *computer aided design of experiments*) [92], propuesto por Kennard & Stone [93], entre otros, para asegurar que exista suficiente representatividad del conjunto de muestras en el primer subconjunto. El subconjunto de entrenamiento (o calibración) se utiliza para desarrollar el modelo de aprendizaje. Una vez creado, le sigue la etapa de validación del modelo, diferenciando entre (i) validación interna cruzada (*cross-validation*); y (ii) validación externa. La (i) se realiza sobre el conjunto de entrenamiento, que se divide en segmentos y utiliza unos para generar el modelo y otros para validarlo; existen diferentes estrategias para ello. Mientras que en la (ii) validación externa se aplica el modelo ya desarrollado al subconjunto de validación, con el objetivo de verificar que el modelo funciona correctamente cuando se aplica a muestras no empleadas en el entrenamiento [71]. A partir de los resultados de entrenamiento y de validación, y comparando con el conocimiento previo (variable Y o valores verdaderos), se calculan parámetros indicativos de la calidad del modelo. Éstos serán distintos según el modelo sea de clasificación o cuantificación [94,95]. A cada parámetro le debe ir asignado un valor crítico que marcará el límite de aceptación de los resultados, es decir, si son o no satisfactorios para el fin buscado. Dicho valor crítico o límite se establecerá en función del escenario y el objetivo perseguido con el desarrollo del modelo, p.ej. un modelo de clasificación multivariable para el desarrollo de un método de cribado podría considerarse satisfactorio con un valor de especificidad superior a 0.5 sobre 1, y un valor de sensibilidad superior a 0.95 sobre 1. Estos valores se establecen de forma previa al desarrollo del modelo por parte del analista. Si los resultados de estos parámetros son satisfactorios, indica que los datos analíticos

[92] Chu, X.; Huang, Y.; Yun, Y.H.; Bian, X. Method of Selecting Calibration Samples. In *Chemometrics Methods in Analytical Spectroscopy Technology*, Springer: Singapore, 2022, pp 297-308.

[93] Kennard, R.W.; Stone, L.A. Computer aided design of experiments. *Technometrics* **1969**, *11*, 137-148. DOI: 10.1080/00401706.1969.10490666.

[94] Cuadros-Rodríguez, L.; Pérez-Castaño, E.; Ruiz-Samblás, C. Quality performance metrics in multivariate classification methods for qualitative analysis. *Trends Anal. Chem.* **2016**, *80,* 612-624. DOI: 10.1016/j.trac.2016.04.021.

[95] ASTM International. *Standard Practice for Validation of Empirically Derived Multivariate Calibrations*; ASTM 2617-17; West Conshohocken, U.S., 2017. DOI: 10.1520/E2617-17.

multivariable contienen la información necesaria capaz de predecir el atributo o característica de los objetos estudiados.

Por último, en la denominada etapa de predicción se aplica el modelo a muestras en las que el valor Y es desconocido, donde se calculan los valores predichos de Y mediante interpolación a partir de la función calculada en la primera etapa de calibración [**96**].

La selección del tipo de método supervisado depende del objetivo perseguido, a su vez determinado por la naturaleza de la variable Y. Si se trata de un atributo o cualidad (análisis cualitativo), la elección será un método de clasificación. Mientras que en los casos en los que Y viene definido por un número que determina la cantidad de una característica propia de la muestra, en este caso del alimento, se escogerá un método de cuantificación (análisis cuantitativo). Un caso particular se genera cuando la Y a su vez esta conformada por un vector de valores, cada uno de ellos correspondiente a una propiedad diana; en este caso asistimos a un método multivariable multiparamétrico.

A continuación, se enumeran los métodos más comunes en ambos casos:

❖ Métodos de clasificación: Pueden dividirse a su vez en métodos de análisis discriminante, modelado de clases y aprendizaje profundo (*deep learning*). En el primero se asigna una clase de pertenencia a cada muestra u objeto, mientras que en el segundo puede ocurrir que una muestra se le asigne la pertenencia a varias clases, o incluso ninguna. En definitiva, la clasificación discriminante se centra en las diferencias entre las muestras, y el modelado de clases en capturar las similitudes [**97**]. Por su parte, los algoritmos propios del *deep learning*, entre los que destacan las redes neuronales (ANN, del inglés *artificial neural networks*), tratan de imitar las redes de neuronas del cerebro humano desarrollando modelos de aprendizaje, que pueden ser utilizados tanto para clasificación como para cuantificación.

También es común diferenciar la clasificación multivariable en dos tipos de estrategias, en función de si se pretende modelar una clase (*one-class*) o varias clases (*multi-class*). El análisis discriminante a priori solo puede ser empleado para seguir una estrategia *multi-class*, mientras que el

96. Cuadros-Rodríguez, L.; Jiménez-Carvelo, A.M.; Andrade, J.M. A multivariate approach to Analytical Chemistry. In *Problem-Oriented Analytical Chemistry Driven by Chemometrics,* Cuadros-Rodríguez, L.; Jiménez-Carvelo, A.M.; Andrade, J.M., Eds.; Elsevier: Amsterdam, Netherlands, 2025 [*submitted, in process*].

97. Belvilacqua, 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: Amsterdam, Netherlands, 2013, pp 171-234.

modelado de clases es útil para ambas. A continuación, se enumeran los métodos más comunes para cada estrategia:

- Análisis discriminante: análisis discriminante lineal (LDA), análisis discriminante cuadrático (QDA), análisis discriminante mediante regresión parcial de mínimos cuadrados (PLS-DA), clasificación basada en los k vecinos más cercanos (*k*-NN), sistemas de aprendizaje automático mediante vectores de soporte (SVM).
- Modelado de clases: modelos de clases desiguales (UNEQ), modelado flexible e independiente por analogía de clases (SIMCA) o métodos de función potencial (PFM) [**91,97**]. También se han presentado modificaciones de métodos discriminantes para que sea posible realizar un modelado de clases basado p.ej. en PLS, como el método PLS de una clase (OC-PLS) o basado en SVM, como la descripción del dominio mediante vectores de soporte (SVDD) [**98**].
- Aprendizaje profundo: perceptrón multicapa (MLP), red neuronal convolucional (CNN), red neuronal recurrente (RNN), red adversarial generativa (GAN), modelos basados en el mecanismo de atención (AM) o modelos integrados de aprendizaje profundo [**99**].

Además, existen otro tipo de algoritmos que pueden ser utilizados como método tanto de clasificación como de cuantificación multivariable, y cada vez se van haciendo más hueco en el análisis de alimentos. Ejemplo de ello son el árbol de clasificación y regresión (CART, del inglés *classification and regression tree*) y el bosque aleatorio (RF, del inglés *random forest*). Estos siguen una estrategia de modelado por ensambles (*ensemble modelling*) [**100**].

❖ Métodos de cuantificación: su objetivo es ajustar una función que relacione los dos bloques de variables X e Y, siendo la última en este caso los valores continuos de uno o varios parámetros característicos cuantificables. Según cómo sea la función a ajustar, es posible diferenciar entre modelos de regresión lineal y modelos no-lineales. La más sencilla regresión lineal viene dada por el modelo general:

98. Oliveri, P. Class-modelling in food analytical chemistry: Development, sampling, optimisation and validation issues – A tutorial. *Anal. Chim. Acta* **2017**, *982*, 9-19. DOI: 10.1016/j.aca.2017.05.013.

99. Deng, Z.; et al. Deep learning in food authenticity: Recent advances and future trends. *Trends Food Sci. Technol.* **2024**, *144*, 104344. DOI: 10.1016/j.tifs.2024.104344.

100. Chu, X.; Huang, Y.; Yun, Y.H.; Bian, X. Methods for Improving Prediction Ability of Model. In *Chemometrics Methods in Analytical Spectroscopy Technology*, Springer: Singapore, 2022, pp 399-422.

$$\mathrm{Y} = \mathrm{a} + \mathrm{Xb}$$

donde a y b son los coeficientes (escalares) que se calculan para ajustar los datos contenidos en X e Y. La aplicación matemática una vez estimados estos coeficientes a través de un modelo de mínimos cuadrados que trata de minimizar el error, es calcular la Y estimada ($\widehat{\mathbf{Y}}$) a partir de los datos X. A este cálculo siempre va asociado un error (e), que es el que limita la calidad de predicción del modelo generado a partir de la función matemática.

$$\mathrm{Y} = \widehat{\mathbf{Y}} + \mathrm{e} = \mathrm{Xb} + \mathrm{e}$$

En quimiometría la señal analítica adquirida (X) es multivariable, de forma que cada objeto (muestra) queda definido por un vector, matriz o cubo de datos. Por lo tanto, se calculan tantos coeficientes **b** (b_1, b_2, …, b_p) como número de variables contenga X.

Cabe destacar en este punto que la variable respuesta Y, puede ser a su vez única o múltiple. En el segundo caso hablaríamos entonces de análisis multivariado (o multiparamétrico), donde se modela independientemente cada una de las respuestas contenidas en Y, calculando por lo tanto un coeficiente **b** para cada respuesta:

$$\mathrm{Y} = \widehat{\mathbf{Y}} + \mathrm{E} = \mathrm{XB} + \mathrm{E}$$

Independientemente de si Y es única o múltiple, los coeficientes de regresión definen la relación entre ambas variables X e Y [**101**].

Los métodos de regresión lineal más comunes son: la regresión lineal multivariable (MLR), la regresión por componentes principales (PCR), o la regresión parcial por mínimos cuadrados (PLS) y su variante N-PLS para datos multi-vía. Entre los métodos de cuantificación no lineales se incluyen aquellos basados en la clasificación ya mencionados anteriormente, como el método SVM, los métodos de aprendizaje profundo (*deep learning)* y los algoritmos basados en arboles de decisión [**99,102**].

Cada uno de los métodos tiene un fundamento matemático diferente, de forma que es habitual testear diferentes tipos de métodos para averiguar con cuál de ellos se obtiene un modelo de aprendizaje con mejores capacidades predictivas. Si bien es cierto que los usados por excelencia son PCA para el análisis exploratorio, y PLS tanto para clasificación (PLS-DA) como para regresión (PLSR) [**84**].

[101] Westad, F.; Belvilacqua, M.; Marini, F. Regression. In *Chemometrics in Food Chemistry*, Marini, F., Ed.; Elsevier: Amsterdam, Netherlands, 2013, pp 127-170.

[102] Chu, X.; Huang, Y.; Yun, Y.H.; Bian, X. Nonlinear Calibration Methods. In *Chemometrics Methods in Analytical Spectroscopy Technology*, Springer: Singapore, 2022, pp 255-296.

Como consecuencia del surgimiento de la quimiometría, y gracias a los grandes avances realizados en torno a ella, y en concreto en la metodología de huellas instrumentales a lo largo de los años, es posible desarrollar métodos analíticos con un enfoque no dirigido para el análisis de la calidad y autenticidad de alimentos, basados en técnicas espectrométricas rápidas y no destructivas o poco invasivas, en línea con los principios de la GAC. Para que éstos formen parte del análisis rutinario en los laboratorios, y algún día lleguen a formar parte de los métodos oficiales, es necesaria una investigación profunda y específica. En aras de que un método llegue a sustituir otro ya existente, debe igualar o mejorar a su precedente en todos los aspectos (económico, calidad de resultados, tiempo requerido, etc.). A continuación, se presenta el manuscrito del capítulo de libro 1, que recoge una visión global y actual del enfoque verde aplicado al análisis de alimentos. Dicho capítulo se encuentra enviado a la editorial y está pendiente de publicación.

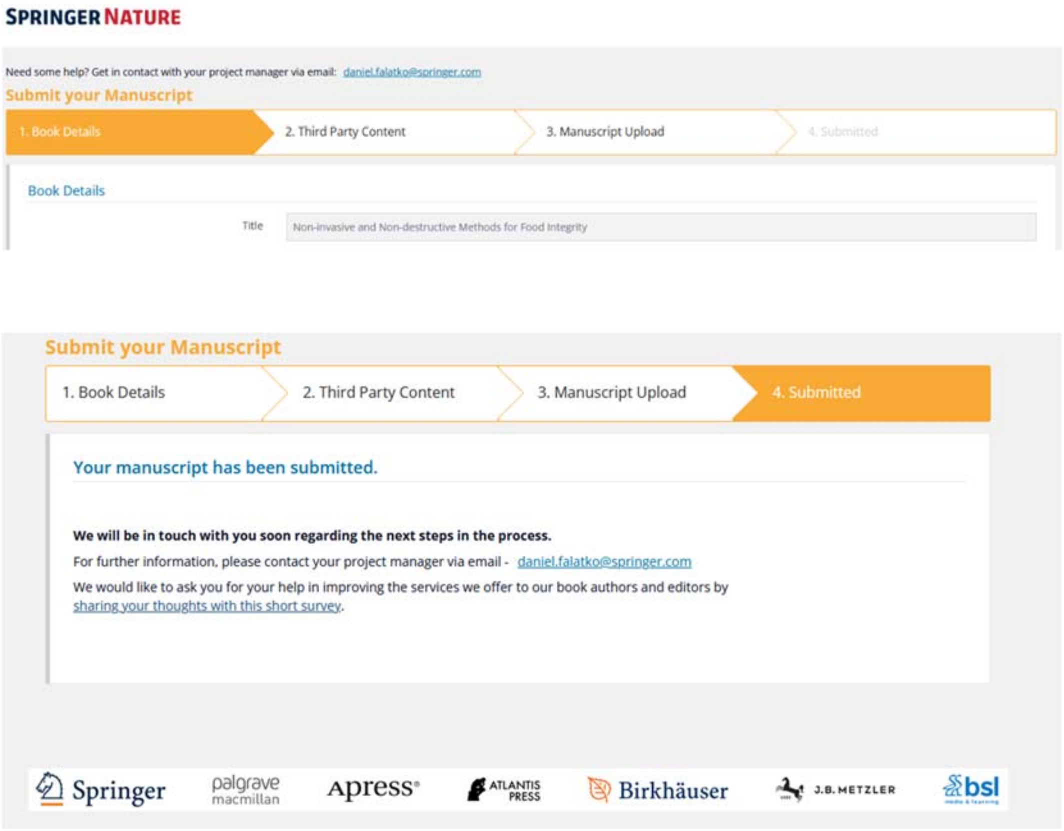
1.5. Capítulo de libro 1

Chapter 1. Food quality assessment from green approach.

Ana M. Jiménez-Carvelo[✉], Alejandra Arroyo-Cerezo, Luis Cuadros-Rodríguez

En el libro: Food Integrity by Non-invasive / Non-destructive Methods, Jiménez-Carvelo, A.M.; Arroyo-Cerezo, A.; Cuadros-Rodríguez, L., Eds. Springer, 2024.

ISBN: 978-3-031-76464-6.



Need some help? Get in contact with your project manager via email: daniel.falatko@springer.com

Submit your Manuscript

Book Details

Title: Non-invasive and Non-destructive Methods for Food Integrity

Submit your Manuscript

Your manuscript has been submitted.

We will be in touch with you soon regarding the next steps in the process.

For further information, please contact your project manager via email - daniel.falatko@springer.com

We would like to ask you for your help in improving the services we offer to our book authors and editors by sharing your thoughts with this short survey.

Capítulo de libro enviado a la editorial Springer (03-08-2024)

[✉] Corresponding author (e-mail: amariajc@ugr.es)

Abstract

Environmentally friendly analytical methods are being developed and implemented, avoiding the use of chemicals and solvents as far as possible, and reducing the time and cost of analysis. For an analytical method to be considered as part of green analytical chemistry, it must meet certain requirements. Therein lies the importance of having tools at one's disposal with which to assess how sustainable an analytical method is according to its intended purpose, in this case food quality assurance. The scientific community is advancing by leaps and bounds in the development of these new analytical methods. The techniques used that can be considered as green analytical techniques are becoming more and more sophisticated. This is a challenge for both industry and analysts, as the results produced by these techniques require the application of artificial intelligence tools. This also makes it possible to generate multivariate analytical methods through the development of machine learning models. This chapter looks at sustainable practices in the field of analytical food chemistry from a generic perspective and highlights current trends in the field.

Keywords:

Green analytical chemistry

Sustainability

Food control

Whiteness evaluation

Multivariate method

Artificial intelligence

Contents

1. Introduction

Sustainability in the food industry has become a global priority. In a constantly evolving world where natural resources are depleting and climate change threatens food security, adopting sustainable practices has become essential to ensure a healthy future for both the planet and future generations. Sustainability in the realm of food quality refers to the production, distribution, assessment, and consumption of food in a manner that minimizes environmental impact, promotes sustainability, and ensures the safety and quality of foodstuffs (Galanakis 2024, El Bilali et al. 2021). Indeed, one of the earliest definitions put forth regarding the concept of sustainability was suggested in Brundtland report in 1987, in which it was defined as development that meets the needs of the present without compromising the ability of future generations to meet their own needs (WCED 1987). Currently, the concept of sustainability has evolved, positioning itself at the core of three interconnected pillars: (i) social, (ii) economic, and (iii) environmental (Purvis et al 2019, Spiliotopoulou and Roseland 2020). In this respect, the current internationally recognised ISO definition of sustainability gathers these three pillars and sets out that it is the state of the global system, including environmental, social, and economic aspects, in which the needs of the present are met without compromising the ability of future generations to meet their own needs (ISO 82:2019). So that, the issue of sustainability has garnered significant traction in contemporary research, particularly given humanity's confrontation with the erosion of agrobiodiversity and ongoing degradation of ecosystems. These phenomena pose potential social, economic, and environmental repercussions, potentially culminating in the collapse of agri-food systems (Domingues 2023).

Sustainability is closely linked to social responsibility, defined as the responsibility of an organization for the impacts of its decisions and activities on society and the environment, through transparent and ethical behaviour (ISO 2010). In fact, organizations are under increasing scrutiny from stakeholders regarding their impact on the environment, as a result of the growing perception of the need to ensure healthy ecosystems. Therefore, one of the main objectives of social responsibility, among others, is to contribute to sustainable development, including health and the welfare of society. In this respect, ISO provides guidance on the application of social responsibility specifically for food chain organizations and is intended to be useful to all types of related entities (ISO/TS 2019).

The United Nations Sustainable Development Program, also known as the 2030 Agenda for Sustainable Development (European Union 2022), is an ambitious action plan adopted by all United Nations Member States in 2015. The 2030 Agenda consists of 17 Sustainable Development Goals (SDGs) addressing a wide range of social, economic, and environmental challenges facing the world today.

The 2030 Agenda is a call to action for all countries, both developed and developing, and recognises the interconnection between different global challenges. The SDGs are designed to be holistically integrated into national policies and development plans of each country, seeking to promote development that is economically, socially, and environmentally sustainable. Therefore, according to current policies, research and innovation in cleaner methodologies for food quality control are crucial today for several fundamental reasons. Firstly, food safety is a constantly evolving global concern, with increasingly rigorous demands in terms of quality, traceability, and sustainability. The development of more efficient and environmentally friendly quality control methods not only ensures the health and well-being of consumers but also contributes to environmental protection by reducing the use of natural resources and minimizing waste generation.

In this framework, the following question could be posed: What is the true meaning of sustainability in the food industry, and why is it so relevant? And particularly, what policies are being applied for its implementation? In the European Union (EU) scope, sustainable methodologies are closely aligned with key policies such as the 'Farm to Fork' strategy (European Commission 2020). This strategy is an integral part of the European Green Deal (European Union 2019), an ambitious initiative launched by the European Commission in December 2019 with the aim of making the European Union (EU) climate-neutral by 2050. This comprehensive plan seeks to transform the EU economy towards a more sustainable model, promoting the transition towards a cleaner and more environmentally friendly society and industry.

The Farm to Fork strategy focuses on aspects such as food security, environmental sustainability, animal welfare, reduction of food waste, and promotion of healthier and more sustainable diets. By aligning with the promotion of more sustainable agricultural practices, reducing the use of pesticides and chemical fertilizers, protecting biodiversity, improving animal welfare, and reducing greenhouse gas emissions throughout the food chain. Additionally, the objective includes improving transparency and traceability in the food supply chain, which contributes to greater food security and consumer protection. By promoting the production and consumption of healthier and more sustainable foods, this strategy not only benefits the environment but also has a positive impact on public health and the resilience of food systems.

Furthermore, competitiveness in the food market is becoming increasingly intense, with consumers demanding high-quality, safe, and ethically produced products. Companies that adopt and implement innovations in their quality control processes not only meet these expectations but can also differentiate themselves in a saturated market, thus gaining a significant competitive advantage. In this context, a multidisciplinary approach is essential for institutions dedicated to research in the field of food quality and safety, as the ultimate tool to effectively address a challenge that affects the world globally, rather than considering it an exception. Therefore, collaboration between academia, industry, and regulatory bodies is essential to ensure that cleaner and more advanced methodologies are effectively implemented and adapted to the specific needs and realities of each context (Brennan 2024). It is also important to note that a large amount of information is now being generated and made available due to advances in analytical technologies. In turn, these technological advances in their idiosyncrasy make it require adopting multivariable analytical approaches in which artificial intelligence methodologies are applied to carry out the analysis of large volumes of data. However, despite its importance and potential positive impact, research and innovation in food quality control methodologies still face several constraints and challenges. These include lack of funding for research, resistance to change by established companies, complexity of the food supply chain, and the need to comply with a number of international regulations and standards.

In this context, this chapter aims to describe current green analytical approaches for evaluating quality and authenticity in the food industry, as well as the application of specific methodologies to artificial intelligence in this field. It should be noted that this chapter is a generic one, providing a prelude to the rest of the chapters in this book. To this end, each of the analytical techniques mainly used for non-invasive/non-destructive analysis of foods, generating sustainable methodologies, as well as the main applications in the food industry for the control of foods such as milk, honey, vegetable oils, among others, are covered.

2. Green analytical approaches in the food field

In 1998, Anastas and Warner proposed the twelve principles of green chemistry, which serve as criteria for assessing the "greenness" of an analytical procedure and are summarized in the acronym PRODUCTIVELY (see Figure 1a) (Anastas and Warner 1998, Tang et al. 2005). Building upon these, Galuszka, Migaszewski, and Namiesnik introduced in 2013 the 12 principles of green analytical chemistry (GAC), which can be abbreviated using the acronym SIGNIFICANCE (see Figure 1b) (Galuszka 2013). These principles involve all steps of the analytical process and, even if not explicitly, refer to sampling, sample preparation, sample analysis, instrument configuration, and supply. GAC in the realm of food quality represents a significant evolution in how food analyses are conducted. This approach is rooted in the integration of sustainability principles and the reduction of environmental impact in analytical practices, aiming to ensure the safety and quality of food products.

One key aspect of GAC is the reduction or elimination of the use of harmful solvents and chemical reagents, opting for more benign alternatives for both the environment and human health (Koel 2016, Nanda et al. 2024). This is achieved through the development and application of analytical techniques that require minimal amounts of solvents or employ less toxic solvents, such as ionic liquids or water instead of conventional organic solvents (Koel 2024, Venkatesan et al. 2024).

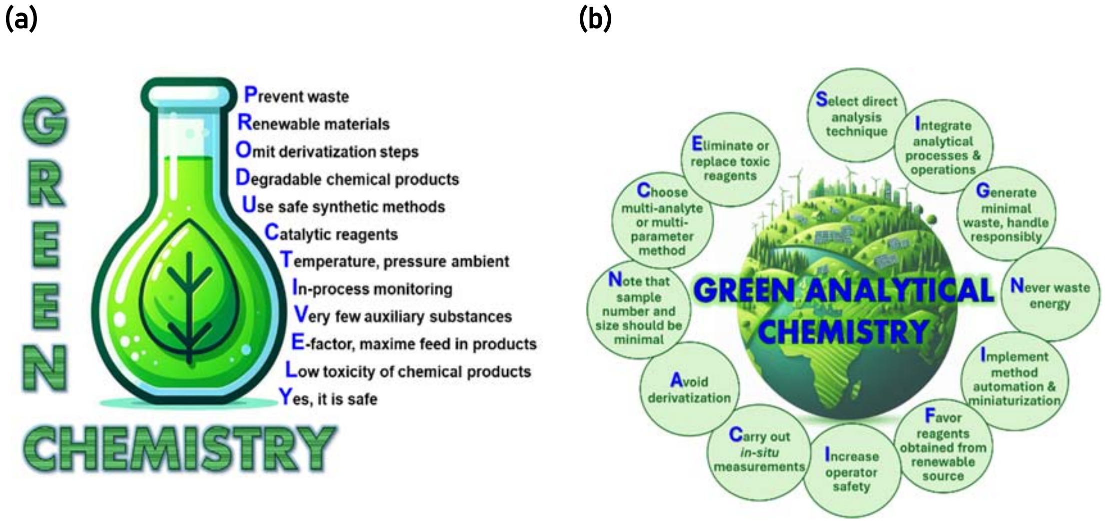

(a)
GREEN
CHEMISTRY
Prevent waste
Renewable materials
Omit derivatization steps
Degradable chemical products
Use safe synthetic methods
Catalytic reagents
Temperature, pressure ambient
In-process monitoring
Very few auxiliary substances
E-factor, maxime feed in products
Low toxicity of chemical products
Yes, it is safe
(b)
GREEN ANALYTICAL CHEMISTRY
Select direct analysis technique
Integrate analytical processes & operations
Generate minimal waste, handle responsibly
Never waste energy
Implement method automation & miniaturization
Favor reagents obtained from renewable source
Increase operator safety
Carry out in-situ measurements
Avoid derivatization
Note that sample number and size should be minimal
Choose multi-analyte or multi-parameter method
Eliminate or replace toxic reagents

Figure 1.9. (a) Principles of green chemistry and (b) Principles of green analytical chemistry (GAC).

Additionally, spectrometric analytical techniques that generate less waste and consume less energy are promoted, such as infrared spectroscopy or nuclear magnetic resonance (NMR), enabling the analysis of multiple components in a sample without the need for large amounts of solvents. These approaches not only contribute to environmental protection by reducing the quantity of hazardous chemicals released during analyses but can also enhance the efficiency and speed of analytical processes, resulting in improved capability to detect and quantify adulterants and/or contaminants in food. Ultimately, GAC plays a crucial role in strengthening food quality control systems, aligning them with sustainability and food safety objectives in the European Union and beyond.

In the same direction, various proposals have emerged for evaluating the so-called "greenness" of an analytical method, that is, its safety in terms of user health and environmental impact (Sajid and Plotka-Wasylka 2022, Krankovic 2023). This involves the constant development of new algorithms to assess "greenness", including both qualitative and quantitative approaches. In 2007, Keith et al. (Keith et al. 2007) suggested evaluating analytical methods using pictograms proposed by the National Environmental Methods Index (NEMI)

(National Environmental Methods Index 2002). Following them, in 2011, Guardia and Armenta proposed the evaluation in terms of operational risk, energy consumption, reagents, and volume of generated waste de la Guardia M, Armenta 2011). Similarly, Galuszka et al. in 2012 introduced the so-called Eco-Scale (Galuszka et al. 2012), which is a scale where different variables of the analytical process are assessed, assigning a score ranging from 0 to 100, with the latter value corresponding to an ideal situation. A method is considered excellent if it scores between 100 and 75; acceptable if between 50 and 75; and inadequate if below 50. This evaluation was later revised and modified by Armenta et al. in 2015 (Armenta et al. 2015a, Armenta et al. 2015b), in which the authors proposed classifying analytical methods through the "Green Certificate". This scale is based on the application of letters (from A to G) and colours (from green to red), with the A classification being the "greenest". Each score of the evaluated methods is associated with the number of penalty points assigned based on the use of reagents, the quantity used of them, their hazardous nature, energy consumption, potential occupational hazards, and generated waste. Even though the green metrics are a simple tool of ease application, only few reports have contemplated them to evaluate the analytical procedure sustainability. For instance, Espino et al. employed the Green Certificate to evaluate the methodology for extracting phenolic compounds in medicinal plants (Espino et al. 2018). Cina et al. employed the Green Certificate along with a variant applied through the software called AGREE for estimating the "greenness" of the analytical procedure for quantifying Ochratoxin A in coffee infusions (Cina et al. 2022). Similarly, Mateu et al. assessed the ranking of the analytical method process for determining copper and mancozeb in pesticide formulations within the Green Certificate ranking (Gallart-Mateu et al. 2016). El-Maghrabey et al. carried out a comprehensive study comparing and evaluating different chromatography-based analytical methods for the use of aldehydes as chemical markers of food quality using the AGREE software. The main conclusions were that it is crucial to avoid derivatization to the extent possible or select an appropriate non-toxic derivatizing agent. Additionally, the use of microextraction methodologies such as liquid-liquid microextraction (µLLE) and micro-solid phase extraction (µSPE), dispersive liquid-liquid microextraction (µDLLE), online solid-phase microextraction (online-SPµE), and gas diffusion microextraction are highly recommended for enhancing the ecological fingerprint of sample treatment steps by reducing the need for reagents and chemicals and minimizing sample volumes (El-Maghrabey 2024). In addition, the authors identified a gap in the AGREE software as certain derivatizing reagents were not included, rendering their evaluation impossible. This suggests that it is not only necessary to establish metrics or software, but also to continuously update them as new reagents and solvents emerge, and as methods must adapt to evolving changes and analytical methodologies.

Another type of metrics to evaluate analytical procedures, including the green analytical procedure index (GAPI) (Plotka-Wasylka 2018), complementary green analytical procedure index (ComplexGAPI) (Plotka-Wasylka and Wojnowski 2021), RGB color-method evaluation (Nowak and Kóscielniak 2019) or HEXAGON algorithm (Ballester-Caudet et al. 2019), have been proposed by other authors. All of them are based on the principles of GAC with slight modifications, such as incorporating energy consumption, or to be used when choosing which analytical method would be more sustainable, estimating the metric in each analytical procedure/method and carrying out a comparison among them.

Parallelly, other authors proposed the white analytical chemistry concept (WAC) as a complement and extension of the GAC principles and based on the RGB metric (Nowak et al 2021). This new concept is based on integrating the principles of GAC with additional considerations for analytical efficiency and practical/economic factors. The 12 WAC principles consist of 4 "green" rules (G1-G4), 4 "red" principles (R1-R4) related to analytical efficiency, and 4 "blue" principles (B1-B4) related to practical and economic criteria. While all principles and colours are deemed equally important in the concept of WAC, the actual significance of each may vary depending on specific circumstances. Analytical methods are evaluated based on these principles, with emphasis placed on those most crucial for a given application. The compliance of an analytical method with the WAC principles is quantified using a parameter called "whiteness", which measures how well the analytical method aligns with the intended application. This new approach has been discussed by other authors with the aim of bolstering its utility within the realms of food analytical chemistry and bioanalytics. Consequently, the authors have implemented the principles of WAC on illustrative cases within these domains, culminating in the principal finding that WAC principles facilitate the advancement of environmentally friendly, hygienic, safe, and sustainable analytical and bioanalytical methodologies (Hussain et al. 2023).

It is worth noting that, so far, all metrics proposals cited above to assess the sustainability of an analytical procedure have been based on ex-post evaluation, that is, an evaluation once the method has been developed and applied. However, what is interesting from a practical and sustainable point of view would be to carry out an ex-ante evaluation, that is, before developing the method. In this way, it would be possible to select the most sustainable methodology for the proposed purpose with clear understanding of the study objective and before applying one analytical technique or another (whenever possible using different techniques), or with the same technique, optimising resources. Accordingly, Jiménez-Carvelo et al. have proposed the use of WAC principles through a modification of the RGB metric, for ex-ante evaluating different analytical procedures in order to select the most sustainable one that meets the application requirements before conducting the analysis of the food samples under study (Jiménez-Carvelo et al. 2024).

At this juncture and continuing with the evaluation of analytical procedures from a sustainable perspective, it is worth mentioning the well-known approach termed the 'Analytical Procedure Lifecycle' (APC), as denoted by the United States Pharmacopeia (USP) (Martin et al. 2017), which stems from life cycle assessment (LCA). At the same time, this is derived from the ISO 14040 series which provides principles and a framework encompassing the entirety of the analytical methodology process (ISO 2006). This cycle refers to the set of stages that encompass from the planning and design of the analytical procedure to its implementation, execution, monitoring, and continuous improvement. It is in this final improvement stage where modifications of reagents and solvents are included, along with the reduction of chemicals used through automation and advanced flow techniques, miniaturization, and even the elimination of sampling through the measurement of analytes in situ, online, or in the field.

Considering the studies discussed above, it can be assumed that the scientific community's attention towards environmentally-friendly analytical practices has surged significantly. This is apparent from the continual rise in the volume of scientific publications addressing GAC issues.

3. Artificial intelligence in food quality assessment

Artificial Intelligence (AI) refers to the simulation of human intelligence processes by finely programmed computer systems. These processes include learning (the acquisition of information and guidelines for managing it), reasoning (using decision rules to reach approximate or definite conclusions), and self-correction and feedback (Xu et al. 2021). They achieve this through algorithms and massive amounts of data, often taking advantage of methodologies related to machine learning (ML) and deep learning (DL). It comprises employing diverse mathematical, statistical, and computational algorithms to detect patterns, correlations, and anomalies within the datasets.

Note that, numerous mathematics and statistics methods for analyzing extensive datasets have emerged, often bearing different names. However, within the food sector, terms such as 'data mining' and 'machine learning' have gained notoriety (Ayres et al. 2021, Jiménez-Carvelo and Cuadros-Rodríguez 2021, Tseng et al. 2023). 'Data mining' refers to the extraction of meaningful knowledge from useful but non-evident information, which is hidden within large datasets, while machine learning term is identified with the computational methods involved in dealing with vast data in the most intelligent fashion (by developing algorithms) to derive actionable insights (Analytics Vidhya). Additionally, 'deep learning' has been proposed as a subset of machine learning techniques, which employ deep artificial neural networks (ANN) to refine the multilevel architecture. In this chapter, 'data mining' and 'machine learning' terms are used interchangeably to denote the multivariate analysis methods applied within the food industry.

In recent years, AI has been playing an increasingly important role in the food sector, transforming the way food is produced, distributed, and consumed. From agricultural production to delivery to the end consumer, AI is being applied in various areas to improve efficiency, food safety, sustainability, and consumer confidence (Chandra et al. 2024, Yu et al. 2024). For example, through image analysis, sensors, and data, AI systems can quickly identify defects in food, such as bacterial contamination or spoilage, allowing for an immediate response to minimize risks to public health and maintain high quality standards. Furthermore, thanks to AI, it is possible to optimise the entire document management system. Proof of this has been the emergence of blockchain technology (Bosona and Gebresenbet 2023). In short, the blockchain is a chronological data structure in which transactions are grouped into groups or blocks, which are then recorded identically in a computer network (Casino et al. 2019). That is, instead of being warehoused in one specific location by a particular server, for instance in an Excel spreadsheet, the information is stored by generating many identical copies that are stored on several computers called nodes that are distributed over a network. However, it is noteworthy that blockchain has a high degree of complexity, verification mechanism and cost of implementation. Moreover, the confidentiality and data protection are some of the issues that need to be overcome in order to optimally implement this technology in the food chain.

In the realm of food quality, when discussing AI, it is undoubtedly necessary to consider three widely known approaches/strategies: Quality by Testing (QbT), Quality by Design (QbD) and Process Analytical Technology (PAT), all of which make use of AI. This handling can be either to monitor in real-time or to develop multivariable models using machine learning methods to predict properties/characteristics of the analyzed food product. These approaches can be summarized as (Pérez-Beltrán et al. 2023, Hitzmann et al. 2015, Breitkreitz 2021):

- QbT refers to the practice of ensuring product quality. This involves identifying and correcting defects, verifying that the product meets customer requirements and expectations, and ensuring its optimal performance before market release.
- QbD is a systematic, risk-based approach to product and process development. It focuses on understanding how variations in processes and materials affect the quality of the final product. Instead of detecting and correcting quality issues during later stages of the process, QbD aims to

design quality into the product from the outset. This is achieved by identifying and controlling Critical Process Parameters (CPPs) and Critical Quality Attributes (CQAs) from the beginning of product development. QbD involves a multidisciplinary approach involving teams of experts from various fields of knowledge.

❖ PAT is an approach that employs online and real-time analytical technologies to monitor and control manufacturing processes. PAT aims to understand and control key process variables in real-time, rather than solely relying on offline sample inspection and analysis. By integrating analytical instrumentation directly into the manufacturing process, PAT enables a deeper understanding of processes, quicker deviation identification, and overall improvement in production quality and efficiency. This technique aids in waste reduction, resource optimization, and ensures consistency in product quality.

Note that, the fundamental difference between QbT and QbD lies in when and how quality is addressed and incorporated into the production process. While QbT focuses on detecting and correcting defects after production, QbD aims to prevent defects by proactively designing and controlling the manufacturing process. As an add-on, PAT defines a process monitoring and control strategy that is developed in a successful QbD scenario.

AI-driven data analysis is commonly employed for guiding decisions in overseeing food processing and quality control. This involves extracting insights from complex datasets generated by analytical instruments. These instruments utilize various measurement approaches, including in line, on line, at line, and off-line methods, to monitor and control processes. In-line and on line acquisition provide real-time data, whereas at line and off line strategies involve sampling for subsequent analysis either on-site or in a laboratory setting. In line and on line acquisition are often applied for food processing, where AI can be used to control/evaluate steps such as extraction, separation, dehydration, drying, filtration, and piece-packaging, among others. Highlighting the fact that, one of the key operations in which AI methods are being widely used is in the drying stage of food production. In the realm of drying operations, ANN lead highest as the top AI method utilized. This complex method serves as a potent tool for the modelling, prediction, and optimization of critical aspects such as heat and mass transfer phenomena. Moreover, it also facilitates the monitoring of product quality spoilage, a critical concern in industrial processes (Jiménez-Carvelo et al. 2022).

Note that, despite the number of analytical technologies available for performing the monitoring the food processing, only a few assemble the specific requirements for application in food processing. These requirements include ease of use, non-invasive, rapid acquisition, preferably in-line or on-line capabilities, high informativeness, reliability, and suitability for industrial environments (Ropodi et al. 2016). In this context, on the one hand vibrational spectroscopy (near-infrared (NIR), mid-infrared (MIR), Raman, etc.) and machine vision (hyperspectral imaging (HSI), thermal imaging, etc.) (Saha and Manickavasagan 2021), and on the other hand, a particular type of electrochemical sensors such as electronic sensing devices (e-tongue and e-nose) can be used (Tan and Zu 2020).

As stated above at-line and off-line strategies involve further analysis of the raw data in the laboratory. The common data flow encompasses two stages (Jiménez-Carvelo et al. 2021): (i) data pre-processing and (ii) model development (see Figure 2).

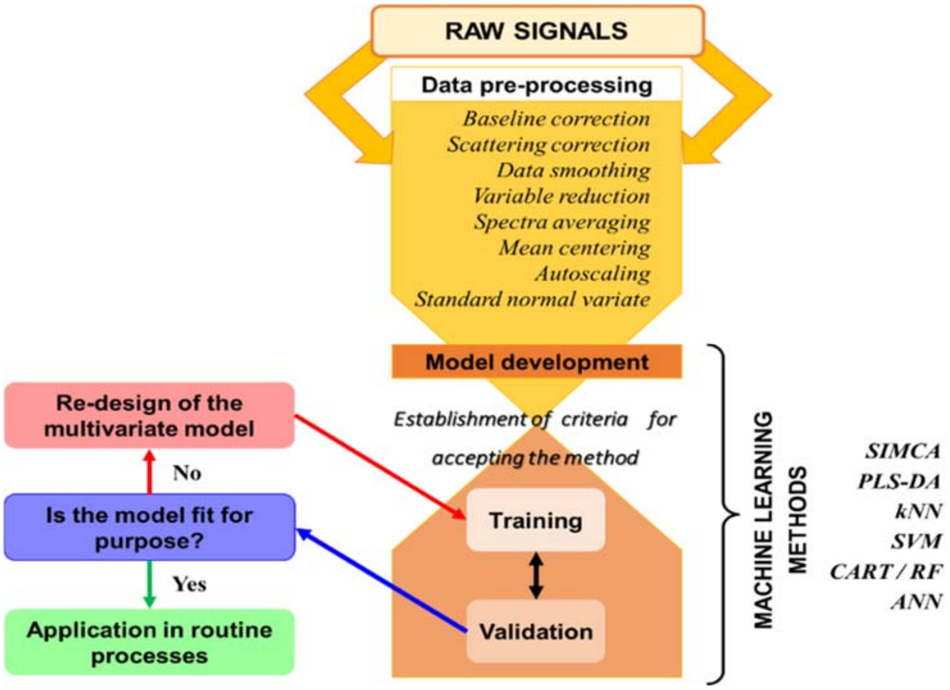


Figure 1.10. Workflow of the development of the multivariate model using ML/DP methods.

Before applying any AI algorithm, it is critical to perform a preparation step whose aim is to get an overview of the process data and select the most appropriate data samples for modelling. This step named as 'pre-processing' of the raw data can comprise the use of different pre-processing methods in order to linearise the responses and eliminate extraneous sources of variability that would otherwise compromise the predictability of the multivariate model. The commonly applied pre-processing methods include baseline correction, scattering correction, data smoothing, variable reduction, averaging, mean centring, autoscaling, or normalising (i.e., standard normal variate), among others.

Once the data have been pre-processed, appropriate AI methods are applied to develop the multivariate models. Given the differences among data collection environments, spectroscopic or imaging analytical techniques as well as quality evaluation targets, the appropriate processing methods for specific applications are often different. Thus, choosing an appropriate AI method is the first step when facing food quality evaluation. Usually, before the development of the predictive model, an exploratory/screening data analysis (unsupervised) is conducted to observe if there are natural groupings of objects and, by further, detect objects (samples) with anomalous behaviour that could be suspected of being outliers. Among the most prominent methods are principal component analysis (PCA) or cluster analysis (CA). Once this unsupervised study is completed, the establishment of the predictive or explanatory model proceeds. This involves fundamentally two stages: the first stage is based on model training, for which a large number of samples are used and whose features or classes are known. Following this first stage, the model validation stage is carried out. For the model validation stage, different strategies can be employed. On one hand, internal cross-validation can be conducted, and on the other hand, external validation should further be performed (Cuadros-Rodríguez et al. 2016). The difference between both strategies lies in the samples used for validation: while in the first, the same samples used in the model training stage are employed, in the second, samples different from those used in the model training stage are utilized. The choice between one strategy or another depends on the quantity of samples available for model development, although the second strategy is more realistic and reliable than the first one also avoiding what is known as model overfitting.

In addition to the above-mentioned, there is another critical aspect to consider regarding the validation of multivariable models or methods, which is not common when developing such AI-based methodologies, nevertheless is well established when validating conventional analytical methods focused on a univariate approach, where AI is not applied. According to ISO/IEC 17025 the validation step is defined as the provision of objective evidence that a given item fulfils specified requirements, where the specified requirements are adequate for an intended use (ISO/IEC 2017). Usually for validating the multivariate methods, several quality parameters such as sensitivity, specificity, precision, efficiency, among others, based on the results collected as a contingency table. After determining this quality metrics, researchers/analysts rule whether their method is considered acceptable or not. However, surprisingly, it is not common to establish previously the acceptable criteria based on the intended application goal of the method. Since, is any value of sensitivity and precision acceptable?

Is a method with a sensitivity and precision of 0.8 and 0.7 considered suitable for the proposed purpose? Such questions are resolved by establishing a priori acceptance criteria based on the intended application objective of the multivariable method. In this regard, Cuadros-Rodríguez et al. (Cuadros-Rodríguez et al. 2020) proposed the use of different indices named, error index (IERROR), saving index (ISAVING), penalty index (IPENALTY) and loss index (ILOSS) depending on the method's application scenario: trade/marketing or conformity assessment. Once these indices are established, it is possible to determine the minimum sensitivity and precision required for the developed multivariable method to be suitable for the intended purpose. Consequently, in order to make the decision on the validity of a multivariate method, the applicability indicators can be considered as a mandatory pre-step to be carried out before the validation process is performed.

The most commonly employed machine learning methods in the food field include k nearest neighbours (kNN), soft independent modelling of class analogy (SIMCA), partial least squares-discriminant analysis (PLS-DA), support vector machine (SVM), decision tree methods such as classification and regression tree (CART) and random forest (RF), and artificial neural network (ANN) and its derivatives such as convolutional neural networks (CNN), deep learning neural network (DLNN), radial basis artificial neural network (RBANN), among others (Belvilacqua et al. 2013, Oliveri and Downey 2012, Deng et al. 2024). Note that, SIMCA is not usually considered as an ML/DP method. Conversely, SIMCA has an advantage over other ML/DP methods since it allows the development of a classification/qualification model considering only a single input class. In other words, there is no need to train the model using samples from various classes, resulting in reduced economic expenses. Additionally, the food sector might find value in this efficient approach, utilizing it for final product oversight. For instance, ensuring product specifications, which could be seamlessly managed through a one-class classification approach. In this case, constructing a model exclusively from compliant products would serve the purpose of categorising any item deviating from the model as out-of-specifications. In this regard, some authors have argued against the inappropriate use of discriminant analysis in this food authentication scenario (Rodionova et al. 2016).

Aiming at food authentication, the above-mentioned classification methods have proven to be very useful for this purpose. However, quality assurance often involves complying with limits of certain critical parameters. For these cases, it is necessary to develop quantitation models to predict the value of a particular target feature, i.e., (i) the concentration/quantity of the constituent of interest (analyte) in the sample, (ii) determine method-defined quality parameters or (iii) estimate chemical proficiency indices for a total characterisation of the food product (Baena and Valcárcel 2003, Simonet et al. 2006). Among the most commonly used ML/DL methods are multivariate linear regression (MLR), principal component regression (PCR), or partial least squares regression (PLS) and its variant N-PLS for multi-way data. It is also possible to make use of the same algorithms discussed above for qualitative purposes, namely: SVM, deep learning methods (ANN, CNN, DLNN, RBANN) and algorithms based on decision trees (CART and RF) (Westad et al. 2013, Chu et al. 2022). The performance metrics used to evaluate multivariate quantification models differ from classification models. Following the recommendations published in the ASTM Standard Practice, at least: coefficient of determination of the model (R2), standard error of validation (SEV) (also known as the root mean square error of validation (RMSEV)), standard deviation of validation residuals (SDV), and validation bias (also known as mean absolute error of validation (MAEV)) should be considered for this evaluation (ASTM 2017). Note that, except for R2 which is a dimensionless parameter, the other metrics are expressed in the same units as the target feature to be quantified. As discussed above, critical values for the performance metrics should be established prior to model development, and these limits will depend on the application scenario (Jiménez-Carvelo et al. 2024).

Please note that in the last section of this book you will find in more detail about the use of ML/DP methods to carry out the analysis of different foods, using non-invasive/non-destructive analytical techniques.

4. Conclusions

The implementation of sustainable methodologies in the food industry is crucial to ensure the long-term viability of food production and evaluation, minimizing environmental impact, and promoting responsible practices. As evidenced by studies reported in the literature, analytical methodologies are being adapted to measure and improve sustainability at all stages, from production to distribution. However, challenges persist, such as the real-world implementation of non-invasive/non-destructive analytical methodologies, which go hand in hand with the application of artificial intelligence methods. All of this requires a paradigm shift from prevailing conventional methodologies. Additionally, it is necessary to have qualified staff in data processing to train personnel involved in food analysis activities using these new approaches. Thus, the application of AI in the food industry posed a challenge for manufacturers and analysts alike, as they strive to enhance all processes within their facilities to meet the required standards of quality control, production efficiency, risk assessment, and cost savings.

Acknowledgements

AMJC acknowledges the Grant (RYC2021-031993-I) funded by MCIN/AEI/501100011033 and "European Union NextGeneration EU/PRTR". **AAC** gratefully acknowledges the Spanish Ministry of Universities for a pre-doctoral fellowship FPU (FPU20/04711, Formación del Profesorado Universitario, *Training of University Lecturer*).

References

Analytics Vidhya. https://www.analyticsvidhya.com/ Accessed 08 May 2024

Anastas PT, Warner JC (eds) (1998) Green chemistry: theory and practice. University Press, Oxford.

Armenta S, de la Guardia M, Namiesnik J (2015a). Green microextraction. In Valcárcel M, Cárdenas S, Lucena R (eds) Analytical microextraction techniques, Betham Science Publishers, Sharjah, p 3-27.

Armenta S, Garrigues S, de la Guardia M (2015b) The role of green extraction techniques in green analytical chemistry. Trends Anal Chem 71: 2-8. https://doi.org/10.1016/j.trac.2014.12.011

ASTM E2617-17 (2017) Standard Practice for Validation of Empirically Derived Multivariate Calibrations, ASTM International, West Conshohocken.

Ayres LB, Gomez FJV, Linton JR, Silva MF, Garcia CD (2021) Taking the leap between analytical chemistry and artificial intelligence: A tutorial review. Anal Chem Acta 1161:338403. https://doi.org/10.1016/j.aca.2021.338403

Baena JR, Valcárcel M (2003) Total indices in analytical sciences. Trends Anal. Chem 22:641-646. https://doi.org/10.1016/S0165-9936(03)01101-4

Ballester-Caudet A, Campís-Falcó P, Pérez B, Sancho R, Lorente M, Sastre G, González C (2019) A new tool for evaluating and/or selecting analytical methods: Summarizing the information in a hexagon. Trends Anal Chem 118:538-547. https://doi.org/10.1016/j.trac.2019.06.015

Belvilacqua M, Bucci R, Magrì AD, Magrì AL, Nescatelli R, Marini F (2013) Classification and Class-Modelling. In: Marini F (ed) Chemometrics in Food Chemistry. Data Handling in Science and Technology, vol 28. Elsevier, Amsterdam, p 171-234.

Bosona T, Gebresenbet G (2023) The role of blockchain technology in promoting traceability systems in agri-food production and supply chains. Sensors 23:5342. https://doi.org/10.3390/s23115342

Breitkreitz MC (2021) Analytical quality by design. Braz J Anal Chem 8(2):1-5. https://doi.org/10.30744/brjac.2179-3425.editorial.mcbreitkreitz.N32

Brennan CS (2024) Regenerative food innovation: the role of agro-food chain by-products and plant origin food to obtain high-value-added foods. Foods 13(3):427. https://doi.org/10.3390/foods13030427

Casino F, Dasaklis TK, Patsakis C (2019) A systematic literature review of blockchain-based applications: Current status, classification and open issues. Telemat Inform 36:55-81. https://doi.org/10.1016/j.tele.2018.11.006

Chandra Natha P, Kumar Mishra A, Sharma R, Bhunia B, Mishra B, Tiwari A, Kumar Nayak P, Sharma M, Bhuyan T, Kaushal S, Kishore Mohanta Y, Sridhar K (2024) Recent advances in artificial intelligence towards the sustainable future of agri-food industry. Food Chem 447:138945. https://doi.org/10.1016/j.foodchem.2024.138945

Chu X, Huang Y, Yun YH, Bian X (2022) Nonlinear Calibration Methods. In: Chemometrics Methods in Analytical Spectroscopy Technology, Springer, Singapore, Ch 8, p 255-296.

Cina M, del Valle Ponce M, Fernandez L, Cerutti S (2022) A green approach for Ochratoxin A determination in coffee infusions. J Food Compos Anal 114:104777. https://doi.org/10.1016/j.jfca.2022.104777

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

Cuadros-Rodríguez L, Valverde-Som L, Jiménez-Carvelo AM, Delgado-Aguilar M (2020) Validation requirements of screening analytical methods based on scenario-specified applicability indicators. Trends Anal Chem 122:115705. https://doi.org/10.1016/j.trac.2019.115705

de la Guardia M, Armenta S (2011) Green analytical chemistry. In Barceló D (ed) Handbook of Comprehensive Analytical Chemistry, vol 57. Elsevier, Amsterdam.

Deng Z, Wang T, Zheng Y, Zhang W, Yun YH (2024) Deep learning in food authenticity: Recent advances and future trends. Trends Food Sci Technol 144:104344. https://doi.org/10.1016/j.tifs.2024.104344

Domingues GB (2023) Sustainability implications and relevance of using omics sciences to investigate cheeses with protected designation of origin. J Sci Food Agric. https://doi.org/10.1002/jsfa.13403

El Bilali H, Stassener C, Hassen TB (2021) Sustainable agri-food systems: environment, economy, society, and policy. Sustainability 13:6260. https://doi.org/10.3390/su13116260

El-Maghrabey MH, Hashem HM, El Hamd MA, El Shaheny R, Kishikawa N, Kuroda N, Magdy G (2024) Comprehensive greenness evaluation of the reported chromatographic methods for aldehydes determination as clinical biomarkers and food quality indicators. Trends Anal Chem 171:117548. https://doi.org/10.1016/j.trac.2024.117548

Espino M, Fernández MA, Gómez FJV, Boiteux J, Silva MF (2018) Green analytical chemistry metrics: Towards a sustainable phenolics extraction from medicinal plants. Microchem J 141:438-443. https://doi.org/10.1016/j.microc.2018.06.007

European Commission, A farm to fork strategy for a fair, healthy and environmentally-friendly food system, COM/2020/381 final, Document 52020DC0381, Brussels, 2020.

European Union. Communication (EU) 2019/640 from the Commission of 11 December 2019 on The European Green Deal. Document 52019DC0640.

European Union. Decision (EU) 2022/591 of the European Parliament and of the Council of 6 April 2022 on a General Union Environment Action Programme to 2030. Official Journal of the European Union L 114/22, 12 April 2022.

Galanakis CM (2024) The future of food. Foods 13:506. https://doi.org/10.3390/foods13040506

Gallart-Mateu D, Armenta S, de la Guardia M (2016) Green near-infrared determination of copper and mancozeb in pesticide formulations. Anal Bioanal Chem 408:1259-1268. https://doi.org/10.1007/s00216-015-9235-8

Galuszka A, Konieczka P, Migaszewski ZM, Namiesnik J (2012) Analytical EcoScale for assessing the greenness of analytical procedures. Trends Anal Chem 37: 61-72. https://doi.org/10.1016/j.trac.2012.03.013

Galuszka A, Migaszewski Z, Namiesnik J (2013) The 12 principles of green analytical chemistry and the SIGNIFICANCE mnemonic of green analytical practices. Trends Anal Chem 50:78-84. http://dx.doi.org/10.1016/j.trac.2013.04.010

Hitzmann B, Hauselmann R, Niemoeller A, Sangi D, Traenkle J, Glassey J (2015) Process analytical technologies in food industry – challenges and benefits: A status report and recommendations. Biotechnol J 10:1095-1100. https://doi.org/10.1002/biot.201400773

Hussain CM, Hussain CG, Keçili (2023) White analytical chemistry approaches for analytical and bioanalytical techniques: Applications and challenges. Trends Anal Chem 159:116905. https://doi.org/10.1016/j.trac.2022.116905

ISO 14040:2006. Environmental management – Life cycle assessment – Principles and framework. International Organization for Standardization, Geneva.

ISO 26000:2010. Guidance on social responsibility. International Organization for Standardization, Geneva.

ISO Guide 82:2019. Guidelines for addressing sustainability in standards. International Organization for Standardization, Geneva.

ISO/IEC 17025:2017, General requirements for the competence of testing and calibration laboratories. International Organization for Standardization, Geneva, 2017.

ISO/TS 26030:2019. Social responsibility and sustainable development – Guidance on using ISO 26000:2010 in the food chain. International Organization for Standardization, Geneva.

Jiménez-Carvelo AM, Arroyo-Cerezo A, Cuadros-Rodríguez L (2024) Evaluating the whiteness of spectroscopy-based non-destructive analytical methods – Application to food analytical control. Trends Anal Chem 170:117463. https://doi.org/10.1016/j.trac.2023.117463

Jiménez-Carvelo AM, Cruz CM, Cuadros-Rodríguez L, Koidis A (2022) Machine learning techniques in food processing. In: Tarafdar A, Pandey A, Sirohi R, Dussp CG, Soccol CR (eds) Current developments in biotechnology and bioengineering, 1st edn. Elsevier, Amsterdam, p 333-350.

Jiménez-Carvelo AM, Cuadros-Rodríguez L (2021) Data mining / machine learning methods in foodomics. Curr Opin Food Sci 3:76-82. https://doi.org/10.1016/j.cofs.2020.09.008

Jiménez-Carvelo AM, Martín-Torres S, Cuadros-Rodríguez L, González-Casado A (2021) Nontargeted fingerprinting approaches. In: Galanakis CM (ed) Food authentication and traceability, 1st edn. Elsevier, Amsterdam, p 163-193.

Keith LH, Gron LU, Young JL (2007) Green analytical methodologies. Chem Rev 107:2695-2708. https://doi.org/10.1021/cr068359e

Koel M (2016) Do we need green analytical chemistry. Green Chem 18:923-931. https://doi.org/10.1039/c5gc02156a

Koel M (2024) Developments in analytical chemistry initiated from green chemistry. SCENV 5:100078. https://doi.org/10.1016/j.scenv.2024.100078

Krankovic M (2023) Green chemical analysis: main principles and current efforts towards greener analytical methodologies. Anal Methods 15:6631. https://doi.org/10.1039/d3ay01644g

Martin GP, Barnett KL, Burgess C, Curry PD, Ermer J, Gratzl GS, Hammond JP, Herrmann J, Kovacs E, LeBlond DJ, LoBrutto R, McCasland-Keller AK, McGregor PL, Nethercote P, Templeton AC, Thomas DP, Weitzel MLJ, Pappa H (2017) Proposed new USP general chapter: The analytical procedure lifecycle ⟨1220⟩, Pharma Forum 43(1): 1-9. Accessed 26 Apr 2024

Nanda BP, Chopra A, Kumari Y, Narang RK, Bhatia R (2024) A comprehensive exploration of diverse green analytical techniques and their influence in different analytical fields. Sep Sc. plus 7:2400004. https://doi.org/10.1002/sscp.202400004

National Environmental Methods Index (NEMI) 2002. http://www.nemi.gov/home/. Accessed 24 Apr 2024

Nowak PM, Kóscielniak P (2019) What color is your method? adaptation of the RGB additive color model to analytical method evaluation. Anal Chem 91:10343-10352. https://doi.org/10.1021/acs.analchem.9b01872

Nowak PM, Wietecha-Posluszny R, Pawliszyn J (2021) White Analytical Chemistry: An approach to reconcile the principles of Green Analytical Chemistry and functionality. Trends Anal Chem 138:116223. https://doi.org/10.1016/j.trac.2021.116223

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

Pérez-Beltrán CG, Jiménez-Carvelo AM, Torrente-López A, Navas NA, Cuadros-Rodríguez (2023) QbD/PAT – State of the art of multivariate methodologies in food and food-related biotech industries. Food Eng Rev 15:24-40. https://doi.org/10.1007/s12393-022-09324-0

Plotka-Wasylka J (2018) A new tool for the evaluation of the analytical procedure: Green Analytical Procedure Index. Talanta 181:204-209. https://doi.org/10.1016/j.talanta.2018.01.013

Plotka-Wasylka J, Wojnowski W (2021) Complementary green analytical procedure index (ComplexGAPI) and software. Green Chem 23:8657-8665. https://doi.org/10.1039/d1gc02318g

Purvis B, Mao Y, Robinson D (2019) Three pillars of sustainability: in search of conceptual origins. Sustainability Sci 14:681–695. https://doi.org/10.1007/s11625-018-0627-5

Rodionova OY, Titova AV, Pomerantsev AL (2016) Discriminant analysis is an inappropriate method of authentication. Trends Anal Chem 78:17-22. https://doi.org/10.1016/j.trac.2016.01.010

Ropodi AI, Panagou EZ, Nychas GJE (2016) Data mining derived from food analyses using non-invasive/nondestructive analytical techniques; determination of food authenticity, quality & safety in tandem with computer science disciplines. Trends Food Sci Technol 50:11-25. https://doi.org/10.1016/j.tifs.2016.01.011

Saha D, Manickavasagan A (2021) Machine learning techniques for analysis of hyperspectral images to determine quality of food products: a review. Current Res Food Sci 4:28-44. https://doi.org/10.1016/j.crfs.2021.01.002

Sajid M, Plotka-Wasylka J (2022) Green analytical chemistry metrics: A review. Talanta 238:123046. https://doi.org/10.1016/j.talanta.2021.123046

Simonet BM, Lendl B, Valcárcel M (2006) Method-defined parameters: measurands sometimes forgotten. Trends Anal Chem 25:520-527. https://doi.org/10.1016/j.trac.2005.09.007

Spiliotopoulou M, Roseland M (2020) Urban sustainability: from theory influences to practical agendas. Sustainability 12(8):7245. https://doi.org/10.3390/su12187245

Tan J, Zu J (2020) Applications of electronic nose (e-nose) and electronic tongue (e-tongue) in food quality-related properties determination: A review. Artif Intell Agric 4:104-115. https://doi.org/10.1016/j.aiia.2020.06.003

Tang SML, Smith RL, Poliakoff M (2005) Principles of green chemistry: PRODUCTIVELY. Green Chem 7:761-762. https://doi.org/10.1039/b513020b

Tseng YJ, Chuang PJ, Appel M (2023) When machine learning and deep learning come to the big data in food chemistry. ACS Omega 8:15854:15864. https://doi.org/10.1021/acsomega.2c07722

Venkatesan K, Sundarababu J, Anandan SS (2024) The recent developments of green and sustainable chemistry in multidimensional way: current trends and challenges. Green Chem Lett Rev 17(1):2312848. https://doi.org/10.1080/17518253.2024.2312848

WCED (1987) Report of the World Commission on Environment and Development: Our Common Future. United Nations.

Westad F, Belvilacqua M, Marini F (2013) Regression. In: Marini F (ed) Chemometrics in Food Chemistry. Data Handling in Science and Technology, vol 28. Elsevier, Amsterdam, p 127-170.

Xu Y, Liu X, Cao X, Huang C, Liu E, Qian S, Liu X, Wu Y, Dong F, Qiu CW, Qiu J, Hua K, Su W, Wu J, Xu H, Han Y, Fu C, Yin Z, Liu M, Roepman R, Zhang J (2021) Artificial intelligence: A powerful paradigm for scientific research. The Innovation 2:100179. https://doi.org/10.1016/j.xinn.2021.100179

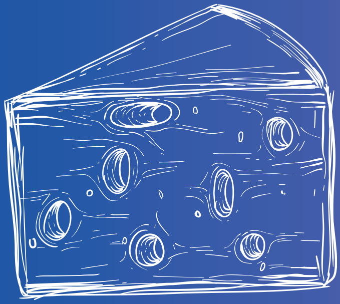
Yu Q, Zhang M, Mujumdar AS, Li J (2024) AI-based additive manufacturing for future food: Potential applications, challenges and possible solutions. Innov Food Sci Emerg Technol 92:103599. https://doi.org/10.1016/j.ifset.2024.103599

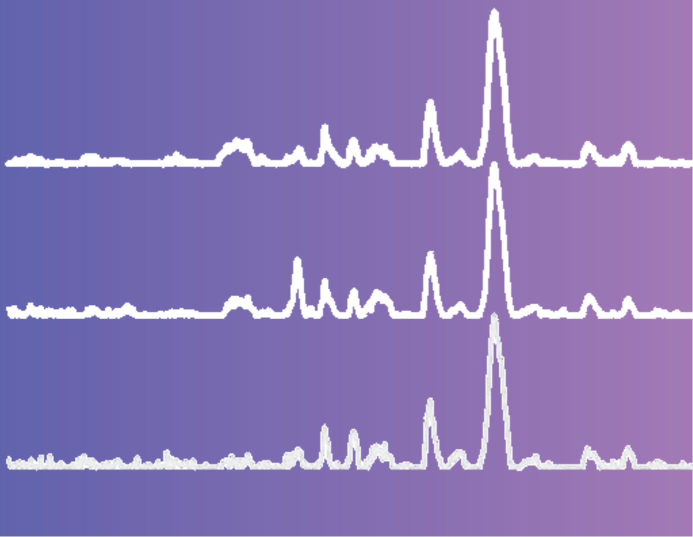
1.6. Contribuciones a congresos

1. A. Arroyo Cerezo, A.M. Jiménez Carvelo, L. Cuadros Rodríguez. **Nuevas técnicas de análisis no invasivo y su papel en la agricultura sostenible.** [Oral 15']. *RIARES 3º Seminario "Sistemas de Producción Agrícola y Sostenibilidad, Cosecha y Poscosecha. Almería (España), marzo 2023.*
2. A. Arroyo Cerezo, A.M. Jiménez Carvelo, L. Cuadros Rodríguez. **Hacia el desarrollo de métodos analíticos verdes para el control de la calidad de alimentos.** [Oral 5']. *IV Congreso / VI Jornadas de Investigadores en Formación Fomentando la Interdisciplinariedad. Granada (España), junio 2023.*

CAPÍTULO 2

Autentificación no invasiva de productos lácteos





2.1. Presentación

En este capítulo se presentan los resultados derivados de la aplicación de la técnica analítica de espectrometría Raman con compensación espacial (SORS, del inglés *spatially offset Raman spectroscopy*) como alternativa rápida y no invasiva para el control analítico de la calidad de dos tipos de productos lácteos envasados en plástico: margarinas y quesos, junto con la metodología de huellas instrumentales y el empleo de herramientas de minería de datos y aprendizaje automático.

La espectrometría Raman es una técnica vibracional basada en la medida del fenómeno de dispersión inelástico que tiene lugar tras irradiar un material en la región infrarroja del espectro electromagnético. Este fenómeno fue observado experimentalmente por primera vez en 1928 por C.V. Raman y K.S. Krishnan [**1**]. La señal adquirida en forma de espectro contiene información no específica sobre los enlaces químicos de los compuestos presentes en el material, dando lugar a una huella instrumental característica de su composición. Su uso en el campo del análisis de alimentos ha sido ampliamente explorado [**2**,**3**].

A lo largo de los años, se han desarrollado alternativas más avanzadas y sofisticadas basadas en esta espectrometría, con la intención de mejorarla abordando los inconvenientes o desventajas que presenta. La modalidad SORS es una de estas alternativas, que nació en 2005 en Gran Bretaña, concretamente en el laboratorio Rutherford Appleton, y fue propuesta por Pavel Matousek y su grupo de investigación [**4**]. La técnica SORS se desarrolló inicialmente para su uso en el ámbito biomédico y farmacéutico. La principal diferencia con la técnica convencional radica en la adquisición de la señal, ya que ésta se recoge en un punto desplazado respecto al punto de incidencia del láser, a diferencia de la convencional que recoge la señal en el mismo punto de incidencia, como muestra la Figura 2.1 [**5**].

Este fundamento convierte a la técnica SORS en una técnica no invasiva y, por ende, no destructiva, ya que permite realizar las medidas del material incluso a

[1] Smith, E.; Dent, G. *Modern Raman Spectroscopy: A Practical Approach,* 2nd ed.; John Wiley & Sons: Chichester, U.K., 2019, pp. 1-20.

[2] Xu, Y.; et al. Raman spectroscopy coupled with chemometrics for food authentication: A review. *Trends Anal. Chem.* **2020**, *131*, 116017. DOI: 10.1016/j.trac.2020.116017.

[3] Wang, K.; Li, Z.; Li, J.; Lin, H. Raman spectroscopic techniques for nondestructive analysis of agri-foods: A state-of-the-art review. *Trends Food Sci. Technol.* **2021**, *118*, 490-504. DOI: 10.1016/j.tifs.2021.10.010.

[4] Matousek, P.; et al. Subsurface Probing in Diffusely Scattering Media Using Spatially Offset Raman Spectroscopy. *Appl. Spectrosc.* **2005**, *59*, 393-400. DOI: 10.1366/0003702053641450.

[5] Mosca, S.; Conti, C.; Stone, N.; Matousek, P. Spatially Offset Raman Spectroscopy. *Nat. Rev. Methods Primers* **2021**, *1*, 1-21. DOI: 10.1038/s43586-021-00019-0.

través de una superficie, p.ej., a través del envase de las materias primas en el ámbito farmacéutico. Pocos años después de su surgimiento, se desarrollaron los primeros equipos portátiles basados en esta técnica para su uso en industrias farmacéuticas en el control de materias primas, o en aeropuertos para la detección de explosivos y drogas [**6**].

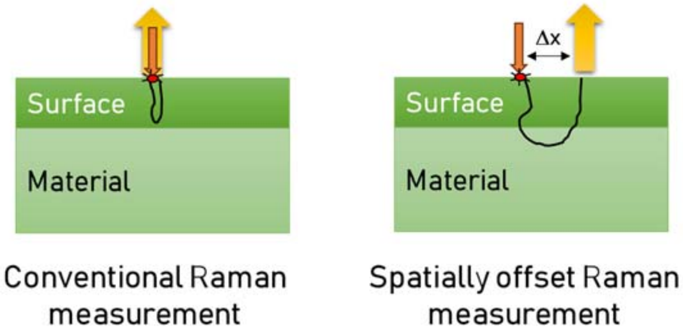

Surface
Material
Conventional Raman measurement
Δx
Surface
Material
Spatially offset Raman measurement

Figure 2.1. Diferencias en la adquisición de la señal Raman entre la técnica convencional y la modalidad SORS.

La disponibilidad de estos espectrómetros portátiles que proporcionan la posibilidad de realizar medidas *in situ*, rápidas y no invasivas, a través del envase de los productos, y sumado al potencial uso demostrado de los espectros Raman para el análisis de la calidad y autentificación de alimentos, hace que esta técnica cumpla con todos los requisitos necesarios para proveer de una alternativa totalmente "verde" en línea con los 12 principios de la química analítica verde (GAC, del inglés *green analytical chemistry*) [**7**] para el desarrollo de métodos de análisis de alimentos. Sin embargo, en el ámbito de la química alimentaria aún son pocas las aplicaciones publicadas, en comparación con otros campos de aplicación [**8**].

Por ello, este capítulo tiene como objetivo principal explorar la técnica analítica SORS en el desarrollo de métodos analíticos "verdes" para la autentificación de alimentos. De éste, derivan los siguientes objetivos específicos:

i. Recopilar un banco de muestras representativo de la variedad encontrada en el mercado de los alimentos bajo estudio: margarinas y quesos.

6. Matousek, P. Spatially offset Raman spectroscopy for non-invasive analysis of turbid samples. *Trends Anal. Chem.* **2018**, *103*, 209-214. DOI: 10.1016/j.trac.2018.04.002.

7. Gałuszka, A., Migaszewski, Z., & Namieśnik, J. The 12 principles of green analytical chemistry and the SIGNIFICANCE mnemonic of green analytical practices. *Trends Anal. Chem.* **2013**, *50*, 78-84. DOI: 10.1016/j.trac.2013.04.010.

8. Arroyo-Cerezo, A.; et al. Deep (offset) non-invasive Raman spectroscopy for the evaluation of food and beverages – A review. *LWT - Food Sci. Technol.* **2021**, *149*, 111822. DOI: 10.1016/j.lwt.2021.111822.

ii. Adquirir la huella instrumental característica de los productos recopilados mediante la técnica analítica SORS.

iii. Analizar los datos espectrales adquiridos (huellas instrumentales) a través del uso de herramientas quimiométricas.

iv. Desarrollar modelos supervisados totalmente validados para el análisis de la calidad y/o autenticidad de las muestras de alimentos recopiladas.

La consecución de estos objetivos ha derivado en la publicación de dos artículos científicos en revistas de alto impacto, cuyas referencias son:

1. Rapid and non-destructive spatially offset Raman spectroscopic analysis of packaged margarines and fat-spread products. *Microchem. J.* **2022**, *178,* 107378. DOI: 10.1016/j.microc.2022.107378.
2. The potential of the spatially offset Raman spectroscopy (SORS) for implementing rapid and non-invasive in-situ authentication methods of plastic-packaged commodity foods – Application to sliced cheeses. *Food Control* **2023**, *146,* 109522. DOI: 10.1016/j.foodcont.2022.109522.

El primero versa sobre el uso de SORS para el análisis cualitativo y cuantitativo de un conjunto de 62 margarinas comerciales adquiridas en diferentes países. Para el análisis cualitativo, se desarrollaron diversos modelos de clasificación basados en el origen geográfico y en la presencia o ausencia de ciertos ingredientes característicos de este producto. Los métodos quimiométricos de clasificación explorados, con el objetivo de evaluar cuál de estos algoritmos es capaz de extraer con mejores resultados la información implícita en los datos espectrales, fueron: análisis discriminante mediante regresión parcial de mínimos cuadrados (PLS-DA), sistemas de aprendizaje automático mediante vectores de soporte (SVM) y modelado flexible e independiente por analogía de clases (SIMCA). Por otro lado, el contenido total de grasa declarado en el etiquetado de los productos fue usado para desarrollar un modelo de cuantificación mediante el método de regresión parcial por mínimos cuadrados (PLS). Además, se realizó en paralelo el mismo procedimiento de adquisición de medidas, tratamiento y análisis de los datos y desarrollo de los modelos supervisados mediante un espectrómetro Raman convencional, con el fin de comparar los resultados obtenidos.

Por otro lado, el segundo articulo recoge los resultados de analizar 80 muestras comerciales de quesos en lonchas que diferían entre ellos en el origen animal de la leche empleada para la fabricación de los mismos. Tras el tratamiento de los datos, se llevó a cabo un análisis de la similitud de las señales espectrales adquiridas. Asimismo, empleando la estrategia *one input-class* (1iC), se desarrollaron modelos de clasificación mediante un modelado de clases con el método SIMCA, y modelos de cuantificación del contenido total en grasas y proteínas mediante el uso de PLS.

En ambos casos las medidas se llevaron a cabo a través del envoltorio de plástico de forma no invasiva, sin abrir los envases originales de los productos, que quedaban totalmente intactos tras la medida. Esto proporciona una gran ventaja y demuestra el potencial uso de esta técnica en el ámbito del control de la calidad de producto terminando.

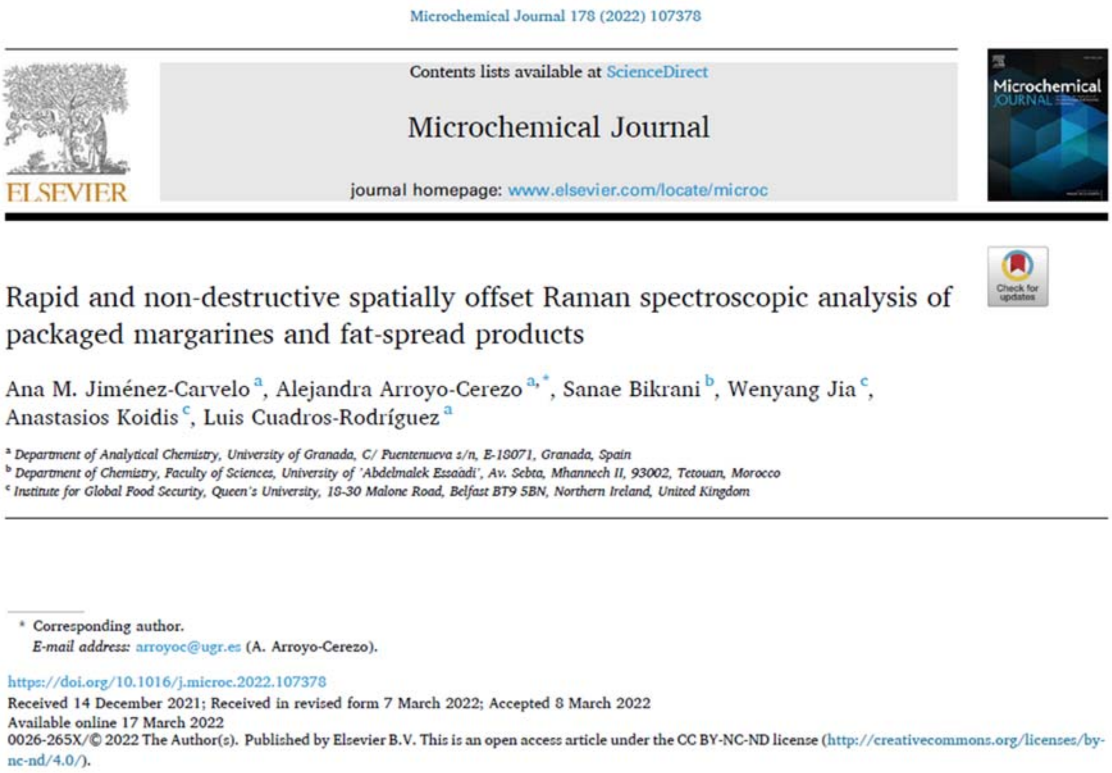
Los siguientes apartados presentan ambas publicaciones, que describen en detalle los estudios realizados, resultados y conclusiones obtenidos.

2.2. Artículo científico 1

Rapid and non-destructive spatially offset Raman spectroscopic analysis of packaged margarines and fat-spread products.

Publicado en 2022 en la revista Microchemical Journal 178, 107378.

DOI: 10.1016/j.microc.2022.107378.



Microchemical Journal 178 (2022) 107378

Contents lists available at ScienceDirect

Microchemical Journal

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

Rapid and non-destructive spatially offset Raman spectroscopic analysis of packaged margarines and fat-spread products

Ana M. Jiménez-Carvelo [a], Alejandra Arroyo-Cerezo [a,*], Sanae Bikrani [b], Wenyang Jia [c], Anastasios Koidis [c], Luis Cuadros-Rodríguez [a]

[a] Department of Analytical Chemistry, University of Granada, C/ Fuentenueva s/n, E-18071, Granada, Spain
[b] Department of Chemistry, Faculty of Sciences, University of 'Abdelmalek Essaâdi', Av. Sebta, Mhannech II, 93002, Tetouan, Morocco
[c] Institute for Global Food Security, Queen's University, 18-30 Malone Road, Belfast BT9 5BN, Northern Ireland, United Kingdom

[*] Corresponding author.
E-mail address: arroyoc@ugr.es (A. Arroyo-Cerezo).

https://doi.org/10.1016/j.microc.2022.107378
Received 14 December 2021; Received in revised form 7 March 2022; Accepted 8 March 2022
Available online 17 March 2022

Highlights:

- In-pack measurements in margarines (n=62) using spatially offset Raman spectroscopy (SORS).
- Chemometric models from SORS data provided better classification compared to conventional Raman data.
- SIMCA models for margarines containing phytosterols, olive oil or linseed oil, exhibit very high predictability.
- SORS can bypass fluorescence, a major drawback when analysing food samples with conventional Raman spectroscopy.

Keywords:

Spatially offset Raman spectroscopy (SORS)

Non-destructive analytical techniques

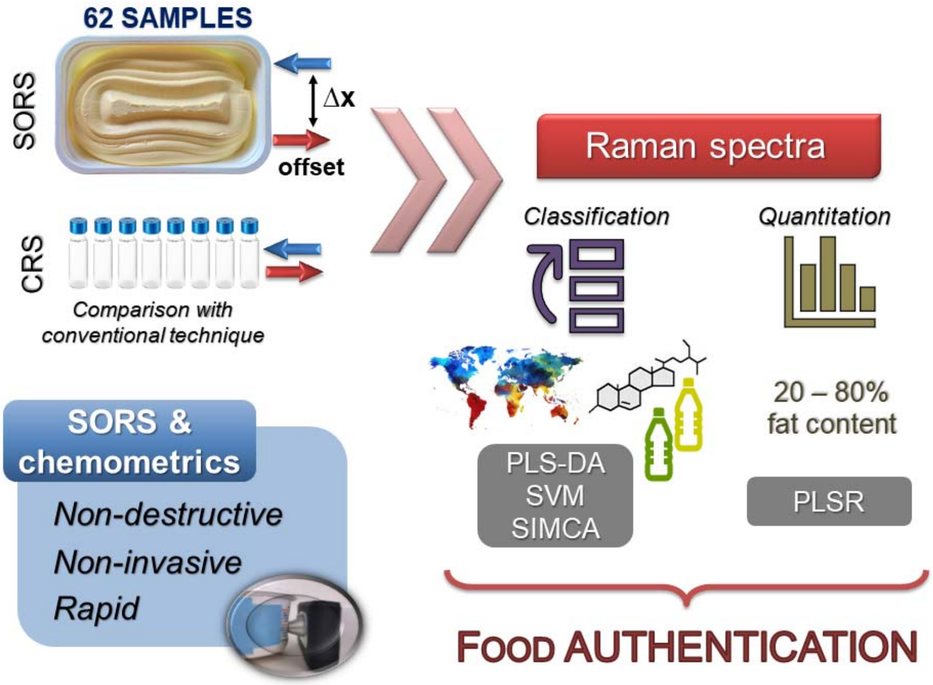
Chemometrics and data mining

In-pack measurement

Food quality and authenticity

Margarines and fat-spreads.

Graphical abstract



Abstract

Spatially offset Raman spectroscopy (SORS) is a novel technique capable of measuring samples through the original packaging and recovering the spectra without the contribution of surface layers. Here, a portable SORS equipment was used to measure 62 samples of margarines and fat spreads through the original plastic container. Chemometric tools were used to analyse the data obtained. A total of 25 classification models were developed based on: (i) geographical origin, (ii) vegetable oils and (iii) some significant minor constituents present in the samples. Partial least squares-discriminant analysis (PLS-DA), support vector machine (SVM) and soft independent modelling of class analogy (SIMCA) were used for model classification. Quantitative analysis using the partial least squares regression (PLSR) method was also performed to determine the total fat content. In parallel, a benchtop conventional Raman spectrometer was used to analyse the same samples, develop the models with the same training and validation sets in order to compare the results. The calculated classification performance metrics showed better classification models from SORS data than conventional Raman spectroscopy (CRS), highlighting the one-class SIMCA models for margarines containing phytosterols, olive oil or linseed oil. These models exhibited very high predictability (performance parameters with values equal to or higher than 0.8, 0.9 and 1, respectively). The quantitation model developed from SORS exhibited a higher R2 than from CRS data, and prediction errors below 5% from SORS versus errors between 5 and 13% from CRS data.

These results reveal the ability of SORS to avoid the influence of fluorescence, a major drawback when analysing Raman spectra, but also the potential of the technique as a fast, non-destructive and non-invasive analytical technique in the field of food analysis. In conclusion, the tandem 'SORS-chemometrics' has been shown to be a potential tool in the food quality and food authentication fields. Thus, it is necessary to perform further investigations in this field in order to advance the knowledge of this technique and to be able to develop new methods of rapid analysis.

1. Introduction

Advanced Raman spectroscopy techniques such as spatially offset Raman spectroscopy (SORS) have demonstrated the ability to overcome some disadvantages of conventional Raman spectroscopy (CRS). When SORS is applied, Raman signal is acquired at a certain distance from the laser incidence (some millimetres) and the collected signal provides spectral information of both the outer and inner layers of the measured material due to the spatially shift. This shift allows the deep layers photons to be emitted from a laterally shifted point of the incidence region while the surface photons are emitted from the same incidence point. The pathway followed by the interior photons is randomised, so it is more likely to retrieve this interior emission from a shifted point than from the same point of incidence [1]. For this, SORS becomes even more relevant by being able to collect photon emission from the inside of diffusely scattering materials, since most samples are either opaque or contained in opaque materials, while allowing measurements to be performed without damaging the sample, making SORS a remote, non-invasive and non-destructive technique [2,3].

Ensuring the stability of the sample during measurements as well as properly selecting the optimal offset are important parameters for measurement success. The optimal offset is characteristic for each type of sample as the intensity of spectral bands and the depth range to be reached depends on it [1,4]. Conventional instruments to perform SORS measurements are custom design builds, primarily composed of a laser source, a charge-coupled device (CCD) camera as a sensor and a fibre bundle or fibre optic. Moreover, filters are used to supress the fluorescence radiation emitted by the measured sample that could interfere with the low intensity Raman signals. Usually, wavelengths excitation in SORS measurements are 785, 830 and 1064 nm. The latter has demonstrated a greater ability to remove fluorescence [2].

However, since 2017, it is possible to acquire Raman signals with branded portable and handheld SORS equipment [5]. These instruments were manufactured for particular applications, making easier fast and effective analysis, e.g., for hazardous substances security screening or raw materials control in pharmaceutical industry. Commercial equipment still has some drawbacks inherent to the Raman spectroscopy, such as the inability to perform measurements through fully opaque package materials as metal materials, including tetrabrik. Minor sensitivity and lower signal-to-noise ratio in collected Raman signals have also been reported using commercial instruments [2].

Note that when Raman spectroscopy is employed as an analytical technique to evaluate the authenticity or quality of foodstuffs, the signal obtained is considered as non-specific, in which all the chemical and structural information of the sample is collected. In other words, this signal is an instrumental fingerprint and therefore it is necessary to apply chemometric/data mining tools in order to extract the information of interest which is not shown in an evident form [6]. The tandem composed by fingerprinting methodology and chemometrics is focused on the development of multivariate (qualitative or quantitative) models using proper mining/machine learning algorithms [5], so-called pattern recognition methods. The aim is to establish the belonging to one class or another of a set of samples with a characteristic feature (e.g., origin, ingredients, manufacturing, differentiated quality claims, etc.) or to carry out the quantitation of one or more feature-related parameters [8]. Chemometric pattern recognition methods are usually divided into two categories:

unsupervised and supervised methods. Unsupervised methods show the intrinsic data pattern and are typically used to exploratory purposes. The most common unsupervised approaches are hierarchical cluster analysis (HCA) and principal components analysis (PCA). Supervised methods consider the belonging class of the sample, and the development of model involves a training step followed by a validation step which can be performed using new samples different from those used in the training step (external validation) or using the same samples (cross-validation). Some of the supervised methods include k-nearest neighbours (kNN), soft independent modelling by class analogy (SIMCA), partial least squares-discriminant analysis (PLS-DA), or support vector machine (SVM). Furthermore, as far as quantitative analysis is concerned, the most widely used is by partial least squares regression (PLSR) [9].

SORS measurements in combination with chemometric tools have been used for multiple applications in different fields. To date, the most explored with industrial application is pharmaceutical industry for the identification of raw materials, drug detection through packaging or control the adulteration of drugs [2]. Other fields of application are the non-invasive analysis of artworks [10], the detection of explosives in liquids in the context of the security [11], and biomedical applications [12] or some in the food and beverage sector [2,4]. However, despite the potential of this analytical technique to be used in the food quality control, food safety or food authentication fields among others, applications of SORS-chemometrics are still limited in the scientific literature and as far as is known, they are not yet applied at the industry level.

Margarines and related fat-spread products could be an appropriate target for SORS research. The container for these products is usually made of plastic, a material that is permeable to the laser allowing acquisition of Raman spectra of the sample contained inside. Historically, margarine was created to replace butter as a lower cost option. It is a water-in-oil solid emulsion composed mainly of vegetable fats (such as sunflower, rapeseed, palm and olive oil) and rarely animal fats up to a maximum of 3% [13]. Unlike butter which must derive only from milk, other ingredients are also permitted in margarine manufacturing such as phytosterols, vitamins, minerals or sugars as well as additives, including colouring agents, emulsifiers, stabilisers or antioxidants [14]. Fat-spreads are classified according to the total fat percentage by Codex Alimentarius. Commonly, for a product to be properly called margarine, it must have at least 80% fat, otherwise it is called fat-spread (with less than 80%) [15]. However, the EU legislation establishes more categories: (i) margarine, if fat content is between 80% and 90%, (ii) three-quarter-fat margarine, if fat content is between 60% and 62%, (iii) half-fat margarine if fat content is between 39% and 41% and (iv) fat spreads X% for the rest of fat content percentages [16]. The legislation of other counties may be different. For instance, Moroccan law provides that margarine is any edible fat other than butter and lard. This legislation does not specify a minimum fat percentage required to call the product margarine, and also allows for the addition of up to 10% milk fat, either from milk or whipped cream [17].

Most traditional analytical techniques, such as high-performance liquid chromatography (HPLC) or gas chromatography (GC), employed to analyse this type of product involve several sample pre-treatments, such as previous extraction or isolation of the required compound or compounds family as well as need long time to perform the analysis [18]. However, vibrational spectroscopy techniques such as Fourier transform-near infrared (FT-NIR) provide the possibility to conduct rapid, simple and non-destructive analysis of margarines and fat-spreads in terms to assess proper quality control and food authentication [19]. Among these techniques, SORS has already been applied to detect the presence of margarine in butter, i.e., to detect butter adulteration [20]. Thus, a food such as margarine could be a good product to be analysed with the SORS technique in a fast, non-invasive and non-destructive form, for authenticating this food product.

The present paper aims the use of SORS to extract Raman spectra of a set of margarines and fat-spread products measured through the original packaging using a recently commercialised handheld instrument. Different multivariate models have been developed using unsupervised (PCA) and supervised (PLS-DA, SVM, SIMCA) methods for prior screening and qualitative classification, and PLSR to quantify the fat contents of the samples. To further improve the reliability of the measurements performed, a comparison with CRS has been also carried out.

2. Material and methods

2.1. Margarines and spreads samples

A total of 62 samples of margarines and fat-spreads from different geographic origins of production were analysed. Table 2.1 shows the different geographic origins of manufacture and the number of samples from each of them. The samples were purchased in different local grocery shops and supermarkets in Spain, France, United Kingdom and Morocco. As for the samples whose manufacturing origin is different (Belgium, Germany and Holland), they were purchased in retail in Spain. In addition to the geographic origins of manufacture, the samples differ in composition (ingredients) and amount of macronutrients (carbohydrates, fats and proteins). After purchase, the samples were stored under conditions similar to those of retail sale, i.e., refrigerated at 4°C.

For CRS measurements, a small portion of each sample (approximately 10 g) was transferred to a vial for analysis. For the SORS measurements, this was not necessary as they were measured through the original container.

Table 2.1. Geographical origin of the margarine samples included in the study.

Geographical origin		Number of samples
Europe	Spain	19
	France	13
	United Kingdom	12
	Belgium	4
	Germany	1
	The Netherlands	1
Morocco		12
Total number of samples		**62**

2.2. SORS measurements

The commercial SORS equipment used for measuring margarines and fat-spreads products through the original packaging was the Vaya Raman (Agilent Technologies, Santa Clara, CA, USA). This device uses a laser with an excitation wavelength of 830 nm, which allows fluorescence to be suppressed. The spectral range was 350-2000 cm^{-1}, while the maximum power of the laser was 450 mW (100% power), which is user adjustable.

The equipment performs two measurements: with zero offset (equivalent to the CRS spectrum) and with spatial offset, namely with an offset of 0.7 cm from the point of incidence of the laser to the collection point. After internal signal processing, the SORS spectrum is obtained: a Raman spectrum of the sample without the influence of the container layers. Figure 2.2 shows an example of the spectra obtained. Figure 2.2A shows the raw spectra of the container (spectrum without offset, light green line) and of one of the samples without offset (dark green line) and with offset (blue line). Note that the contribution of the container seen in the spectrum (dark green line), coinciding with the spectrum of the container, is completely subtracted from SORS measurement. Figure 2.2B shows the final spectrum of the same sample pre-processed and normalised. The 'final spectrum' obtained from SORS were the data used to carry out the multivariate data treatment.

Measurements were taken directly from the original packaging of the 62 margarine or fat-spread samples. The measurement time for each sample was between 30 s and 2 min, while the exposure time of the samples to the laser was 0.5 to 2 s.

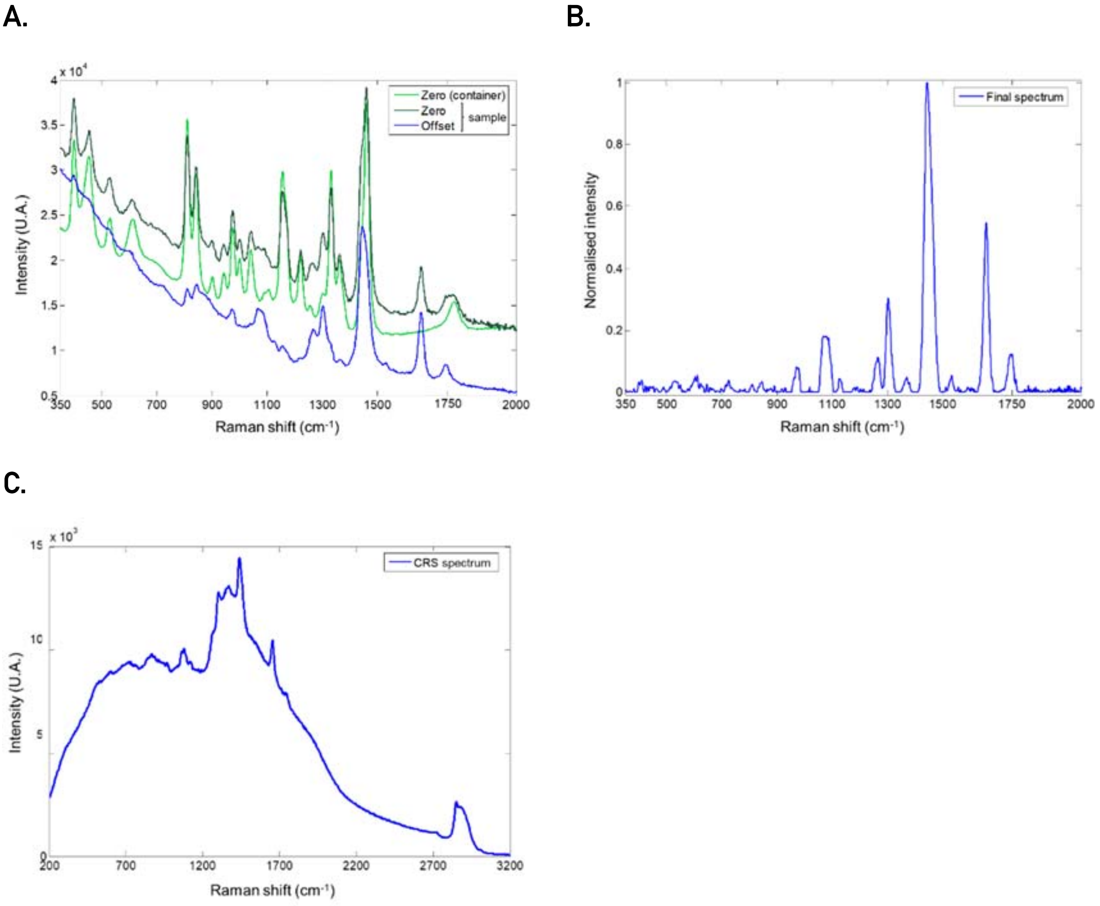


Figure 2.2. Raw Raman spectra of the container (zero offset) and one of the samples measured through packaging without and with offset using SORS (**A**), final Raman spectrum of the same sample after pre-processing and normalise using SORS (**B**), Raman spectrum of the same sample using conventional Raman spectroscopy (CRS) (**C**).

2.3. CRS measurements

The measurements were performed with the IDRAMAN Reader (Ocean Optics, Oxford, UK), equipped with laser wavelength at 785 nm with 100 mW laser power. The scattered light was collected on a 2048-element NIR-enhanced CCD array with thermoelectric cooling to -10 °C. The samples were heated to 40 °C in a water bath until fully mixed, and a 2 ml liquid was placed into a sealed glass vial. For the spectra acquisition, 2 mL vial analysed sample was used and each sample was obtained three times. The spectral range was 200–3200 cm^{-1} with an integration time of 20 s, and all Raman measurements were conducted at room temperature. Figure 2.2C shows an example of the obtained spectra of one of the samples using CRS.

2.4. Multivariate data treatment

Raman and SORS raw data were exported from CSV format (comma-separated vectors) to .mat using MATLAB (Mathworks, Massachusetts, USA, version R2013a).

All multivariate data treatment was carried out using PLS_Toolbox (Eigenvector Research Inc. MA, USA, version 7.5.0) working under the MATLAB framework. Chemometric tools applied were (i) PCA as a non-supervised pattern exploratory method, (ii) PLS-DA, SVM and SIMCA as supervised pattern recognition methods to build different classification models and (iii) PLSR to develop quantitation models to estimate fat percentage. All the above methods were applied to both Raman and SORS data and full comparisons are presented in the results section.

For the classification models, a plot with training and validation set will be presented in the next section. The discrimination threshold was established at 0.5 for all models and a range of ±0.1 was assigned to designate an area of inconclusive results. In addition, a cut-off was also set at ±1 (i.e., 1.5 above and -0.5 below) and samples falling in these areas, both at the upper and lower limits, were assigned to the inconclusive results group when calculating quality parameters.

The most appropriate pre-processing was chosen according to the multivariate method to be used. Different pre-processing methods were tested and the best results were obtained by applying mean center or autoscaling, as discussed in the next section.

A total of 50 classifications models were developed according to different characteristics of the samples. Firstly, 8 models related to the different geographical origin were developed, 4 for each of the techniques used (CRS and SORS), differentiating between samples manufactured in Morocco, Spain, France and United Kingdom. Then, 42 models (21 with the data obtained for each technique) based on the different ingredients that constitute the margarines analysed in order to differentiate between those that have a particular ingredient from those that do not. The ingredients consider as 'target class' were sunflower oil, olive oil, linseed oil, palm oil or fat, buttermilk, phytosterols and lecithin. Different classification performance metrics, such as sensitivity, specificity, positive and negative predictive value or efficiency, among others, were determined according to the tutorial published by Cuadros *et al.* [21]. For the four quantitation models developed, the root mean square error of validation (RMSEV), the mean absolute error of validation (MAEV), the median absolute error of validation (MdAEV), the standard deviation of validation residuals (SDV) [22] and the coefficient of determination (R^2) were used to evaluate prediction accuracy. A comprehensive layout is shown in Figure 2.3.

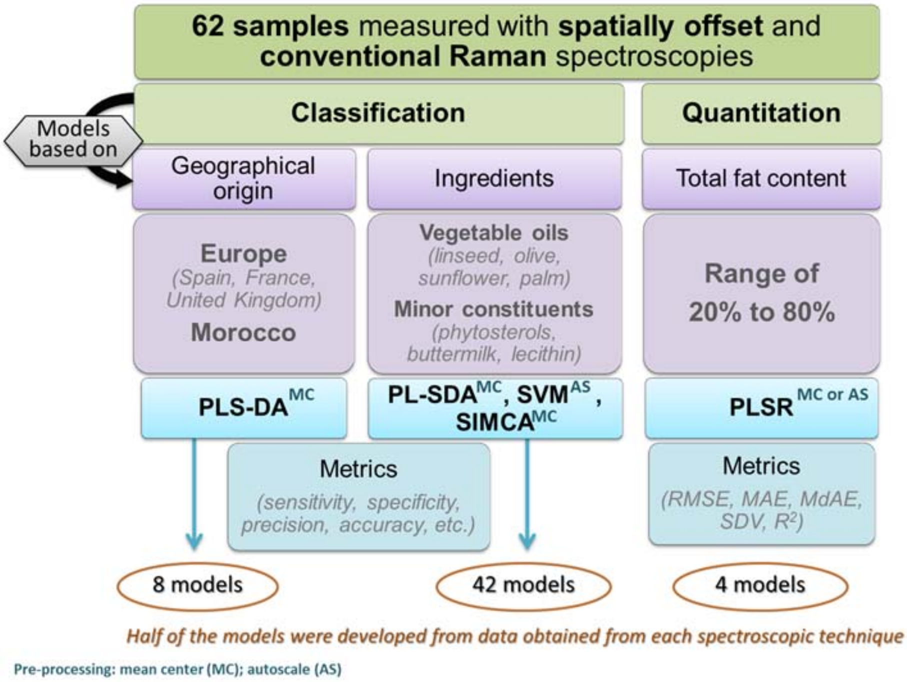

62 samples measured with spatially offset and conventional Raman spectroscopies
Models based on
Classification
Quantitation
Geographical origin
Ingredients
Total fat content
Europe (Spain, France, United Kingdom) Morocco
Vegetable oils (linseed, olive, sunflower, palm)
Minor constituents (phytosterols, buttermilk, lecithin)
Range of 20% to 80%
PLS-DA MC
PL-SDA MC, SVM AS, SIMCA MC
PLSR MC or AS
Metrics (sensitivity, specificity, precision, accuracy, etc.)
Metrics (RMSE, MAE, MdAE, SDV, R2)
8 models
42 models
4 models
Half of the models were developed from data obtained from each spectroscopic technique
Pre-processing: mean center (MC); autoscale (AS)

Figure 2.3. Comprehensive layout of the strategy used to perform the classification and quantitation models.

Note that for classification purposes, samples were divided into two general classes. For the geographical origin models, one class was the country in which the samples were manufactured and the other class consisted of the rest of the samples manufactured in another country (e.g. for 'Spain / no Spain' model, the samples of the 'no Spain' class were the samples manufactured in the UK, France, Belgium, Germany and Netherlands). And for the ingredient models, one class is the 'target class' (samples containing the ingredient in its composition) and the other class is composed of the samples that do not have that ingredient.

3. Results and discussion

The Raman spectra peak at 1750 cm^{-1} (see Figure 2.2B) corresponds to the ester bonds related to the presence of triglycerides [23]. The band observed at 3050-3090 cm^{-1} (see Figure 2.2C) may be attributable to lipids, corresponding to the peaks observed in different vegetable oils in the study of Dyminska *et al.* [24]. Peaks around 875 cm^{-1} may be attributed to phosphodiester symmetric stretching of the phospholipids present in the margarine samples [25]. These peaks only showed in either SORS or CRS (Table 2.2), therefore, these three bands were excluded for the model development. The rest of the peaks correspond to margarine samples can be found below. The band around 1654 cm^{-1}, belongs to the aromatic ring stretch, were related to oil content [26,27]. The peak around 1440 cm^{-1} is related to CH2 deformations [28], while the peak observed around 1260 cm^{-1} is assigned to the (C–H) bending vibration at the cis double bond in R–HC=CH–R [29]. And the band at 1061 cm^{-1} originate from the (C–C) stretching [30].

Table 2.2. Band assignment of the SORS and CRS spectra.

Raman shift from SORS (cm^{-1})	Raman shift from CRS (cm^{-1})	Molecule	Group	Vibration
1750	—	RC=OOR	C=O	stretching
1656	1654	*cis* RCH=CHR	C=C	stretching
1442	1440	– CH_2	C–H	deformation
1264	1252	*cis* RCH=CHR	=C–H	deformation
1067	1061	– $(CH_2)_n$ –	C–C	stretching
—	875	– $(CH_2)_n$ –	C–C	Stretching

"—": peak not shown in the spectrum. *Table adapted from [23].*

3.1. Exploratory analysis

PCA was employed to study the natural grouping of the 62 margarine samples. The pre-processing chosen was mean centring for both CRS and SORS data. Five and eight principal components (PCs) explained 86.91% and 99.96% of the cumulative variance (CV) for the SORS and CRS data, respectively. Figure 2.4 shows the scores obtained for each sample in the first two PCs of the SORS and CRS data. It is evident that with both techniques is possible to establish a grouping of the samples according to their geographical origin of production. Figure 2.4A and 2.3B illustrate a clear separation between European or Moroccan origin, and in Figure 2.4C and 2.4D, slight differences and natural groupings can be seen, differentiating those of European origin between Spain, France, United Kingdom and other countries.

For the data obtained from CRS measurements, the positive part of PC1 is related to the samples coming from Morocco and the negative contribution of the same PC to the European samples. However, regarding the SORS data, the difference between the two origins is related to the negative contribution of PC2 for the Morocco origin and the positive contribution of the same PC for the Europe origin.

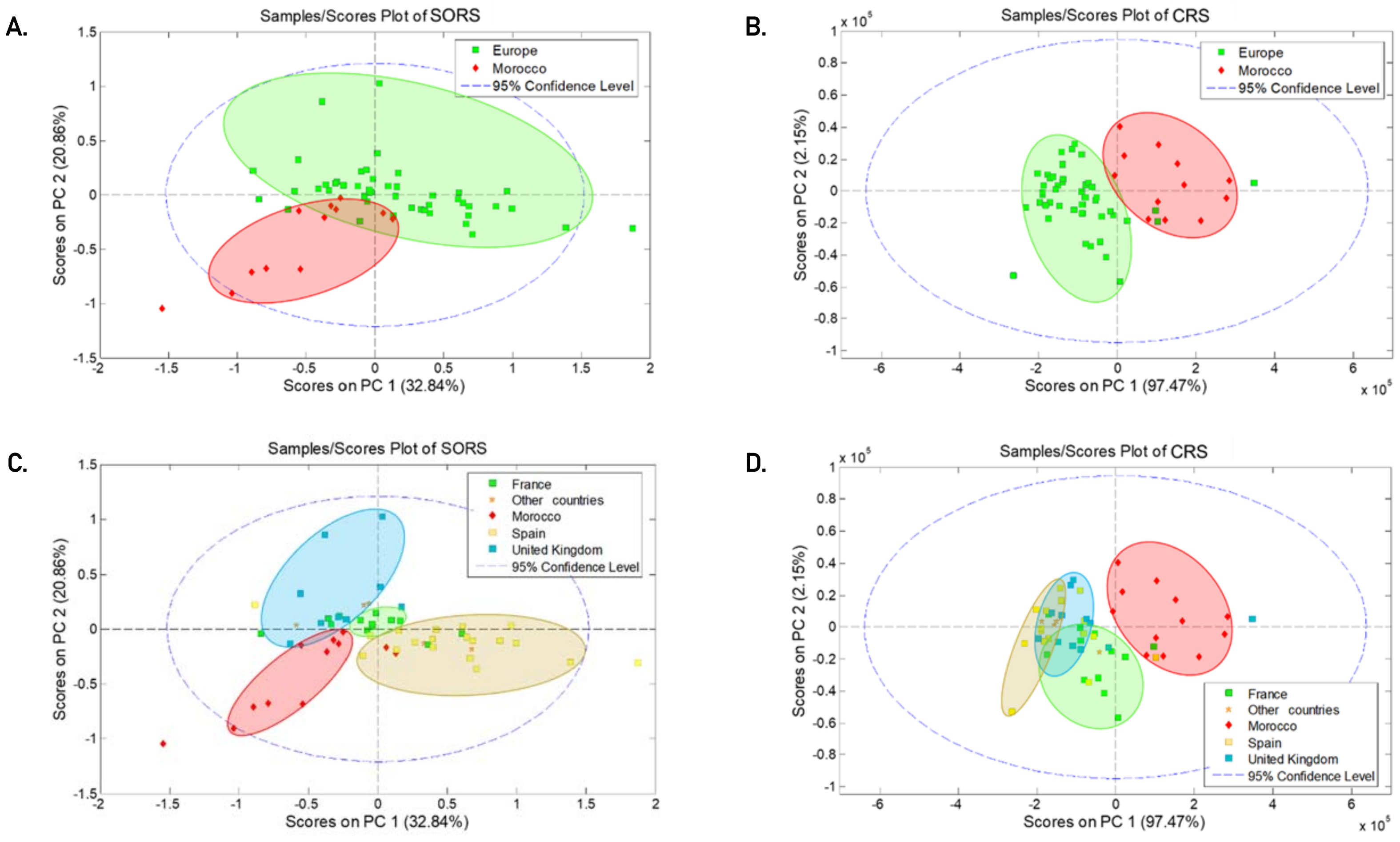


Figure 2.4. PCA scores plot of Raman spectra of margarines samples with data obtained from SORS data (**A,C**) and CRS (**B,D**). The samples are differentiated according to their geographical origin of fabrication following the legend.

With respect to the different countries of origin of the European samples, more separation between the groups from Spain and United Kingdom can be observed when comparing the SORS (Figure 2.4C) with respect to the CRS data (Figure 2.4D), where a greater overlap between the groups is evident.

3.2. Classification by geographical origin

According to the results of the exploratory analysis, different classification models were developed to classify the samples on the basis of geographic origin namely: 'Europe / Morocco' model (50 / 12 samples, respectively), 'Spain / no Spain' with membership in the 'Europe' class (19 / 31 samples, respectively), 'France / no France' with membership in the 'no Spain' class (13 / 18 samples, respectively) and 'United Kingdom / no United Kingdom' model with membership in the 'no France' class (12 / 6 samples, respectively). Pre-processing was mean center for both SORS and CRS data.

PLS-DA method was used to perform the classification models according to geographical origin. Figure 2.5 shows an example of the classification plot obtained for 'Europe / Morocco' model for both SORS data (Figure 2.45) and CRS data (Figure 2.5B). The models were built with a training set composed of 43 samples and an external validation set composed of 19 samples. Five and six latent variables (LVs) explaining 85.53% and 99.95% of the CV for SORS and CRS data respectively, were selected for the development of the models.

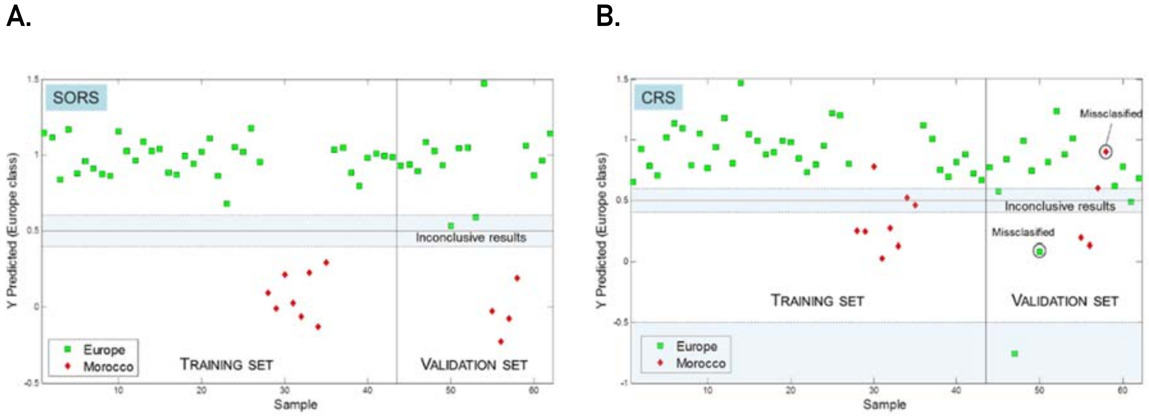


Figure 2.5. Classification plots obtained for the 'Europe / Morocco' models developed by applying PLS-DA to the SORS data (**A**) and CRS data (**B**).

Table 2.3 shows the results of the four classification models: (i) 'Europe / Morocco', (ii) 'Spain / no Spain' (seven LVs explaining 85.73% CV for SORS data and three LVs explaining 99.53% CV for CRS data), (iii) 'France / no France' (six LVs explaining 82.68% CV for SORS data and two LVs explaining 99.63% CV for

CRS data) and (iv) 'United Kingdom / no United Kingdom' (six LVs explaining 90.21% CV for SORS data and four LVs explaining 99.56% CV for CRS data).

Table 2.3. Results of the assigned class to the validation sets of the classification models developed according to the geographical origin of the samples, both for SORS and CRS data.

			SORS			CRS		
			Reference class		TNA	Reference class		TNA
Europe / Morocco			Europe	Morocco		Europe	Morocco	
	Assigned class	Europe	13	0	13	11	1	12
		Morocco	0	4	4	1	2	3
		Inconclusive	2	0		3	1	
		TNB	15	4		15	4	
Spain / no Spain			Spain	no Spain		Spain	no Spain	
	Assigned class	Spain	4	0	4	2	2	4
		No Spain	0	8	8	2	7	9
		Inconclusive	2	1		2	0	
		TNB	6	9		6	9	
France / no France			France	no France		France	no France	
	Assigned class	France	3	0	3	2	0	2
		No France	0	4	4	1	3	4
		Inconclusive	0	1		0	2	
		TNB	3	5		3	5	
UK / no UK			UK	no UK		UK	no UK	
	Assigned class	UK	2	0	2	2	1	3
		No UK	0	2	2	0	0	0
		Inconclusive	2	0		2	1	
		TNB	4	2		4	2	

TNA: total number of samples assigned to the class; TNB: total number of samples belonging to the class; UK: United Kingdom.

The corresponding classification plots for the last three models can be found in the supplementary material (Figures 2.S1-2.S3). The training sets were composed of 35, 23 and 12 samples and the external validation sets of 15, 8 and 6 samples for the 'Spain / no Spain' model, the 'France / no France' model and the 'United Kingdom / no United Kingdom' model, respectively.

Some of the most relevant classification performance metrics calculated for the geographical origin classification models are shown in Table 2.4. The values correspond to the average calculated after estimating the parameters for the two classes that constitute each model. It should be noted that the best results were obtained for both 'Europe / Morocco' and 'France / no France' models, and especially with the data obtained with the SORS. This may be related to the influence of fluorescence on the spectra obtained with the CRS or with the applied pre-processing. Both aspects are discussed in section 3.6.

Table 2.4. Classification performance metrics for PLS-DA geographical origin models developed from both types of Raman techniques.

	Europe / Morocco		Spain / no Spain		France / no France		UK / no UK	
	SORS	CRS	SORS	CRS	SORS	CRS	SORS	CRS
Sensitivity	0.89	0.68	0.80	0.60	0.88	0.63	0.67	0.33
Specificity	0.97	0.55	0.76	0.51	0.93	0.64	0.83	0.17
Positive PV (Precision)	1.00	0.86	1.00	0.67	1.00	0.84	1.00	—
Negative PV	1.00	0.72	1.00	0.61	1.00	0.91	1.00	—
Efficiency (Accuracy)	0.89	0.68	0.80	0.60	0.88	0.63	0.67	0.33
AUC (CCR)	0.93	0.62	0.78	0.56	0.90	0.63	0.75	0.25
Matthews CC	0.93	0.45	0.77	0.23	0.89	0.55	0.71	—
Kappa coeff.	0.75	0.33	0.65	0.25	0.77	0.37	0.50	0.00

PV: predictive value; AUC: area under curve; CCR: correctly classified rate; CC: correlation coefficient; "—": the parameter could not be determined because the number of samples assigned to the class was 0.

3.3. Classifications by vegetable oil types

As mentioned above, legislation allows margarines and fat spreads to include different types of vegetable oils or fats in their composition. This aspect has been used to perform different classification models according to some of the types of oil present in the analysed samples. PLS-DA, SVM and SIMCA models based on different samples parameters were built with data obtained both from SORS and CRS measurements. Linseed oil, olive oil, sunflower oil and palm oil were the four selected vegetable oil types to perform the classification models. PLS-DA and SIMCA classification models were developed applying mean center as pre-processing, while SVM classification models applying autoscaling. This was applicable for the data obtained by both techniques (SORS and CRS) to be compared later.

SIMCA models were developed following the one-class strategy (OC-SIMCA). This approach has been highlighted several times in the literature for food authentication and it is especially relevant when classifying foods that have a particularity (in this study, an ingredient) from the rest that do not [31]. For this purpose, the target input class with which the model is trained are only samples that include this ingredient in their composition. The plots to be presented for the models developed with OC-SIMCA represent Hotteling's T^2 versus Q-residuals of the target class at a confidence level of 95%. To consider a sample as a target class, i.e. 'within the model', the values of these two statistics (T^2 and Q) must be below 1, within the bounded square in the corresponding plot. The OC-SIMCA plots shown represent the values obtained only for the validation set and are zoomed to see the confidence area.

'Linseed / no linseed' models are shown in Figure 2.6. PLS-DA models were built with eight LVs each, explaining 79.53% and 99.99% of the CV, and OC-SIMCA models were built with seven and two PCs explaining 88.27% and 99.81% of the CV, respectively for SORS and CRS data in both cases. PLS-DA and SVM models were developed with a training set of 43 samples and 19 samples for the external validation set. The OC-SIMCA models were performed with a training set of 18 samples of the target class and the validation set including all samples (62 in total).

Best results were obtained with PLS-DA models compared to SVM, especially for SORS data. Furthermore, with respect to the OC-SIMCA model, notice that better results were obtained with SORS data compared to CRS despite that the same samples were analysed. This can be seen in Table 2.5 which gives the quality parameters of the six models built to differentiate the samples containing linseed oil. For this case, the OC-SIMCA model developed from the SORS data is the one that best classifies the samples according to this parameter, all the classification performance metrics are above 0.8.

Note that negative value of the Matthews correlation coefficient of the OC SIMCA model built from CRS data indicate a negative correlation. However, negative value of the Kappa coefficient indicates the model's prediction to be worse than a random prediction. This occurs in a few more models throughout the text.

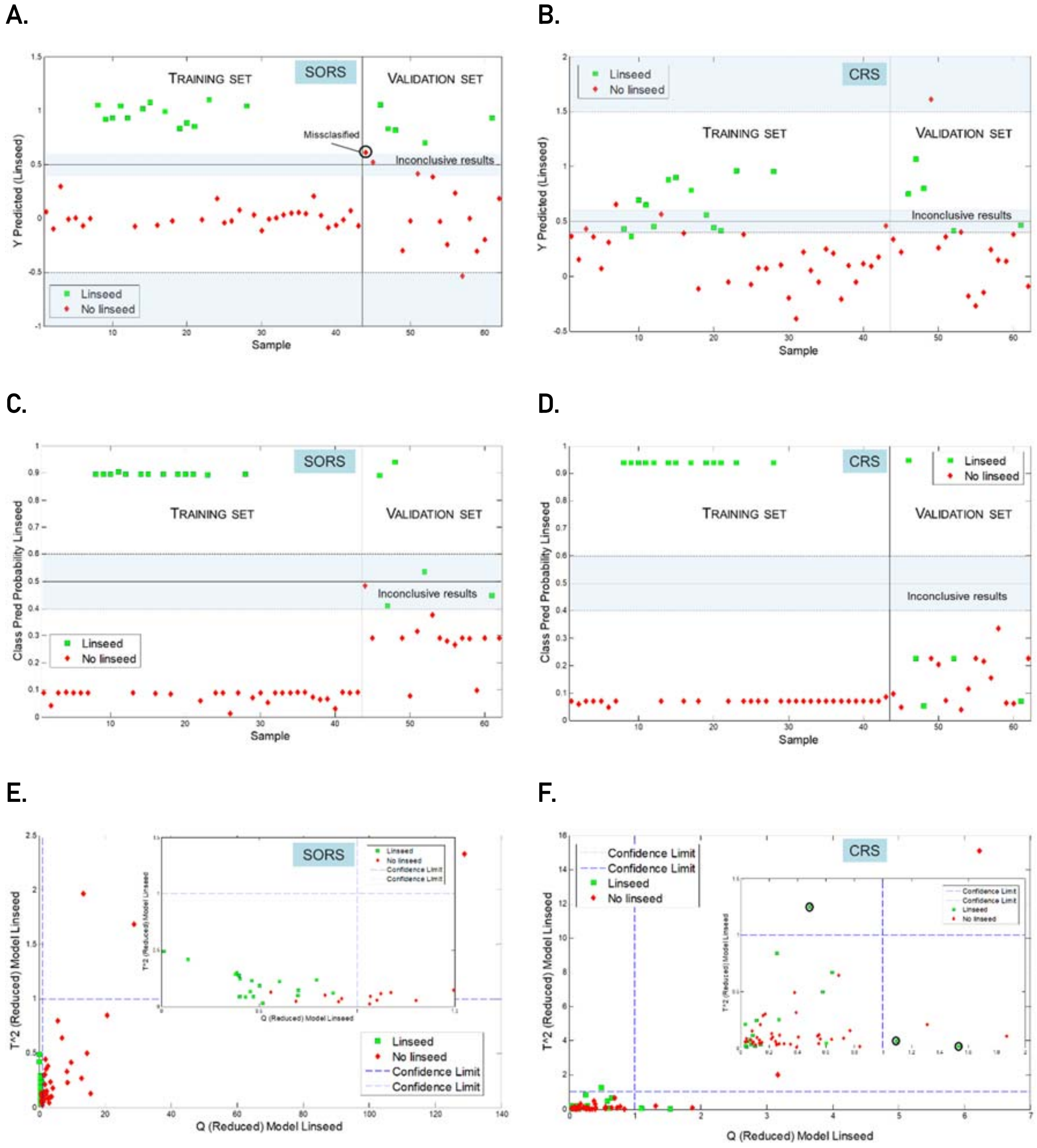


Figure 2.6. Classification plots obtained for the 'linseed / no linseed' models developed by applying PLS-DA to the SORS data (**A**) and CRS data (**B**), SVM-C to the SORS data (**C**) and CRS data (**D**), and OC-SIMCA to the SORS data (**E**) and CRS data (**F**).

The corresponding classification plots for the others vegetable oils models (olive, sunflower and palm) can be found in the supplementary information (Figures 2.S4-2.S6), as well as the parameters selected to carry out the PLS-DA and OC-SIMCA models, explained in the figure caption. The quality parameters of these models are shown in Tables 2.S2-2.S4.

Table 2.5. Classification performance metrics for PLS-DA, SVM and OC-SIMCA 'linseed / no linseed' models developed from both types of techniques.

	PLS-DA model		SVM model		OC-SIMCA model	
	SORS	CRS	SORS	CRS	SORS	CRS
Sensitivity	0.79	0.79	0.79	0.79	0.92	0.31
Specificity	0.92	0.67	0.54	0.41	0.97	0.62
Positive PV (Precision)	0.96	1.00	1.00	0.84	0.94	0.48
Negative PV	0.88	1.00	1.00	0.94	0.85	0.36
Efficiency (Accuracy)	0.79	0.79	0.79	0.79	0.92	0.31
AUC (CCR)	0.86	0.73	0.66	0.60	0.94	0.46
Matthews CC	0.77	0.72	0.61	0.39	0.83	-0.11
Kappa coeff.	0.60	0.57	0.55	0.27	0.82	-0.05

PV: predictive value; AUC: area under curve; CCR: correctly classified rate; CC: correlation coefficient.

For the 'olive / no olive' models developed with SVM the quality parameters were not calculated (Table 2.S1), because these models did not provide good results, not being able to distinguish samples with and without olive oil (Figure 2.S4C and 2.S4D). This may be due to the limited number of samples containing olive oil in their composition (6 samples out of 62). The same occurs with the PLS-DA models, which were not able to correctly discriminate the samples of the validation set containing olive oil. However, it is again remarkable the results obtained with the OC-SIMCA model with the SORS data, for which all quality parameters are above 0.9.

Regarding the other models ('sunflower / no sunflower' and 'palm / no palm'), the results were generally not as satisfactory, perhaps because the amount of these ingredients in the margarines is relatively small, or because the spectra do not offer sufficiently relevant chemical information to detect these ingredients compared to other components. Even so, the quality parameters of the 'sunflower / no sunflower' SVM model developed from the data obtained with CRS (Figure 2.S5C, Table 2.S2) and the PLS-DA 'palm / no palm' model from the SORS data (Figure 2.S6B, Table 2.S3) stand out.

3.4. Classification by other ingredients

In addition to the models based on the type of vegetable oil or fat included in some margarines, classification models based on other minor constituents were also developed: buttermilk, sterol esters (called phytosterols because they are of vegetable origin) and lecithin (some margarines contain soybean lecithin, others sunflower lecithin and others do not specify the type of lecithin, so they were divided into two groups according to whether or not they contain lecithin without specifying the origin). The same data pre-processing strategy as in the previous section was used. The results obtained for the 'phytosterols / no phytosterols' classificatory models are presented below, while the rest can be found in the supplementary information (Figures 2.S7 y 2.S8).

Figure 2.7 shows the results for PLS-DA, SVM and OC-SIMCA 'phytosterols / no phytosterols' models. PLS-DA models were built with nine and eight LVs explaining 83.09% and 99.99% of CV, and OC-SIMCA models were built with three PCs each, explaining 89.26% and 99.83% of CV, respectively for SORS and CRS data in both cases.

Despite the scarce number of samples including phytosterols in their composition (6 samples out of 62), the developed models demonstrated excellent results for classifying margarines according to this parameter. As shown in Table 2.6, classification performance metrics show a perfect OC-SIMCA model developed with SORS data (all parameters with a value of 1.00), the same as the OC-SIMCA model built from CRS data, for which all quality parameters are above 0.87.

Table 2.6. Classification performance metrics for PLS-DA, SVM and OC-SIMCA 'phytosterols / no phytosterols' models developed from both types of techniques.

	PLS-DA model		SVM model		OC-SIMCA model	
	SORS	CRS	SORS	CRS	SORS	CRS
Sensitivity	0.95	0.89	0.95	0.84	1.00	0.98
Specificity	0.99	0.55	0.55	0.10	1.00	1.00
Positive PV (Precision)	1.00	1.00	1.00	0.80	1.00	0.99
Negative PV	1.00	1.00	1.00	0.09	1.00	0.87
Efficiency (Accuracy)	0.95	0.89	0.95	0.84	1.00	0.98
AUC (CCR)	0.97	0.72	0.75	0.47	1.00	0.99
Matthews CC	0.97	0.69	0.71	-0.08	1.00	0.92
Kappa coeff.	0.78	0.56	0.73	-0.08	1.00	0.91

PV: predictive value; AUC: area under curve; CCR: correctly classified rate; CC: correlation coefficient.

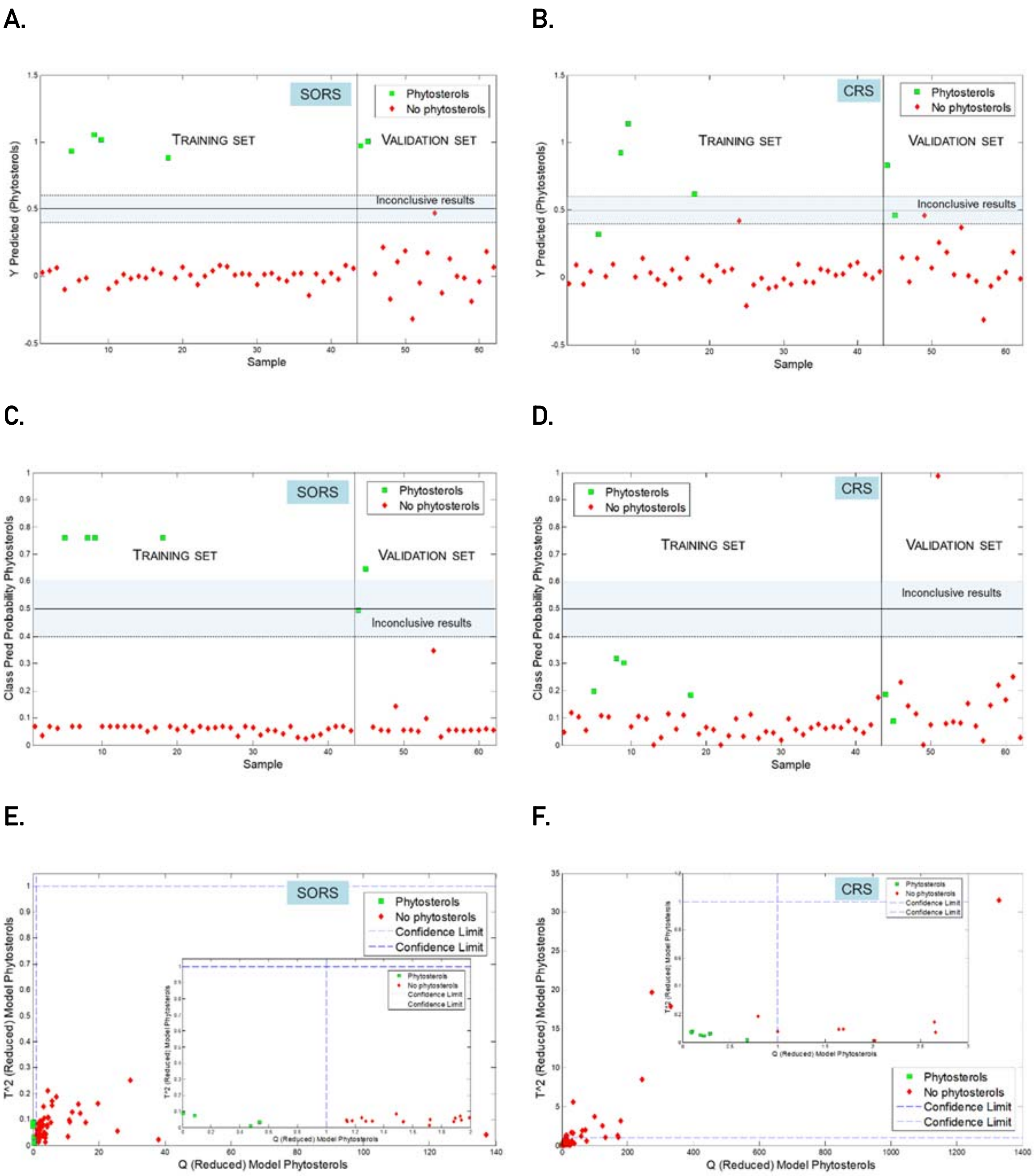


Figure 2.7. Classification plots obtained for the 'phytosterols / no phytosterols' models developed by applying PLS-DA to the SORS data (**A**) and CRS data (**B**), SVM-C to the SORS data (**C**) and CRS data (**D**), and OC-SIMCA to the SORS data (**E**) and CRS data (**F**).

When differentiating samples containing buttermilk or lecithin (classification performance metrics for these models can be found in supplementary material), the results obtained were not as good as in the case of phytosterols. This is probably due to the same reasons stated in the previous section for the 'sunflower / no sunflower' and 'palm / no palm' models. In fact, some quality parameters were not calculated like those of the 'buttermilk / no buttermilk' model developed applying SVM from SORS data (Table 2.S4) or the 'lecithin / no lecithin' model developed with SVM (Table 2.S5), because these models were not able to distinguish between the two classes.

3.5. Quantitation of fat content

Once the potential of the SORS technique to perform qualitative analysis was established, a quantitative analysis to determine the total fat content was performed. This quantitative method is just proposed to be applied as a pre-screening method to verify the fat content stated on the label declared by the manufacturer.

In line with the EU regulation [**36,37**], the samples were divided into different groups, namely: margarine (80% fat), fat spread 70%, three-quarter-fat margarine (60%), fat spread 50%, half-fat-margarine (40%) and fat spread 32%. It should be noted that the percentage is an average of the total number of samples included in that group (see Table 2.7). Also note that this percentage refers to the total fat content in grams per 100 grams of product. The samples were then split into a training set and an external validation set, with the exception of six samples whose total fat content was not stated and which were reserved for a final step as test data in the model development.

Table 2.7. Fat content of each analysed sample and groups established to carry out the quantitation models.

Group	Fat % range	Fat % average	Number of samples		
			Total	Training set	Validation set
Margarine*	90 – 80	80	2	2	0
Fat spread 70%	79 – 63	70	12	8	4
Half-fat margarine*	62 – 60	60	13	9	4
Fat spread 50%	59 – 42	50	19	13	6
Three-quarter-fat margarine*	41 – 39	40	6	4	2
Fat spread 32%	< 38	32	4	3	1
Unknown	—	—	6	—	—

*Designation according to legislation.

As means of a fair comparison between both Raman techniques, two PLSR quantitation models for each dataset (SORS and CRS) were built applying both mean center and autoscaling as pre-processing. The models performed from SORS data were built by choosing ten and eight LVs, explaining 89.28% and 54.83% of the CV, after using the mean center and autoscaling as pre-processing, respectively. The CRS models were built with eight LVs each, explaining 99.99% and 99.85% of the CV, respectively. The internal validation for all the PLSR models was venetian blinds with 6 splits and 1 sample per split.

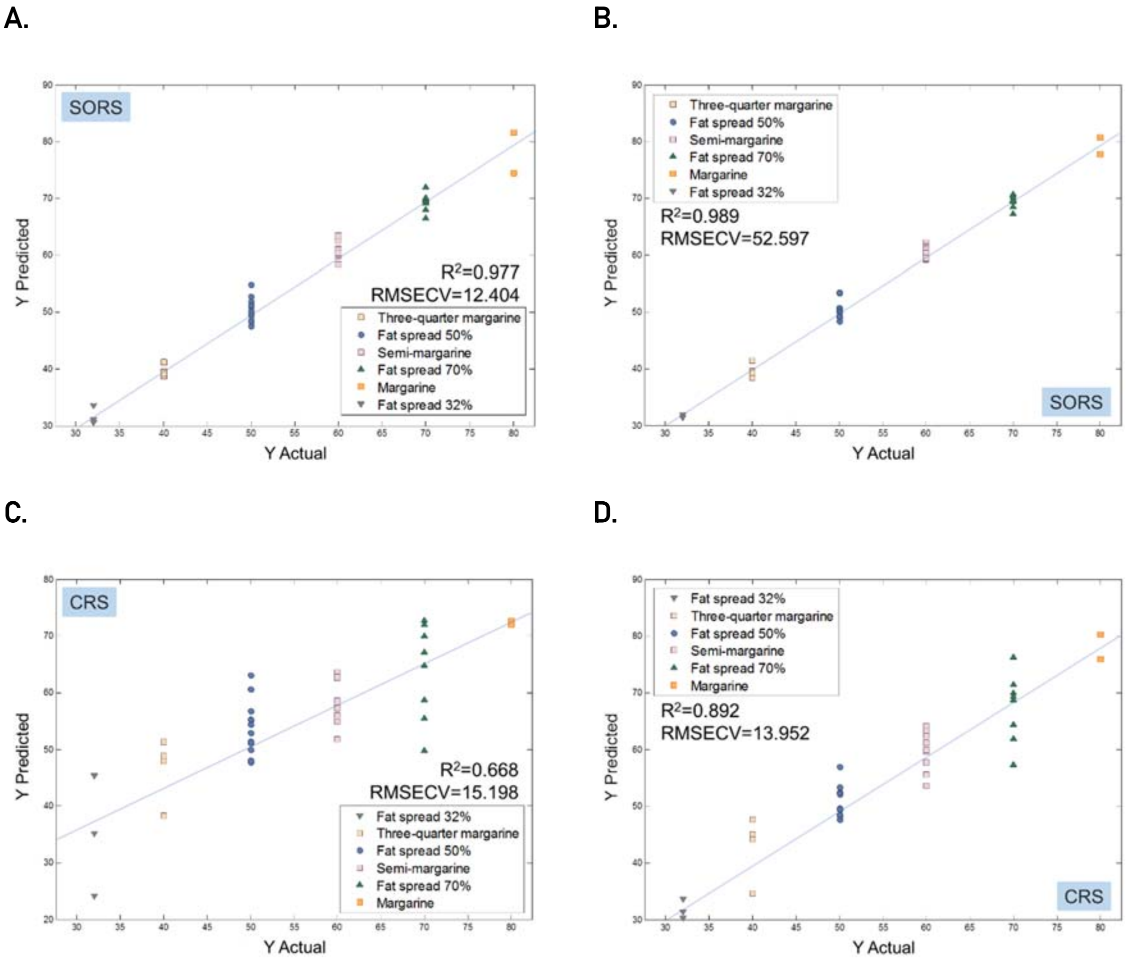


Figure 2.8. Total fat content of margarine samples predicted by the PLSR model for the training set using data obtained by SORS with mean center (**A**) and autoscaling as pre-processing (**B**) and CRS data with mean center (**C**) and autoscaling as pre-processing (**D**) versus total fat content reported.

Figure 2.8 shows the plots obtained by representing the total fat content predicted by the different models against the total fat content declared in the labelling of the samples. These plots correspond to the training set. The coefficient of determination (R^2) and the corresponding root mean square error of cross validation (RMSECV) are indicated in each figure. It is worth noting that the best results when building the models were obtained by using autoscaling as pre-processing for the data obtained by both techniques (Figure 2.8B and 2.8D). However, when comparing the two techniques, the R^2 values from SORS data were much higher with both types of pre-processing.

The predicted values of the total fat content of the samples from the validation set obtained with the different models developed by PLSR (see Table 2.S6) were compared with the original total fat content declared on the labelling of the samples to calculate the quantitation performance metrics shown in Table 2.8. Based on these results, it can be inferred that the best models were obtained using autoscaling as pre-processing for the data obtained from CRS data and mean center for the SORS data. The results from the CRS showed quantitation performance metrics below 10%, except for the SDV with a value of 10.6%. However, with the spectral data obtained after measuring the samples through their original packaging, i.e. from the SORS, errors were even lower, below 5%. The R^2 confirms that the best predicting model for total fat content is the developed from SORS data using mean center as a pre-processing.

Table 2.8. Quantitation performance metrics for the predicted total fat content results of the validation set from the different PLSR models developed.

	SORS		CRS	
	Mean center	Autoscaling	Mean center	Autoscaling
Root Mean Square Error (RMSE, %)	5.0	10.4	12.3	10.0
Mean Absolute Error (MAE, %)	4.1	6.0	6.2	5.6
Median Absolute Error (MdAE, %)	4.0	7.6	9.8	7.3
Standard Deviation of Validation Residuals (SDV)	5.0	10.1	12.3	10.6
Coefficient of determination (R^2)	0.7	0.2	0.1	0.3

Finally, evaluation of the models developed was carried out with six of the samples analysed that did not declare the total fat content on their labelling, so their content was unknown. Table 2.9 shows the results predicted by each model developed for these samples, while the obtained plots can be seen in supplementary material (Figure 2.S9). Different results can be observed between the values obtained with the models developed from both SORS and CRS data.

The values predicted of the total fat content from the CRS measurements are inconsistent because they are always largely higher than 100%. However, the predicted values from SORS data could potentially make sense, as they are ranging between 60 – 80%. This is the case with both models developed from SORS data, i.e., with the two pre-processing methods that were applied (mean center and autoscaling).

Table 2.9. Predicted results of the total fat content of margarines, whose value was unknown, by the different PLSR quantitation models developed.

Sample	Predicted total fat content (g/100 g product)			
	SORS		CRS	
	Mean center	Autoscaling	Mean center	Autoscaling
1	67.4	64.5	152.6	122.1
2	60.0	67.9	167.0	138.8
3	79.0	81.5	155.6	130.1
4	68.3	73.8	149.3	122.6
5	66.8	72.2	134.9	122.4
6	70.2	77.8	142.1	125.4

3.6. Comparison between CRS and SORS

The results shown in the previous sections comparing the data obtained by the two techniques, both for qualitative and quantitative analysis of margarine and fat spreads samples, highlight the great potential of SORS to perform the authentication of these products, considering different characteristics (geographical origin, different oil types, different minor constituents and total fat content). Evidence of this is the high-performance parameters calculated for the models developed.

When comparing the Raman spectra without any pre-processing method of the same sample in the same spectral range (350-2000 cm^{-1}) with both techniques, it can be seen that the chemical information provided is the same, as shown in Figure 2.9 in the regions marked in green and orange. In this figure it is also observed that the influence of the fluorescence is higher in the CRS spectrum (area marked in blue), while the non-preprocessed SORS spectrum exhibits a lower influence, due to the fact that the equipment is already programmed for this purpose. In addition, the figure shows the processed spectrum obtained with the same equipment, where the influence of fluorescence is eliminated, obtaining a completely clean spectrum but with the same relevant chemical information. This could be the reason why the models seem to perform better from SORS data than from CRS data. Here, it is demonstrated that despite the use of a portable SORS, the spectral resolution does not appear to be lower than that obtained from a benchtop CRS.

It should therefore be noted that with the SORS technique it is possible to obtain the spectrum of the margarines by measuring through their original packaging, and also with a higher resolution, i.e. enriched structural information, of the spectrum. This can also be seen in the intensity scale, which is an order of magnitude higher in the spectrum obtained using SORS.

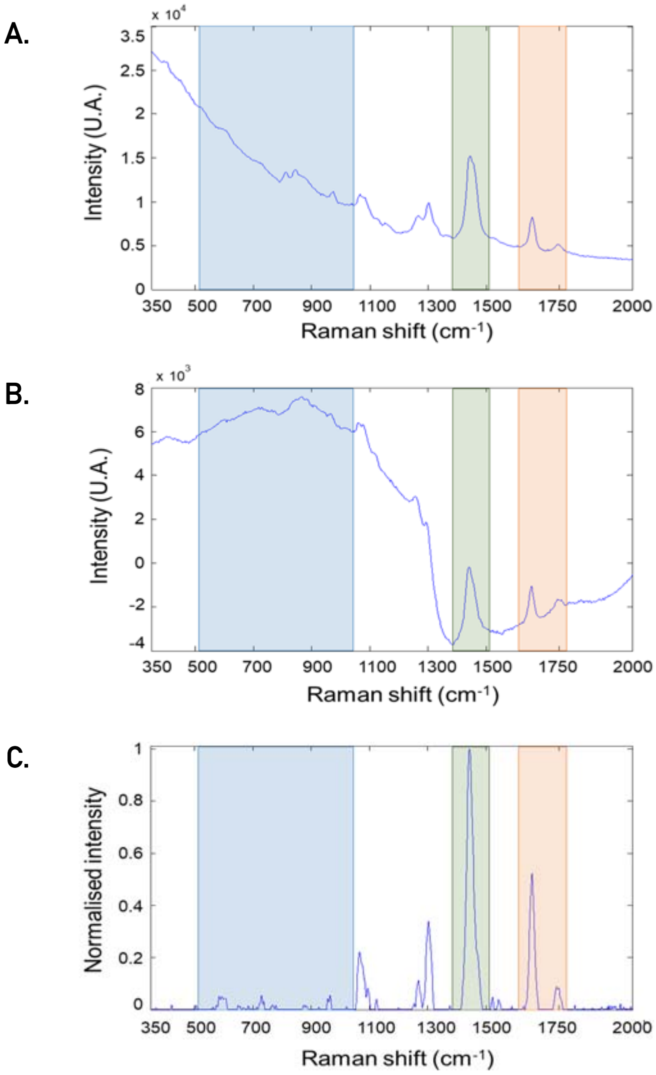


Figure 2.9. Comparison of the non-preprocessed spectrum of sample 1 from SORS (**A**), non-preprocessed spectrum of sample 1 from CRS (**B**) and pre-processed spectrum of sample 1 from SORS (**C**). The spectrum from CRS has been cut to be in the same range as the SORS spectrum for comparison (350-2000 cm^{-1}).

Loadings plots of all developed models were also examined. Most relevant ones are included in supplementary material as examples. According to the band assignment (see Table 2.2) the peak around 1654 cm^{-1} stands out, which is related to oil content [25] and more specifically to sunflower oil content [32]. The information from this peak seems to be relevant for the development of the PLSDA 'sunflower / no sunflower' and PLSR models from SORS (see Figures 2.S10-2.S12). Similar for CRS the peaks highlighted in the loading plot of the PLSR model are related to the presence of unsaturated and saturated fatty acids respectively [30]. However, the loading plot of the 'linseed / no linseed' model seems to be influenced by the fluorescence perturbation, which may explain why the results of this model were not so satisfactory.

In short, the chemical information provided by the data from SORS, especially related to oils and fatty acids, is responsible for developing classification and quantitation models with better results than those from CRS.

4. Conclusions

The results presented in this paper demonstrate the ability of SORS to recover the Raman spectra of margarines and fat spread products in a rapid, non-invasive and non-destructive way after measuring them through the original packaging. The chemometric tools allowed the extraction of the relevant chemical information from these spectra for qualitative and quantitative analyses. High-quality multivariate classification models were developed to distinguish samples according to their geographical origin of production and according to some of the relevant ingredients of their composition, namely linseed oil, olive oil and phytosterols. It is worth highlighting the case of both 'phytosterols / no phytosterols' and 'olive / no olive' models, which have obtained very acceptable classification results, as precision values above 0.8 in these cases, and most around 1. Despite not finding significant differences between the spectra of samples containing this ingredient and those that do not, the use of chemometric tools to treat the data as an instrumental fingerprint characteristic of each product has proven to be able to extract this relevant and non-evident information. Furthermore, the quantitation model performed with PLSR allowed predicting the total fat content of the samples with errors below 5%, as well as predicting the fat content of unknown samples, with values within the limit of what can be expected. The proposed method could be therefore applied as a pre-screening method to straightforwardly verify some of the claims stated on the label declared by the manufacturer.

Comparison with CRS measurements revealed the potential of SORS to avoid the main problem associated with the use of this technique, namely the influence of fluorescence. Moreover, a lower spectral resolution was not observed compared to CRS and, in fact, the results of the classification and quantitation models were better from the SORS data. These facts are consistent with the literature on the use of the 'SORS-chemometrics' tandem and open the door to further research into the use of SORS in the area of analytical chemistry for food quality and authentication control.

Acknowledgments

This work was partially supported by University of Granada (Spain) within the framework of the funding corresponding to program 'pre-competitive research projects for young researchers'.

AMJC wish to acknowledge the Department of Economic Transformation, Industry, Knowledge and Universities belong to Regional Andalusia Government (Spain) for the Postdoctoral fellowship (DOC_00121). In addition, **AAC** wants to express their sincere gratitude to the Spanish Ministry of Sciences, Innovation and Universities for a pre-doctoral fellowship FPU (FPU20/04711, Formación del Profesorado Universitario).

References

[1] S. Mosca, P. Dey, M. Salimi, B. Gardner, F. Palombo, N. Stone, P. Matousek, Spatially offset Raman spectroscopy – How deep?, Anal. Chem. 93 (2021) 6755−6762. DOI: 10.1021/acs.analchem.1c00490.

[2] S. Mosca, C. Conti, N. Stone, P. Matousek, Spatially offset Raman spectroscopy, Nat. Rev. Meth. Primers 1 (2021) 21. DOI: 10.1038/s43586-021-00019-0.

[3] P. Matousek, I.P. Clark, E.R.C. Draper, M.D. Morris, A.E. Goodship, N. Everall, M. Towrie, W.F. Finney, A.W. Parker, Subsurface probing in diffusely scattering media using spatially offset Raman spectroscopy, Appl. Spectrosc. 59 (2005) 393–400. DOI: 10.1366/0003702053641450.

[4] F. Martelli, T. Binzoni, A. Pifferi, L. Spinelli, A Farina, A. Torricelli, There's plenty of light at the bottom: statistics of photon penetration depth in random media, Sci. Rep. 6 (2016) 27057. DOI: 10.1038/srep27057.

[5] A. Arroyo-Cerezo, A.M. Jiménez-Carvelo, A. González-Casado, A. Koidis, L. Cuadros-Rodríguez, Deep (offset) non-invasive Raman spectroscopy for the evaluation of food and beverages – A review, LWT - Food Sci. Technol. 149 (2021) 111822. DOI: 10.1016/j.lwt.2021.111822.

[6] Y. Xu, P. Zhong, A. Jiang, X. Shen, X. Li, Z. Xu, Y. Shen, Y. Sun, H. Lei, Raman spectroscopy coupled with chemometrics for food authentication: A review, Trends Analyt. Chem. 131 (2020) 116017. DOI: 10.1016/j.trac.2020.116017.

[7] A.I. Ropodi, E.Z. Panagou, G.-J.E. Nychas, Data mining derived from food analyses using non-invasive/non-destructive analytical techniques; determination of food authenticity quality & safety in tandem with computer science disciplines, Trends Food Sci. Technol. 50 (2016) 11-25. DOI: 10.1016/j.tifs.2016.01.011.

[8] C. Berghian-Grosan, D.A. Magdas, Raman spectroscopy and machine-learning for edible oils evaluation, Talanta 218 (2020) 121176. DOI: 1016/j.talanta.2020.121176.

[9] A.M. Jiménez-Carvelo, A. González-Casado, M.G. Bagur-González, L. Cuadros-Rodríguez, Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review, Food Res. Int. 122 (2019) 25-39. DOI: 10.1016/j.foodres.2019.03.063.

[10] C. Conti, C. Colombo, M. Realini, P. Matousek, Subsurface analysis of painted sculptures and plasters using micrometre-scale spatially offset Raman spectroscopy (micro-SORS), J. Raman Spectrosc. 46 (2015) 476-482. DOI: 10.1002/jrs.4673.

[11] C. Eliasson, N.A. Macleod, P. Matousek, Noninvasive Detection of cocaine dissolved in beverages using displaced Raman spectroscopy, Anal. Chim. Acta 607 (2008) 50-53. DOI: 10.1016/j.aca.2007.11.023.

[12] M. Z. Vardaki, C.G. Atkins, H.G. Schulze, D.V. Devine, K. Serrano, M.W. Blades, R.F.B. Turner, Raman spectroscopy of stored red blood cell concentrate within sealed transfusion blood bags, Analyst 143 (2018) 6006. DOI: 10.1039/c8an01509k.

[13] M. Arellano, I.T. Norton, P. Smith, Specialty oils and fats in margarines and low-fat spreads, in: G. Talbot (Ed.), Specialty Oils and Fats in Food and Nutrition – Properties, Processing and Applications, Woodhead Publishing / Elsevier, Cambridge, 2015, ch. 10, pp. 241–270. DOI: 10.1016/B978-1-78242-376-8.00010-7.

[14] N.W.G. Young, P. Wassell, Margarines and spreads, in: G.L. Hasenhuettl, R.W. Hartel (Eds.), Food Emulsifiers and their Applications, third ed., Springer Nature, Cham, 2019, ch. 13, pp. 379–405. DOI: 10.1007/978-3-030-29187-7_13.

[15] Codex Stan 256-2007, Standard for fat spreads and blended spreads, Codex Alimentarius FAO/WHO, 2007.

[16] Regulation (EU) No 1308/2013 establishing a common organization of the markets in agricultural products, OJ02013R1308-EN-003.001-223 (consolidated version 01.08.2017), European Commission, Brussels, 2017.

[17] Decree No 1153-66 on the regulations implementing the Dahir of 14 October 1914 on the suppresion of fraud in the manfacture and sale of margarine, Official Gazette No 2988, p. 214, Rabat, Morocco, 1970.

[18] M.D. Guillén, M.L. Ibargoitia, P. Sopelana, Margarines: composition and analysis, in: B. Caballero, P.M. Finglas, F. Toldrá (Eds.), Encyclopedia of Food and Health, vol. III, Academic Press / Elsevier, Oxford, 2016, pp. 646–653. DOI: 10.1016/B978-0-12-384947-2.00446-3.

[19] A. Rácz, M. Fodor, K. Héberger, Development and comparison of regression models for the determination of quality parameters in margarine spread samples using NIR spectroscopy, Anal. Methods 10 (2018) 3089. DOI: 10.1039/c8ay01055b.

[20] S. Lohumi, H. Lee, M.S. Kim, J. Qin, B.K. Cho, Through-packaging analysis of butter adulteration using line-scan spatially offset Raman spectroscopy, Anal. Bioanal. Chem. 410 (2018), 5663-5673. DOI: 10.1007/s00216-018-1189-1.

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

[22] ASTM International E2617-17. Standard practice for validation of empirically derived multivariate calibrations, 2017.

[23] V. Baeten, P. Hourant, M.T. Morales, R. Aparicio, Oil and Fat Classification by FT-Raman Spectroscopy, J. Agric. Food Chem. 46 (1998) 2638-2646. DOI: 10.1021/jf9707851.

[24] L. Dymińska, M. Calik, A.M.M. Albegar, A. Zając, K. Kostyń, J. Lorenc, J. Hanuza, Quantitative determination of the iodine values of unsaturated plant oils using infrared and Raman spectroscopy methods, Int. J. Food Prop. 20 (2017) 2003-2015. DOI: 10.1080/10942912.2016.1230744.

[25] A. Nedeljković, P. Rösch, J. Popp, J. Miočinović, M. Radovanović, P. Pudja, Raman spectroscopy as a rapid tool for quantitative analysis of butter adulterated with margarine, Food Anal. Methods 9 (2016) 1315-1320. DOI: 10.1007/s12161-015-0317-1.

[26] P. Bock, N. Gierlinger, Infrared and Raman spectra of lignin substructures: Coniferyl alcohol, abietin, and coniferyl aldehyde, J. Raman Spectrosc. 50 (2019) 778-792. DOI: 10.1002/jrs.5588.

[27] G. Yang, Q. Wang, C. Liu, X. Wang, S. Fan, W. Huang, Rapid and visual detection of the main chemical compositions in maize seeds based on Raman hyperspectral imaging, Spectrochim. Acta A Mol. Biomol. Spectrosc. 200 (2018) 186-194. DOI: 10.1016/j.saa.2018.04.026.

[28] F. Adar, Introduction to Interpretation of Raman Spectra Using Database Searching and Functional Group Detection and Identification, in: Spectroscopy Solutions for Materials Analysis. Molecular Spectroscopy Workbench: The 2016 Collection, 31 (2016) pp. 18-23.

[29] I.H. Boyaci, H.T. Temiz, H.E. Geniş, E.A. Soykut, N.N. Yazgan, B. Güven, R.S. Uysal, A.G. Bozkurt, K. İlaslan, O. Toruna and F.C.D. Şeker, Dispersive and FT-Raman spectroscopic methods in food analysis, RSC Adv 5 (2016) 56606-56624. DOI: 10.1039/c4ra12463d.

[30] J.M. Benevides, S.A. Overman, G.J. Thomas, Raman, polarized Raman and ultraviolet resonance Raman spectroscopy of nucleic acids and their complexes, J. Raman Spectrosc. 36 (2005) 279-299. DOI: 10.1002/jrs.1324.

[31] S. Medina, R. Perestrelo, P. Silva, J.A.M. Pereira, J.S. Câmara, Current trends and recent advances on food authenticity technologies and chemometric approaches, Trends Food Sci. Technol. 85 (2019) 163-176. DOI: 10.1016/j.tifs.2019.01.017.

[32] M. Varnasseri, H. Muhamadali, Y. Xu, P.I.C. Richardson, N. Byrd, D.I. Ellis, P. Matousek, R. Goodacre, Portable through Bottle SORS for the Authentication of Extra Virgin Olive Oil, Appl. Sci. 11 (2021) 8347. DOI: 10.3390/app11188347.

SUPPLEMENTARY INFORMATION *(Artículo científico 1)*

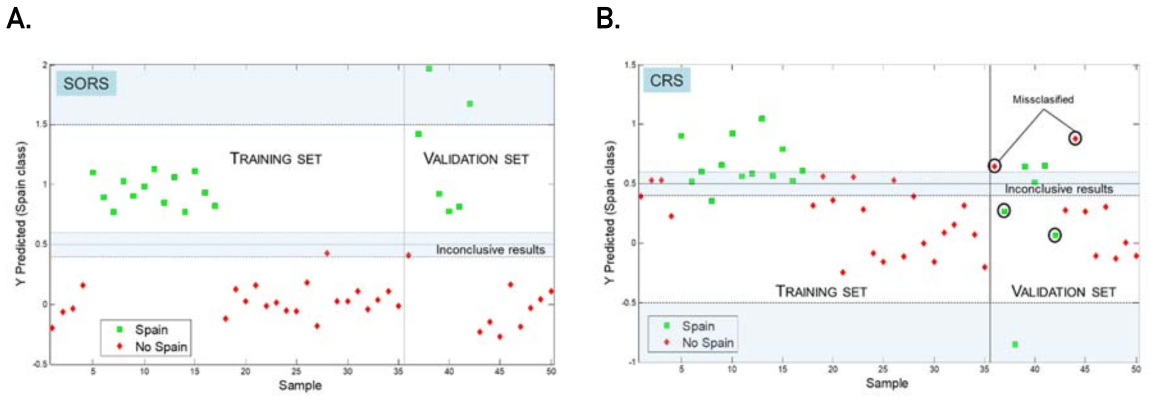


Figure 2.S1. Classification plots obtained for the 'Spain / no Spain' models developed by applying PLS-DA to the SORS data (**A**) and CRS data (**B**).

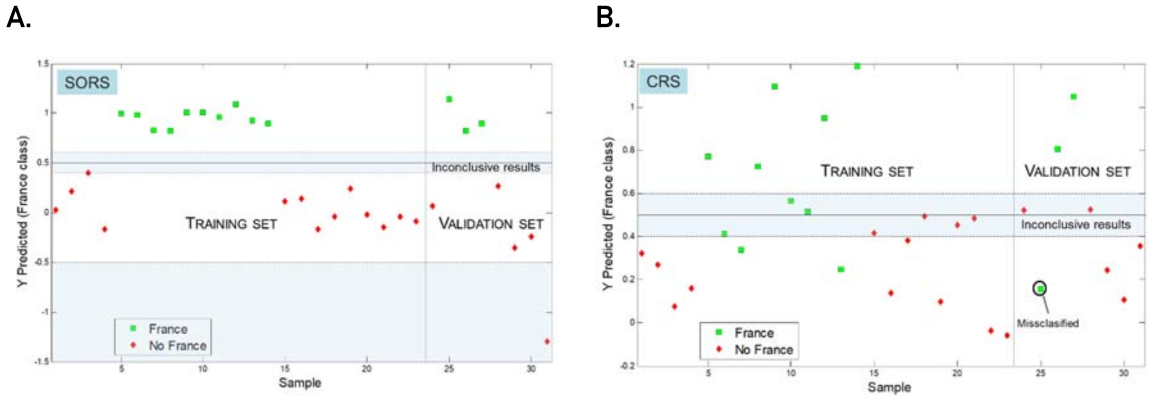


Figure 2.S2. Classification plots obtained for the 'France / no France' models developed by applying PLS-DA to the SORS data (**A**) and CRS data (**B**).

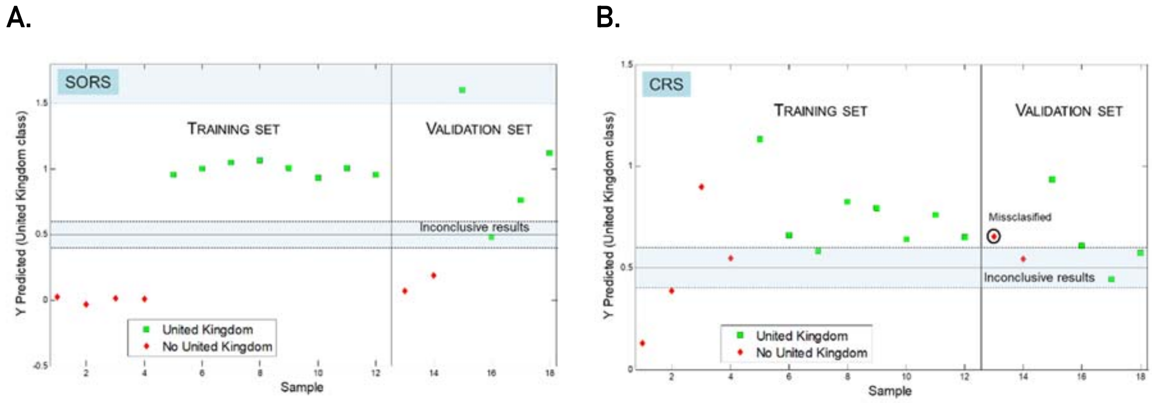


Figure 2.S3. Classification plots obtained for the 'United Kingdom / no United Kingdom' models developed by applying PLS-DA to the SORS data (**A**) and CRS data (**B**).

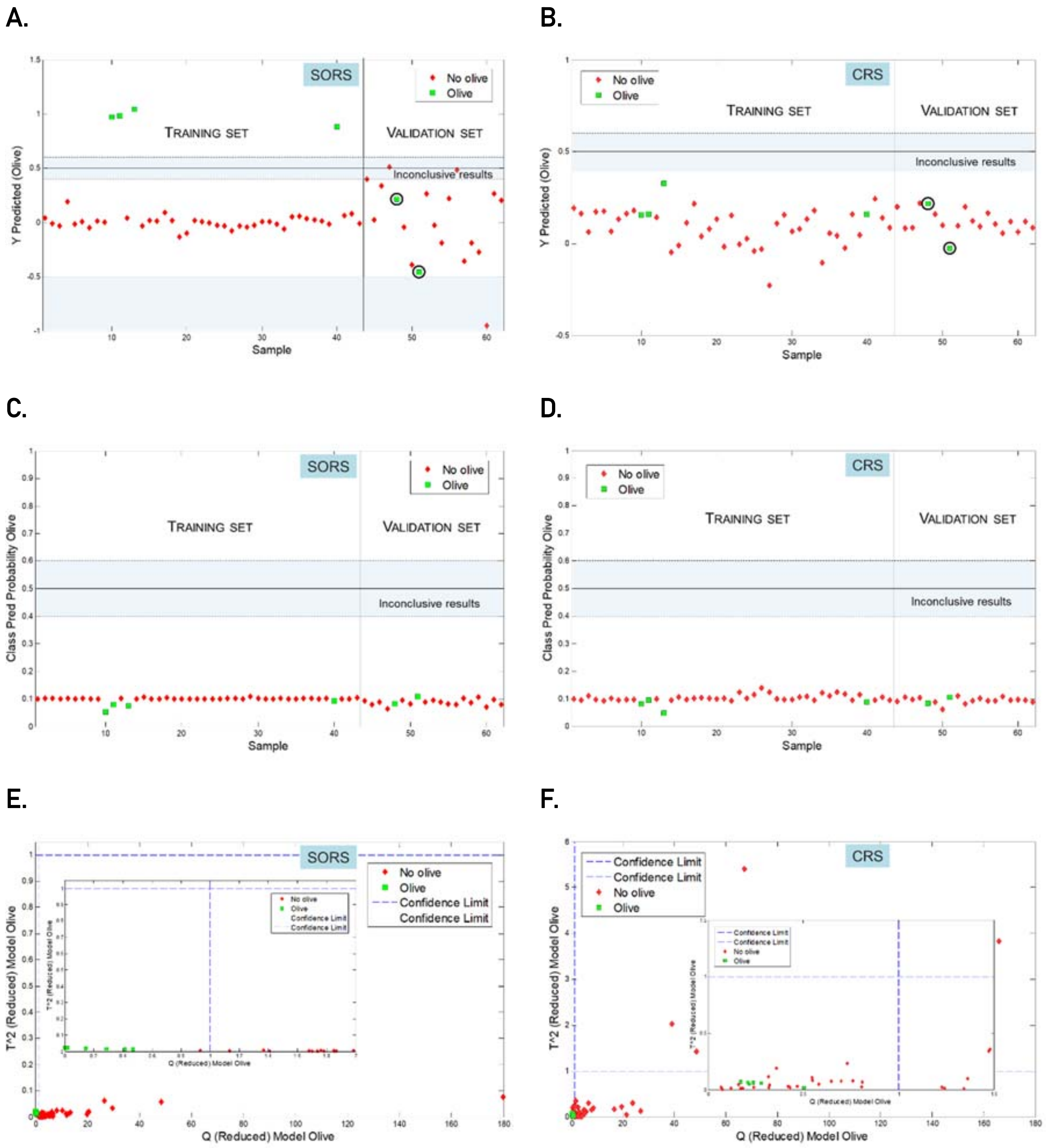


Figure 2.S4. Classification plots obtained for the 'olive / no olive' models developed by applying PLS-DA to the SORS data (**A**) and CRS data (**B**), built with nine and five LVs explaining 83.18% and 99.95% of the cumulative variance (CV), respectively; by SVM-C to the SORS data (**C**) and CRS data (**D**); and OC-SIMCA to the SORS data (**E**) and CRS data (**F**) built with four and three PCs explaining 95.80% and 99.86% of CV, respectively.

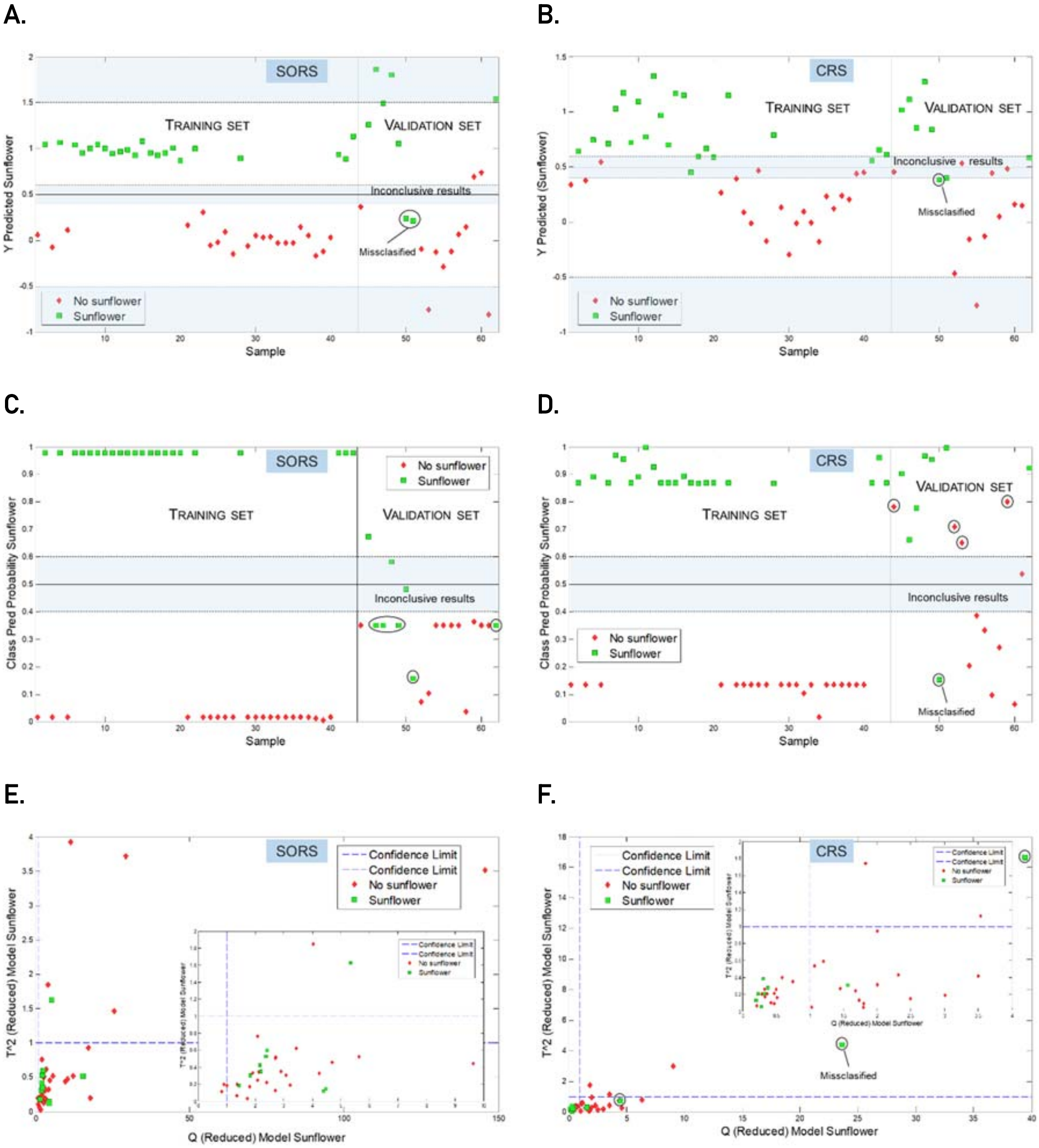


Figure 2.S5. Classification plots obtained for the 'sunflower / no sunflower' models developed by applying PLS-DA to the SORS data (**A**) and CRS data (**B**), built with eight and five LVs explaining 80.75% and 99.92% of CV, respectively; SVM-C to the SORS data (**C**) and CRS data (**D**); and OC-SIMCA to the SORS data (**E**) and CRS data (**F**) built with six and four PCs explaining 81.20% and 99.97% of CV, respectively.

Table 2.S1. Classification performance metrics for PLS-DA, SVM and OC-SIMCA 'olive / no olive' models developed from both types of techniques.

	PLS-DA model		SVM model		OC-SIMCA model	
	SORS	CRS	SORS	CRS	SORS	CRS
Sensitivity	0.74	0.89	N/A	N/A	0.98	0.61
Specificity	0.09	0.11	N/A	N/A	1.00	0.96
Positive PV (Precision)	—	—	N/A	N/A	0.99	0.92
Negative PV	—	—	N/A	N/A	0.87	0.28
Efficiency (Accuracy)	0.74	0.89	N/A	N/A	0.98	0.61
AUC (CCR)	0.41	0.50	N/A	N/A	0.99	0.79
Matthews CC	—	—	N/A	N/A	0.92	0.34
Kappa coeff.	-0.07	0.00	N/A	N/A	0.91	0.21

PV: predictive value; AUC: area under curve; CCR: correctly classified rate; CC: correlation coefficient; "—": the parameter could not be determined because the number of samples assigned to the class was 0; N/A: not applicable.

Table 2.S2. Classification performance metrics for PLS-DA, SVM and OC-SIMCA 'sunflower / no sunflower' models developed from both types of techniques.

	PLS-DA model		SVM model		OC-SIMCA model	
	SORS	CRS	SORS	CRS	SORS	CRS
Sensitivity	0.53	0.58	0.63	0.68	0.71	0.62
Specificity	0.49	0.59	0.49	0.74	0.22	0.61
Positive PV (Precision)	0.70	0.92	0.82	0.76	0.57	0.71
Negative PV	0.60	0.94	0.87	0.73	0.18	0.45
Efficiency (Accuracy)	0.53	0.58	0.63	0.68	0.71	0.62
AUC (CCR)	0.51	0.59	0.56	0.71	0.47	0.61
Matthews CC	0.27	0.54	0.29	0.46	-0.13	0.19
Kappa coeff.	0.23	0.38	0.25	0.42	-0.09	0.18

PV: predictive value; AUC: area under curve; CCR: correctly classified rate; CC: correlation coefficient.

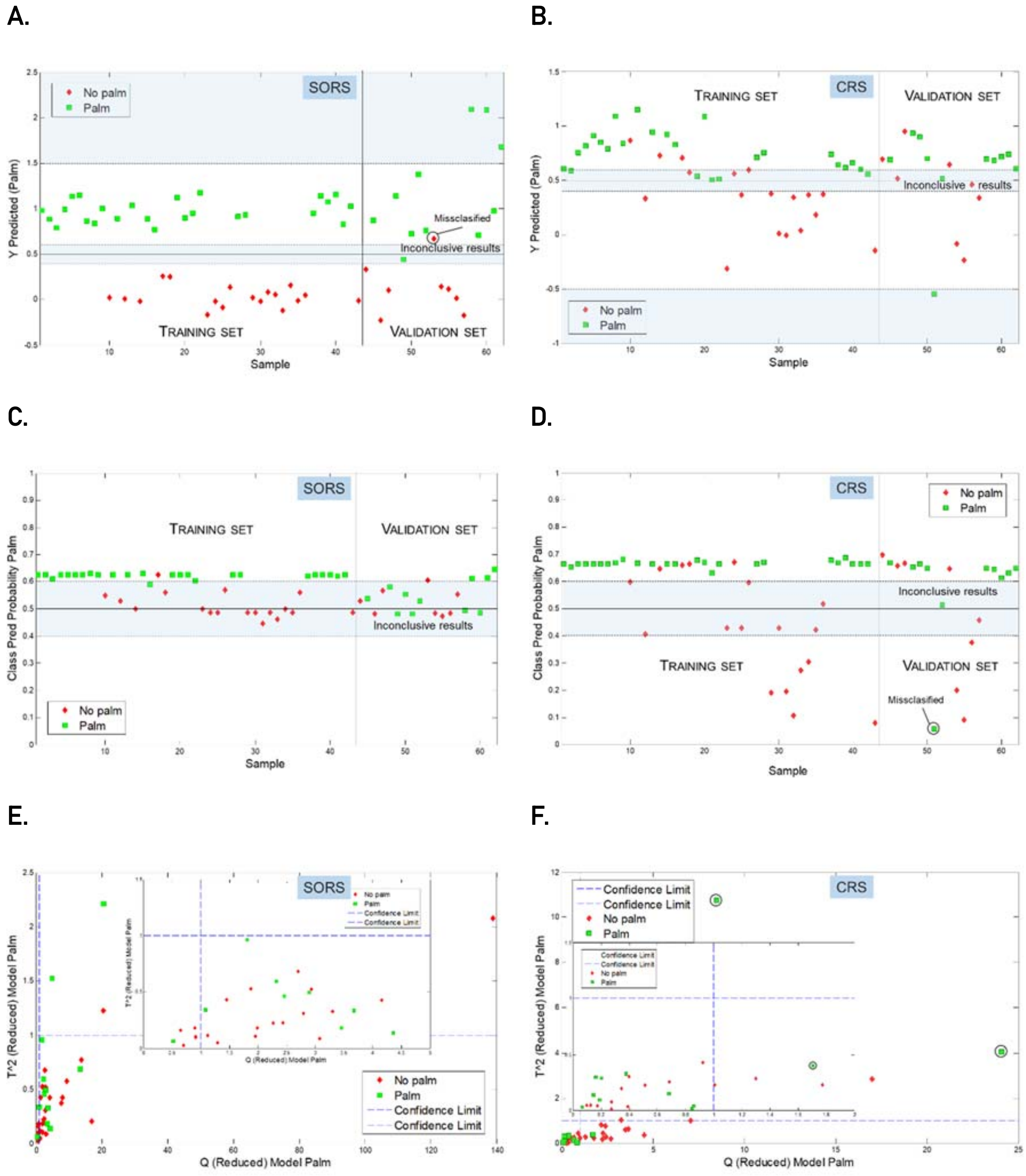


Figure 2.S6. Classification plots obtained for the 'palm / no palm' models developed by applying PLS-DA to the SORS data (**A**) and CRS data (**B**), built with eight and six LVs explaining 85.95% and 99.97% of CV, respectively; SVM-C to the SORS data (**C**) and CRS data (**D**); and OC-SIMCA to the SORS data (**E**) and CRS data (**F**) built with seven and four PCs explaining 85.57% and 99.95% of CV, respectively.

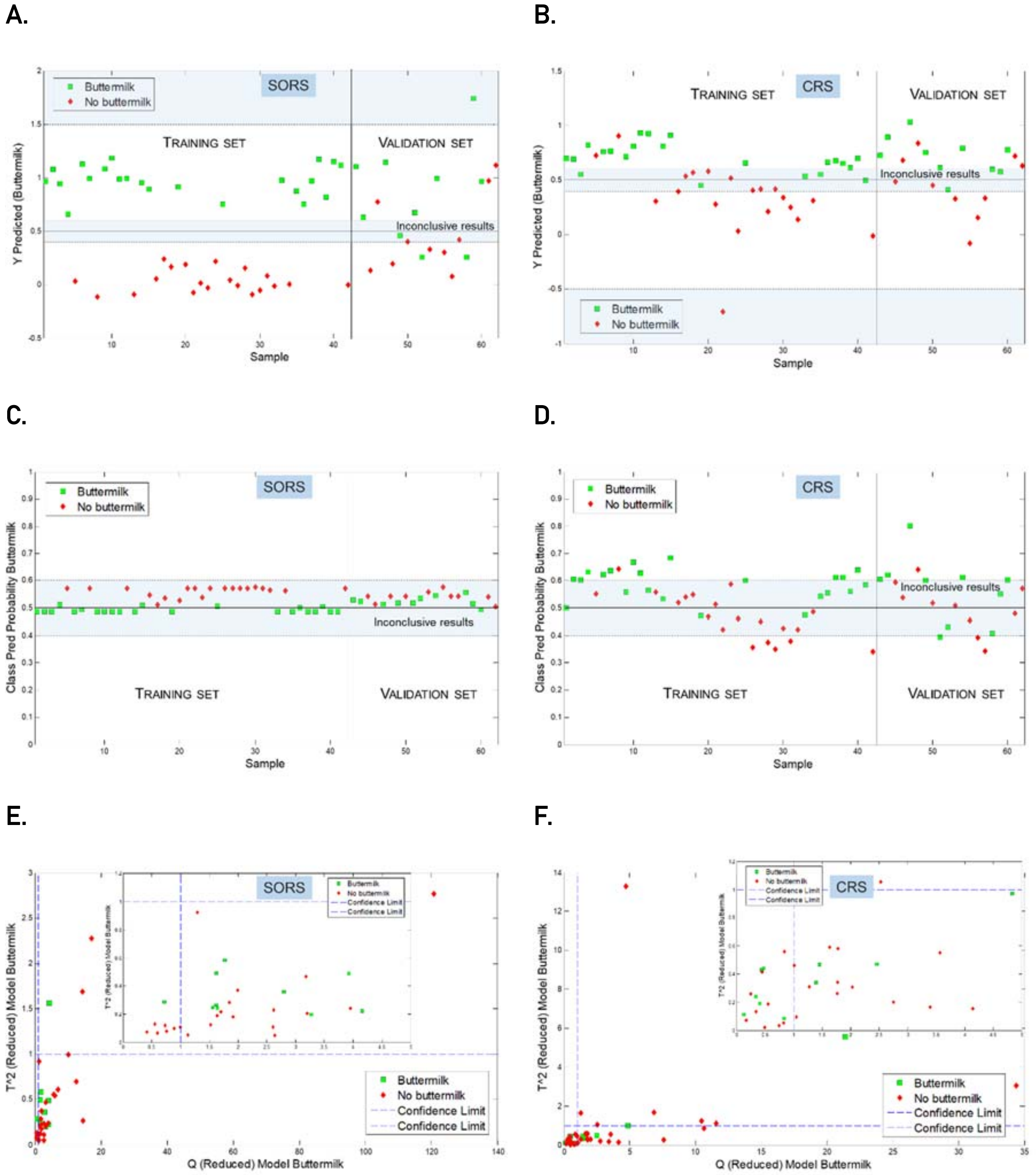


Figure 2.S7. Classification plots obtained for the 'buttermilk / no buttermilk' models developed by applying PLS-DA to the SORS data (**A**) and CRS data (**B**), built with eight and five LVs explaining 84.84% and 99.95% of CV, respectively; SVM-C to the SORS data (**C**) and CRS data (**D**); and OC-SIMCA to the SORS data (**E**) and CRS data (**F**) built with six and four PCs explaining 85.44% and 99.97% of CV, respectively.

Table 2.S3. Classification performance metrics for PLS-DA, SVM and OC-SIMCA 'palm / no palm' models developed from both types of techniques.

	PLS-DA model		SVM model		OC-SIMCA model	
	SORS	CRS	SORS	CRS	SORS	CRS
Sensitivity	0.74	0.63	0.16	0.63	0.58	0.63
Specificity	0.77	0.56	0.11	0.56	0.31	0.70
Positive PV (Precision)	0.93	0.86	—	0.72	0.50	0.71
Negative PV	0.95	0.89	—	0.73	0.32	0.57
Efficiency (Accuracy)	0.74	0.63	0.16	0.63	0.58	0.63
AUC (CCR)	0.76	0.60	0.14	0.60	0.45	0.66
Matthews CC	0.70	0.48	—	0.34	-0.14	0.30
Kappa coeff.	0.56	0.35	0.04	0.28	-0.13	0.28

PV: predictive value; AUC: area under curve; CCR: correctly classified rate; CC: correlation coefficient; "—": the parameter could not be determined because the number of samples assigned to the class was 0.

Table 2.S4. Classification performance metrics for PLS-DA, SVM and OC-SIMCA 'buttermilk / no buttermilk' models developed from both types of techniques.

	PLS-DA model		SVM model		OC-SIMCA model	
	SORS	CRS	SORS	CRS	SORS	CRS
Sensitivity	0.55	0.55	N/A	0.35	0.63	0.63
Specificity	0.55	0.55	N/A	0.35	0.28	0.52
Positive PV (Precision)	0.69	0.82	N/A	0.75	0.59	0.66
Negative PV	0.69	0.82	N/A	0.75	0.29	0.47
Efficiency (Accuracy)	0.55	0.55	N/A	0.35	0.67	0.63
AUC (CCR)	0.55	0.55	N/A	0.35	0.46	0.58
Matthews CC	0.30	0.42	N/A	0.21	-0.15	0.14
Kappa coeff.	0.25	0.28	N/A	0.16	-0.31	0.14

PV: predictive value; AUC: area under curve; CCR: correctly classified rate; CC: correlation coefficient; N/A: not applicable.

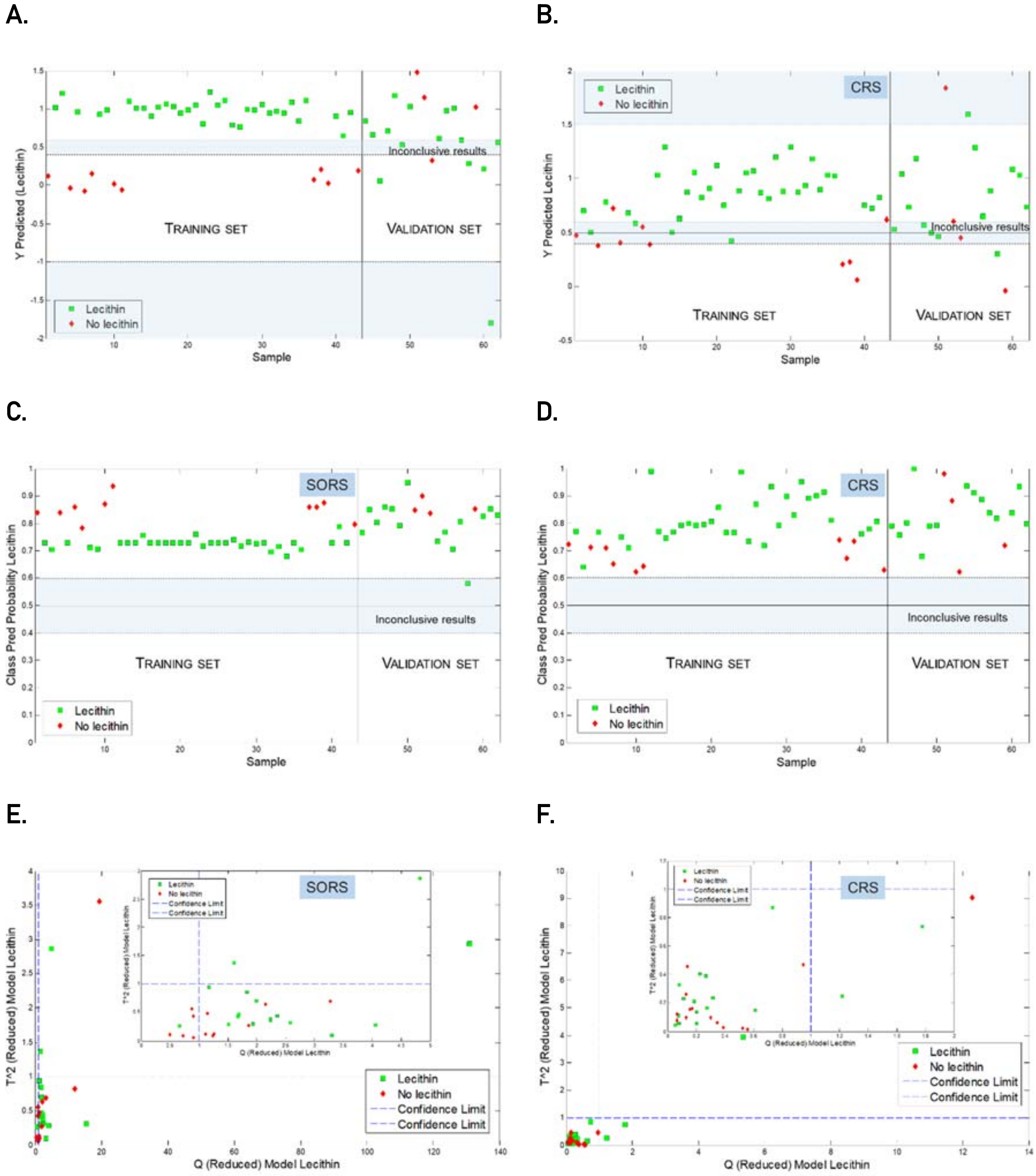


Figure 2.S8. Classification plots obtained for the 'lecithin / no lecithin' models developed by applying PLS-DA to the SORS data (**A**) and CRS data (**B**), built with eight LVs each, explaining 79.59% and 99.99% of CV, respectively; SVM-C to the SORS data (**C**) and CRS data (**D**); and OC-SIMCA to the SORS data (**E**) and CRS data (**F**) built seven six and four PCs explaining 85.30% and 99.93% of CV, respectively.

Table 2.S5. Classification performance metrics for PLS-DA, SVM and OC-SIMCA 'lecithin / no lecithin' models developed from both types of techniques.

	PLS-DA model		SVM model		OC-SIMCA model	
	SORS	CRS	SORS	CRS	SORS	CRS
Sensitivity	0.47	0.53	N/A	N/A	0.32	0.48
Specificity	0.31	0.32	N/A	N/A	0.38	0.41
Positive PV (Precision)	0.63	0.82	N/A	N/A	0.25	0.40
Negative PV	0.35	0.58	N/A	N/A	0.27	0.37
Efficiency (Accuracy)	0.47	0.53	N/A	N/A	0.32	0.48
AUC (CCR)	0.39	0.43	N/A	N/A	0.35	0.45
Matthews CC	-0.02	0.23	N/A	N/A	-0.38	-0.16
Kappa coeff.	-0.06	0.16	N/A	N/A	-0.28	-0.11

PV: predictive value; AUC: area under curve; CCR: correctly classified rate; CC: correlation coefficient; N/A: not applicable.

Table 2.S6 Predicted total fat content (TFT) of margarines (g/100 g product) from the validation set by the different PLSR quantitation models developed.

Sample	TFT declared	TFT predicted by SORS		TFT predicted by CRS	
		Mean center	Autoscaling	Mean center	Autoscaling
1	40	48.5	42.6	46.0	43.8
2	60	60.1	54.4	64.9	59.4
3	39	35.2	46.5	46.1	41.8
4	60	55.2	60.8	61.5	60.6
5	70	65.8	56.1	48.6	53.1
6	35	36.6	40.0	38.3	36.0
7	60	65.1	30.0	58.1	55.4
8	70	58.1	52.1	71.3	61.4
9	60	55.2	58.5	49.2	53.5
10	70	63.8	63.3	53.6	60.0
11	55	57.9	61.4	49.1	47.5
12	52	56.1	58.8	78.9	76.3
13	50	55.4	53.8	34.8	48.6
14	52	52.4	53.3	56.8	47.3
15	49	51.8	52.5	63.0	51.2
16	42	43.5	48.0	48.2	57.1
17	70	70.6	60.1	51.5	46.9

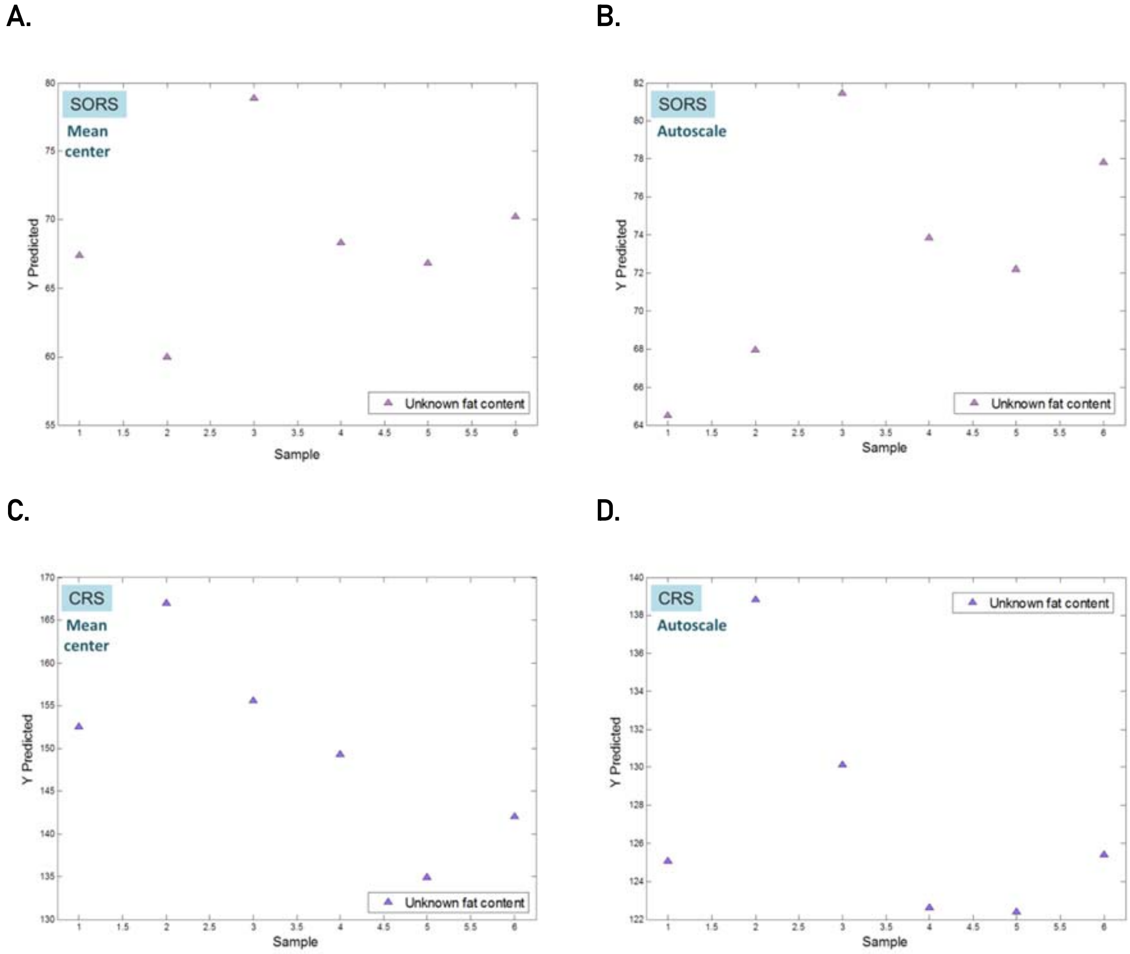


Figure 2.S9. Predicted results of the total fat content of margarines, whose value was unknown, by PLSR quantitation models developed from SORS data applying mean center (**A**) and autoscaling as pre-processing (**B**) and CRS data applying mean center (**C**) and autoscaling (**D**).

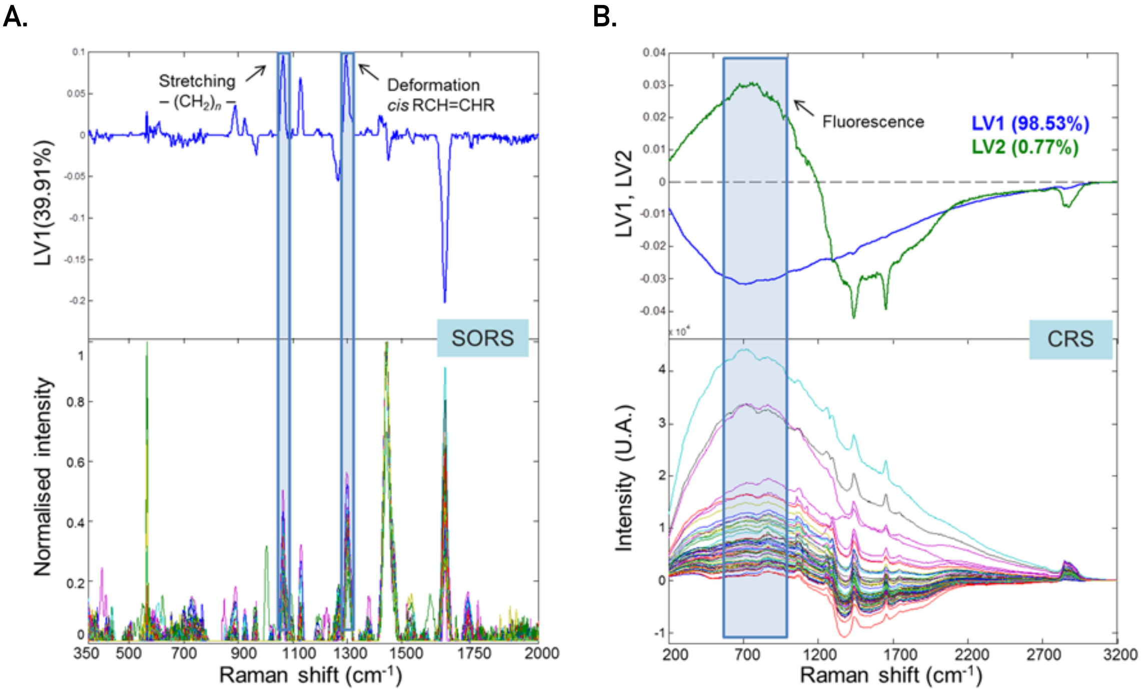


Figure 2.S10. Loading plots of the PLSDA 'linseed / no linseed' models and superposed SORS (A) and CRS spectra (B).

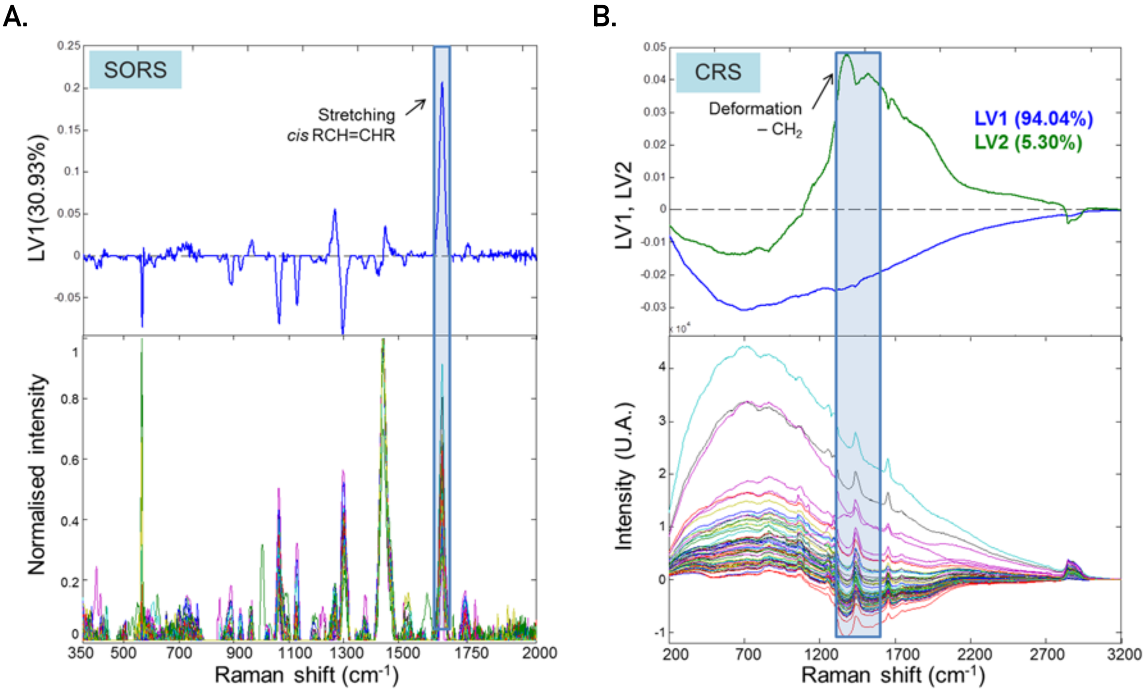


Figure 2.S11. Loading plots of the PLSDA 'sunflower / no sunflower' models and superposed SORS (A) and CRS spectra (B).

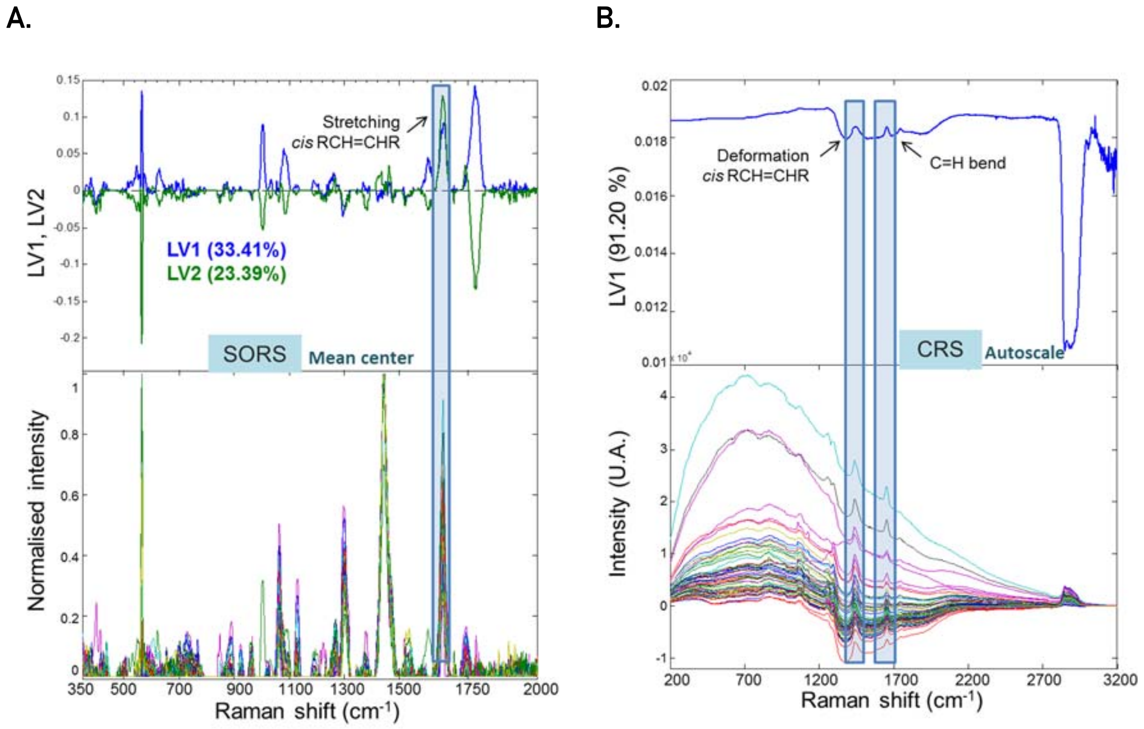


Figure 2.S12. Loading plots of the PLSR quantitation models and superposed SORS (**A**) and CRS spectra (**B**).

.

2.3. Artículo científico 2

The potential of the spatially offset Raman spectroscopy (SORS) for implementing rapid and non-invasive in-situ authentication methods of plastic-packaged commodity foods – Application to sliced cheeses.

Publicado en 2023 en la revista Food Control 146, 109522.

DOI: 10.1016/j.foodcont.2022.109522.

Food Control 146 (2023) 109522

Contents lists available at ScienceDirect

Food Control

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

The potential of the spatially offset Raman spectroscopy (SORS) for implementing rapid and non-invasive in-situ authentication methods of plastic-packaged commodity foods – Application to sliced cheeses

Alejandra Arroyo-Cerezo [a], Ana M. Jiménez-Carvelo [a,*], Antonio González-Casado [a], Itziar Ruisánchez [b], Luis Cuadros-Rodríguez [a]

[a] Department of Analytical Chemistry, University of Granada, C/ Fuentenueva s/n, E-18071, Granada, Spain
[b] Chemometrics, Qualimetrics and Nanosensors Group, Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo s/n, E-43007, Tarragona, Spain

* Corresponding author.
E-mail address: amariajc@ugr.es (A.M. Jiménez-Carvelo).

https://doi.org/10.1016/j.foodcont.2022.109522
Received 2 July 2022; Received in revised form 17 November 2022; Accepted 19 November 2022
Available online 21 November 2022

Highlights:

- Non-invasive in-situ measurements through sliced cheese packaging by SORS.
- Similarity study on cheese samples from different animal origin was performed.
- One input-class strategy was used to develop SIMCA-based classification methods.
- PLSR models were developed to quantify both fat and protein contents.

Keywords:

Spatially offset Raman spectroscopy

Non-targeted spectrometric fingerprinting

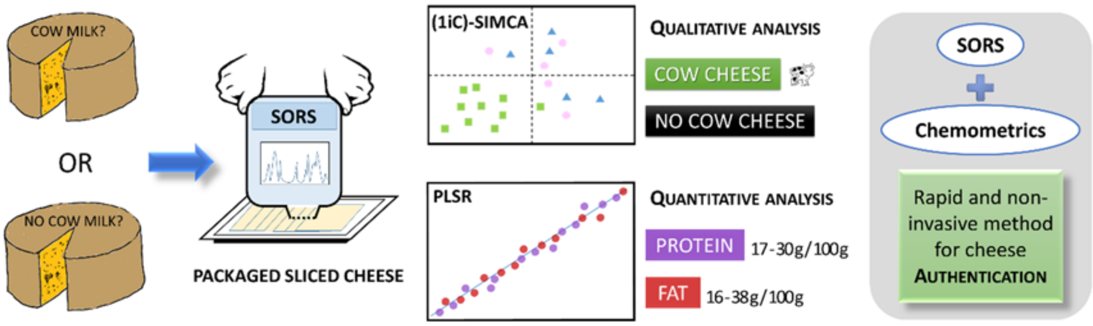
Chemometrics/data mining

Cheese milk-based authentication

Similarity analysis

Multivariate classification and quantitation analytical methods.

Graphical abstract



Abstract

Key points for the authentication of foodstuffs are the verification of the stated labelling, as well as checking its nature and verifying the absence of fraud. Portable spatially offset Raman spectroscopy (SORS) allows the rapid and straightforward acquisition of Raman spectra of packaged foods without the package opening and can therefore be considered as a non-invasive analytical technique par excellence since it enables in-situ analytical controls even directly on the grocery and supermarket shelves. Cheese is a dairy product available on the market which can be classified depending mainly on two characteristic features: (i) the animal origin of the milk from which it is produced and (ii) the manufacturing processes. This study aims to authenticate plastic-packaged sliced cheeses according to the origin of the milk stated by animal species (cow, sheep and/or goat). The NEAR similarity index was previously applied to verify that cheeses manufactured from the same animal species milk have a high similarity (NEAR > 0.8), higher than those found among cheeses produced from milk of different animal species. From the acquired spectra, the one input-class SIMCA approach was used to develop multivariate classification models in order to distinguish the origin of the milk by animal species. The performance parameters of the developed classification models show values

close to 0.9 for sensitivity and precision among others. The PLSR method was used to develop quantification models capable of predicting the total protein and fat contents (major nutrients) in cheese samples having errors of less than 2%. These advances open the door to the development of new rapid and non-invasive analytical methods for in-situ packaged food authentication, demonstrating the potential of SORS for food authentication supported by proper chemometrics.

1. Introduction

Food authenticity is currently a field of great importance for researchers, producers and consumers throughout the food production-consumption chain (**Shao *et al.*, 2019**). Moreover, there is no doubt that ensuring the authenticity of a product in order to detect any type of fraud has become a social concern and a problem which, beyond its economic implications, it could become a public health problem in certain cases (**Medina *et al.*, 2019**). "Authenticating/assuring a food" relies on checking what is declared on its label, verifying the actual nature of the food itself, e.g. the species or variety from which it comes, its geographical origin, as well as the absence of any adulterations (**Abbas *et al.*, 2018**). In this sense, products whose animal origin can determine a higher or lower quality are one of those that must be authenticated, e.g. meat or dairy products.

Cheese is a dairy product obtained by ripening the curd of animal milk. It is a food rich in fats and proteins and micronutrients such as calcium, phosphorus, sodium, magnesium, zinc, fat-soluble vitamins and B vitamins. Today there is a wide variety of cheese types commercially available (Brie, Cheddar, Edam, Emmental, Gouda, Mascarpone, Mozzarella, Parmesan, etc.) commonly grouped into soft, semi-soft or semi-hard, hard or extra-hard (**Farkye, 2004; Johnson, 2017; Zhen et al., 2021**) and this is mainly due to differences in the manufacturing processes used (**DKS 28-1, 2014**; **Dinkçi *et al.*, 2011**). In the same way, cheeses made from cow's, sheep's and goat's milk or from two or all three are often found on the market today. The main difference among the milk of these three species is the content and the nature of the major components (fats and proteins) and the minor components (minerals such as calcium and phosphorus) (**Coppa *et al.*, 2011**). Cheeses made from goat's or sheep's milk are usually more expensive than cow's milk, and in turn more expensive than mixed milk cheeses. This leads to one of the most common frauds by adulteration being the substitution of ingredients with lower cost ingredients, such as vegetable fat, which is added during the mixing process (**Dankowska *et al.*, 2015**), or the partial substitution of higher-value milk with cheaper milk of animal origin (**Genis *et al.*, 2021**). Currently, the strategy used for the authentication of milk and dairy products in the framework of the European Union is based on the analysis of the protein (casein) fraction using isoelectric focusing (**Regulation (EU) 2018/150**). However, protein analysis has a major drawback which lies in the complexity of the pre-treatment procedure of the sample to isolate and break down the caseins and carry out the subsequent characterisation of the caseins by isoelectrofocusing.

To date, studies on cheese authentication have been based on the quantitation of triglycerides and fatty acids, in other words, the targeted approach has been applied. Nevertheless, note that the greater or lesser proportion of these compounds does not imply that cheese has been adulterated or that a certain type of milk has been used for its production, since the differences in the composition of these compounds in the milk are affected by different factors, including the animal's diet, as well as a fat fortification of cow's milk to mimic the fat composition of more expensive milks (goat or sheep). Thus, the targeted approach could not be convenient to authenticate the animal origin of the milk used in the production of the cheese, being more useful to employ an untargeted approach, e.g. the application of well-known spectrometric fingerprinting methodology. In this context, spatially offset Raman spectroscopy (SORS) technique, which has been scarcely applied in the area of food authenticity (**Arroyo-Cerezo *et al.*, 2021**), is presented as a promising candidate for the development of screening analytical methods since it allows measurements by applying the laser directly on the surface of the product in its original packaging without the need to open it and without affecting the results, obtaining a distinctive cheese fingerprint.

In fact, the tandem Raman measurements and chemometrics was revealed by Yazgan et al. (**Yazgan *et al.*, 2020**) as a rapid analytical methodology to assess the animal origin of a total of 56 milk samples (cow, sheep and goat) with 100% classification accuracy values. Therefore, the application of this tandem could be reliable to perform analytical processes in the most sustainable way possible, eliminating the sample treatment step, the replacement of toxic reagents and the reduction of waste generated. However, to date there is only one study combining SORS and chemometric tools for ensuring the cheese authenticity. This study performed by Ostovar Pour *et al.* (**Ostovar Pour *et al.*, 2021**), employed SORS together with principal component analysis-discriminant analysis (PCA-DA) to characterise Cheddar, Manchego and Pecorino Romano cheeses by developing a multivariate descriptive model of these cheeses but no predictive models able to differentiate between the different types of cheese analysed were reported. By the contrast, some studies using conventional Raman with chemometrics focused on (i) the characterization of cheeses of different commercial brands using the spectral regions of specific compounds such as carbohydrates or proteins (**Zhang, 2020**), (ii) evaluation of the type of fat in cheese and detection of adulterations (**Genis *et al.*, 2021**), (iii) differentiation between whole cheeses and skimmed cheeses (**Oliveira *et al.*, 2016**), (iv) characterization of protected designation of origin cheeses such as Parmigiano Reggiano (**Vigni *et al.*, 2020**) or (v) analysis of cheese microstructure using Raman imaging (**Smith *et al.*, 2016**) have been reported. Although, none of them based on the authentication of the milk origin by animal species of the cheese.

In this context, the study presented here aims at applying the SORS technique employing the non-targeted fingerprinting methodology in conjunction with chemometric tools to authenticate plastic-packaged sliced cheese samples. Previously, a similarity analysis using the nearness index (NEAR) (**Valverde-Som *et al.*, 2018**) has been applied to assess the differences and similarities between cheeses according to their animal origin. Then, one input-class soft independent modelling of class analogies (1iC SIMCA) approach has been used (**Jimenez-Carvelo *et al.*, 2017**), training the model with representative samples of genuine cow's milk cheese (target class). Note that class modelling methods such as SIMCA are considered the optimal for food authentication purposes by some authors (**Rodionova, Titova, & Pomerantsev, 2016**), compared to discriminant classification methods. Finally, quantitative models of protein and fat contents were performed using a partial least-squares regression (PLSR) method to verify what is declared on the label.

2. Materials and methods

2.1. Samples

A set of 80 samples of different cheeses was used for this study. All samples were in thin transparent plastic containers, which are common in similar marketed products in supermarkets. Products were purchased in supermarkets located in Granada, Malaga and Almería (Southern Spain). Mainly these were sliced cheeses, but some cheese wedges were also included. The main difference among the samples was the milk origin by animal species, namely: 35 samples only from cow's milk, 10 samples only from goat's milk, 6 samples only from sheep's milk and 29 samples from a mixture of two or all three milks (see Table 2.10). Note that the proportion of milk stating on the label of the cheeses made from a mixture of two or more animal milks is the minimum percentage in the product, so that the sum may not reach 100%.

Manufacturing process of the studied cheeses was also different, identifying many different types of cheese, such as cheddar, edam, soft, semi-cured, among others. This is also relevant to understand the different ranges of macronutrients found in the sample set: 17 - 30 g protein / 100 g cheese and 16 - 38 g fat / 100 g cheese.

Table 2.10. Number of cheese samples included in the study divided according to the milk composition by animal species.

	Milk animal species origin (%)			Number of
Groups	Cow	Goat	Sheep	samples
Cow	100	0	0	35
Goat	0	100	0	10
Sheep	0	0	100	6
Two-milks mixture	80	5	0	6
Three-milks mixture	60-65	10-17	10-15	16
	68	3	14	2
	32	33	20	1
	29	21	35	1
	N/S	N/S	N/S	3
			TOTAL	**80**

N/S: only the type of milk is stated but the % content is not specified.

2.2. SORS measurement

A portable equipment was used to measure the cheese samples: Vaya Raman (Agilent Technologies, Santa Clara, CA, USA). The excitation wavelength was 830 nm. This equipment is manufactured to avoid the influence of fluorescence, a major drawback in conventional Raman measurements. In addition, the spatial offset (in this case, 0.7 cm) allows measuring through the original packaging of the samples.

Each sample was measured without opening the plastic container from the bottom side, performing non-invasive measurements. The equipment performs a first measurement without offset (i.e., a conventional Raman measurement where the collection point is the same as the point of incidence of the laser), a second measure with offset (0.7 mm shift) and finally performs a correction of both spectra to eliminate all possible influence of top layers, the plastic packaging in this case. This provides a (0-1) normalized Raman spectrum of the sample material in the container without the influence of the plastic packaging in the range of 350-2000 cm−1 Raman shift, because the zero offset spectrum has a major contribution from the surface area of the sample, the plastic packaging, and the spectrum collected with offset has deeper contribution, from the sub-layers, the cheese in this case. This fact has already been verified by the authors in a recently published study on the application of SORS in the authenticity testing of margarines (**Jimenez-Carvelo et al., 2022**). The measurement time was between 30 s and 2 min for each sample, while the exposure time of the samples to the laser varied between 0.5 and 2 s for each sample.

2.3. Multivariate analysis

SORS raw data files from each spectrum were exported in 'comma separated value' (CSV) format, and then converted to MATLAB format (ver. R2017b, Mathworks, MA, USA). All multivariate analysis was carried out using PLS_Toolbox (ver. 8.2.0, Eigenvector Research Inc. MA, USA) working under the MATLAB environment. The data vector collected was composed of 1651 variables. The data pre-processing was done with a home-programmed MATLAB script. This implemented several algorithms from the MATLAB Bioinformatics Toolbox™. The pre-processing steps were: (i) grouping and overlay of the spectra, (ii) removing of the noise, (iv) correction of the baseline, and (vi) autoscaling of the data set.

SIMCA was used as classification method for the development of the multivariate models, for which the one input-class (1iC) approach was adopted, considering samples of cheese produced from 100% cow's milk as the target class. For validation purposes, the samples were split into three external validation sets considering two general classes: 'cow' class and 'non-cow' class. Within the latter class, different types of cheeses were available according to the milk origin by animal species. Detailed information on the number and type of samples used for the validation of the model is shown in Table 2.11. In all cases the performance of the classification methods has been evaluated by obtaining proper quality parameters, (**Cuadros-Rodríguez, Pérez-Castaño, & Ruiz-Samblás, 2016**).

Table 2.11. Information related to the type and number of cheese samples used for the validation of the SIMCA model developed.

External validation set	Number and type of samples
1	10 (100% cow milk) 10 (100% goat milk) 6 (100% sheep milk)
2	10 (100% cow milk) 12 (mixtures of cow and sheep milk)
3	10 (100% cow milk) 10 (100% goat milk) 6 (100% sheep milk) 12 (mixtures of cow and sheep milk) 17 (mixtures of cow, goat and sheep milk)

** Notice that the training set consisted of 25 cheese samples produced from cow's milk only.*

PLSR was used to build two quantitation multivariate models according to (i) the protein content and (ii) the fat content stated on the labelling of the cheese samples. The total set of 80 samples was divided into a training set and a validation set.

3. Results and discussion

Figure 2.10 shows the Raman spectra of three cheese samples from milk origin by animal species (cow's, sheep's and goat's milk). These SORS spectra show a typical shape of conventional Raman spectra. It is known that generally weak spectral signals below 800 cm^{-1} correspond to amino acids, and more intense signals from 1000 cm^{-1} onwards to lipids. (**Ostovar Pour et al., 2021**). The highest peak for all three types of cheese is obtained at 1446 cm^{-1}, a typical instrumental fingerprint reflecting a high fat content, especially from cholesterol (**Smith et al., 2016**). The next highest peaks at 1300 and 1063 cm^{-1} correspond to phospholipids and other lipids respectively (**Ostovar Pour et al., 2021**). Finally, the peak at 1128 cm^{-1} is related to the presence of saturated fatty acids (**Amjad et al., 2018**).

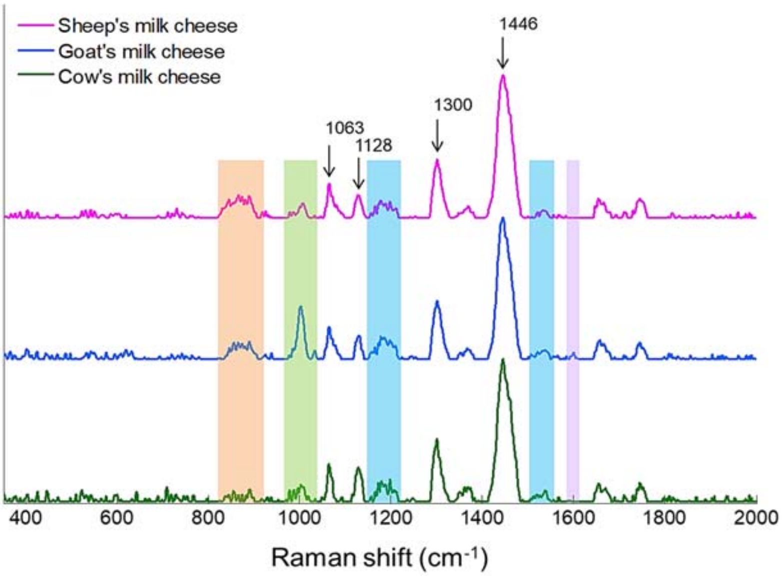


Figure 2.10. Final Raman spectrum of three random cheese samples from different origin animal after pre-processing and normalization measured by SORS.

Slight differences can be identified between the three spectra, which are highlighted in Figure 2.10. In the orange-coloured region (between 844 and 890 cm^{-1}), a higher intensity is observed in the spectrum of cheese made from sheep's milk than in that of goat's milk, which is in turn higher than that of cow's milk. The concentration of fats and proteins in sheep's milk is known to be higher than in goat's and cow's milk, in the same way that sheep's and goat's milk has a higher proportion of short-chain fatty acids. Also noteworthy is the peak at

1002 cm^{-1} (green region), notably more intense in the cheese made from goat's milk than the other two, commonly associated with the presence of phenylalanine. Similarly, a peak at 1600 cm^{-1} (purple region) is only detectable in the spectrum of cheese made from goat's milk. The blue zones indicate bands where other slight differences in the spectra are observed, probably also attributable to compositional differences (**Yazgan et al., 2020**).

These small dissimilarities found in the spectra show that cheeses could be differentiated according to the milk origin by animal species. However, they can sometimes be difficult to detect visually, especially when there is a large set of samples. Therefore, a multivariate data approach was used to analyse the data and try to discriminate the cheese samples on the basis of this compositional difference. It should be highlighted that the differences affect the full spectrum as a whole and not specific peaks or regions. This is why a multivariate approach is essential to mine the significative information from spectrometric fingerprints, which is fundamentally different from a lineal combination of several univariate events (or single measurements) simultaneously considered. This is the distinctive feature of non-targeted analytical methods, and any attempt to rationally justify their application by searching for differences in chemical composition is usually destined to fail.

3.1. Similarity study

A comprehensive similarity study was carried out considering all the cheese samples concerned. Performing a similarity analysis is crucial in order to obtain reliable results when multivariate classification methods are applied to problems similar to this one, i.e., problems in which it is intended to differentiate classes consisting of products (in this case, cheeses) showing differences due to features others than those to be classified, e.g., physical appearance, ripening, manufacturing or technological processes, or even presence of other components.

For this case, the similarity analysis is performed following a similar setup to the analysis of variance. Indeed, the similarity between cheeses produced from the milk of different animal (between-milks similarity) species is contrasted with the inherent lack of similarity of the different cheeses having a common dairy origin (within-milk similarity).

For this purpose, the nearness index (NEAR) was applied which may be calculated by the following equation:

$$\mathrm{NEAR}\,(\mathbf{X}_1, \mathbf{X}_2) \;=\; 1 - \left[\sqrt{\frac{(\mathbf{X}_1 - \mathbf{X}_2)\;\times\;(\mathbf{X}_1 - \mathbf{X}_2)^{\mathrm{T}}}{(\mathbf{X}_1 + \mathbf{X}_2)\;\times\;(\mathbf{X}_1 + \mathbf{X}_2)^{\mathrm{T}}}} \;\right]$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are the characteristic vectors defining each of the spectra to be compared (the superscript T is denoting the transposed matrix). Note that NEAR is calculated for each pair of spectra thus the term in square brackets represents the (0-1) normalized Euclidean distance between the two vectors (**Valverde-Som et al., 2018**). Unlike other similarity indices, such as the cosine of the angle between the two vectors (cos) or the correlation coefficient (R), NEAR depends linearly on the degree of similarity, making it possible to establish comparisons in an intuitive way (**Pérez-Robles et al., 2017**). In this way, both $\text{NEAR}_{\text{within}}$ and $\text{NEAR}_{\text{between}}$ values were calculated for all comparisons. A layout of both experimental design and similarity results is displayed in Figure 2.11.

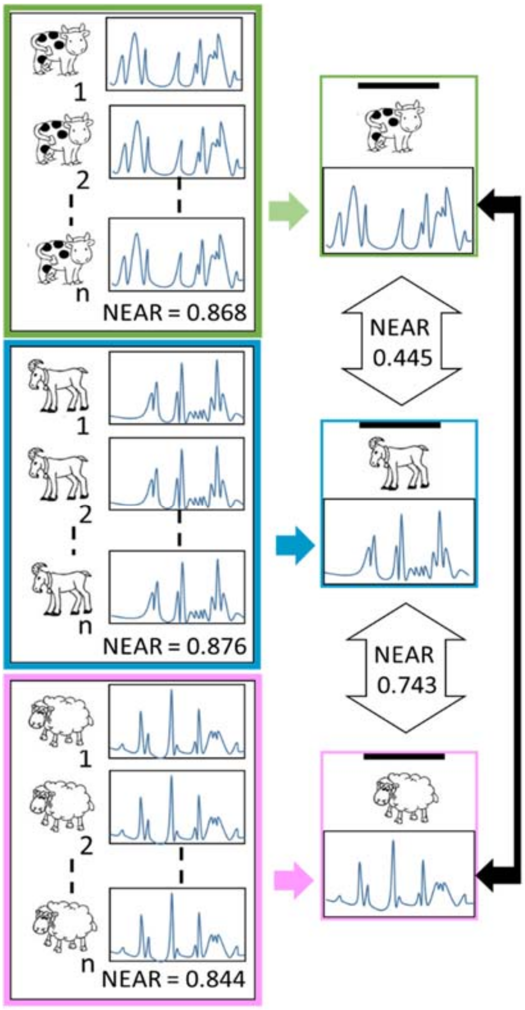


Figure 2.11. Workflow of the similarity experimental design showing the values of the NEAR index for both within-milk and between-milks similarities.

For each pair of spectra of cheese samples produced from milk of the same animal species, a single $\text{NEAR}_{\text{within}}$ value is obtained. A representative value for each group of cheese is then obtained by calculating the mean of all the values in concern, which are referred to as 'overall $\mathrm{NEAR}_{\mathrm{within}}$'. In addition, the corresponding standard deviation values are also calculated. The results are shown in Table 2.12.

Table 2.12. Overall $\mathrm{NEAR}_{\mathrm{within}}$ values obtained after evaluating the similarity between cheese samples produced from milk of the same type of animal species.

	Pairwise comparison		
Parameter	within-cow	within-goat	within-sheep
Overall $\mathrm{NEAR}_{\mathrm{within}}$	0.87	0.88	0.84
$\mathrm{NEAR}_{\mathrm{within}}$ standard deviation	0.03	0.04	0.03

In all cases, the overall $\mathrm{NEAR}_{\mathrm{within}}$ value is significantly high, above 0.8. This corroborates that the cheeses produced with the same type of milk by animal species show a high degree of within-group similarity. In addition, the small values obtained for the $\mathrm{NEAR}_{\mathrm{within}}$ standard deviations, ranged from 0.03 to 0.04, demonstrate a low dispersion of the similarity results, which indicates that most of the NEAR indices of the samples tends to be clustered close to their mean.

Next, the similarity of cheese samples produced from milks of different animal species was evaluated with each other. For this purpose, an average vector for each group of cheese samples, according to the milk origin by animal species, was obtained and finally these mean vectors were compared with each other. In two cases (cow-goat and cow-sheep) the $\mathrm{NEAR}_{\mathrm{between}}$ index is significantly less than 1, indicating a low between-group similarity of the averaged spectrometric fingerprints. However, it should be noted that a no-negligible similarity was found between cheeses made from goat's and sheep's milk. It was therefore decided to perform a classification using SIMCA, as described in section 2.3, in order to authenticate cow's milk cheeses from the other ones.

3.2. Authentication according to animal origin

As mentioned in previous sections, in order to authenticate the different cheeses analysed, 1iC SIMCA approach was applied for the development of one multivariate model, considering three different set for the validation step. SIMCA is widely known, and in general terms is based on building a principal component analysis (PCA) model for each class, which describes the structure of that class as well as possible. The optimal number of principal components should be chosen for each model separately. Then, with the information obtained from the PCA model for each class, SIMCA performs a spatial distribution in which it defines each of the classes. Finally, the assignment of each unknown sample to a particular class is based on the nearest distance to the corresponding regions established by the PC model.

In this study, only one input-set composed by 25 cheese samples made exclusively with cow's milk was considered for the model training purpose. Note that these samples were chosen by applying the Kennard-Stone algorithm (**Kennard, & Stone, 1969**). 10 principal components (PCs) were enough to explain 70.66% of the variance for the cow class. Then, the model was validated considering three different external validation sets (see Table 2.11), whose classification plots and classification results in terms of quality parameters are shown in Figure 2.12 and Table 2.13, respectively.

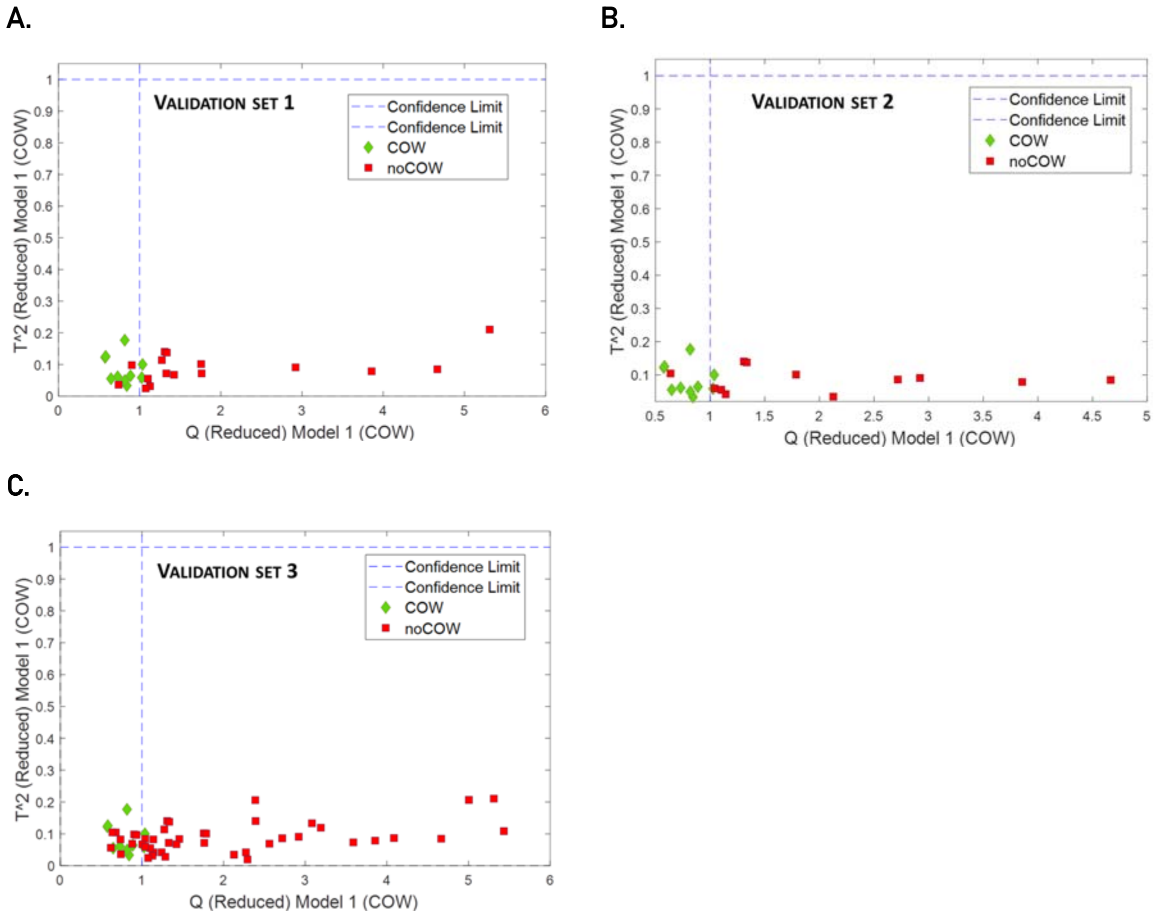


Figure 2.12. Classification plots obtained for the three external validations performed on the (1iC)-SIMCA model. (**A**) external validation set 1, (**B**) external validation set 2 and (**C**) external validation set 3.

Note that validation set 1 is composed only of samples whose animal origin is 100% cow, 100% sheep and 100% goat, validation set 2 is made with 100% cow and cow-sheep mixture samples and finally, validation 3 is composed of 100% cow, 100% sheep, 100% goat samples and all the mixtures analysed.

The classification of the samples was performed from both Q and Hotelling T^2-statistics. The classification region for the cow class was established according to Q and T^2 normalized values equal to 1, meaning that a sample must take values lower than 1 to be classified in the cow class. As can be seen in Figure 2.12, in all the validation sets there are only two cheese samples whose milk origin is 100% cow's milk that are misclassified and whose difference between them is due to the ripening and manufacturing process, one of them being a semi-cured cheese and the other a soft cheese. In addition, when evaluating the misclassified samples in the three sets belonging to the 'non-cow' class, it was observed that in all cases they corresponded to samples whose cow's milk content is higher than 65% and, in some cases, reaching 80%.

Table 2.13. Overall classification performance metrics for the 1iC SIMCA model considering three validation sets.

Metrics	External validation set		
	1	2	3
Sensitivity	0.84	0.85	0.84
Specificity	0.82	0.84	0.81
Positive predictive value (Precision)	0.84	0.86	0.87
Negative predictive value	0.95	0.93	0.99
Efficiency (Accuracy)	0.84	0.86	0.84
AUC (Correctly classified rate)	0.83	0.85	0.82

AUC: area under curve

However, it is worth noting that, despite these errors in classification, the model obtained in all cases values of classification performance metrics above 0.8 for sensitivity, specificity and precision (see Table 2.13), indicating that the 1iC SIMCA model is an efficient tool for discriminating samples of cheeses made from 100% cow's milk from cheeses made from other types of milk (goat, sheep and mixtures). Note that with similar techniques, such as near infrared spectroscopy coupled with chemometrics, result for degree of ripening prediction was 100%; but only obtained 50% correct classification for the animal origin of the cheese (**Soto-Barajas *et al.,* 2013**). Therefore, it is proved that developed 1iC SIMCA classification method is useful for the reliable evaluation of the authenticity of the animal-origin of the milk used in cheese production.

Moreover, the advantages of this method over the official method of casein detection by isoelectric focusing are evident, as rapid, non-invasive and in-situ authentication. Other alternatives can be found in scientific literature, i.e. immunochemical methods such as lateral flow immunochromatographic tests and enzyme-linked immunosorbent assays (ELISA) that are fast and useful for the detection of cow's milk in pure sheep or goat cheeses, since they detect bovine immunoglobulin G (**Amaral *et al.,* 2018**). Also, with DNA-based techniques such as real-time PCR good results are obtained for the authentication of the animal origin of the milk used for cheese (**Golinelli *et al.,* 2014**; **Seçkin *et al.,* 2017**; **Guo *et al.*, 2019**). But the methodology shown here, based on SORS-1iC SIMCA, would offer advantages over the latter in terms of speed of analysis and no sample treatment or manipulation needed.

3.3. Verification of the protein and fat contents

Proteins and fats are the main nutrients present in cheeses and their proportion in this food varies according to the type of milk used for its production (taking into account the compositional variations of each milk mentioned above). Thus, verifying the content of both nutrients is in accordance with the labelling is also an important point to ensure the quality of this product. Two quantitation models were developed to predict protein and fat content using the PLSR, which is a well-known multivariate quantitation method in food analysis.

For both models, a training set of 68 cheese samples was selected by applying the Kennard-Stone algorithm (**Kennard, & Stone, 1969**), and a validation set of 12 samples. 7 latent variables (LVs) were used to build both models (protein and fat) explaining 33.51% and 33.46% of the variance, respectively. These criteria were enough to develop the models with a high fitting goodness, as an R^2 of 0.991 and 0.992 was obtained for the training set of the protein and fat models. The root mean square error of cross validation (RMSECV) was also calculated. The development of the quantitation models was concluded with the external validation and the predicted values obtained are shown in Table 2.14. The reliability of the built models was established on the basis two criteria: (i) the suitability and goodness of fit of the quantification model in the training stage by means of the determination coefficient (R^2), and (ii) the errors of quantification (external validation errors) via root mean square error of prediction (RMSEP), mean absolute error (MAE), median absolute error (MdAE) and standard deviation of validation residuals (SDV) (**ASTM E2617-17, 2017**) and the results can be seen in Figure 2.13.

Clarify that, the RMSEP, MAE, MdAE and SVD values are successful (less than 2%). This indicates that the value predicted by the multivariate model is very close to the reference value, i.e. the value indicated on the food label.

Ma et al. (**Ma *et al.*, 2019**) reported similar results using a portable near infrared spectroscopy equipment and PLSR. The RMSEP values obtained in the prediction of the total protein content of cheeses are comparable, although the R^2 obtained in the present study was closer to 1. In addition, their proposed method required the sample to be removed from its original packaging for analysis. Another advantage of the method proposed in this study with respect to others found in the literature is that it allows verification of the total % of fat in a non-invasive way, without the need to carry out a compound extraction stage and subsequent analysis by conventional techniques such as gas chromatography (**González-Martín *et al.,* 2017**; **Eisenstecken *et al.,* 2020**).

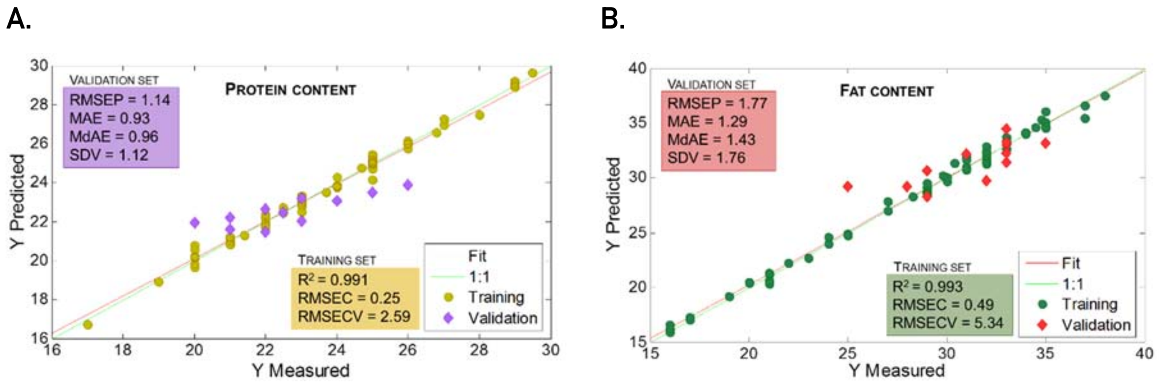


Figure 2.13. Predicted values by the PLSR models for the training (68 samples) and validation (12 samples) sets of the (**A**) protein and (**B**) fat content.

Table 2.14. Values stated on the label and predicted results by the developed PLSR quantitation models for the total protein and fat contents of the cheese samples included in the external validation set.

Sample	Protein content (g/100 g cheese)		Fat content (g/100 g cheese)	
	Stated	Predicted	Stated	Predicted
1	21.0	21.6	33.0	33.0
2	22.5	22.5	32.0	29.7
3	22.0	22.6	28.0	29.2
4	25.0	23.5	29.0	30.7
5	24.0	23.1	33.0	31.4
6	20.0	21.9	35.0	33.1
7	21.0	22.2	33.0	32.2
8	23.0	22.0	33.0	34.4
9	26.0	23.9	33.0	33.2
10	24.0	23.1	25.0	29.2
11	23.0	23.2	29.0	28.2
12	22.0	21.5	31.0	32.2

Note also that the prediction errors of the protein quantitation model are significantly better than those for the fat quantitation model. This could be due to the fact that protein bonds generate more specific Raman signals than fat's bonds (**Moros, Garrigues, & de la Guardia, 2007**). Characteristic bands of Raman spectra other than those mentioned above are the amide bands of proteins, namely amide I at 1650-1670 cm^{-1}, amide II at 1550-1560 cm^{-1} (blue zone on the right of Figure 2.10) and the amide III at 1300 cm^{-1}. The last one is often attributed to lipids, but also an overlap with the protein band at 1270 cm^{-1} (**Smith *et al.*, 2016**).

4. Conclusions

The research study has presented the development of an analytical screening methodology to authenticate the animal origin of the milk used in the cheese production, based on the combination of spatially offset Raman spectroscopy (SORS) and chemometrics. Moreover, a comprehensive similarity study has been performed, which has revealed the difference between the Raman signals of cheeses of different animal origin. All this has argued the application of multivariate classification techniques for the development of a model for assessing the authenticity according to the animal origin of the different cheeses. In addition, as a complementary action, the possibility of verifying the total fat and protein content declared on the label of the cheeses was also proved, obtaining prediction errors of less than 2% for the validation set.

In this sense, it should be noted that the potential of the SORS technique in the field of food authenticity has been demonstrated, as well as the instrumental fingerprinting methodology as a promising screening alternative to be applied as a preliminary to the traditional characterisation of the caseins by the isoelectrofocusing approach to evaluate the type of milk used in cheese production. This SORS-based methodology is fast, inexpensive, non-invasive, and allows for potential on-line/at-line implementations and can be straightforwardly implemented in routine quality control laboratories. The use of one input-class approach to develop the multivariate authentication model for cheese is highly recommended not only for the industry but also for regulatory organizations, since they would only need to consider genuine samples from the class to be authenticated in order to train the model, as well as being able to quantify the total protein and fat content in a simple way.

Acknowledgments

This work was partially supported by University of Granada (Spain) within the framework of the funding corresponding to program 'pre-competitive research projects for young researchers' (PPJIA2021-09). **AAC** wants to express their sincere gratitude to the Spanish Ministry of Sciences, Innovation and Universities for a pre-doctoral fellowship FPU (FPU20/04711, Formación del Profesorado Universitario). In addition, **AMJC** wish to acknowledge the Department of Economic Transformation, Industry, Knowledge and Universities belong to Regional Andalusia Government (Spain) for the Postdoctoral fellowship (DOC_00121).

References

Abbas, O., Zadravec, M., Baeten, V., Mikus,T., Lesic, T., Vulic, A., Prpic, J., Jemersic, L., & Pleadin, J. (2018). Analytical methods used for the authentication of food of animal origin. *Food Chem., 246,* 6-17. https://doi.org/10.1016/j.foodchem.2017.11.007

Amaral, J. S., Mafra, I., Pissard, A., Fernández Pierna, J. A., & Baeten, V. (2018). Milk and milk products. *FoodIntegrity Handbook. A Guide to Food Authenticity Issues and Analytical Solutions; Morin, J.F., Lees, M.*, Eds, 3-25. https://doi.org/10.32741/fihb.1.milk

Amjad, A., Ullah, R., Khan, S., Bilal, M., & Khan, A. (2018). Raman spectroscopy based analysis of milk using random forest classification. *Vib. Spectrosc.*, *99*, 124-129. https://doi.org/10.1016/j.vibspec.2018.09.003

Arroyo Cerezo, A., Jiménez Carvelo, A.M., González Casado, A., Koidis, A., & Cuadros Rodríguez, L. (2021). Deep (offset) non-invasive Raman spectroscopy for the evaluation of food and beverages – A review. *LWT-Food Sci. Technol., 149,* Article 111822. https://doi.org/10.1016/j.lwt.2021.111822

ASTM E2617-17. (2017). Standard practice for validation of empirically derived multivariate calibrations, ASTM International.

Regulation (EU) 2018/150 amending Implementing Regulation (EU) 2016/1240 as regards methods for the analysis and quality evaluation of milk and milk products eligible for public intervention and aid for private storage. *OJEU,* L26/14- 47.

Coppa, M., Ferlay, A., Monsallier, F., Verdier-Metz, I., Pradel, P., Didienne, R., Farruggia, A., Montel, M.C., & Martin, B. (2011). Milk fatty acid composition and cheese texture and appearance from cows fed hay or different grazing systems on upland pastures. *J. Dairy Sci., 94,* 1132-1145. https://doi.org/10.3168/jds.2010-3510

Cuadros Rodríguez, L., Pérez Castaño, E., & Ruiz Samblás, C. (2016). Quality performance metrics in multivariate classification methods for qualitative analysis. *Trend Anal. Chem., 80*, 612-624. http://dx.doi.org/10.1016/j.trac.2016.04.021

Dankowska, A., Małecka, M., & Kowalewski, W. (2015). Detection of plant oil addition to cheese by synchronous fluorescence spectroscopy. *Dairy Sci Technol., 95*, 413-424. https://doi.org/10.1007/s13594-015-0218-5

Dinkçi, N., Kesenkaş, H., Seçkin, A., Kınık, Ö., & Gönç, S. (2011). Influence of a vegetable fat blend on the texture, microstructure and sensory properties of kashar cheese. *Grasas y Aceites, 62*, 275-283. https://doi.org/10.3989/gya.091810

DKS 28-1. (2014). Cheese Specification. Part 1: General. Nairobi: Kenya Bureau of Standards.

Eisenstecken, D., Stanstrup, J., Robatscher, P., Huck, C.W., & Oberhuber, M. (2021). Fatty acid profiling of bovine milk and cheese from six European areas by GC-FID and GC-MS. *Int. J. Dairy Technol. 74*, 215-224. https://doi.org/10.1111/1471-0307.12749

Farkye, N.Y. (2004). Cheese technology. *Int. J. Dairy Technol., 57,* 91-98. https://doi.org/10.1111/j.1471-0307.2004.00146.x

Genis, D.O., Sezer, B., Durna, S., & Boyaci, I.H. (2021). Determination of milk fat authenticity in ultra-filtered white cheese by using Raman spectroscopy with multivariate data analysis. *Food Chem., 336,* Article 127699. https://doi.org/10.1016/j.foodchem.2020.127699

Golinelli, L.P., Carvalho, A.C., Casaes, R.S., Lopes, C.S.C., Deliza, R., Paschoalin, V.M.F., & Silva, J.T. (2014). Sensory analysis and species-specific PCR detect bovine milk adulteration of frescal (fresh) goat cheese. *J. Dairy Sc., 97*, 6693-6699. https://doi.org/10.3168/jds.2014-7990

González-Martín, M. I., Palacios, V. V., Revilla, I., Vivar-Quintana, A. M., & Hernández-Hierro, J. M. (2017). Discrimination between cheeses made from cow's, ewe's and goat's milk from unsaturated fatty acids and use of the canonical biplot method. *J. Food Compos. Anal.*, *56*, 34-40. http://dx.doi.org/10.1016/j.jfca.2016.12.005

Guo, L., Ya, M., Hai, X., Guo, Y.S., Li, C.D., Xu, W.L., Liao, C.S., Feng, W. & Cai, Q. (2019). A simultaneous triplex TaqMan real-time PCR approach for authentication of caprine and bovine meat, milk and cheese. *Int. Dairy J., 95*, 58-64. https://doi.org/10.1016/j.idairyj.2019.03.004

Jiménez Carvelo, A.M., Pérez Castaño, E., González Casado, A., & Cuadros Rodríguez, L. (2017). One input-class and two input-class classifications for differentiating olive oil from other edible vegetable oils by use of the normal-phase liquid chromatography fingerprint of the methyl-transesterified fraction. *Food Chem., 221*, 1784-1791. http://dx.doi.org/10.1016/j.foodchem.2016.10.103

Jiménez Carvelo, A.M., Arroyo Cerezo, E., Bikrani, S., Jia, W., Koidis, A., & Cuadros Rodríguez, L. (2022). Rapid and non-destructive spatially offset Raman spectroscopic analysis of packaged margarines and fat-spread products. *Microchem. J., 178,* Article 107378. https://doi.org/10.1016/j.microc.2022.107378

Johnson, M.E. (2017). A 100-Year review: Cheese production and quality. *J. Dairy Sci.*, *100*, 9952-9965. https://doi.org/10.3168/jds.2017-12979

Kennard, R.W., & Stone, L.A. (1969). Computer aided design of experiments. *Technometrics*, *11,* 137-148. https://doi.org/10.1080/00401706.1969.10490666

Ma, Y.B., Babu, K.S., & Amamcharla, J.K. (2019). Prediction of total protein and intact casein in cheddar cheese using a low-cost handheld short-wave near-infrared spectrometer. *LWT, 109*, 319-326. https://doi.org/10.1016/j.lwt.2019.04.039

Medina, S., Perestrelo, R., Silva, P., Pereira, J.A., & Câmara, J.S. (2019). Current trends and recent advances on food authenticity technologies and chemometric approaches. *Trends Food Sci. Tech., 85,* 163-176. https://doi.org/10.1016/j.tifs.2019.01.017

Moros, J., Garrigues, S., & de la Guardia, M. (2007). Evaluation of nutritional parameters in infant formulas and powdered milk by Raman spectroscopy. *Anal. Chim. Acta, 593,* 30-38. https://doi.org/10.1016/j.aca.2007.04.036

Oliveira, K.S., Callegaro, L.S., Stephani, R., Almeida, M.R., & Cappa de Oliviera, L.F. (2016). Analysis of spreadable cheese by Raman spectroscopy and chemometric tools. *Food Chem., 194,* 441-446. http://dx.doi.org/10.1016/j.foodchem.2015.08.039

Ostovar Pour, S., Afshari, R., Landry, J., Pillidge, C., Gill, H., & Blanch, E. (2021). Spatially offset Raman spectroscopy: A convenient and rapid tool to distinguish cheese made with milks from different animal species. *J. Raman Spectrosc., 52*, 1-7. https://doi.org/10.1002/jrs.6179

Pérez Robles, R., Navas, N., Medina Rodríguez, S., & Cuadros Rodríguez, L. (2017). Method for the comparison of complex matrix assisted laser desorption ionization-time of flight mass spectra – Stability of therapeutical monoclonal antibodies. *Chemometr. Intell. Lab. Syst. 170*, 58-67. https://doi.org/10.1016/j.chemolab.2017.09.008

Rodionova, O.Y., Titova, A.V., & Pomerantsev, A.L. (2016). Discriminant analysis is an inappropriate method of authentication. *Trend Anal. Chem., 78*, 17-22. https://doi.org/10.1016/j.trac.2016.01.010

Seçkin, A.K., Yilmaz, B., & Tosun, H. (2017). Real-time PCR is a potential tool to determine the origin of milk used in cheese production. *LWT, 77*, 332-336. https://doi.org/10.1016/j.lwt.2016.11.065

Shao, B., Li, H., Shen, J., & Wu, Y. (2019). Nontargeted detection methods for food safety and integrity. *Ann. Revi. Food Sci. T., 10,* 429-455. https://doi.org/10.1146/annurev-food-032818-121233

Smith, G.P.S., Holroyd, S.E., Reid, D.C.W., & Gordon, K.C. (2016). Raman imaging processed cheese and its components. *J. Raman Spectrosc.*, *48*, 374-383. https://doi.org/10.1002/jrs.5054

Soto-Barajas, M.C., González-Martín, M.I., Salvador-Esteban, J., Hernández-Hierro, J.M., Moreno-Rodilla, V., Vivar-Quintana, A.M., Revilla, I., Lobos Ortega I., Morón-Sancho, R. & Curto-Diego, B. (2013). Prediction of the type of milk and degree of ripening in cheeses by means of artificial neural networks with data concerning fatty acids and near infrared spectroscopy. *Talanta, 116*, 50-55. https://doi.org/10.1016/j.talanta.2013.04.043

Valverde Som, L., Ruiz Samblás, C., Rodríguez García, F.P., & Cuadros Rodríguez, L. (2018). Multivariate approaches for stability control of the olive oil reference materials for sensory analysis – part I: framework and fundamentals. *J. Sci. Food Agri., 98,* 4237-4244. https://doi.org/10.1002/jsfa.8948

Vigni, M.L., Durante, C., Michelini, S., Nocetti, M., & Cocchi, M. (2020). Preliminary assessment of Parmigiano Reggiano authenticity by handheld Raman spectroscopy. *Foods*, *9*, Article 1563. http://dx.doi.org/10.3390/foods9111563

Yazgan, N.N., Genis, H.E., Bulat, T., Topcu, A., Durna, S., Yetisemiyen, A., & Boyaci, I.H. (2020). Discrimination of milk species using Raman spectroscopy coupled with partial least squares discriminant analysis in raw and pasteurized milk. *J. Sci. Food Agric.*, *100*, 4756-4765. https://doi.org/10.1002/jsfa.10534

Zhang, Z.Y. (2020). Rapid discrimination of cheese products based on probabilistic neural network and Raman spectroscopy. *J. Spectrosc.*, *2020*, Article 896535. https://doi.org/10.1155/2020/8896535

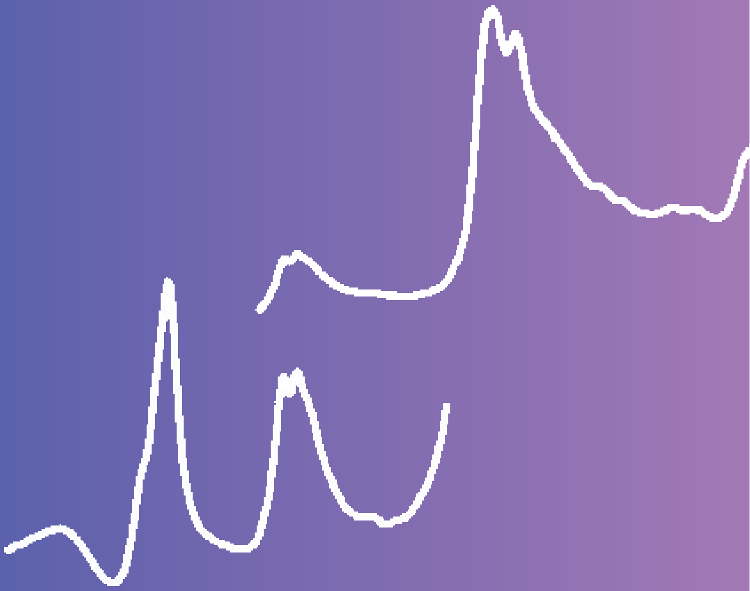
Zheng, X., Shi, X. & Wang, B. (2021). A Review on the general cheese processing technology, flavor biochemical pathways and the influence of yeasts in cheese. *Front. Microbiol.*, *12*, Article 703284. https://doi.org/10.3389/fmicb.2021.703284

2.4. Contribuciones a congresos

1. A. Arroyo Cerezo, A.M. Jiménez Carvelo, L. Cuadros Rodríguez. **Detección de aceites vegetales en muestras de materias grasas para untar.** [Póster]. *EXPOLIVA XX Simposio Científico-Técnico. Jaén (España), septiembre 2021.*
2. A. Arroyo Cerezo, A.M. Jiménez Carvelo, L. Cuadros Rodríguez. **Authentication of margarines and related fat-spread products by spatially offset Raman spectroscopy.** [Póster]. *First SensorFINT International Workshop. Oporto (Portugal), septiembre 2021.*
3. A. Arroyo Cerezo, A.M. Jiménez Carvelo, A. González Casado, L. Cuadros Rodríguez. **Raman through original food packaging – Authentication of the milk origin by animal species of sliced cheeses.** [Oral 15']. *First SensorFINT International Conference. Izola (Eslovenia), mayo 2022.*
4. A. Arroyo Cerezo, A.M. Jiménez Carvelo, A. González Casado, L. Cuadros Rodríguez. **Espectroscopia Raman con compensación espacial para el control no invasivo de productos alimenticios**. [Póster]. *XXIII Reunión de la Sociedad Española de Química Analítica (SEQA 2022). Oviedo (España), julio 2022.*
5. A. Arroyo Cerezo, A.M. Jiménez Carvelo, L. Cuadros Rodríguez. **Non-invasive Raman and machine learning algorithms for qualitative and quantitative authentication of margarines**. [Póster]. *XI Simposio de Investigación en Ciencias Experimentales*. *Almería (España), noviembre 2022.*

CAPÍTULO 3

Evaluación rápida de la calidad de aceites de oliva



3.1. Presentación

El aceite de oliva es uno de los alimentos de origen vegetal de mayor valor en el mercado, y especialmente en su categoría de máxima calidad: aceite de oliva virgen extra. Es por ello que asegurar la calidad y autenticidad de este producto alimenticio se vuelve aún más crucial. La normativa actualmente vigente establece los límites de diversos criterios de calidad y pureza que deben satisfacer las diferentes categorías de aceite de oliva para ser comercializados. Ello supone la determinación de hasta más de 20 parámetros que debe realizarse siguiendo los métodos de análisis propuestos por el Consejo Oleícola Internacional (COI) y reconocidos por la Unión Europea en los controles oficiales de conformidad de este producto, suponiendo el uso de diversas técnicas analíticas: cromatografía de gases que implica métodos distintos para cada analito, espectrofotometría UV, y otras determinaciones llevadas a cabo mediante métodos clásicos no instrumentales [**1**,**2**]. Además, es habitual, aunque no obligatorio, que en el control rutinario de la calidad del aceite de oliva se apliquen dichos métodos antes de que el producto sea comercializado, con el objeto de asegurar el cumplimiento de la legislación. Los citados métodos requieren por lo general periodos de tiempo de análisis relativamente largos, etapas de tratamiento de muestra con el consecuente uso de productos químicos y generación de residuos perjudiciales para el medio ambiente, habilidades por parte del analista, etc. En definitiva, quedan lejos de seguir los principios de la química analítica verde (GAC) [**3**].

Todo ello pone de manifiesto la necesidad del desarrollo de métodos analíticos capaces de sustituir, o al menos apoyar los métodos existentes, que aborden todas las mencionadas desventajas, y que vayan de la mano de los avances científicos de los últimos años.

Siguiendo esta línea, se presenta en este capítulo una publicación científica que recoge la propuesta de una metodología para la evaluación *ex-ante* de nuevos métodos analíticos no destructivos, junto a un ejemplo describiendo la comparación de tres métodos desarrollados para el análisis de la calidad de

1. Commission Delegated Regulation (EU) 2022/2104 of 29 July 2022 supplementing Regulation (EU) No 1308/2013 of the European Parliament and of the Council as regards marketing standards for olive oil, and repealing Commission Regulation (EEC) No 2568/91 and Commission Implementing Regulation (EU) No 29/2012. Official Journal of the European Union L 284/1, 2022.

2. Commission Implementing Regulation (EU) 2022/2105 of 29 July 2022 laying down rules on conformity checks of marketing standards for olive oils and methods of analysis of the characteristics of olive oil. Official Journal of the European Union L 284/23, 2022.

3. de la Guardia, M.; Garrigues, S. The concept of Green Analytical Chemistry. In *Handbook of Green Analytical Chemistry.*; de la Guardia, M.; Garrigues, S., Eds.; John Wiley & Sons: Chichester, U.K., 2012, pp 3-16.

aceite de oliva. De esta forma, se pretende destacar la importancia de evaluar un método analítico desarrollado en un entorno científico antes de ser transferido e implementado en un laboratorio de control.

La referencia de este manuscrito publicado en forma de artículo científico en una revista de alto impacto es la siguiente:

1. Evaluating the whiteness of spectroscopy-based non-destructive analytical methods – Application to food analytical control. *TrAC-Trend Anal. Chem.* **2024**, *170*, 117463. DOI: 10.1016/j.trac.2023.117463.

En consonancia con ello, se recogen a continuación dos estudios de investigación encaminados a desarrollar métodos analíticos de cribado rápidos y multiparamétricos para la evaluación de la calidad de aceites de oliva, con el objetivo principal de transferir un método capaz de agilizar el trabajo en los laboratorios de control.

El primero de los estudios se basa en explorar la espectrometría de infrarrojo cercano (NIR) mediante el uso de equipos miniaturizados para la adquisición de datos espectrales de aceites vegetales, siguiendo una estrategia no destructiva y con la ausencia total de productos químicos. Además, dicho estudio se ejecuta como parte fundamental del trabajo llevado a cabo durante la estancia predoctoral realizada en la Università degli Studi di Padova (Italia) entre los meses de abril y junio de 2022, enmarcada dentro de la COST Action 19145 (SensorFINT) que posibilitó su realización mediante la concesión de una ayuda para la movilidad de investigadores (STSM, *Short-Term Scientific Mission*).

La espectrometría NIR es una espectrometría óptica basada en el estudio de la interacción entre la luz en la región del infrarrojo cercano (780 – 2500 nm) y la materia irradiada. Concretamente, estudia las transiciones vibracionales a nivel molecular que tienen lugar al hacer incidir una radiación electromagnética (REM) sobre un material a una determinada longitud de onda [**4**] incluida en el intervalo antes definido. El descubrimiento de esta radiación data del 1800 por parte de W. Herschel, aunque no fue hasta mediados del siglo XX cuando aparecieron las primeras aplicaciones industriales. No obstante, se le atribuye un gran mérito por destacar el potencial de esta técnica a principios de los años 50 a investigadores como K.H. Norris [**5**].

[4] Pavia, D.L.; Lampman, G.M.; Kriz, G.S.; Vyvyan, J.R. Infrared Spectroscopy. In *Introduction to Spectroscopy*, 4th ed.; Brooks/Cole: Cham, U.S., 2009, pp 15-104.

[5] Siesler, H.W. Basic Principles of Near-Infrared Spectroscopy. In *Handbook of Near-Infrared Analysis*, 3rd ed.; Burns, D.A.; Ciurczak, E.W., Eds.; CRC Press: Florida, U.S., 2008, pp 7-19.

Desde entonces, NIR se ha convertido en una de las técnicas analíticas más estudiadas en el ámbito del análisis de alimentos [**6**,**7**]. A nivel instrumental, es probablemente una de las espectrometrías que menor coste supone. Además, por sus características cumple con prácticamente todos los requisitos de la GAC, siendo una técnica rápida y no destructiva, que requiere una mínima o nula preparación de muestra, ya que incluso se pueden realizar medidas directamente sobre el material a analizar. A su vez, permite llevar a cabo análisis multiparamétricos, ya que el espectro recogido puede ser considerado una huella instrumental característica del material bajo estudio [**8**].

A su vez, cabe destacar la tendencia de los últimos años por del desarrollo de equipos NIR miniaturizados, que abren la posibilidad a llevar a cabo medidas *in-situ*, reduciendo aún más el coste económico y el consumo de energía. Es por ello que cada vez son más los estudios publicados en los que se investigan las amplias posibilidades del uso de estos equipos portátiles para el control de la calidad y autenticidad de alimentos [**9**,**10**,**11**].

Este primer estudio se recoge en forma de publicación científica publicada en una revista de alto impacto en colaboración internacional, y cuya referencia es:

2. Assessment of extra virgin olive oil quality by miniaturized Near Infrared instruments in a rapid and non-destructive procedure. *Food Chem.* **2024**, *430*, 137043. DOI: 10.1016/j.foodchem.2023.137043.

Por otro lado, el segundo estudio presentado forma parte del proyecto de Colaboración Público-Privada (CPP2021-008672), en el marco del Plan Estatal de Investigación Científica y Técnica y de Innovación, titulado "Implantación de la resonancia magnética nuclear de baja frecuencia de campo (LF-NMR) en

6. 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.

7. Moghaddam, H.N.; Tamiji, Z.; Lakeh, M.A.; Khoshayand, M.R.; Mahmoodi, M.H. Multivariate analysis of food fraud: A review of NIR based instruments in tandem with chemometrics. *J. Food Compos. Anal.* **2022**, *107*, 104343. DOI: 10.1016/j.jfca.2021.104343.

8. Yang, X.; Berzaghi, P. Near-Infrared spectroscopy. In *Food Integrity by Non-invasive / Non-destructive Methods;* Jiménez-Carvelo, A.M.; Arroyo-Cerezo, A.; Cuadros-Rodríguez, L., Eds.; Springer Nature: Berlin, Germany, 2024 [*submitted, in process*].

9. Müller-Maatsch, J.; et al. The spectral treasure house of miniaturized instruments for food safety, quality and authenticity applications: A perspective. *Trends Food Sci. Technol.* **2021**, *110,* 841-848. DOI: 10.1016/j.tifs.2021.01.091.

10. McVey, C.; et al. Portable spectroscopy for high throughput food authenticity screening: Advancements in technology and integration into digital traceability systems. *rends Food Sci. Technol.* **2021**, *118,* 777-790. DOI: 10.1016/j.tifs.2021.11.003.

11. Squeo, G.; et al. Considerations about the gap between research in Near Infrared spectroscopy and official methods and recommendations of analysis in foods. *Curr. Opin. Food Sci.* **2024**, 101203. DOI: 10.1016/j.cofs.2024.101203 laboratorios de control para estudios cuantitativos y de clasificación de productos alimenticios y de otros sectores industriales", cuyo objetivo principal es el de desarrollar métodos analíticos basados en la espectrometría de resonancia magnética nuclear de baja frecuencia de campo (LF-NMR del inglés *low-field nuclear magnetic resonance*). En dicho estudio se presentan los resultados preliminares de un trabajo que a día de hoy sigue en desarrollo.

La espectrometría NMR mide la interacción entre un pulso de una onda de radiofrecuencia y los protones de los núcleos que componen el material a analizar que se somete a un campo magnético estático, dando lugar a un espectro. Según la frecuencia de campo empleada se puede diferenciar entre:

- NMR de alta frecuencia de campo y alta resolución (> 250 MHz)
- NMR de baja frecuencia de campo y alta resolución (20 – 100 MHz)
- NMR de baja frecuencia de campo y baja resolución (2 – 20 MHz)

Entre ellas, la más utilizada es la primera cuya abreviatura es HF-NMR (del inglés *high-field NMR*). Esta técnica proporciona ventajas notables en términos de calidad y precisión de resultados en comparación con otras, ya que es capaz de proporcionar cuantificaciones de analitos que se encuentran incluso en concentraciones a nivel de traza, es decir, menores al 1% [**12**]. Sin embargo, a pesar de estas ventajas, el uso de HF-NMR en el análisis de alimentos es limitado. Esto se atribuye principalmente a ciertas desventajas que supone en términos económicos, requerimientos de espacio e infraestructura, y el mantenimiento instrumental necesario. Todo ello dificulta su implementación en laboratorios de control de alimentos rutinario.

Estas limitaciones pueden ser abordadas con el advenimiento de equipos LF-NMR de sobremesa operando en el rango 20 - 100 MHz, en los que se prima la mejora de la sensibilidad y estabilidad para proporcionar una suficientemente alta resolución con respecto a los predecesores que trabajan en el rango 2 - 20 MHz. Además, estos equipos permiten realizar medidas sin la necesidad de utilizar disolventes deuterados o incluso en ausencia total de disolvente, haciendo por tanto la LF-NMR más respetuosa con el medio ambiente, en consonancia con los principios GAC [**13**].

Los núcleos de interés para el análisis de alimentos son fundamentalmente hidrógeno (1H) y carbono (13C), y en menor medida flúor (19F), fosforo (31P) y silicio (29Si). Los espectros proporcionados pueden ser tratados como huella

12. Kamal, G.M.; et al. Nuclear Magnetic Resonance Spectroscopy in Food Analysis. In *Techniques to Measure Food Safety and Quality: Microbial, Chemical, and Sensory*, Khan, M.S.; Rahman, M.S., Eds.; Springer: Cham, Switzerland, 2021, pp. 137-168.

13. Yu, H.Y.; Myoung, S.; Ahn, S. Recent applications of benchtop nuclear magnetic resonance spectroscopy. *Magnetochemistry* **2021**, *7*, 121. DOI: 10.3390/magnetochemistry7090121.

instrumental para realizar análisis no dirigidos y desarrollar métodos multiparamétricos. En el ámbito alimentario son muchos los estudios que exploran el potencial de NMR para su uso en la calidad y autenticidad de aceites vegetales con resultados prometedores [**14,15**]. La llegada al mercado de un equipo de sobremesa que trabaja en el máximo del rango LF-NMR (100 MHz) y la evidencia científica que respalda el potencial del uso de NMR, catalizaron el comienzo del proyecto anteriormente mencionado (CPP2021-008672) para la implementación de LF-NMR en laboratorios de control. Y con ello, el estudio que más adelante se presenta basado en el uso del citado equipo de sobremesa para el control de calidad de aceite de oliva.

14. Maestrello, V.; Solovyev, P.; Bontempo, L.; Mannina, L.; Camin, F. Nuclear magnetic resonance spectroscopy in extra virgin olive oil authentication. *Compr. Rev. Food Sci. Food Saf.* **2022**, *21*, 4056-4075. DOI: 10.1111/1541-4337.13005.

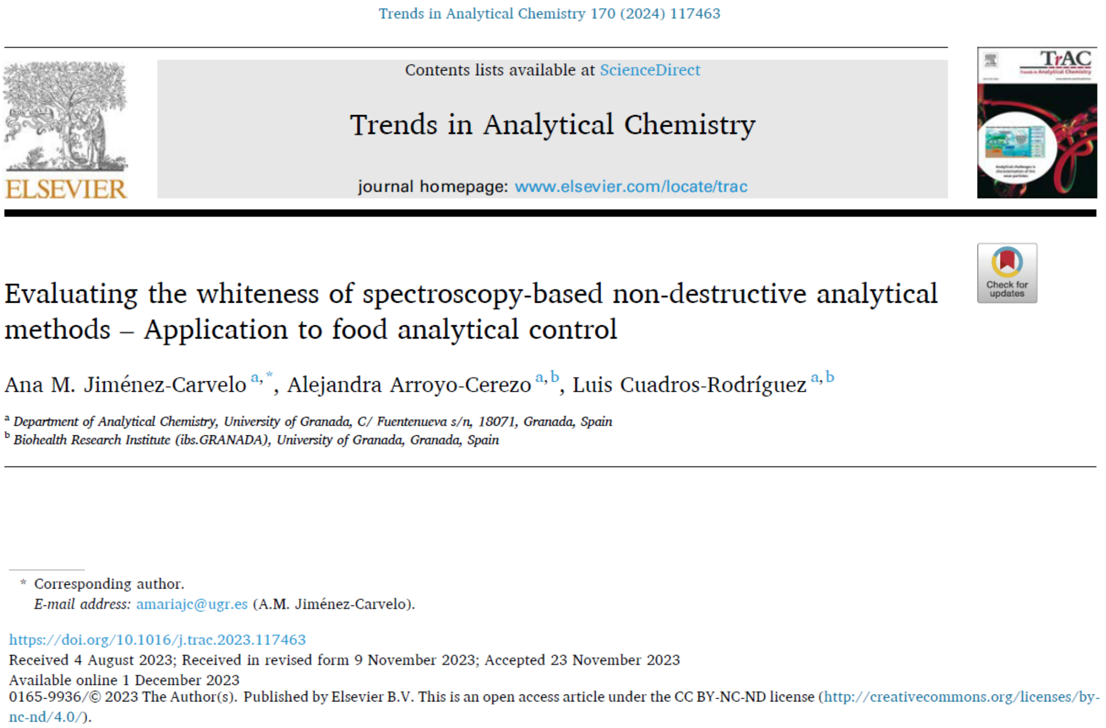
15. Calò, F.; Girelli, C.R.; Wang, S.C.; Fanizzi, F.P. Geographical origin assessment of extra virgin olive oil via NMR and MS combined with chemometrics as analytical approaches. *Foods* **2022**, *11*, 113. DOI: 10.3390/ foods11010113.

3..2. Artículo científico 3

Evaluating the whiteness of spectroscopy-based non-destructive analytical methods – Application to food analytical control.

Publicado en 2024 en la revista Trends in Analytical Chemistry 170, 117463.

DOI: 10.1016/j.trac.2023.117463.



Trends in Analytical Chemistry 170 (2024) 117463

Contents lists available at ScienceDirect

Trends in Analytical Chemistry

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

Evaluating the whiteness of spectroscopy-based non-destructive analytical methods – Application to food analytical control

Ana M. Jiménez-Carvelo [a,*], Alejandra Arroyo-Cerezo [a,b], Luis Cuadros-Rodríguez [a,b]

[a] *Department of Analytical Chemistry, University of Granada, C/ Fuentenueva s/n, 18071, Granada, Spain*
[b] *Biohealth Research Institute (ibs.GRANADA), University of Granada, Granada, Spain*

[*] Corresponding author.
E-mail address: amariajc@ugr.es (A.M. Jiménez-Carvelo).

https://doi.org/10.1016/j.trac.2023.117463
Received 4 August 2023; Received in revised form 9 November 2023; Accepted 23 November 2023
Available online 1 December 2023

Highlights:

- *Ex-ante* evaluation of non-destructive analytical methods is presented.
- A novel modified RGB methodology is proposed to evaluate new methods.
- Evaluable sub-items for each colour parameter are included.
- Whiteness assessment of three non-destructive methods for food analytical control.

Keywords:

Non-destructive method

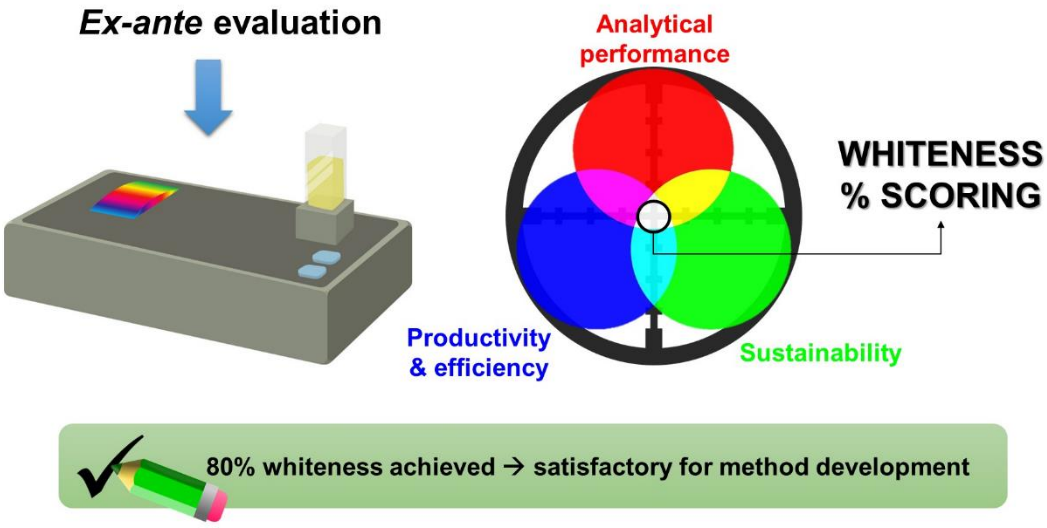
Whiteness method

Ex-ante evaluation

Green analytical chemistry

RGB methodology

Graphical abstract


Ex-ante evaluation
Analytical performance
WHITENESS % SCORING
Productivity & efficiency
Sustainability
80% whiteness achieved → satisfactory for method development

Abstract

Recent advancements in analytical chemistry in the food quality field have emphasized ecofriendly analytical techniques eschewing chemicals and solvents. Various methodologies exist for assessing the sustainability of analytical methods, however none has provided guidance for appraising non-destructive methods, especially pre-development. Among these, the RGB approach stands out, evaluating method colour via three main criteria: analytical performance, environmental impact, and practical efficiency. This framework offers a comprehensive evaluation, aiming for a "white" colour denoting excellence across all three categories.

This article introduces an adapted RGB method for *ex-ante* evaluation of new non-destructive analytical methods pre-development. It outlines key steps for evaluating method "whiteness". As a guiding example, the approach was applied to three analytical methods focussed on quality and authenticity control of edible vegetable oils utilizing solvent-free spectroscopic techniques. Results underscored a priori feasibility assessment value, aligning evaluative objectives with intended method goals.

1. Introduction

Ensuring food quality and safety is crucial along the entire food chain: from raw materials through all production stages to the final product reaching consumers. Food fraud is a scourge as old as food production itself. And just as food technology is advancing by leaps and bounds, the ways of committing food fraud are becoming increasingly sophisticated. The most common acts of food fraud encompass counterfeiting, mislabelling, misrepresentation, dilution, unauthorized enhancement, or adulteration; this information can be expanded by referring to literature, as the recent review by Bannor *et al*. [**1**]. However, the analytical methods used to ensure proper food quality control and authentication are not keeping pace with this growth. Instrumental techniques such as liquid or gas chromatography coupled to mass spectrometry are nowadays commonly used for these types of analysis. Those analytical techniques undoubtedly provide accurate results. Nevertheless, there are some disadvantages associated with these techniques. These include time-consuming analysis, several sample preparation steps (separation and/or dissolution processes), allowing off-line analysis only (some instances also on-line analysis), the need to use chemical solvents, which can be environmentally hazardous and pose risks to analysts, and the requirement for expertise due to their complexity [**2**].

Against this background, there is undoubtedly an urgent demand for analytical methods involving the use of alternative techniques to the widely employed ones mentioned above. The most important requirement is to enable rapid, non-destructive analysis and in-situ if possible. In such a way that it should be non-destructive, including the total or partial absence of chemical solvents and minimising waste as much as possible, thus fostering a more sustainable analytical chemistry, also known as green chemistry. Current research in the field of food analysis does follow this trend from several years ago [**3**]. Many studies can be found in the scientific literature exploring the potential of different alternatives to conventional analytical techniques, providing the possibility of rapid and non-destructive analysis [**4**]. Moreover, some advanced techniques even enable non-invasive analysis [**5**]. The difference between a non-destructive technique and a non-invasive one is worth noting. In the former, the sample remains in the same physical and chemical state after measurement. It requires only a minimal amount of sample, resulting in hardly any wasted product and allowing for potential future re-measurement. Whereas with a non-invasive technique, the measurement is performed directly on the food product, i.e., no sampling is necessary. The tested foodstuff remains unaltered, as the non-destructive, but also remains fully intact, e.g., the packaging is not opened to measure the food in the case of final product analysis, and it could even be consumed. So, any non-invasive technique is in turn a non-destructive technique, but not vice versa. A general example of what a destructive (conventional), non-destructive and non-invasive analytical technique would consist of is illustrated in Figure 3.1, taking as an example the analytical measurement of a vegetable oil at the final stage of production, i.e., already packaged.

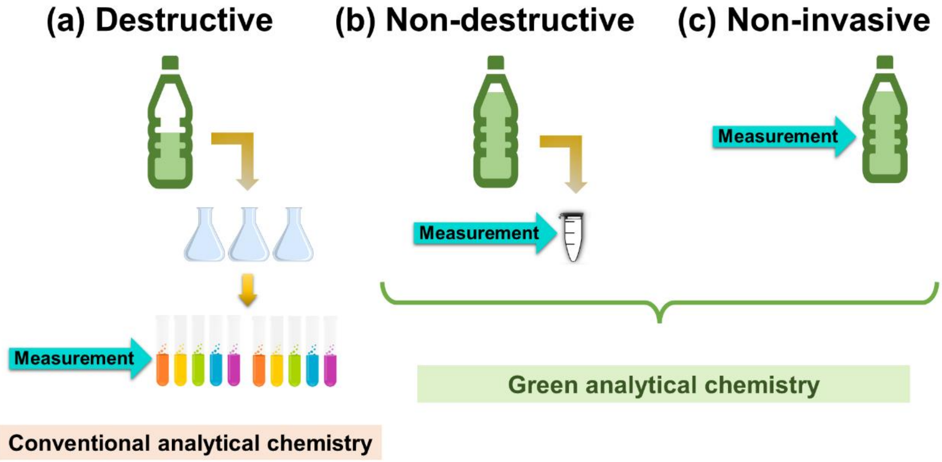


Figure 3.1. Illustrative example of an analytical measurement of a packaged vegetable oil by using (a) destructive, (b) non-destructive and (c) non-invasive analytical technique.

Generally, the most commonly used techniques that meet the aforementioned requirements include spectroscopic techniques, sensing techniques (such as e-nose or e-tongue) and techniques based on image analysis (e.g., computer vision, multi- or hyper-spectral imaging) or nanotechnology [**2**,**6**]. Among them, the spectroscopic techniques such as infrared [**7**], Raman [**8**,**9**] or nuclear magnetic resonance (NMR) [**10**,**11**] spectroscopies are the most investigated. All these techniques yield substantial chemical information by generating corresponding spectra, enabling rapid acquisition of results. The utilization of solvents is dispensable, except in the case of NMR, although low-field benchtop equipment is now accessible to facilitate solvent-free measurements [**12**].

In this regard, these analytical techniques would adhere to the principles of sustainability and green chemistry mentioned earlier. However, guaranteeing this alignment requires the availability of appropriate tools for their assessment. For this purpose, some developed indexes can be found in scientific literature. These include the Life Cycle Assessment (LCA) or the National Environmental Methods Index (NEMI), both have been applied to assess analytical chemistry methods, even though they are not exclusive to the field [**13**,**14**]. Other tools have been developed specifically for analytical chemistry, such as the analytical Eco-scale, which is intended to rank any analytical procedure as: excellent, acceptable, or inadequate from the green analysis standpoint [**15**]. Also, the Green Analytical Procedure Index (GAPI) and the ComplexGAPI evaluate the green profile of an analytical method based on several criteria, involving the sample preparation and analysis procedures [**16**]. Another recent index is the Analytical GREEnness Metric Approach (AGREE), which responds to the 12 principles of the Green Analytical Chemistry (GAC) [**17**,**18**]. Nevertheless, all the mentioned metrics only assess the eco-friendly aspect of the methods, based on the 12 GAC principles, overlooking other important aspects such as cost, time, or analytical performance.

Consequently, new methodologies for the comprehensive assessment of analytical methods have recently been proposed. For instance, HEXAGON algorithm, an extension of the Eco-scale which not only assess the green profile, but also the characteristics of the analytical method such as the technical reliability, and the economic cost [**19**]. Despite, some limitations are presented such as not considering key sampling aspects, thus it has not been widely employed by the scientific community [**16**]. A group of algorithms named Multi-Criteria Decision Analysis (MCDA) [**20**] has also been applied for this purpose [**21**], although it primarily emphasizes decision-making rather than the evaluation of the analytical method itself. Lastly, Nowak *et al.* [**22**] developed the novel RGB algorithm based on the three primary colours, each embodying a set of criteria to be evaluated: analytical performance, safety and environmental friendliness, and practical efficiency and productivity. This methodology provides the capacity to assess an analytical method both qualitatively and quantitatively. The authors subsequently re-versioned this algorithm to the new RGB 12, including the 12 GAC principles by grouping them into 4, alongside other aspects, leading to the 12 principles of White Analytical Chemistry (WAC) [**23**].

Within this framework and to the best of our knowledge, RGB algorithm has not been applied to assess an analytical method based on a non-destructive technique for food quality control purposes. Besides, all the above-mentioned approaches have been founded on an *ex-post* evaluation, i.e., assessment conducted after the method development. However, it would be advantageous to have the possibility to perform an *ex-ante* evaluation of the method, enabling the decision-making process regarding its development before initiation. Therefore, the aim of this paper is to establish the key points on which to base the *ex-ante* evaluation of a new analytical method based on non-destructive/non-invasive techniques for food control, although could also be useful for non-food analytical chemistry fields. For this purpose, a proposal based on the RGB methodology is presented together with a practical example, as it was considered the most suitable of the indexes listed above for evaluating this type of analytical method, due to its flexibility and the ability to evaluate not only the ecology of the method, but also other critical and important points.

2. Analytical method whiteness

The RGB algorithm is based on the three primary colours: red, green, and blue. Each of one represents a set of evaluable items categorised into three groups: RED to assess the analytical performance; GREEN to evaluate the sustainability (greenness) of the method; and BLUE for the productivity and practical efficiency assessment. The mixture of these three primary colours at 100% scoring results in the colour WHITE. This is where the concept of the "white analytical method" arises, representing the optimal solution. Consequently, the RGB algorithm appraises the method whiteness, a concept initially introduced by the authors responsible for devising the RGB 12 algorithm [**23**].

Note that, the assessment of the analytical method may result in different colours depending on the scores obtained in each of the three RGB colour parameters, spanning from black (0% scoring) to white (100% scoring). To quantitatively assess the colour score, the authors devised an index termed Method Brilliance (MB) in the RGB algorithm and whiteness in the RGB 12 algorithm, both of which are estimated differently. More details about RGB and RGB 12 algorithms can be found in the source proposal, in references [**22**] and [**23**] respectively.

Given the aforementioned considerations, the RGB algorithm was deemed the most appropriate methodology for assessing new analytical method based on non-destructive/non-invasive analytical techniques for food analysis. On one hand, this methodology offers the advantage of assessing not only the method greenness but also encompasses other crucial issues, including its practicality concerning time, cost, and performance, as well as result reliability. On the other hand, it provides flexibility in assigning weights to each evaluable item according to the pursued objective, in contrast to the rigidity of the RGB 12 algorithm.

In this context, it is crucial to emphasize a significant aspect: thus far, the RGB algorithm has been employed *ex-post*, meaning it has been applied only after the analytical method development. However, a more reasonable approach would involve assessing the method prior its development. In other words, defining acceptable and satisfactory limits for each evaluable item, and assigning the expected scores based on the intended procedure. Once the outcome is determined, the analyst should decide whether the candidate analytical method is valid to achieve the proposed goal and develop it or whether it is preferable to seek a better alternative.

Considering this matter, this paper proposes to use the RGB algorithm in a novel way, aimed at defining the minimum whiteness scoring that the candidate analytical method should have before developing it. In addition, intermediate values for acceptable and satisfactory results of the main items evaluated for each colour have been set at 50% and 80% respectively, instead of the 33.3% and 66.6% used so far. The proposal is also based on establishing several secondary items (or sub-items) within main item comprising each of the three colour parameters. It should be borne in mind that *ex-ante* evaluation of a non-destructive/non-invasive analytical method requires careful selection of the evaluable items. In the following section the main evaluable items and sub-items proposed here will be presented and discussed.

3. Practical applications

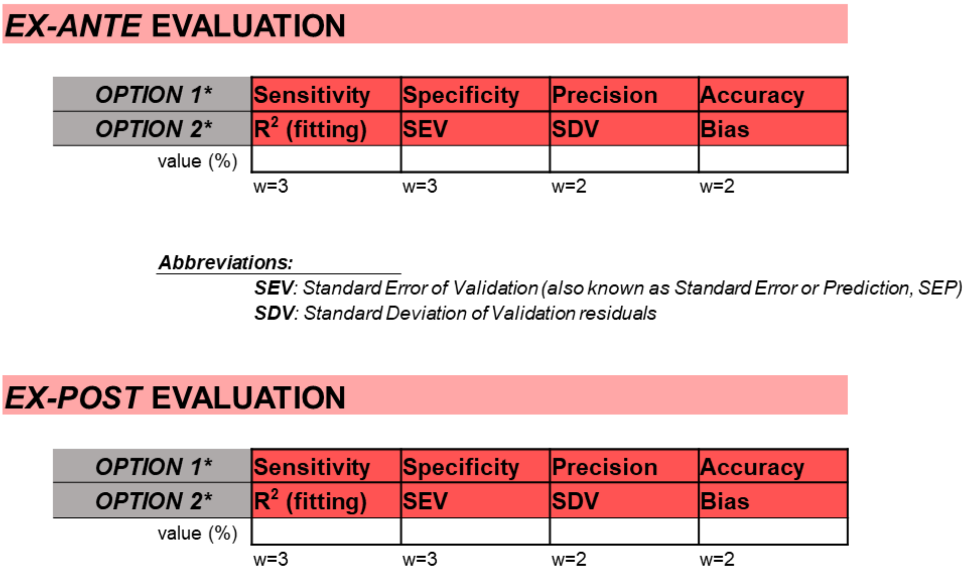
As a practical instance, three analytical methods based on non-destructive/non-invasive techniques for analysing food quality and authenticity have been selected for *ex-ante* evaluation using a proposed RGB-modified algorithm. The main aim of these methods is to serve as a screening method based on a multivariate classification for identification purposes. The difference among the three lay in the analytical technique involved on each method: near infrared spectroscopy (NIR), spatially offset Raman spectroscopy (SORS) and low-field nuclear magnetic resonance (LF-NMR). NIR was selected for its widespread use and versatility in the field of food analysis [**24**]. SORS is an advanced mode of conventional Raman spectroscopy that allows through-container measurements, making it a non-invasive technique [**25**]. And LF-NMR was selected for its increasing interest in recent times, which in the low frequency field mode allows for non-destructive solvent-free measurements [**26**].

All three analytical methods pursued the same aim: the quality and/or authentication control of virgin olive oils. Therefore, the target material to be analysed was vegetable oils at the final stage of the production process, i.e., already marketed. The reader should take this as a practical example of how the algorithm should be applied using the approach described here, to serve as an implementation guideline but it could be applied to any other foodstuff. Moreover, it should be emphasized that the three spectroscopic techniques provide a large amount of chemical information, especially in food products which are chemically very complex materials. Thus, there was no doubt that data treatment should be addressed from a non-targeted approach by using data mining chemometric tools [**27**,**28**,**29**], also in line with green chemistry [**30**].

In order to perform an *ex-ante* comprehensive evaluation of an analytical method being developed, an estimate of the scores for each of the three colour parameters, ranging from 0 to 100, is estimated. However, redness cannot be properly assessed, as depends on the final results. To overcome this drawback, two alternatives can be considered. The first, and easier to apply, considers the value of 80% as the starting point for all evaluable sub-items. Moreover, according to the published scientific literature, the three spectroscopic techniques coupled with chemometrics obtain high performance metrics in the analysis of vegetable oils [**31**,**32**,**33**]. On the other hand, from a more analytical standpoint, the critical values of the performance metrics defining the fitness for purpose of the analytical method under study, which had to be previously established in order to be applied for validation, could now be considered.

Considering a method applying machine learning/chemometrics can be oriented toward developing either a classification or quantitation model, each evaluated by different analytical metrics, two options are proposed to evaluate the redness, shown in Figure 3.2. Option one is designed for classification/discrimination models and four main evaluable items are proposed, where each score represents the 0-1 scaled metric value expressed as a percentage. The metrics are sensitivity, specificity, precision (positive predictive value) and accuracy or efficiency, although more metrics could be added [**34**]. While option two intends to evaluate quantitation models, by using four alternative metrics as main evaluable items: determination coefficient of the model (R^2), standard error validation (SEV), standard deviation of validation residuals (SDV) and relative bias. These quantitation metrics agree with the recommendations outlined in the ASTM Standard Practice [**35**].

Depending on the scenario, the critical values of the performance metrics to be considered may be different. For instance, in a screening method based on a multivariate classification regarding an official conformity evaluation scenario, values of 0.95, 0.50, 0.95 and 0.90 could be set for sensitivity, specificity, precision and accuracy, respectively [**36**]. Likewise, regarding a method of quantifying an adulterant (e.g., a cheaper vegetable oil) in extra virgin olive oil, values of 0.95, 0.90, 0.90 and 0.95 could be set for R^2, SEV, SDV and bias, respectively [**37**,**38**]. Practitioners are encouraged to review recent literature to ascertain the performance of the candidate analytical technique on the particular sample being analysed in order to define consistent critical values for *ex-ante* evaluation.



***EX-ANTE* EVALUATION**

*OPTION 1**	Sensitivity	Specificity	Precision	Accuracy
*OPTION 2**	R^2 (fitting)	SEV	SDV	Bias
value (%)				
	w=3	w=3	w=2	w=2

Abbreviations:

***SEV**: Standard Error of Validation (also known as Standard Error or Prediction, SEP)*
***SDV**: Standard Deviation of Validation residuals*

***EX-POST* EVALUATION**

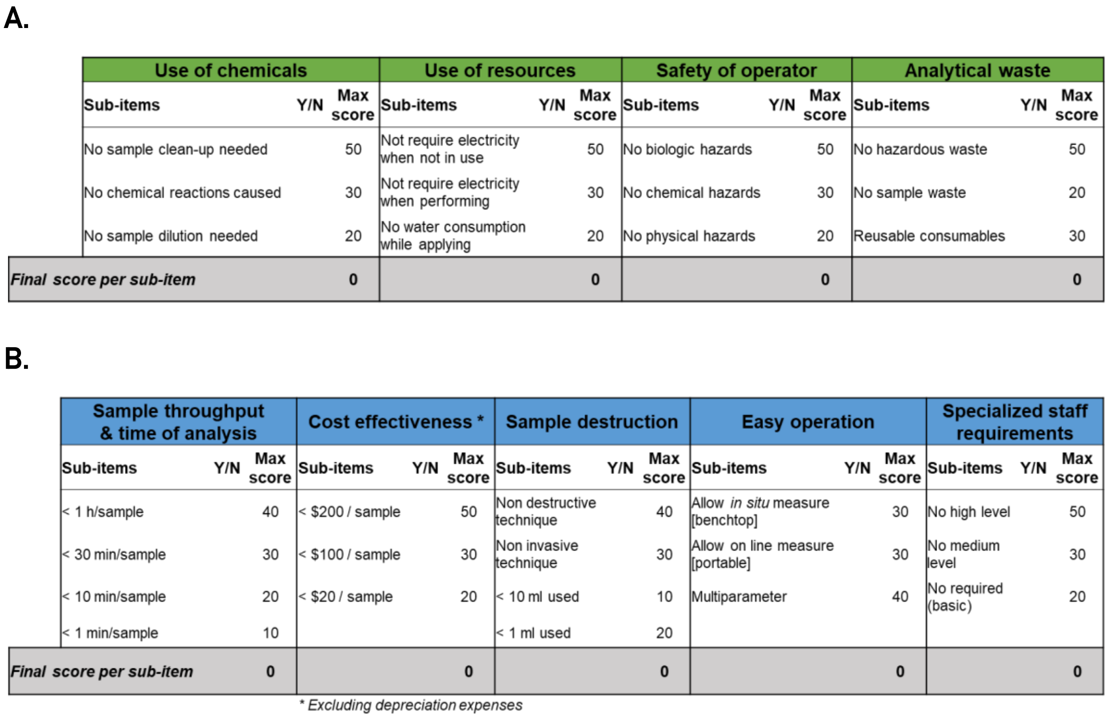
*OPTION 1**	Sensitivity	Specificity	Precision	Accuracy
*OPTION 2**	R^2 (fitting)	SEV	SDV	Bias
value (%)				
	w=3	w=3	w=2	w=2

Figure 3.2. Proposed items for redness rating of an analytical method according to its main purpose: classification or quantitation.

**Note that OPTION 1 addresses the metrics to be considered in classification/discrimination methods for identification or detection purposes and OPTION 2 in quantitation methods for prediction purposes.*

Further on, specific evaluable items for assessing greenness and blueness are proposed. Within each of them, several sub-items that should be considered have been included. Upon meeting these sub-items, a score up to 100 is assigned for each item, contributing to the overall evaluation of each colour parameter. Note that the proposal is based on a simple "yes" or "no" (Y/N column) response for each evaluable sub-item, as shown in Figure 3.3, thus rendering the tool user-friendly. The template for colour evaluation to be filled in, along with brief instructions can be found in the supplementary material (note that the sub-items of cost-effectiveness and sample destruction are adapted to the objective stated here, and may be modified at the analyst's request).

The selected evaluable items for greenness do not differ widely from their predecessor RGB methods [22,23], while the considered sub-items have been carefully chosen according to what the authors find most critical in requiring an analytical method to achieve 100% greenness scoring. As for blueness, it was deemed appropriate to include a main item considering the analyst's specialization requirements. Besides, it is worth noting that the sub-items within the cost-effectiveness shown in Figure 3.3B have been adapted to the specific case of the full analysis of a virgin olive oil. It is understood that a method pursuing this goal will be cost-effective when is at least cheaper than the total cost ($200/sample). Observe that, this information was provided from an official laboratory dedicated to the analysis of olive oil. Our recommendation to the reader is either seek a reliable source knowing this information (cost of the official or recognized method/s normally used for the food or sample under study) or estimate the cost as close to reality as possible. The datum will be well-known in the case of an industry aiming to evaluate a new method to perform an analysis already carried out by another conventional analytical method that is destined to be substituted.



A.

Use of chemicals			Use of resources			Safety of operator			Analytical waste		
Sub-items	Y/N	Max score	Sub-items	Y/N	Max score	Sub-items	Y/N	Max score	Sub-items	Y/N	Max score
No sample clean-up needed		50	Not require electricity when not in use		50	No biologic hazards		50	No hazardous waste		50
No chemical reactions caused		30	Not require electricity when performing		30	No chemical hazards		30	No sample waste		20
No sample dilution needed		20	No water consumption while applying		20	No physical hazards		20	Reusable consumables		30
Final score per sub-item		0			0			0			0

B.

Sample throughput & time of analysis			Cost effectiveness *			Sample destruction			Easy operation			Specialized staff requirements		
Sub-items	Y/N	Max score	Sub-items	Y/N	Max score	Sub-items	Y/N	Max score	Sub-items	Y/N	Max score	Sub-items	Y/N	Max score
< 1 h/sample		40	< $200 / sample		50	Non destructive technique		40	Allow *in situ* measure [benchtop]		30	No high level		50
< 30 min/sample		30	< $100 / sample		30	Non invasive technique		30	Allow on line measure [portable]		30	No medium level		30
< 10 min/sample		20	< $20 / sample		20	< 10 ml used		10	Multiparameter		40	No required (basic)		20
< 1 min/sample		10				< 1 ml used		20						
Final score per sub-item		0			0			0			0			0

** Excluding depreciation expenses*

Figure 3.3. Proposed items and sub-items for the greenness (A) and blueness (B) rating of a non-destructive/non-invasive analytical method.

The following sections provide a breakdown of the scores obtained for each of the three methods selected to evaluate their whiteness with our proposal as a practical example. Some aspects to be considered for the application of this proposal are discussed. In accordance with the pursued aim of the three proposed methods under consideration (screening method), the aforementioned critical values for multivariate classification metrics have been established in the redness assessment [**36**]. Note that, based on the objective of the analytical methods evaluated, the authors have established a minimum whiteness requirement of 80% scoring. If this criterion is met, i.e., the whiteness result exceeds 80% once completing the *ex-ante* evaluation, the next step is to proceed with the method's development in order to conduct a comprehensive *ex-post* evaluation, and compare the results with the *ex-ante* evaluation.

3.1. The colour of a mNIR-based method

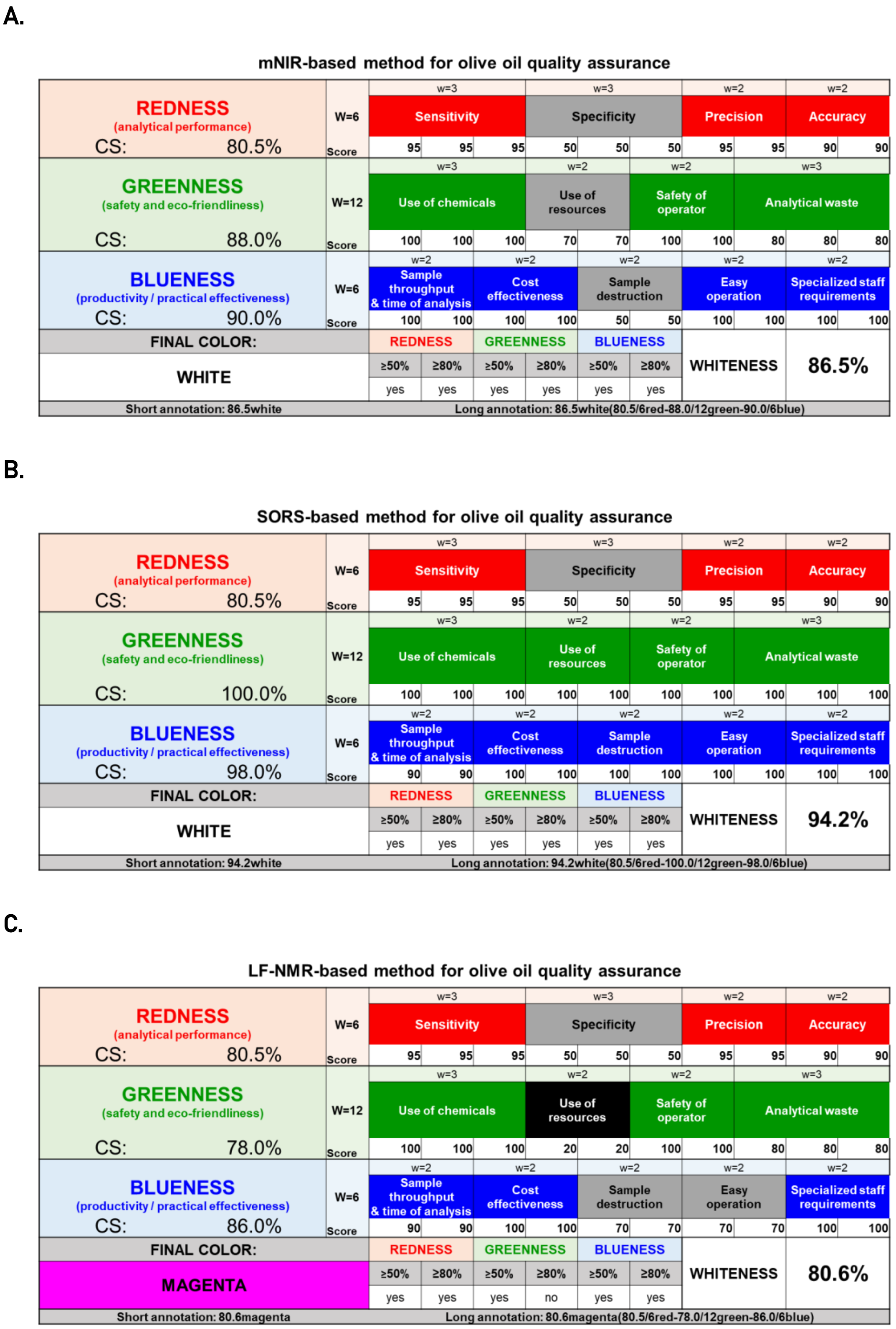
NIR spectroscopy instruments have come a long way in recent years, to the point where miniaturised, portable, cost-effective instruments are now available that can be used to measure anywhere, including in-line production analysis [**39**,**40**]. In this method, a miniaturised portable NIR equipment (mNIR) is used to measure vegetable oil samples placed in glass vials. The samples are measured directly and at room temperature, so there is no sample treatment, and they remain unaffected after measure, in a non-destructive way. Following these premises, the template developed to establish the greenness and blueness of this method has been filled in, and some details are given below.

Regarding the greenness scoring of this method, 100% was not achieved. The items not scoring full goals were the use of resources and analytical waste. The 70% scoring on the first item was related to the need to be plugged into the electricity supply to be able to measure, even though it is a portable device. Besides, it is a non-destructive technique, but invasive since the vegetable oil bottle must be opened to collect a small sample. Consequently, it lacks 100% blueness and is not compliant with the "no sample waste" sub-item (see Figure 3.3A). Nevertheless, considering all the evaluable items as a whole by calculating the weighted average, white was the colour achieved (see Figure 3.4A), and the whiteness scoring of this method would be at least 86.5%, assuming that the established critical values will be satisfied or exceeded after developing the method in the *ex-post* evaluation of the redness.

3.2. The colour of a SORS-based method

The emergence of portable and handheld Raman measurement equipment has opened up the possibility of using this technique for at-line production control and exploring its potential in food analysis [**41**]. In addition, portable equipment based on the non-invasive SORS technique is already commercially available and currently used for the control of raw materials in the pharmaceutical industry [**42**], so its use in the food industry is becoming more real and nearer [**25**]. Following the same methodology, an *ex-ante* evaluation was performed on this method, achieving the 100% greenness scoring.

This is coherent given that SORS allows to measure through the original container, therefore the sample is not altered in any way as long as the inside food product is insensitive to laser irradiation.



A.

mNIR-based method for olive oil quality assurance

		w=3			w=3			w=2		w=2	
REDNESS (analytical performance) CS: 80.5%	W=6	Sensitivity			Specificity			Precision		Accuracy	
	Score	95	95	95	50	50	50	95	95	90	90
		w=3			w=2		w=2		w=3		
GREENNESS (safety and eco-friendliness) CS: 88.0%	W=12	Use of chemicals			Use of resources		Safety of operator		Analytical waste		
	Score	100	100	100	70	70	100	100	80	80	80
		w=2		w=2		w=2		w=2		w=2	
BLUENESS (productivity / practical effectiveness) CS: 90.0%	W=6	Sample throughput & time of analysis		Cost effectiveness		Sample destruction		Easy operation		Specialized staff requirements	
	Score	100	100	100	100	50	50	100	100	100	100
FINAL COLOR:		REDNESS		GREENNESS		BLUENESS		WHITENESS		86.5%	
WHITE		≥50%	≥80%	≥50%	≥80%	≥50%	≥80%				
		yes	yes	yes	yes	yes	yes				
Short annotation: 86.5white		Long annotation: 86.5white(80.5/6red-88.0/12green-90.0/6blue)									

B.

SORS-based method for olive oil quality assurance

		w=3			w=3			w=2		w=2	
REDNESS (analytical performance) CS: 80.5%	W=6	Sensitivity			Specificity			Precision		Accuracy	
	Score	95	95	95	50	50	50	95	95	90	90
		w=3			w=2		w=2		w=3		
GREENNESS (safety and eco-friendliness) CS: 100.0%	W=12	Use of chemicals			Use of resources		Safety of operator		Analytical waste		
	Score	100	100	100	100	100	100	100	100	100	100
		w=2		w=2		w=2		w=2		w=2	
BLUENESS (productivity / practical effectiveness) CS: 98.0%	W=6	Sample throughput & time of analysis		Cost effectiveness		Sample destruction		Easy operation		Specialized staff requirements	
	Score	90	90	100	100	100	100	100	100	100	100
FINAL COLOR:		REDNESS		GREENNESS		BLUENESS		WHITENESS		94.2%	
WHITE		≥50%	≥80%	≥50%	≥80%	≥50%	≥80%				
		yes	yes	yes	yes	yes	yes				
Short annotation: 94.2white		Long annotation: 94.2white(80.5/6red-100.0/12green-98.0/6blue)									

C.

LF-NMR-based method for olive oil quality assurance

		w=3			w=3			w=2		w=2	
REDNESS (analytical performance) CS: 80.5%	W=6	Sensitivity			Specificity			Precision		Accuracy	
	Score	95	95	95	50	50	50	95	95	90	90
		w=3			w=2		w=2		w=3		
GREENNESS (safety and eco-friendliness) CS: 78.0%	W=12	Use of chemicals			Use of resources		Safety of operator		Analytical waste		
	Score	100	100	100	20	20	100	100	80	80	80
		w=2		w=2		w=2		w=2		w=2	
BLUENESS (productivity / practical effectiveness) CS: 86.0%	W=6	Sample throughput & time of analysis		Cost effectiveness		Sample destruction		Easy operation		Specialized staff requirements	
	Score	90	90	100	100	70	70	70	70	100	100
FINAL COLOR:		REDNESS		GREENNESS		BLUENESS		WHITENESS		80.6%	
MAGENTA		≥50%	≥80%	≥50%	≥80%	≥50%	≥80%				
		yes	yes	yes	no	yes	yes				
Short annotation: 80.6magenta		Long annotation: 80.6magenta(80.5/6red-78.0/12green-86.0/6blue)									

Figure 3.4. Overall whiteness scoring obtained by *ex-ante* evaluation of a: (**A**) mNIR-based method, (**B**) SORS-based method, and (**C**) LF-NMR-based method for virgin olive oils quality and/or authenticity analytical assessment.

The only evaluable item not reaching full scoring was the analysis time (blueness group) as it requires around 2 min for spectrum acquisition, although it is still a fast technique compared to conventional ones and is reflected with almost 100% blueness scoring. As outcome, the SORS-based analytical method resulted in a 94.2% whiteness scoring. The overall result can be seen in Figure 3.4B.

3.3. The colour of an LF-NMR-based method

The advent of current LF-NMR equipment opens up the possibility of NMR implementation in control laboratories for non-destructive, fast, and much more affordable measurements than conventional high-field NMR, which also requires higher maintenance costs and takes up more space than benchtop LF-NMR equipment [**43**]. Consequently, the *ex-ante* evaluation of the method based on LF-NMR yielded an 80.6% whiteness scoring (see Figure 3.4C). Even so, on this occasion the resulting colour was not white, and this is due to the requirement of these devices to be connected to electricity continuously, even if not measuring. Thus, the method scored low on one of the greenness evaluable items, however it should be noted that weight assigned to greenness was higher than for blueness and redness, which may explain the non-white final colour of this proposed method. Nevertheless, it is always possible to modify this based on the user's aim. Given that the NMR specifications yield good performance, it is likely that better analytical performance metrics will be obtained and therefore redness scoring would be larger than with other analytical techniques. If this is the goal of the method, it is recommended to increase the redness weighting.

4. Final remarks

An *ex-ante* evaluation of a new method should be mandatory before addressing an analytical method development and implementation milestone. On the one hand it allows to identify risks, and on the other hand it helps to improve efficiency especially in the use of resources, since if the method fails to give the expected results after development, a loss of time and certainly economic loss may have been avoided by performing the *ex-ante* evaluation and making decisions according to what has been achieved. It is crucial not to be swayed by the analyst preferences, and to be as objective as possible. The authors consider that by following this simple evaluation method, in which a yes or no answer is easily given, it is possible to evaluate the whiteness of a new non-destructive method.

Through practical application to three analytical methods pursuing the same aim, it has been demonstrated that this approach is sufficiently objective, and amply flexible to be applied in *ex-ante* evaluation under specific user requirements. The proposal described in this paper represents a step forward, applicable in any field of analytical chemistry as it does not only consider GAC principles, but goes beyond them.

Acknowledgments

AMJC acknowledges the Grant (RYC2021-031993-I) funded by MCIN/AEI/501100011033 and "European Union NextGenerationEU/PRTR". In addition, **AAC** gratefully acknowledges the Spanish Ministry of Universities for a pre-doctoral fellowship FPU (FPU20/04711, Formación del Profesorado Universitario).

Funding

This study is part of the public-private collaboration project (ref.: CPP2021-008672), funded by MCIN/AEI/501100011033 (Spanish Ministry of Science and Innovation) and "European Union NextGenerationEU/PRTR". Funding for open access charge: Universidad de Granada / CBUA.

References

[1] R.K. Bannor, K.K. Arthur, D. Oppong, H. Oppong-Kyeremeh, A comprehensive systematic review and bibliometric analysis of food fraud from a global perspective, J. Agric. Food Res. (2023) 100686. https://doi.org/10.1016/j.jafr.2023.100686.

[2] Q. Chen, H. Lin, J. Zhao, Advanced Nondestructive Detection Technologies in Food, Springer Nature, Singapore, 2021.

[3] D. Ballesteros-Vivas, B. Socas-Rodríguez, J.A. Mendiola, E. Ibanez, A. Cifuentes, Green food analysis: Current trends and perspectives, Curr. Opin. Green Sustain. Chem. 31 (2021) 100522. https://doi.org/10.1016/j.cogsc.2021.100522.

[4] R. Aslam, S.R. Sharma, J. Kaur, A.S. Panayampadan, O.I. Dar, A systematic account of food adulteration and recent trends in the non-destructive analysis of food fraud detection, J. Food Meas. Charact. 17 (2023) 3094-3114. https://doi.org/10.1007/s11694-023-01846-3.

[5] E.E. Okere, E. Arendse, H. Nieuwoudt, O.A. Fawole, W.J. Perold, U.L. Opara, Non-invasive methods for predicting the quality of processed horticultural food products, with emphasis on dried powders, juices and oils: A review, Foods 10 (2021) 3061. https://doi.org/10.3390/foods10123061.

[6] M.K.I. Khan, Advances in Noninvasive Food Analysis, CRC Press, New York, 2019.

[7] H.N. Moghaddam, Z. Tamiji, M.A. Lakeh, M.R. Khoshayand, M.H. Mahmoodi, Multivariate analysis of food fraud: A review of NIR based instruments in tandem with chemometrics, J. Food Compost. Anal. 107 (2022) 104343. https://doi.org/10.1016/j.jfca.2021.104343.

[8] M. Petersen, Z. Yu, X. Lu, Application of Raman spectroscopic methods in food safety: a review, Biosensors 11 (2021) 187. https://doi.org/10.3390/bios11060187.

[9] L. Wu, X. Tang, T. Wu, W. Zeng, X. Zhu, B. Hu, S. Zhang, A review on current progress of Raman-based techniques in food safety: From normal Raman spectroscopy to SESORS, Food Res. Int. 169 (2023) 112944. https://doi.org/10.1016/j.foodres.2023.112944.

[10] K. Fan, M. Zhang, Recent developments in the food quality detected by non-invasive nuclear magnetic resonance technology, Critical Rev. Food Sci. Nutr. 59 (2019) 2202-2213. https://doi.org/10.1080/10408398.2018.1441124.

[11] R. Cao, X. Liu, Y. Liu, X. Zhai, T. Cao, A. Wang, J. Qiu, Applications of nuclear magnetic resonance spectroscopy to the evaluation of complex food constituents, Food Chem. 342 (2021) 128258. https://doi.org/10.1016/j.foodchem.2020.128258.

[12] H.Y. Yu, S. Myoung, S. Ahn, Recent applications of benchtop nuclear magnetic resonance spectroscopy, Magnetochemistry 7 (2021) 121. https://doi.org/10.3390/magnetochemistry7090121.

[13] J. Pryshlakivsky, C. Searcy, Life cycle assessment as a decision-making tool: practitioner and managerial considerations, J. Clean. Prod. 309 (2021) 127344. https://doi.org/10.1016/j.jclepro.2021.127344.

[14] L.H. Keith, L.U. Gron, J.L. Young, Green analytical methodologies, Chem. Rev. 107 (2007) 2695-2708. https://doi.org/10.1021/cr068359e.

[15] A. Gałuszka, Z.M. Migaszewski, P. Konieczka, J. Namieśnik, Analytical Eco-Scale for assessing the greenness of analytical procedures, Trends Anal. Chem. 37 (2012) 61-72. https://doi.org/10.1016/j.trac.2012.03.013.

[16] M. Locatelli, A. Kabir, M. Perrucci, S. Ulusoy, H.I. Ulusoy, I. Ali, Green profile tools: current status and future perspectives, Adv. Sample Prep. 6 (2023) 100068. https://doi.org/10.1016/j.sampre.2023.100068.

[17] A. Gałuszka, Z. Migaszewski, J. Namiesnik, The 12 principles of green analytical chemistry and the SIGNIFICANCE mnemonic of green analytical practices, Trends Anal. Chem. 50 (2013) 78-84. https://doi.org/10.1016/j.trac.2013.04.010.

[18] F. Pena-Pereira, W. Wojnowski, M. Tobiszewski, AGREE—Analytical GREEnness metric approach and software, Anal. Chem. 92 (2020) 10076-10082. https://dx.doi.org/10.1021/acs.analchem.0c01887.

[19] A. Ballester-Caudet, P. Campíns-Falcó, B. Pérez, R. Sancho, M. Lorente, G. Sastre, C. González, A new tool for evaluating and/or selecting analytical methods: Summarizing the information in a hexagon, Trends Anal. Chem. 118 (2019) 538-547. https://doi.org/10.1016/j.trac.2019.06.015.

[20] M. Bystrzanowska, M. Tobiszewski, How can analysts use multicriteria decision analysis?, Trends Anal. Chem. 105 (2018) 98-105. https://doi.org/10.1016/j.trac.2018.05.003.

[21] P.M. Nowak, P. Kościelniak, M. Tobiszewski, A. Ballester-Caudet, P. Campíns-Falcó, Overview of the three multicriteria approaches applied to a global assessment of analytical methods, Trend Anal. Chem. 133 (2020) 116065. https://doi.org/10.1016/j.trac.2020.116065.

[22] P.M. Nowak, P. Kościelniak, What color is your method? Adaptation of the RGB additive color model to analytical method evaluation, Anal. Chem. 91 (2019) 10343-10352. https://doi.org/10.1021/acs.analchem.9b01872.

[23] P.M. Nowak, R. Wietecha-Posłuszny, J. Pawliszyn, White Analytical Chemistry: An approach to reconcile the principles of Green Analytical Chemistry and functionality, Trends Anal. Chem. 138 (2021) 116223. https://doi.org/10.1016/j.trac.2021.116223.

[24] H.N. Moghaddam, Z. Tamiji, M.A. Lakeh, M.R. Khoshayand, M.H. Mahmoodi, Multivariate analysis of food fraud: A review of NIR based instruments in tandem with chemometrics, J. Food Compost. Anal. 107 (2022) 104343. https://doi.org/10.1016/j.jfca.2021.104343.

[25] A. Arroyo-Cerezo, A.M. Jimenez-Carvelo, A. González-Casado, A. Koidis, L. Cuadros-Rodríguez, Deep (offset) non-invasive Raman spectroscopy for the evaluation of food and beverages – A review, LWT-Food Sci. Technol. 149 (2021) 111822. https://doi.org/10.1016/j.lwt.2021.111822.

[26] A.P. Sobolev, F. Thomas, J. Donarski, C. Ingallina, S. Circi, F.C. Marincola, D. Capitani, L. Mannina, Use of NMR applications to tackle future food fraud issues, Trends Food Sci. Technol. 91 (2019) 347-353. https://doi.org/10.1016/j.tifs.2019.07.035.

[27] A.I. Ropodi, E.Z. Panagou, G.-J.E. Nychas, Data mining derived from food analyses using non-invasive/non-destructive analytical techniques; determination of food authenticity, quality & safety in tandem with computer science disciplines, Trends Food Sci. Technol. 50 (2016) 11-25. https://doi.org/10.1016/j.tifs.2016.01.011.

[28] N. Mialon, B. Roig, E. Capodanno, A. Cadiere, Untargeted metabolomic approaches in food authenticity: a review that showcases biomarkers, Food Chem. 398 (2022) 133856. https://doi.org/10.1016/j.foodchem.2022.133856.

[29] A. Hassoun, S. Jagtap, G. Garcia-Garcia, H. Trollman, M. Pateiro, J.M. Lorenzo, M. Trif, A.V. Rusus, R.M. Aadil, V. Šimat, J. Cropotova, J.S. Câmara, Food quality 4.0: From traditional approaches to digitalized automated analysis, J. Food Eng. 337 (2022) 111216. https://doi.org/10.1016/j.jfoodeng.2022.111216.

[30] K. Kalinowska, M. Bystrzanowska, M. Tobiszewski, Chemometrics approaches to green analytical chemistry procedure development, Curr. Opin. Green Sustain. Chem. 30 (2021) 100498. https://doi.org/10.1016/j.cogsc.2021.100498.

[31] J.F. García Martín, Potential of near-infrared spectroscopy for the determination of olive oil quality, Sensors 22 (2022) 2831. https://doi.org/10.3390/s22082831.

[32] I.H. Barros, L.S. Paixão, M.H. Nascimento, V. Lacerda Jr, P.R. Filgueiras, W. Romão, Use of portable Raman spectroscopy in the quality control of extra virgin olive oil and adulterated compound oils, Vib. Spectrosc. 116 (2021) 103299. https://doi.org/10.1016/j.vibspec.2021.103299.

[33] D. Galvan, A.A.C. Tanamati, F. Casanova, E. Danieli, E. Bona, M.H.M. Killner, Compact low-field NMR spectroscopy and chemometrics applied to the analysis of edible oils, Food Chem. 365 (2021) 130476. https://doi.org/10.1016/j.foodchem.2021.130476.

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

[35] ASTM E2617-17, Standard Practice for Validation of Empirically Derived Multivariate Calibrations, ASTM International, West Conshohocken, 2017.

[36] L. Cuadros-Rodríguez, L. Valverde-Som, A.M. Jiménez-Carvelo, M. Delgado-Aguilar, Validation requirements of screening analytical methods based on scenario-specified applicability indicators, Trends Anal. Chem. 122 (2020) 115705. https://doi.org/10.1016/j.trac.2019.115705.

[37] N. Nenadis, M.Z. Tsimidou, Perspective of vibrational spectroscopy analytical methods in on-field/official control of olives and virgin olive oil, Eur. J. Lipid Sci. Technol. 119 (2017) 1600148. https://doi.org/10.1002/ejlt.201600148.

[38] R.T. Giebelhaus, K.T. Carrillo, S.L. Nam, A.P. de la Mata, J.F. Araneda, P. Hui, J. Ma, J.J. Harynuk, Detection of common adulterants in olive oils by bench top 60 MHz 1H NMR with partial least squares regression, J. Food Compos. Anal. 122 (2023) 105465. https://doi.org/10.1016/j.jfca.2023.105465.

[39] K.B. Beć, J. Grabska, C.W. Huck, Miniaturized NIR spectroscopy in food analysis and quality control: Promises, challenges, and perspectives, Foods 11 (2022) 1465. https://doi.org/10.3390/foods11101465.

[40] V. Cortés, J. Blasco, N. Aleixos, S. Cubero, P. Talens, Monitoring strategies for quality control of agricultural products using visible and near-infrared spectroscopy: A review, Trends Food Sci. Technol. 85 (2018) 138-148. https://doi.org/10.1016/j.tifs.2019.01.015.

[41] Y. Sun, H. Tang, X. Zou, G. Meng, N. Wu, Raman spectroscopy for food quality assurance and safety monitoring: A review, Curr. Opin. Food Sci. 47 (2022) 100910. https://doi.org/10.1016/j.cofs.2022.100910.

[42] M.A. Mansouri, P.Y. Sacré, L. Coic, C. De Bleye, E. Dumont, A. Bouklouze, Ph. Hubert, R.D. Marini, E. Ziemons, Quantitation of active pharmaceutical ingredient through the packaging using Raman handheld spectrophotometers: A comparison study, Talanta 207 (2020) 120306. https://doi.org/10.1016/j.talanta.2019.120306.

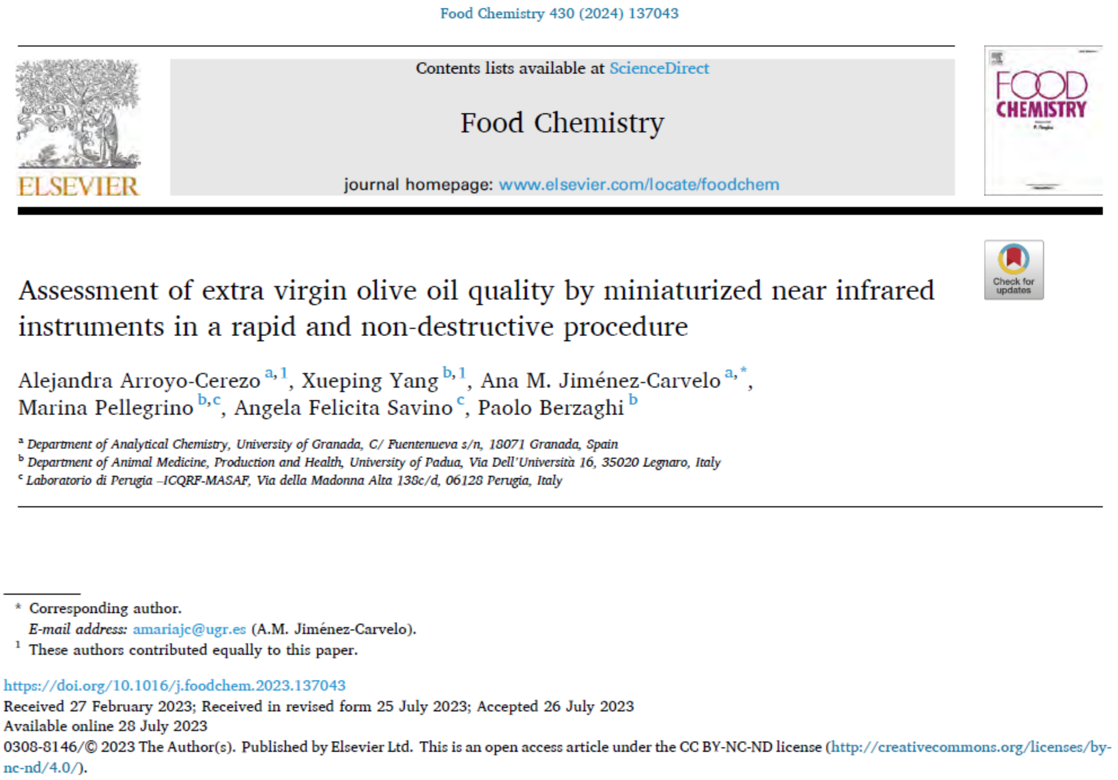
[43] D. Capitani, A.P. Sobolev, V. Di Tullio, L. Mannina, N. Proietti, Portable NMR in food analysis, Chem. Biol. Technol. Agric. 4 (2017) 1-14. https://doi.org/10.1186/s40538-017-0100-1.

3.3. Artículo científico 4

Assessment of extra virgin olive oil quality by miniaturized Near Infrared instruments in a rapid and non-destructive procedure.

Publicado en 2024 en la revista Food Chemistry 430, 137043.

DOI: 10.1016/j.foodchem.2023.137043.



Food Chemistry 430 (2024) 137043

Contents lists available at ScienceDirect

Food Chemistry

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

Assessment of extra virgin olive oil quality by miniaturized near infrared instruments in a rapid and non-destructive procedure

Alejandra Arroyo-Cerezo [a,1], Xueping Yang [b,1], Ana M. Jiménez-Carvelo [a,*], Marina Pellegrino [b,c], Angela Felicita Savino [c], Paolo Berzaghi [b]

[a] Department of Analytical Chemistry, University of Granada, C/ Fuentenueva s/n, 18071 Granada, Spain
[b] Department of Animal Medicine, Production and Health, University of Padua, Via Dell'Università 16, 35020 Legnaro, Italy
[c] Laboratorio di Perugia –ICQRF-MASAF, Via della Madonna Alta 138c/d, 06128 Perugia, Italy

[*] Corresponding author.
E-mail address: amariajc@ugr.es (A.M. Jiménez-Carvelo).
[1] These authors contributed equally to this paper.

https://doi.org/10.1016/j.foodchem.2023.137043
Received 27 February 2023; Received in revised form 25 July 2023; Accepted 26 July 2023
Available online 28 July 2023

Highlights:

- Low-cost portable NIR instruments could be used to screen and assess EVOO quality.
- Three non-EVOO suspects were easily detected based on their spectral information.
- A good prediction accuracy for K232 and fatty acids was achieved.
- An optimal handling study of two miniaturized NIR instruments was performed.

Keywords:

Extra virgin olive oil quality

Chemical parameter

Near-Infrared Spectroscopy

Low-cost instrument

Non-invasive method

Abstract

Food fraud in olive oil is a major concern for consumers and authorities due to the health risks and economic impacts. Common frauds include blending with other cheaper non-olive oils, or misleading labelling. The main issue is that legislation and methods presently used in routine laboratories are not always up to date with current fraudulent practices, making detection difficult, so new analytical methods development is required.

This study focuses on developing an affordable and non-destructive analysis method based on NIR spectroscopy and chemometrics for EVOO quality assessment, specifically by monitoring 7 parameters of interest in EVOO measured by official methods and used to develop calibrations through NIR data. For this, two NIR low-cost portable instruments were employed, studied in-depth and compared with a NIR benchtop instrument. Calibration results enabled detection of atypical olive oils and excellent accuracy, especially for palmitic and oleic acid predictions, demonstrating the potential of the instruments.

1. Introduction

The detection of food fraud is undoubtedly one of the most worrying issues for consumers and authorities. From a public point of view, the concern is mainly linked to health and misleading. And it is no less important in the economic field, given that expert estimates of food crime expenditure situate this cost at around $40 million per year, taking into account everything from direct costs (victims and the judicial system) to intangible and market costs **(Cox et al., 2020)**. Food fraud is becoming increasingly sophisticated, making it harder to detect. Unfortunately, the methods used in routine laboratories and the official food control methods outlined in legislation are not always keeping pace with current food fraud practises.

One of the most common sources of food fraud is the olive oil, according to the European Commission's Knowledge Centre for Food Fraud and Quality (KC-FFQ). In fact, olive oil is a target of food fraud worldwide for years **(Yan et al., 2020)**.

Several types of fraud have been described that affect not only the economic level, but also public health, such as the intoxication of more than 20,000 people in Spain in 1981 by the illegal sale of rapeseed oil as olive oil. Due to the high economic value of this product in its highest quality category (extra virgin olive oil, EVOO) for its characteristics and attributes, this product is one of the most common sources of food fraud in Europe, and some cases have been detected infringing the legislation that protects and differentiates EVOO from other edible oils. This type of fraud involves mainly the adulteration of EVOO with vegetable oils of lower quality or other vegetable oils, leading to mislabelling as to the commercial category and consumer deception **(Lozano-Castellón et al., 2022)**. Therefore, ensuring the authenticity and quality of this product is essential.

Current olive oil legislation in Europe is based on the recent implementing regulation **(Regulation (EU) 2022/2105)**, and on the repeals of the previous ones **(Regulation (EU) 2022/2104)**. These regulations establish the limits of 8 quality characteristics (acidity, peroxide index, K232, K268 or K270, Delta-K, organoleptic evaluation, and fatty acid ethyl esters) as requirements to commercially classify an olive oil as EVOO, and of the other categories of oil produced from olives. Additionally, the composition ranges of 6 fatty acids and 6 sterols, as well as 7 other fatty acids and water content are defined as purity characteristics. All these quality and purity criteria must be analyzed following the official methods and standards of the International Olive Council (IOC). This relies on the need to individually analyze the chemical parameters that determine the quality and purity characteristics by at least 8 different chemical methods using a targeted approach (one method for one analyte) to ensure that the indicated oil category is the correct one **(García González et al., 2018)**. Moreover, the parameters to be evaluated and official methods of analysis have remained the same for more than 30 years and still continue in the new regulation.

The official methods often require the use of chemicals and are time-consuming. As a result, both official control laboratories and routine control laboratories in the olive oil industries face limitations in analyzing a large number of samples. This becomes particularly concerning in the case of official control, as it takes days or even weeks to obtain the results relating to the correct labelling of EVOO according to the legislation. The possibility of having a single multiparametric method for rapid screening of edible oils and detection of atypical would greatly benefit the food fraud police and officials routine control laboratories **(García Martín, 2022)**. This would increase the efficiency of controls and optimize the overall food control process for olive oils.

Several studies have been published to date proposing various alternatives to traditional chemical methods for the determination of olive oil quality and authentication. A recent review by Zaroual and colleagues gathers a wide variety of analytical techniques that have been used for this purpose, from the most complex because of the need for a trained analyst, such as gas or liquid chromatography, to the simplest, such as spectroscopic techniques, as well as more innovative techniques such as electronic sensing **(Zaroual et al., 2021)**. The goals of the studies covered in that review could be categorized into: (i) geographical authentication, (ii) variety authentication, (iii) adulteration detection, (iv) classification by oil type and (v) prediction of chemical parameters. The least abundant studies in the literature are those related to (iv) and (v), while efforts to develop methods for (i) geographical and (ii) variety authentication and (iii) adulteration detection **(González-Pereira et al., 2021)** are higher, probably because there are no recognized methods for the identification of these specific frauds **(Conte et al., 2020)**.

Most of these proposed methods have some disadvantages as they are not easily transferable to the industry, especially when we talk about producers in the olive oil industry. There is still a lack of low-cost tools which in turn allow quick and easy screening of olive oils without the need for a trained analyst, especially for those olive oil industries that can implement a portable, fast, cost-effective, non-destructive, and simple method at the production site. In this line, spectroscopic techniques are the clear candidates to offer this type of rapid, non-destructive and low-cost solution. And within these, near-infrared (NIR) spectroscopy is the one that stands out the most, especially for quantitative analysis over compound identification **(García Martín, 2022)**. In this sense, some reviews **(Zaroual et al., 2021; García Martín, 2022)** were focused on the calibration models developed for the prediction of chemical parameters using NIR data for EVOO quality assessment. For olive oil quality characteristics, namely acidity, peroxides value, K232 and K270, collected NIR spectra allowed obtaining good performance metrics for prediction **(Manley et al., 2006; Inarejos-García et al., 2013; Willenberg et al., 2019)**. Fatty acid contents were also used as reference to develop calibration models with acceptably low prediction errors **(Özdemir et al., 2018)**. But scarce studies can be found on the feasibility of low-cost instruments for this purpose, to make it more affordable to EVOO producers. Garrido-Varo and colleagues **(Garrido-Varo et al., 2017)** achieved good quantification performance metrics for the most important parameters with the use of low-cost portable instruments, although fatty acids were not included in this study. Fatty acids, especially oleic and linoleic, provide information on the possible adulteration of EVOO with poorer quality vegetable oils.

The chemical information provided by a NIR spectrum is a combination of the chemical constituents found in the system to be analyzed, which are observed in the spectrum in the form of bands or peaks from fundamental vibrations of a chemical bond, which may be overlapped or indicate the presence of several different compounds in the same band. Therefore, it is visually difficult to work with these techniques by identifying individual bands, and the use of supervised chemometric tools is necessary for both qualitative and quantitative analysis, given the volume of data obtained with this technique **(Jiménez-Carvelo et al., 2019)**. The potential offered by chemometric methods using the non-targeted approach in spectroscopic techniques such as NIR is remarkable (**Karunathilaka et al., 2016**). Through this approach, the full non-specific signal is used as an instrumental fingerprint providing relevant chemical information to characterize the material (**Mialon et al., 2023**). Advantages of non-targeted over targeted approaches for food authentication purposes have already been highlighted in literature (**Sarkar et al., 2022; Hassoun et al., 2022**). Although integration into official methods is progressing with publications as ASTM standards and USP guidance (**ASTM, 2018; Pharmacopeia, 2019**), further development is still needed.

In this line, the aim of the present work was to develop an affordable, low-cost, and ready-to-use screening method, based on NIR spectroscopy coupled with chemometrics for the rapid and non-invasive control of EVOO from a non-targeted approach, using NIR spectra of olive oils as characteristic signals of each sample as an instrumental fingerprint. The handling of two low-cost portable instruments was also studied in depth, to stablish the optimal acquisition conditions, and compared with spectral data acquired by a benchtop NIR instrument to evaluate the quality of the results.

2. Materials and methods

2.1. Olive oil samples

A total of 195 olive oils (132 from 2021 and 63 from 2022 harvest) samples were analyzed for this study. The origin of oil samples was Italian and they were provided by the laboratory of the Central Inspectorate of Quality Protection and Fraud Repression of Agri-food Products (ICQRF), from the Ministry of Agricultural Food and Forestry Policies (Perugia, Italy).

Twenty-five samples corresponding to the 2022 harvest were received later, so that measurements were taken on different days. This set (VAL) was used to assess the applicability of the model to new independent samples. All of the previous samples (n. 170) were used for calibration development.

2.2. Portable NIR instruments

Two low-cost portable instruments were used to collect the NIR spectra of the samples in transmission mode (NIR-M-T1 and NIR-M-T11, Innospectra Corp., Taiwan). Both instruments had an integrated halogen tungsten lamp and the detector consists of a single 1 mm InGaAs element. Dimensions of both devices were similar: 92×76×41 and 96×48×38 mm respectively for NIR-M-T1 (NIT1) and NIR-M-T11 (NIT2) and weighed approximately 100 g each. The main difference between the two instruments was the collected wavelength range: 900 – 1700 nm for NIT1 and 1350 – 2150 nm for NIT2.

Both devices are based on the Texas Instrument DLP NIRScan Nano Evaluation Module (DLPNIRNANOEVM, Texas Instruments (TI), Dallas, United States) with a single InGaAs detector and digital micromirror device that can be optimized for number of pixels (equivalent to 128 pixel with no overlapping, or 256 pixels with overlapping) and exposure time in the range of 0.635 – 60.960 ms. The instruments were connected via USB cable to a laptop computer and controlled using ISC Winform v3.77 software (Innospectra corp. TW). As scanning settings are programmable, preliminary tests were performed looking for the best scanning time and resolution.

2.2.1. NIT1 configuration

The chosen configuration for this instrument consisted of 0.635 ms of exposure time, and 7.03 nm pattern width. Each spectrum was obtained averaging 20 scans with a digital resolution of 300 points in the 950-1650 nm range. The total measurement time was around 7.6 s per spectrum collected. The first and the last 50 nm in the scanning range were not considered because of greater spectral noise.

2.2.2. NIT2 configuration

The chosen configuration for the upper-wave range instrument was divided into two sections to amplify the captured absorbance intensity. First section, consisted of the 1350-1600 nm range with a digital resolution of 104 points, was performed with 7.03 nm pattern width and 0.635 ms of exposure time, while the second section, with a resolution of 176 points in the 1602-2150 nm range, had 15.22 nm pattern width and exposure time of 2.54 ms. The total digital resolution was 280 points. Each spectrum was obtained averaging 10 scans and the total measurement time was around 8.6 s.

2.2.3. NIR spectra acquisition

Each olive oil sample was transferred into two vials (VWR n. 548-0042) with a diameter of 8 mm. Each vial was scanned in duplicate on each instrument within 5 seconds.

To perform scans, the instruments were first warmed up to a system temperature of approximately 40 °C. Then a reference scan was taken with an empty vial before starting the analysis. This reference scan was repeated every 20 – 25 min during prolonged scanning sessions. All single scans were exported from the software as a single file in CSV format.

In order to develop the fastest possible method for EVOO control, all samples were measured at room temperature (25 ± 2 °C) and then heated at 50 °C, to be compared in this study with laboratory instruments **(Vanstone et al., 2018)** and to check the necessity of heating the EVOOs before collecting their NIR spectra, as although recommended, this temperature can cause oxidation in olive oils and be prejudicial to quality analysis.

2.3. Reference data: NIR spectra and chemical parameters

The spectra obtained with NIT1 and NIT2 were compared with spectra obtained by an FT-NIR MPA II spectrometer (Bruker Corporation, Billerica, Massachusetts) (FT-NIR) in the 12500 – 4000 cm^{-1} wavelength range with 8 cm^{-1} of resolution (converted to 875 – 2530 nm range) to compare the performances of the two portable low-cost against a laboratory benchtop NIR instrument. FT-NIR was located at ICQRF laboratory and NIR data were not available for the entire VAL set.

For this study, important chemical parameters for defining EVOO quality were determined to develop the models, namely: acidity, peroxides, K232, K268 and fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, linolenic and eicosenoic). Reference data were provided by ICQRF laboratory of Perugia (Italy) and all analyses were performed according to the official methods specified in the regulation **(Regulation (EU) 2022/2105)**.

A statistical analysis of wet chemistry results was performed. Moreover, the methods were validated by ICQRF laboratory, and the standard errors of laboratory (SEL) methods had been determined with 10 replicates of the same sample. Maximum R^2 obtainable in the subsequently development of calibration was calculated with the following equation. Since if the wet chemistry data had errors, these would be carried over to the next step when developing prediction models.

$$\mathrm{R}^2_{\max} = \frac{\mathrm{SD}^2 - \mathrm{SEL}^2}{\mathrm{SD}^2}$$

Where SD: standard deviation of data; SEL: standard error of laboratory.

2.4. Multivariate data treatment

Spectra from all single CSV files were imported and linearly interpolated every 2 nm by an R script using RStudio (RStudio version 2022.2.2.485, PBC, Boston, MA). All spectra were averaged by samples (4 scans), by scanning temperature and by instrument, defining four different sets of data (two instruments at two temperatures).

Replicate scans for each sample within each portable instrument were used to calculate repeatability of the instruments and to compare the quality of the obtained spectra between the two devices (NIT1 and NIT2) using the root mean squared (RMS) **(Xue et al., 2014; De la Roza-Delgado et al., 2017)** calculated according to the following equation:

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{n}(y_{im} - y_{ik})^2}{n}}$$

Where y_{im} = absorbance value of scan m of one sample at a wavelength i; y_{ik} = absorbance value of scan k of the sample at wavelength i; n = number of wavelengths.

Different combinations of spectra pre-processing methods were tested and the following was selected and applied to the data: Savitzky-Golay first derivative (9 filter window and 2nd polynomial order) and standard normal variation (SNV).

Wet chemistry data were evaluated for the detection of outliers, as laboratory reference data may have a great impact on the development of predicting models. For this purpose, a matrix composed of 170 rows (samples) and 11 columns (chemical parameters) was autoscaled to perform a principal component analysis (PCA) and the samples with large Q-residuals were removed.

Partial Least Square (PLS) regression method was used to develop calibration models for all the chemical parameters. The calibration dataset was randomly split into train and test sets in an 80:20 ratio and a 4 groups cross-validation was used to optimize the models. Optimal number of PLS components was selected on basis of the minimum root mean squared error (RMSE) balanced with the minimum possible value of the beta coefficient to avoid overfitting the model **(Stoltzfus, 2011)**. All calibration developments were performed under Python 3.9, using the NumPy package **(Harris et al., 2020)**.

Lastly, the VAL set (see section 2.1) was used to test the predictive capacity of the developed models, comparing prediction and reference values by applying models to samples not used for calibration **(García Martín, 2022)**. For this purpose, it was calculated the standard error of prediction (SEP), Bias, and GH distance **(Williams et al., 2017)**. The latter was calculated as the Mahalanobis distance divided by the number of PLS components **(Garrido-Varo et al., 2019)**. Note that in this study the calibration set is considered the set used to develop the models, randomly split into training and test sets. VAL set was the set of independent samples because they were sampled and analyzed at a different time.

Spectral repeatability was translated into predictive repeatability. For this purpose, the non-averaged spectra of the VAL set (4 spectra per sample) were introduced into the developed models. An analysis of variance (ANOVA) of the predicted parameters was performed and the repeatability for each data set was calculated as the square root of the residual variance divided by the degrees of freedom.

3. Results and discussion

The spectral range obtained by each of the three instruments (NIT1, NIT2 and FT-NIR) is shown in Figure 3.5. Note that FT-NIR spectra were trimmed over 2250 nm, as spectra were saturated and no longer provided useful information, according to García Martín **(García Martín, 2022)** can be discarded without losing important information in olive oil samples. As expected, with the bench spectrometer (FT-NIR) spectra show sharper peaks due to the higher resolution than portable instruments, NIT1 and NIT2.

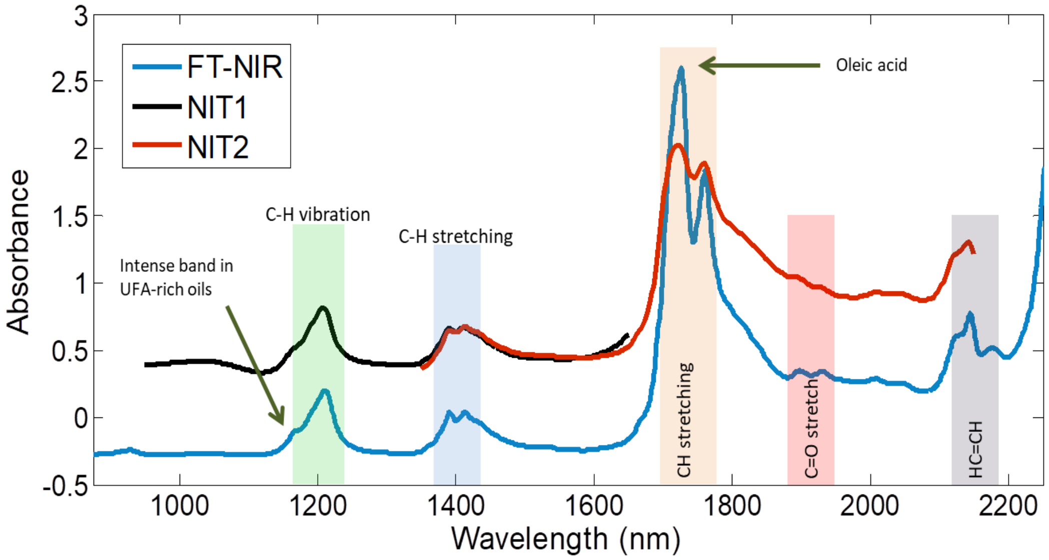


Figure 3.5. EVOO NIR spectra obtained with the two portable instruments (NIT1 and NIT2) and bench instrument (FT-NIR).

The band at 1180 – 1200 nm corresponds to the C-H vibration of CH_2 and CH_3 groups, but this range is not covered by NIT2 instrument. Spectral band at 1165 nm usually appears in oils with high unsaturated fatty acid content and is observed in FT-NIR spectrum, but practically unobservable in NIT1, probably due to the lower spectral resolution. Similarly, the double band between 1890 and 1945 nm, corresponding to C=O stretching, is detectable in FT-NIR spectrum but not in NIT2. The highest absorbance peak (1725 nm) characteristic of oleic acid, specifically triolein **(García-González et al., 2013)**, is not covered by the NIT1 spectral range, which will influence the calibration performances with data from this instrument. Finally, the double bands at 1380-1420 and 1700-1770 nm corresponding to C-H stretching are present in the NIT2 and FT-NIR spectral ranges **(Özdemir et al., 2018; Borghi et al., 2020)**. Pre-processed spectra can be seen in Figure 3.S1 of supplementary material.

3.1. Handling of portable instruments

The spectral repeatability of both portable instruments was calculated by means of the RMS value, the lower the better the repeatability. When raw data were used to calculate it, results were: 7.13×10^{-4}, 1.56×10^{-3}, 2.10×10^{-3}, 2.80×10^{-3} u.A. respectively for olive oils at room temperature and heated using NIT1 and NIT2. Note that this is spectra repeatability, i.e., this was calculated with the duplicate scans measured from one vial per sample. In addition, the repeatability per sample was also calculated, since two vials were measured for each sample in this study, and the results were: 8.54×10^{-3}, 1.02×10^{-2}, 1.04×10^{-2}, 1.67×10^{-2} u.A. These results show the sampling effect, and as expected, spectral repeatability by sample is poorer than repeatability by vials for a factor of 10. It is noteworthy that the repeatability of NIT1 is greater than that of NIT2, as is the case for samples measured at room temperature better than for heated samples. These results would indicate a priori that measurements performed at room temperature with the NIT1 spectrometer were more repeatable than with NIT2 or after heating the samples.

3.2. Wet chemistry data

The detection of outliers through the development of a PCA resulted in 6 samples with high values of Q residuals and Hotelling T^2. However, only three samples with high values of Q-residuals were removed from the dataset. High values of Q-residuals indicate those samples that are not well explained by the model, while high values of Hotelling T^2 correspond to those samples that show deviations. Therefore, in order to have maximum variability when running the model, the three samples with high Hotelling T^2 were not removed. In addition, the variability of these samples, and therefore being detected as outliers, coincides with their chemical values differing from those what would be expected for EVOO. In fact, these results agree with the information provided by ICQRF that, although all three samples were labelled as EVOO, the chemical parameters were very different from those that EVOO should contain. Figure 3.6 shows graphically the differences between the values of the chemical parameters analyzed, representing the ranges of the EVOO samples (n=164, in blue), those suspected non-EVOO (n=3, in red) and the range allowed for EVOO by legislation for each parameter (in green) **(Regulation (EU) 2022/2104)**. For better visualization, the 11 parameters were divided into two groups: high ranges (peroxides, and palmitic, oleic, and linoleic acids) and low ranges (acidity, K232, K268, and palmitoleic, stearic, linolenic and ecosanoic acids). Note that from now on, authentic EVOO samples will be named *typical EVOO*, and the suspected non-EVOO samples will be called *atypical EVOO*, following the Pharmacopoeia guidelines (**Pharmacopoeia, 2019**).

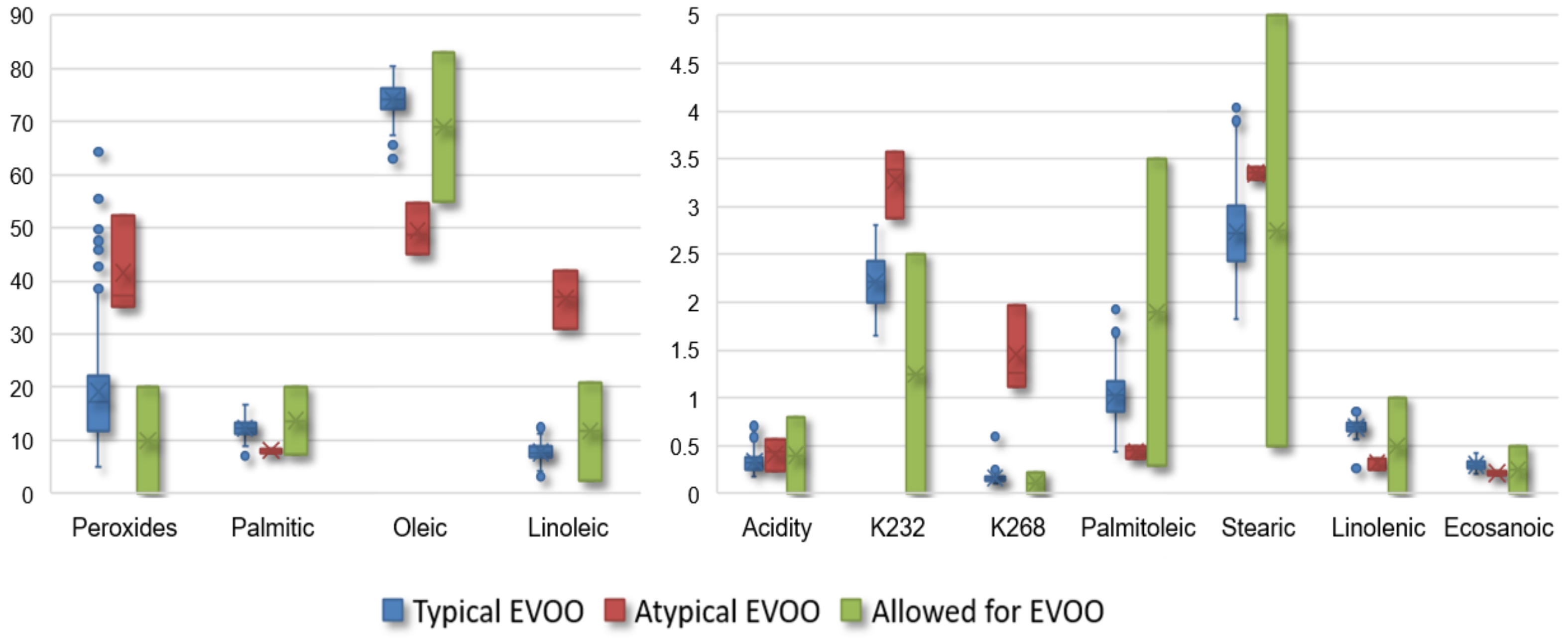


Figure 3.6. Wet chemical data of typical EVOO (authentic) samples and atypical EVOO (suspicious) samples with specified ranges allowed by legislation for EVOO.

Note that for better visualization, the 11 parameters were divided into two groups: high ranges on the left and low ranges on the right.

Note the difference in the K232 and K268 values, which in the atypical EVOO samples were above the permitted value of 2.50 and 0.22, respectively, for EVOO **(Regulation (EU) 2022/2104),** this could be caused by a bad oxidation state on these three oils. Remarkable was the difference found in oleic and linoleic acid, which were also outside the range allowed by legislation and the expected range for EVOO found in the literature **(Ózdemir et al., 2018)** in the three atypical EVOO samples. Further, looking at these lower oleic and higher linoleic values, it could be concluded that these three samples may had been adulterated with other non-olive vegetable oils, as these values were different from typical EVOO **(Aykas et al., 2020; Borghi et al., 2020)**. In addition, it is worth noting the high values in the peroxide index of some typical EVOO samples, above the permitted value, and this could be due to oxidation processes on those samples.

Table 3.1 shows the descriptive statistics of the calibration set after outlier removal, and VAL set, including: number of samples (N), mean, standard deviation, minimum and maximum values, the standard error of laboratory (SEL) of the chemical analysis methods, the maximum R^2 value and the Pearson's correlation coefficients between the chemical parameters. The data corresponding to the measured fatty acids were not available for the whole calibration set of samples, this is the reason for the difference in N value between the parameters.

The maximum R^2 value calculated is quite high and acceptable in most cases (greater than 0.9), especially for K268 and linoleic acid. However, maximum R^2 values for linolenic and eicosenoic acids showed some possible limitations, because reference SELs were large in relation to the variability of the calibration data. Also, it must be noted that in the data corresponding to the VAL set, some ranges were outside the range covered by the calibration data. This was the case of the minimum value of acidity, and maximum values of palmitic and palmitoleic acids. This will be a factor to take into account when evaluating prediction performances by using the VAL set.

Table 3.1. Descriptive statistics of the wet chemistry data of EVOO samples from the calibration (CAL) set and VAL set (N = 25).

	Parameter	Acidity	Peroxides	K232	K268	Palmitic	Palmitoleic	Stearic	Oleic	Linoleic	Linolenic	Ecosanoic
	Units	%	mEq O_2/kg	N/A	N/A	%TFA	%TFA	%TFA	%TFA	%TFA	%TFA	%TFA
CALIBRATION	N	167	166	167	167	141	141	141	141	141	141	141
	Mean	0.34	19.58	2.23	0.19	12.3	1.01	2.73	73.49	8.4	0.69	0.3
	SD	0.10	11.12	0.31	0.18	1.66	0.28	0.49	5.17	4.59	0.09	0.05
	Min	0.18	5.00	1.65	0.10	7.09	0.36	0.08	45.07	3.31	0.25	0.20
	Max	0.70	64.45	3.57	1.98	16.79	1.96	4.04	80.57	41.90	0.86	0.42
	SEL	0.03	2.40	0.06	0.01	0.48	0.07	0.12	0.98	0.24	0.06	0.06
	R^2_{mean}	0.917	0.958	0.962	0.997	0.917	0.938	0.940	0.964	0.997	0.552	-0.648
VALIDATION	Mean	0.28	10.16	1.89	0.13	12.83	1.09	2.55	74.06	7.74	0.69	0.27
	SD	0.10	2.93	0.20	0.01	1.28	0.25	0.43	3.13	1.87	0.04	0.03
	Min	0.08	5.50	1.52	0.10	10.09	0.77	2.03	62.65	4.88	0.62	0.22
	Max	0.45	15.20	2.24	0.16	17.00	2.08	3.64	78.74	13.99	0.76	0.32

N: number of samples; SD: standard deviation; Min: minimum; Max: maximum; SEL: standard error of laboratory; TFA: total fatty acids

3.3. Predictive models

After studying the handling and usability of the data obtained with NIT1 and NIT2 and the wet chemistry reference data, calibrations were developed for 7 parameters with the 5 data sets (NIT1 and NIT2 at room and high temperature and FT-NIR at high temperature, following their protocol for this type of analysis). The VAL set was used to evaluate the predictive capability, except for the calibrations developed with FT-NIR data, since spectral data for all 25 samples were not available. Table 3.2 shows the results obtained for the calibration and VAL sets. In this table, "_RT" refers to olive oils measured at room temperature and "_50" after heating.

As expected, best calibration performance metrics were obtained for FT-NIR data, as spectral range was larger, and resolution was greater than NIT1 and NIT2. However, RMSE values obtained FT-NIR were in most cases not much lower than the portable instruments, and for palmitic and oleic acids, calibrations with NIT1 at room temperature had slightly lower RMSE than using FT-NIR data. It should be noted that all calibration errors were lower than 1%, with the only exception of the peroxides index. This index is a parameter with large fluctuations (see Table 3.1) and is not particularly stable over time, as it in fact determines the deterioration of oil with time. Even so, the calibration R^2 were high, especially for the data obtained with NIT1 at high temperature.

Data acquired at room temperature with both portable instruments showed better calibration and prediction performances (errors and R^2) for K232 and the four fatty acids calibrations. This agrees with the results from Azizian et al. **(Azizian et al., 2007)**, who found that the increase in oil temperature above 40 °C caused classification and quantification errors of the models to be increasingly larger, this could be explained because such temperature could provoke the onset of fatty acids oxidation in olive oils. In addition, better results were also observed with the NIT1 data for fatty acids, both in calibration and in independent set prediction, compared to NIT2. This may be because measurements performed with NIT1 were more repeatable than with NIT2 as discussed in section 3.1 above. In the case of acidity and peroxides, NIT1 data achieved better calibration performances, but prediction performances were better with NIT2, including K232, which could be related to a possible loss of relevant information in the 1652-2150 nm spectral range not covered by NIT1.

Table 3.2. Calibration (CAL) and prediction (VAL) set metrics of developed PLS models

Instrument	CAL				VAL				
	Parameter	PLS f	RMSE	R^2_{cal}	SEP	Bias	GH_{av}	R^2_{pred}	Acc
NIT1_RT	Acidity	5	0.07	0.295	0.09	0.01	1.96	0.208	3.0
	Peroxides	4	3.94	0.845	11.88	11.37	6.01	0.289	5.0
	K232	7	0.19	0.369	0.15	0.00	2.62	0.477	2.5
	Palmitic	13	0.47	0.907	0.66	0.19	3.84	0.815	1.4
	Palmitoleic	10	0.15	0.754	0.14	-0.03	3.96	0.664	2.0
	Oleic	9	0.92	0.922	1.46	0.88	2.97	0.856	1.5
	Linoleic	8	0.48	0.923	0.57	-0.17	7.39	0.921	2.4
NIT1_50	Acidity	5	0.06	0.414	0.08	0.02	2.11	0.305	2.7
	Peroxides	8	3.47	0.880	13.00	12.30	3.19	0.321	5.4
	K232	7	0.20	0.304	0.17	0.04	2.57	0.376	2.8
	Palmitic	10	0.92	0.647	0.85	-0.13	4.35	0.636	1.8
	Palmitoleic	10	0.17	0.688	0.16	-0.05	3.74	0.646	2.3
	Oleic	8	1.78	0.704	1.88	-1.20	2.62	0.780	1.9
	Linoleic	8	0.82	0.774	1.25	1.05	3.78	0.898	5.2
NIT2_RT	Acidity	3	0.08	0.133	0.08	-0.03	2.16	0.328	2.7
	Peroxides	7	4.70	0.781	5.96	5.10	2.94	0.551	2.5
	K232	5	0.16	0.557	0.17	-0.10	1.70	0.528	2.8
	Palmitic	12	0.64	0.832	0.96	-0.09	2.47	0.502	2.0
	Palmitoleic	7	0.15	0.743	0.24	0.09	2.23	0.292	3.4
	Oleic	8	0.98	0.911	2.35	-0.89	2.26	0.584	2.4
	Linoleic	8	0.50	0.915	1.29	0.76	2.80	0.724	5.4
NIT2_50	Acidity	3	0.07	0.304	0.09	0.03	1.46	0.298	3.0
	Peroxides	8	3.86	0.852	12.31	11.70	1.88	0.142	5.1
	K232	5	0.19	0.335	0.18	0.04	1.98	0.193	3.0
	Palmitic	11	0.72	0.786	0.85	-0.30	2.66	0.639	1.8
	Palmitoleic	9	0.19	0.616	0.13	0.01	1.71	0.721	1.9
	Oleic	11	0.92	0.922	1.11	-0.16	2.44	0.873	1.1
	Linoleic	8	0.79	0.792	0.65	-0.39	2.46	0.920	2.7
FT-NIR	Acidity	6	0.08	0.463	N/A				
	Peroxides	11	2.84	0.895					
	K232	6	0.18	0.603					
	Palmitic	8	0.50	0.927					
	Palmitoleic	6	0.12	0.856					
	Oleic	7	0.98	0.976					
	Linoleic	9	0.23	0.999					

PLS f: number of PLS factors used for calibration; RMSE: root mean squared error; SEP: standard error of prediction; Acc: accuracy calculated as SEP/SEL; N/A: not available data

Beside the R^2, a useful criterion for estimating the prediction accuracy of a model is the proximity of the SEP to the standard error of laboratory (SEL). The criteria proposed by **(Shenk & Westerhaus, 1996)** state an excellent accuracy when SEP/SEL value is lower than 1.5; good accuracy for SEP/SEL values < 3; medium accuracy for SEP/SEL values < 4; and low accuracy for SEP/SEL values between 4 and 5. According to this, the predictions performed here for acidity, K232 and fatty acids had good accuracy (excellent for palmitic and oleic, as SEP/SEL was below 2 for NIT1), and low accuracy for peroxides **(García Martín, 2022)**.

The GH parameter was developed to identify spectra not well represented by the calibration data set. It should be noted that GH threshold in this type of application is usually set at 3, so a value above 3 would indicate an outlier within the developed model **(Garrido-Varo et al., 2019)**. Table 3.2 shows the average GH values obtained for the 25 samples of the VAL set. Values below 3 were observed for all predictions with NIT2, however, with NIT1 this only occurred for acidity, K232 and oleic acid. These results could be interpreted in two ways. Having no outliers in the NIT2 VAL set, one could conclude that in the VAL set all samples were typical EVOO. On the other hand, with NIT1 some samples in the VAL set having large (>3) GH values, could raise the suspicion of not being EVOO. However, GH distance is not the appropriate parameter to be used for discriminating being EVOO, but it is useful in the maintenance and updating calibrations to find those samples not well represented in the calibration set, in order to remove them or update the calibration with new samples inclusion.

A more suitable and easy way for the detection of atypical EVOO, could be focused on the predicted value for each chemical parameter and not directly on the spectral data as in the GH parameter. The three atypical EVOO samples (section 3.2) scored Hotelling T^2 values higher than 3 in all cases (FT-NIR, NIT1 and NIT2 at both temperatures) when performing a PCA with the predicted chemical data (all calibration and VAL set samples). The performed prediction also showed that these three samples were out of the allowed range for EVOO (see supplementary material, Figure 3.S2 and 3.S3), especially in K232 and oleic and linoleic acids, although the acidity value was within the allowed range, being one of the official parameters to be analyzed. Therefore, with the developed PCAs, atypical EVOO could be detected when the Hotelling T^2 value is greater than 3. These results show the capability of NIR as a multiparametric tool to be used in olive oil quality control according to J.F. García Marín review **(García Martín, 2022)** as a rapid and non-invasive screening, through the detection and identification of suspicious oils.

It should be noted in the PCA results (Figure 3.S2) that some samples scored too large Q residuals values, especially sample "529". This could be explained by the fact that it was the unique unfiltered olive oil, so the physical state of the oil should be taken into account, or a representative content of such samples should be included in the model. This did not hold true for the PCA developed using the reference chemical data, but it occurred when using the chemical data predicted from the NIR spectra.

3.4. Repeatability of predictions

The calculated spectral repeatability (see section 3.1) was transferred to predictive repeatability. The aim of this predictive repeatability study was to check whether it is necessary to measure the same sample in duplicate vials and also double scan or whether similar results could be obtained without duplicates, making faster analysis. Note that predictive repeatability is a required evaluation for any NIR method development (**Williams et al., 2017**). Table 3.3 shows the results obtained for predictive repeatability. The difference between repeatability per vial and per sample in NIT1 is remarkable. This shows that the effect of performing 2 scans of the same vial does not affect the results to a great extent. However, the fact of measuring two vials per sample due to the sampling and/or vial differences has a much greater effect. To the best of our knowledge, the repeatability per vial had not yet been reported, although this study showed it could influence the acquired NIR signal, and thus the results. Therefore, with NIT1 it is advisable to measure the same sample in two different vials, considering the short analysis time required compared to a more accurate prediction. Predictive repeatability of NIT2 was poorer than NIT1, as it could be expected from spectral repeatability (see section 3.1) and was similar between vials and samples. To obtain more precise predictions when using NIT2, it would be advisable to scan in two vials, both in duplicate.

Table 3.3. Predictive repeatability by ANOVA analysis for oleic and linoleic acid predictions of EVOO measured at room temperature and heated with NIT1 and NIT2.

	Repeatability by scan		**Repeatability by vial**	
	Df = 50		*Df = 25*	
Instrument	**Oleic**	**Linoleic**	**Oleic**	**Linoleic**
NIT1_RT	0.32	0.11	0.71	0.32
NIT1_50	0.34	0.15	0.66	0.31
NIT2_RT	1.52	0.86	1.55	0.95
NIT2_50	1.7	0.76	1.48	0.83

Df = degrees of freedom

These findings were in concurrence with the spectral repeatability assessment, as they substantiated that the outcomes derived from NIT1 exhibited superior spectral repeatability compared to NIT2, and moreover predictive repeatability calculated here was also higher. There was no significant difference between analyzing the olive oils at room or at higher temperature, so for a faster and more affordable analysis they could be measured at room temperature avoiding the heating steps.

4. Conclusions

This study aimed to investigate the potential use of two low-cost portable NIR instruments for rapid and non-destructive quality assessment of extra virgin olive oil (EVOO). The chemical information obtained from the spectral data allowed the development of calibrations for 7 chemical parameters that define EVOO quality, which showed good predictive capabilities. The repeatability of the two instruments was evaluated, showing the short-wavelength instrument (NIT1) to have better spectral repeatability than the upper-wavelength instrument (NIT2). The study also examined the effect of temperature on the oils to be measured, and some quality metrics showed better results in the predictions developed with data from oils at room temperature than heated, consistent with the literature that found higher quantification errors with increasing oil temperature. However, when predictive repeatability study was carried out, no significant differences were found between room temperature and heated oils, so to save time and minimize oxidation of the vegetable oils, it would be advisable to measure at room temperature.

Finally, the calibrations developed proved to detect atypical EVOO and showed good accuracy for most parameters, but more non-EVOO samples would be needed to fully evaluate the performances. The study suggests the use of portable low-cost NIR instruments for rapid and non-destructive analysis to pre-screen EVOO lots directly at production site and retain only the suspect atypical ones for laborious and expensive official analysis. This pre-screening would enable official laboratories to target samples with a greater probability of being adulterated, increasing the effectiveness of controlling EVOO on a territory. Despite the fact that these portable NIR instrument have slightly lower accuracy than the more sophisticated laboratory ones, the accuracy achieved in this study by the portable instruments would enable even small producers to have access to an analytical tool at a limited cost, which would allow the monitoring and improvement of EVOO production even in small operations.

Acknowledgments

AAC gratefully acknowledges the Spanish Ministry of Universities for a pre-doctoral fellowship FPU (FPU20/04711, Formación del Profesorado Universitario) and SensorFINT Cost Action (CA-19145) for the Short Term Scientific Mission (STSM) grant awarded for a research stay at the University of Padova. In addition, **AMJC** acknowledges the Grant (RYC2021-031993-I) funded by MCIN/AEI/501100011033 and "European Union NextGenerationEU/PRTR", and **XY** is grateful to receive funding from the China Scholarship Council program (202006350063).

References

ASTM (2017). E2617-17 Standard Practice for Validation of Empirically Derived Multivariate Calibrations. *ASTM International: West Conshohocken*, PA, USA. https://doi.org/10.1520/E2617-17

ASTM. (2018). E1655-17 Standard Practices for Infrared Multivariate Quantitative Analysis. *ASTM International: West Conshohocken*, PA, USA. https://doi.org/10.1520/E1655-17

Aykas, D.P., Karaman, A.D., Keser, B., & Rodriguez-Saona, L. (2020). Non-targeted authentication approach for extra virgin olive oil. *Foods, 9*, 221. https://doi.org/10.3390/foods9020221

Azizian, H., Kramer, J.K., & Winsborough, S. (2007). Factors influencing the fatty acid determination in fats and oils using Fourier transform near-infrared spectroscopy. *European journal of lipid science and technology, 109*, 960-968. https://doi.org/10.1002/ejlt.200700062

Borghi, F.T., Santos, P.C., Santos, F.D., Nascimento, M.H., Correa, T., Cesconetto, M., Pires, A.A., Ribeiro, A.V.F.N., Lacerda Jr, V., Romao, W., & Filgueiras, P. R. (2020). Quantification and classification of vegetable oils in extra virgin olive oil samples using a portable near-infrared spectrometer associated with chemometrics. *Microchemical Journal, 159,* 105544. https://doi.org/ 10.1016/j.microc.2020.105544

Commission Delegated Regulation (EU) 2022/2104 of 29 July 2022 supplementing Regulation (EU) No 1308/2013 of the European Parliament and of the Council as regards marketing standards for olive oil, and repealing Commission Regulation (EEC) No 2568/91 and Commission Implementing Regulation (EU) No 29/2012, L 284/1. http://data.europa.eu/eli/reg_del/2022/2104/oj

Commission Implementing Regulation (EU) 2022/2105 of 29 July 2022 laying down rules on conformity checks of marketing standards for olive oil and methods of analysis of the characteristics of olive oil, L 284/23. http://data.europa.eu/eli/reg_impl/2022/2105/oj

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

Cox, A., Wohlschlegel, A., Jack, L., & Smart, E. (2020). The cost of food crime. *Food Standard Agency* (Research Project code: FS 301065). https://www.food.gov.uk/research/food-crime/the-cost-of-food-crime

De la Roza-Delgado, B., Garrido-Varo, A., Soldado, A., Arrojo, A.G., Valdés, M.C., Maroto, F., & Pérez-Marín, D. (2017). Matching portable NIRS instruments for in situ monitoring indicators of milk composition. *Food Control, 76*, 74-81. http://dx.doi.org/10.1016/j.foodcont.2017.01.004

García González, D.L., Aparicio, R., & Aparicio-Ruiz R. (2018). Olive oil. In *FoodIntegrity Handbook: A Guide to Food Authenticity Issues and Analytical Solutions* (pp. 335-357). Eurofins Analytics, France. https://doi.org/10.32741/fihb

García Martín, J.F. (2022). Potential of near-infrared spectroscopy for the determination of olive oil quality. *Sensors, 22*, 2831. https://doi.org/10.3390/s22082831

García-González, D.L., Baeten, V., Pierna, J.A.F., & Tena, N. (2013). Infrared, Raman, and fluorescence spectroscopies: Methodologies and applications. In *Handbook of Olive Oil* (pp. 335-393). Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-7777-8_10

Garrido-Varo, A., Garcia-Olmo, J., & Fearn, T. (2019). A note on Mahalanobis and related distance measures in WinISI and The Unscrambler. *Journal of Near Infrared Spectrocopy, 27,* 253-258. https://doi.org/10.1177/0967033519848296

Garrido-Varo, A., Sánchez, M.T., De la Haba, M.J., Torres, I., & Pérez-Marín, D. (2017). Fast, low-cost and non-destructive physico-chemical analysis of virgin olive oils using near-infrared reflectance spectroscopy. *Sensors, 17*, 2642. https://doi.org/10.3390/s17112642

González-Pereira, A., Otero, P., Fraga-Corral, M., Garcia-Oliveira, P., Carpena, M., Prieto, M. A., & Simal-Gandara, J. (2021). State-of-the-art of analytical techniques to determine food fraud in olive oils. *Foods, 10*, 484. https://doi.org/10.3390/foods10030484

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., & Oliphant, T.E. (2020) Array programming with NumPy. *Nature, 585,* 357-362. https://doi.org/10.1038/s41586-020-2649-2

Hassoun, A., Jagtap, S., Garcia-Garcia, G., Trollman, H., Pateiro, M., Lorenzo, J.M., Trif, M., Rusu, A.V., Aadil, R.M., Šimat, V., Cropotova, J., & Câmara, J.S. (2023). Food quality 4.0: From traditional approaches to digitalized automated analysis. *Journal of Food Engineering, 337*, 111216. https://doi.org/10.1016/j.jfoodeng.2022.111216

Inarejos-García, A.M., Gómez-Alonso, S., Fregapane, G., & Salvador, M.D. (2013). Evaluation of minor components, sensory characteristics and quality of virgin olive oil by near infrared (NIR) spectroscopy. *Food Research International, 50*, 250-258. https://doi.org/10.1016/j.foodres.2012.10.029

Jiménez-Carvelo, A.M., González-Casado, A., Bagur-González, M.G., & Cuadros-Rodríguez, L. (2019). Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity–A review. *Food research international, 122*, 25-39. https://doi.org/10.1016/j.foodres.2019.03.063

Karunathilaka, S.R., Kia, A.R.F., Srigley, C., Chung, J.K., & Mossoba, M.M. (2016). Nontargeted, rapid screening of extra virgin olive oil products for authenticity using near-infrared spectroscopy in combination with conformity index and multivariate statistical analyses. *Journal of food science, 81*, C2390-C2397. https://doi.org/10.1111/1750-3841.13432

Lozano-Castellón, J., López-Yerena, A., Domínguez-López, I., Siscart-Serra, A., Fraga, N., Sámano, S., López-Sabater, C., Lamuela-Raventós, R.M., Vallverdú-Queralt, A. & Pérez, M. (2022). Extra virgin olive oil: A comprehensive review of efforts to ensure its authenticity, traceability, and safety. *Comprehensive Reviews in Food Science and Food Safety, 21*, 2639-2664. https://doi.org/10.1111/1541-4337.12949

Manley, M., & Eberle, K. (2006). Comparison of Fourier transform near infrared spectroscopy partial least square regression models for South African extra virgin olive oil using spectra collected on two spectrophotometers at different resolutions and path lengths. *Journal of near infrared spectroscopy, 14*, 111-126. https://doi.org/10.1255/jnirs.597

Mialon, N., Roig, B., Capodanno, E., & Cadiere, A. (2023). Untargeted metabolomic approaches in food authenticity: a review that showcases biomarkers. *Food Chemistry, 398*, 133856. https://doi.org/10.1016/j.foodchem.2022.133856

Özdemir, İ.S., Dağ, Ç., Özinanç, G., Suçsoran, Ö., Ertaş, E., & Bekiroğlu, S. (2018). Quantification of sterols and fatty acids of extra virgin olive oils by FT-NIR spectroscopy and multivariate statistical analyses. *LWT - Food Science and Technology, 91*, 125-132. https://doi.org/10.1016/j.lwt.2018.01.045

Pharmacopoeia, U.S. (2019). Appendix XVIII: Guidance on Developing and Validating Non-targeted Methods for Adulteration Detection. In *US Pharmacopeial Convention: Rockville*, MA, USA.

Sarkar, T., Salauddin, M., Kirtonia, K., Pati, S., Rebezov, M., Khayrullin, M., ... & Lorenzo, J.M. (2022). A review on the commonly used methods for analysis of physical properties of food materials. *Applied Sciences, 12*, 2004. https://doi.org/10.3390/app12042004

Shenk, J., & Westerhaus, M. (1996). Calibration the ISI way. In *Near Infrared Spectroscopy: The Future Waves* (pp. 198-202), Eds. Davies, AMC and Williams. NIR Publications: Chichester, UK.

Stotltzfus, J.C. (2011) Logistic regression: a brief primer. *Academic emergency medicine, 18*, 1099-1105. https://doi.org/10.1111/j.1553-2712.2011.01185.x

Vanstone, N., Moore, A., Martos, P., & Neethirajan, S. (2018). Detection of the adulteration of extra virgin olive oil by near-infrared spectroscopy and chemometric techniques. *Food quality and safety, 2,* 189-198. https://doi.org/10.1093/fqsafe/fyy018

Willenberg, I., Matthäus, B., & Gertz, C. (2019). A new statistical approach to describe the quality of extra virgin olive oils using near infrared spectroscopy (NIR) and traditional analytical parameters. *European Journal of Lipid Science and Technology, 121*, 1800361. https://doi.org/10.1002/ejlt.201800361

Williams, P., Dardenne, P., & Flinn, P. (2017). Tutorial: Items to be included in a report on a near infrared spectroscopy project. *Journal of Near Infrared Spectroscopy, 25*, 85-90. https://doi.org/10.1177/0967033517702395

Xue, J., Yang, Z., Han, L., Chen, L. (2014). Study of the influence of NIRS acquisition parameters on the spectral repeatability for on-line measurement of crop straw fuel properties. *Fuel, 117*, 1027-1083. http://dx.doi.org/10.1016/j.fuel.2013.10.017

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

Zaroual, H., Chénè, C., El Hadrami, E. M., & Karoui, R. (2022). Application of new emerging techniques in combination with classical methods for the determination of the quality and authenticity of olive oil: A review. *Critical Reviews in Food Science and Nutrition, 62*, 4526-4549. https://doi.org/10.1080/10408398.2021.1876624

SUPPLEMENTARY INFORMATION *(Artículo científico 4)*

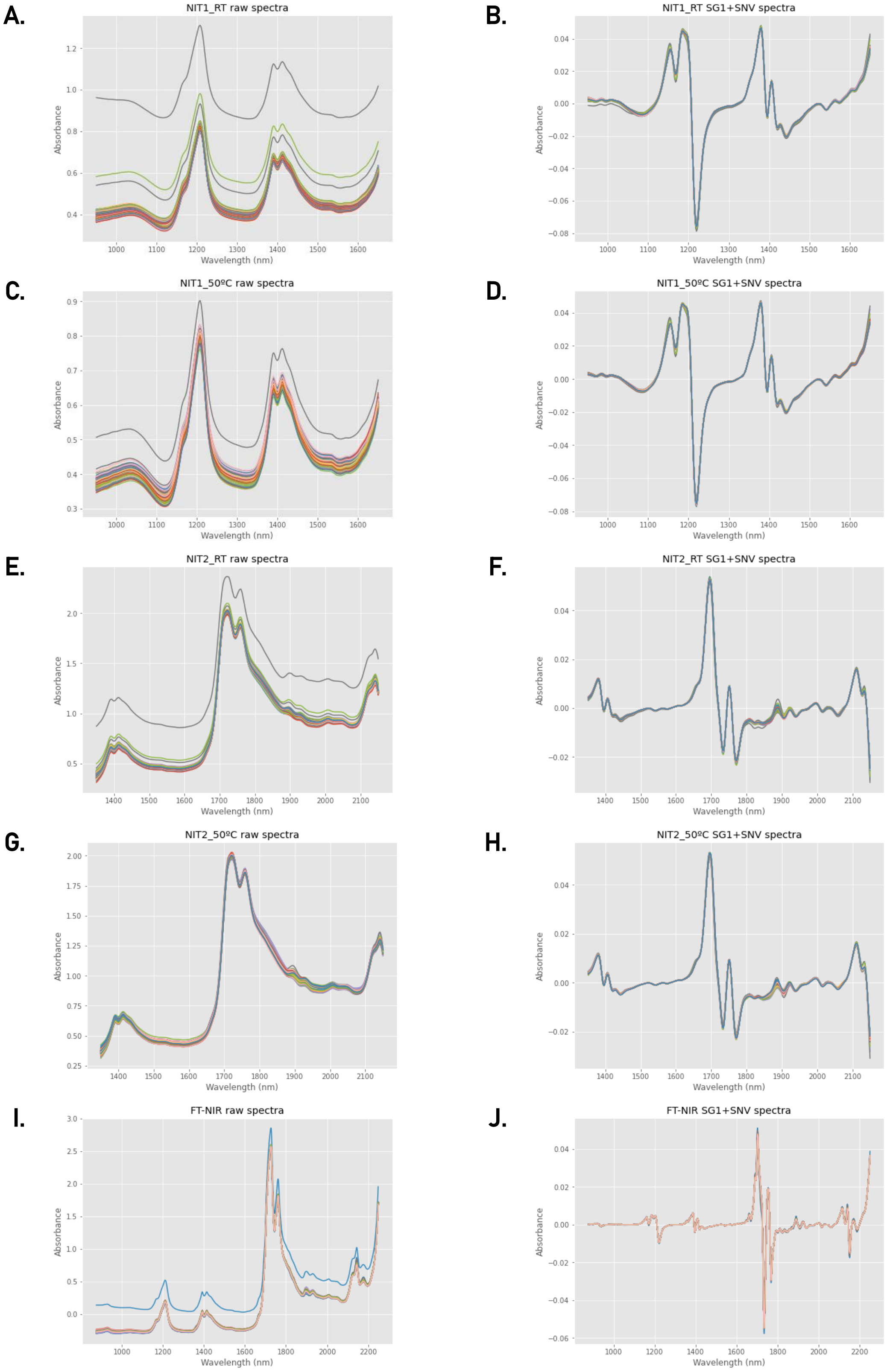


Figure 3.S1. Raw and processed spectra of the calibration set (n=167) EVOO samples acquired with NIT1 and NIT2 at room temperature and heated (A-H), and FT-NIR (I-J)

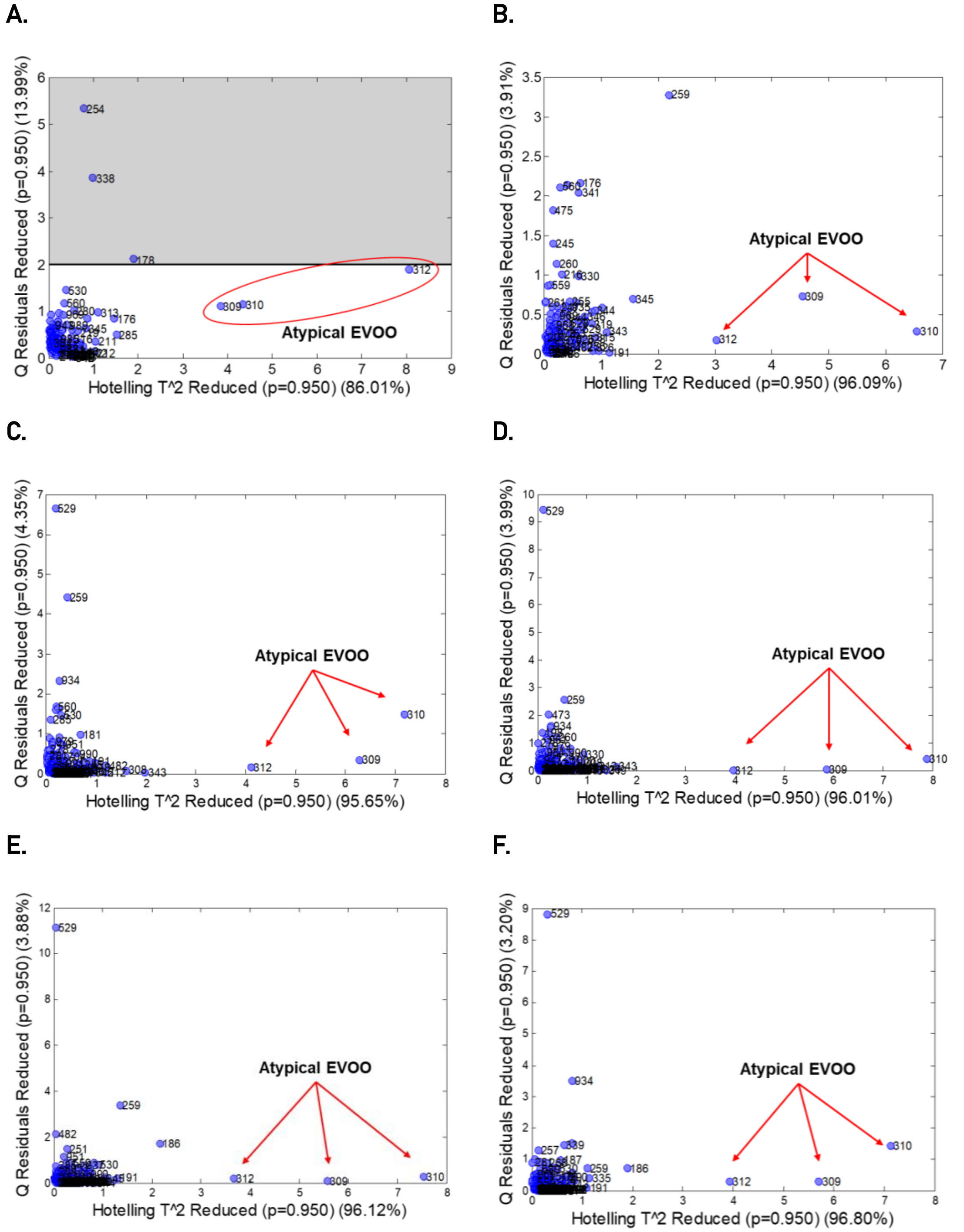


Figure 3.S2. Results of PCA performed with reference wet chemistry data for outliers detection (A) and with predicted chemical data of calibration and prediction (VAL) sets with FT-NIR (B), NIT1 at room and high temperature (C,D) and NIT2 at room and high temperature (E,F)

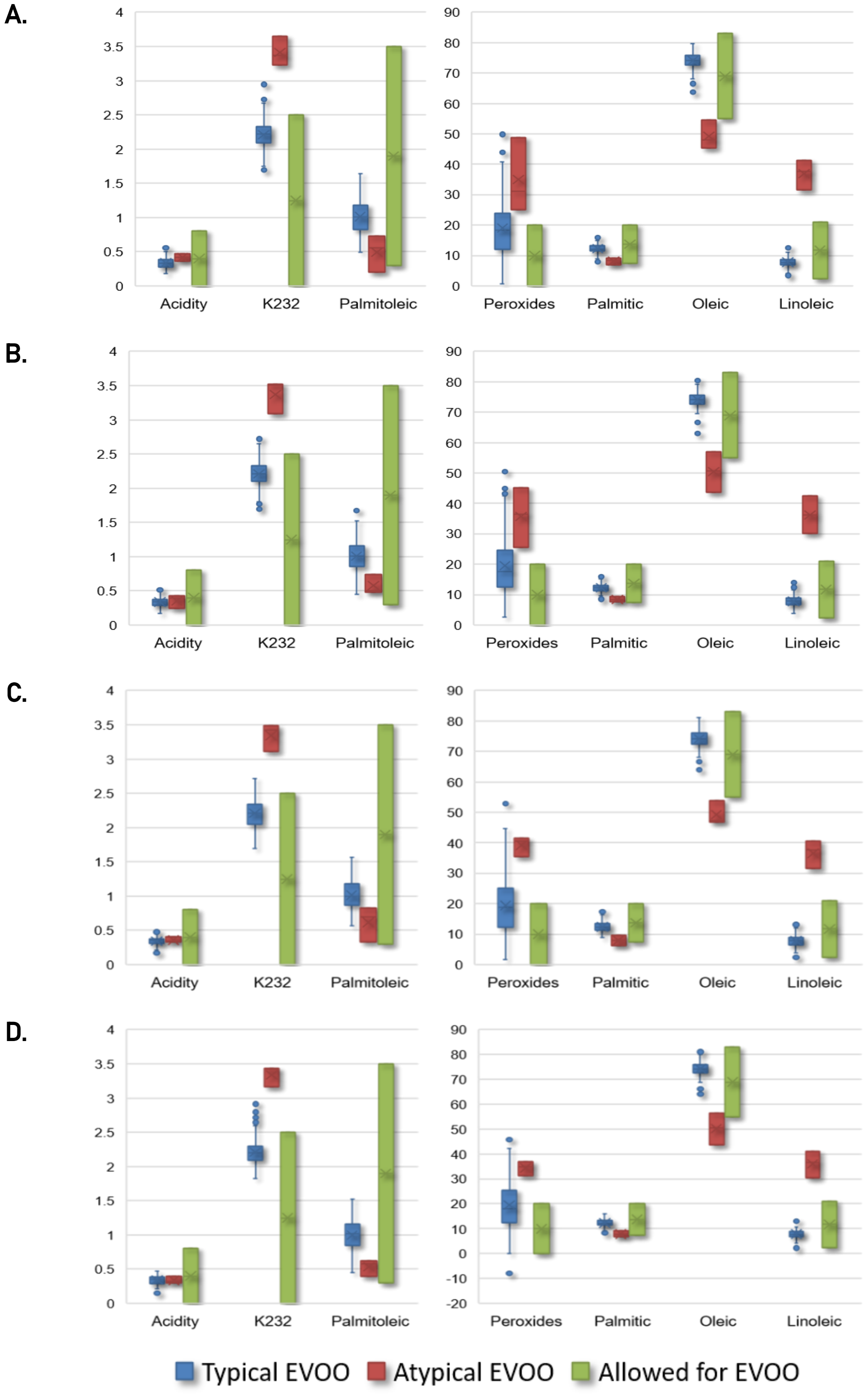


Figure 3.S3. Predicted chemical data of authentic EVOO samples and atypical EVOO samples with specified ranges allowed by legislation for EVOO.

3.4. Estudio 1

Espectrometría LF-NMR y quimiometría aplicada al análisis rápido y poco invasivo de la calidad de aceites de oliva.

Este estudio aborda el uso de la espectrometría de resonancia magnética nuclear de baja frecuencia de campo (LF-NMR) para la adquisición de huellas instrumentales de aceites vegetales y el tratamiento y análisis de los datos siguiendo un enfoque no-dirigido mediante la aplicación de herramientas quimiométricas. Se presenta la metodología empleada y los resultados preliminares obtenidos hasta la fecha de depósito de la presente tesis doctoral. Todo ello forma parte del proyecto de colaboración público-privada CPP2021-008672 titulado "Implantación de la espectrometría de resonancia magnética nuclear de baja frecuencia de campo (LF-NMR) en laboratorios de control para estudios cuantitativos y de clasificación de productos alimenticios y de otros sectores industriales" (NMR-Control) financiado por el Ministerio de Ciencia, Innovación y Universidades.

El contenido presentado en este apartado forma parte de un estudio de investigación que se encuentra en desarrollo y pretende ser enviado para su publicación en forma de artículo científico si así lo merecen los resultados obtenidos.

1. Materiales y métodos

1.1. Muestras y preparación

Un total de 295 aceites vegetales fueron incluidos para este estudio, de los cuales 174 provenían de la campaña 2022-2023 y 121 de la campaña 2023-2024. De entre ellos, 205 eran aceites de oliva de diferentes categorías (virgen, virgen extra, lampante y orujo), y 90 aceites de semilla, en su mayoría de girasol. Todas las muestras fueron proporcionadas por el Laboratorio Tello (Tentamus Group Ltd), formando parte del proyecto de colaboración público-privada CPP2021-008672. Los aceites fueron almacenados en viales de cristal ámbar, bajo refrigeración (4 °C) y oscuridad y encabezadas en nitrógeno hasta el momento de la preparación de la muestra correspondiente.

Se utilizó bromoformo no deuterado 98% estabilizado en etanol (Panreac Quimica SLU, Barcelona, España) como patrón interno (PI).

La preparación de muestras se realizó siguiendo el siguiente procedimiento: se mezcló una alícuota de cada aceite vegetal con bromoformo en una proporción 1:2 (aceite:PI) en un microtubo Eppendorf de 1.5 mL de capacidad. Tras agitación mecánica, se trasvasó un volumen de 700 µL a tubos NMR estándar de 5 mm.

1.2. Medidas LF-NMR

La adquisición de los espectros LF-NMR se llevó a cabo con un espectrómetro Nanalysis Benchtop 100PRO NMR (Nanalysis Corp., Calgary, Canadá) equipado con un automuestreador con capacidad para 25 tubos.

Las condiciones experimentales fueron optimizadas previamente (estudio presentado en el apartado 4.4 del **Capítulo 4** de la presente tesis). Se adquirieron espectros LF-NMR unidimensionales 1H y 13C de todas las muestras de aceite, en un tiempo aproximado de adquisición de 15 segundos y 2.5 minutos respectivamente.

El automuestreador permitió llevar a cabo las medidas en tandas de un máximo de 24 muestras. Al comienzo y finalización de la secuencia, se programó un ajuste de la homogeneidad del campo magnético (conocido como *shimming*), y el mismo ajuste se realizó cada 4-5 nuevos espectros adquiridos. La secuencia consistió en la adquisición del espectro 1H LF-NMR, seguido del 13C LF-NMR de cada muestra, antes de continuar con las medidas de la siguiente muestra. Los espectros se extrajeron directamente del equipo en formato .dx.

1.3. Tratamiento y análisis de datos

En primera instancia, se utilizó el software MestReNova (v14.2.0-26256, Mestrelab Research S.L., Santiago de Compostela, España) que permite importar archivos con extensión .dx, y se realizó una corrección por separado de los espectros 1H y 13C. Esta consistió en lo siguiente:

1. Corrección de fase automática
2. Corrección de línea base con filtro Whittaker
3. Referencia del desplazamiento químico del pico del PI (a 6.91 ppm y a 11.05 ppm respectivamente en los espectros 1H y 13C)

Tras esta corrección, se exportó cada uno de los espectros de forma individual en formato *comma separated values* (CSV).

A continuación, se importaron los archivos en formato .mat utilizando para ello el entorno de lenguaje MATLAB (versión R2022a, Mathworks Inc., Natick, MA, USA) y se aplicó una función programada *in-house* para normalizar los espectros dividiendo cada valor de intensidad por el área del pico del PI.

Tras ello, se armonizó el número de variables de cada espectro para poder construir la matriz de datos de entrada que contiene los espectros de todas las muestras. A continuación, se eliminaron los valores de intensidades correspondientes a los rangos espectrales en los que la información era nula (línea base) o correspondía a la señal del PI. De esta forma, cada espectro contenía solo la huella instrumental característica del aceite vegetal, que se encuentra en el rango 0-6 ppm en el espectro 1H, y en el rango 15-180 ppm. Finalmente, el vector que encerraba cada uno de los espectros contenía un total de 615 y 1401 variables o valores de intensidad respectivamente para 1H y 13C. Esto condujo a una reducción de variables, ya que en ambos casos se partía de 2048 valores de intensidad por cada espectro.

La aplicación de herramientas quimiométricas para el pre-procesado y análisis de los datos espectrales se llevó a cabo utilizando PLS_Toolbox (ver. 9.1, Eigenvector Research Inc. MA, USA) bajo el entorno Matlab.

2. Resultados y discusión

Los aceites vegetales están constituidos en su mayoría por triglicéridos (>95% de su composición). Mientras que el resto de su composición viene dada por componentes minoritarios como ácidos grasos libres, alcoholes, pigmentos, y otros compuestos.

La gran diferencia entre aceites de diferente origen vegetal, p.ej. entre oliva y girasol, viene dada por la conformación de los triglicéridos, es decir, por las cadenas grasas que conforman la molécula. Así, el ácido oleico es el mayoritario en los aceites de oliva.

En los espectros 1H y 13C LF-NMR adquiridos para este estudio se puede observar en la Figura 3.7 un pico mayoritario, que en ambos casos se atribuye a la presencia de cadenas grasas hidrocarbonadas, encontrándose aproximadamente a 1.2 ppm en el espectro 1H y alrededor de 30 ppm en el espectro 13C.

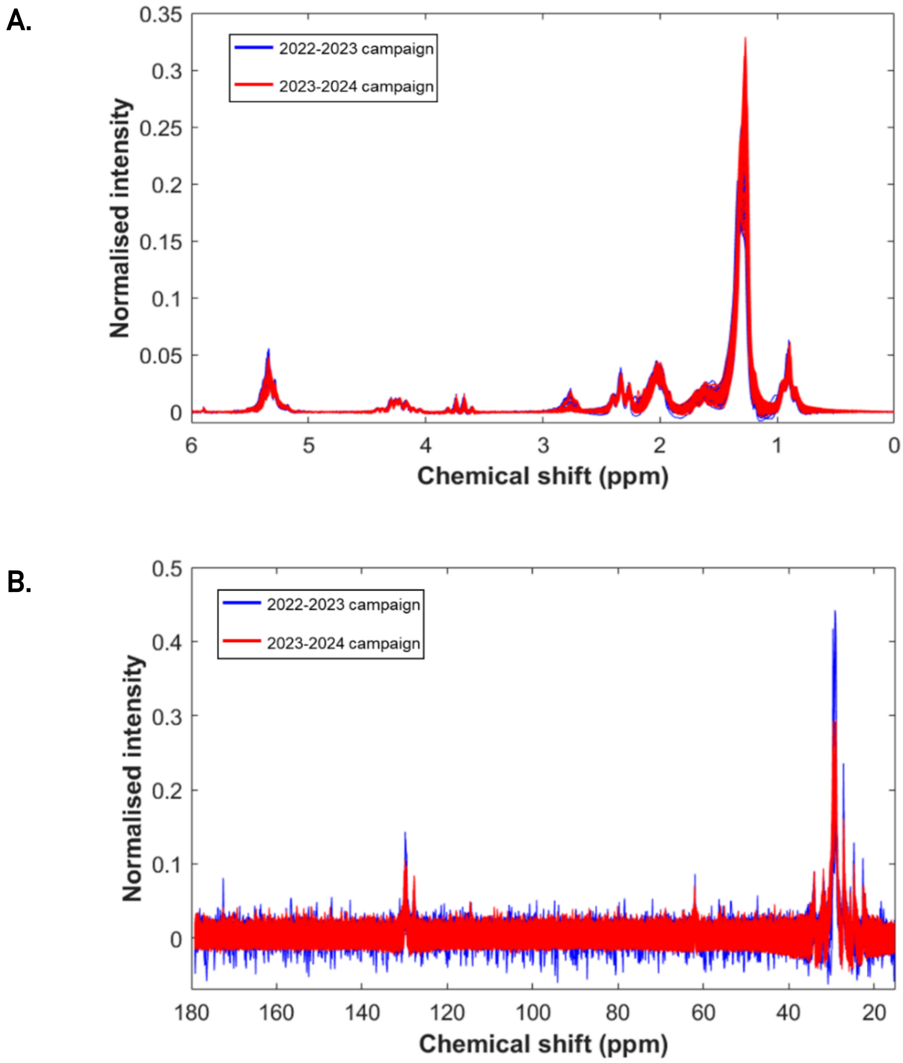
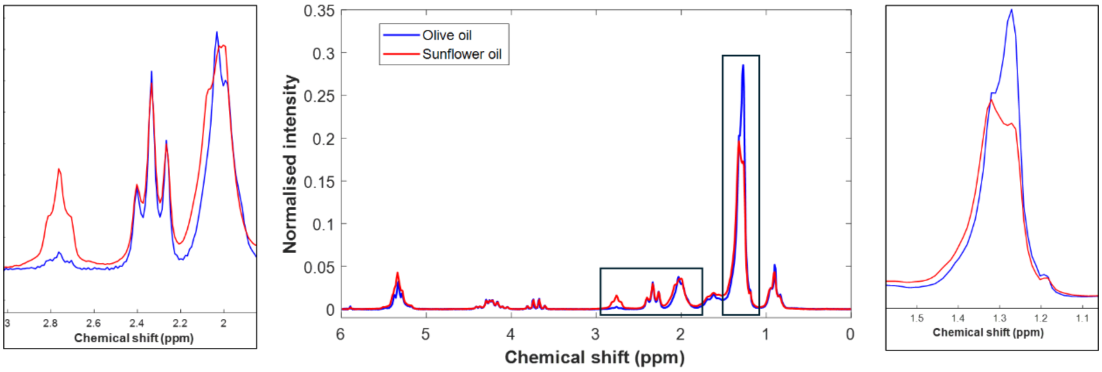


Figura 3.7. Espectros 1H LF-NMR (A) y 13C LF-NMR (B) de muestras de aceites vegetales tras la corrección, normalización y selección del rango de interés. Se muestran con diferentes colores las muestras de la campaña 2022-23 (en azul) y 2023-24 (en rojo).

Por otro lado, también es posible asignar otros picos minoritarios a ácidos grasos insaturados como el linolénico o el linoleico al rango entre 2 y 2.8 ppm en el espectro 1H, o al glicerol que forma parte de los triglicéridos y se observa a 4.2 y 5.2 ppm en el mismo espectro. Por otro lado, los picos situados entre 130 y 140 ppm en los espectros 13C se asignan normalmente a ácidos grasos insaturados, y aquellos que aparecen alrededor de 180 ppm a los triglicéridos [**16**].

La Figura 3.8A ilustra las diferencias significativas observables entre la huella 1H LF-NMR característica de un aceite de oliva y un aceite de girasol. Es destacable una mayor intensidad en los picos correspondientes al ácido linoleico alrededor de 2.8 ppm, dado que el aceite de girasol contiene una mayor cantidad con respecto al de oliva. Asimismo, también destaca la diferencia en la morfología del pico mayoritario correspondiente a las cadenas hidrocarbonadas, alrededor de 1.3 ppm.

A.



Nota. Se amplian las regiones en las que se observan mayores diferencias entre las huellas instrumentales de ambos aceites.

B.

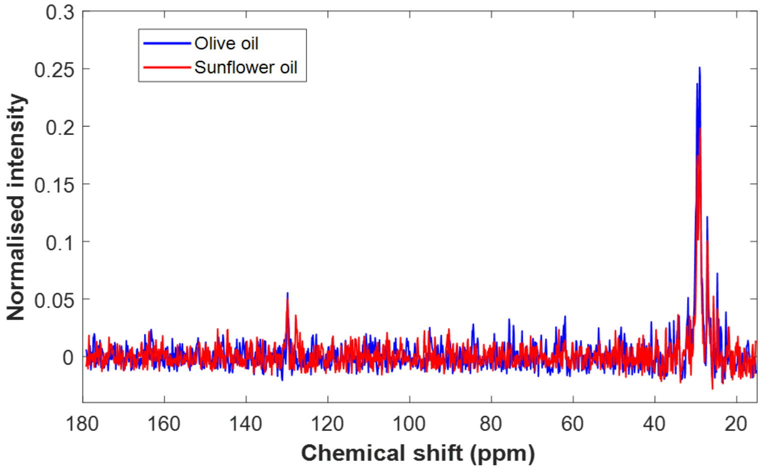


Figura 3.8. Espectros 1H LF-NMR (A) y 13C LF-NMR (B) de dos aceites para su comparación visual: oliva virgen y girasol.

[**16**] Sacchi, R.; Addeo, F.; Paolillo, L. 1H and 13C NMR of virgin olive oil. An overview. *Magn. Reson. Chem.* **1997**, *35,* 133-S145. DOI: 10.1002/(SICI)1097-458X(199712)35:13<S133::AID-OMR213>3.0.CO;2-K.

Sin embargo, las huellas 13C LF-NMR de los mismos aceites no han mostrado tales diferencias (Figura 3.8B), más allá de una mayor intensidad en el pico mayoritario correspondiente a las cadenas hidrocarbonadas del aceite de oliva con respecto al de girasol, hecho que coincide con el caso de la huella instrumental 1H LF-NMR.

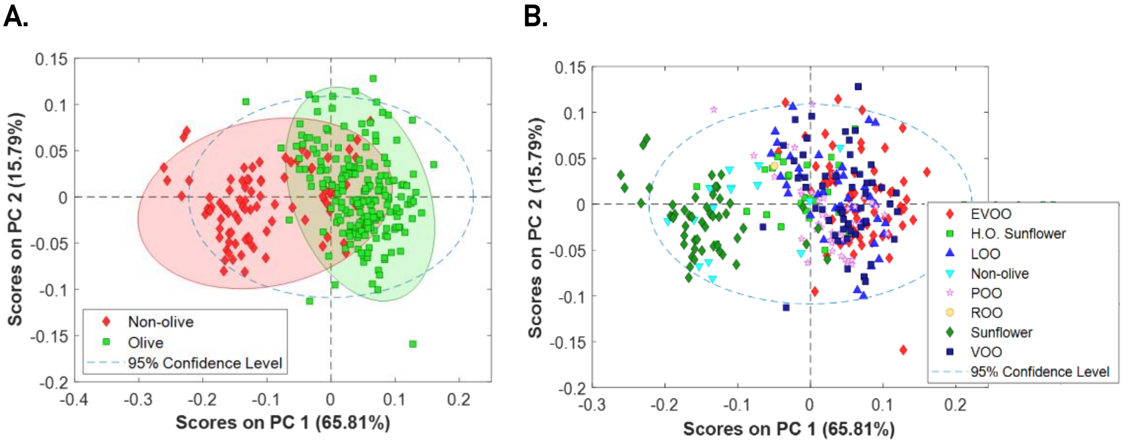
2.1. Análisis exploratorio

Se llevó a cabo un análisis de componentes principales (PCA) a partir de las matrices de datos 1H LF-NMR y 13C LF-NMR generadas para estudiar el agrupamiento natural de las muestras de aceites vegetales.

2.1.1. Huellas 1H LF-NMR

Utilizando como método de pre-procesado centrado en la media (*mean center)*, 6 componentes principales (PCs) explicaban un total 96.76% de la variabilidad en los datos de partida. La selección de los PCs se basó en los eigenvalores.

La Figura 3.9 muestra el gráfico de puntuaciones (*scores plot*) de la PC2 frente a la PC1, tras descartar una muestra detectada como anómala. Se observan dos agrupaciones diferenciadas por las PC1, que se corresponden con los aceites de oliva (puntuaciones positivas de PC1) y aceites no oliva (puntuaciones negativas de PC1). Sin embargo, hay una pequeña agrupación de muestras de aceites no oliva que se encuentran formando parte del grupo oliva. Estas muestras eran en su mayoría aceites de girasol alto oleico, como se puede observar en la Figura 3.9B. Nótese que la única diferencia entre ambas figuras, A y B, radica en las etiquetas de las muestras.



EVOO = extra virgin olive oil; H.O. = high oleic; LOO = lampante olive oil; POO = pomace olive oil; ROO = refined olive oil; VOO = virgin olive oil.

Figura 3.9. Gráfico de puntuaciones de la PC2 frente a la PC1 obtenido del PCA generado a partir de los espectros 1H LF-NMR. Se muestra el mismo gráfico con diferentes etiquetas: para diferenciar entre oliva y no oliva (A), y entre las diversas categorías de aceites vegetales incluidos en el estudio (B).

Cuando se analizaron en conjunto el resto de PCs, no se observaron agrupaciones significativas entre las muestras atendiendo a la información disponible sobre origen vegetal y/o categoría comercial, siendo por tanto la PC1 la que explica la variabilidad en los espectros provocada por esta característica.

2.1.2. Huellas 13C LF-NMR

Se siguió el mismo procedimiento para el análisis exploratorio de los datos 13C LF-NMR. De igual forma, centrado en la media (*mean center*) fue el método de pre-procesado seleccionado, ya que arrojaba mejores resultados en cuanto a la visualización de agrupaciones entre las muestras.

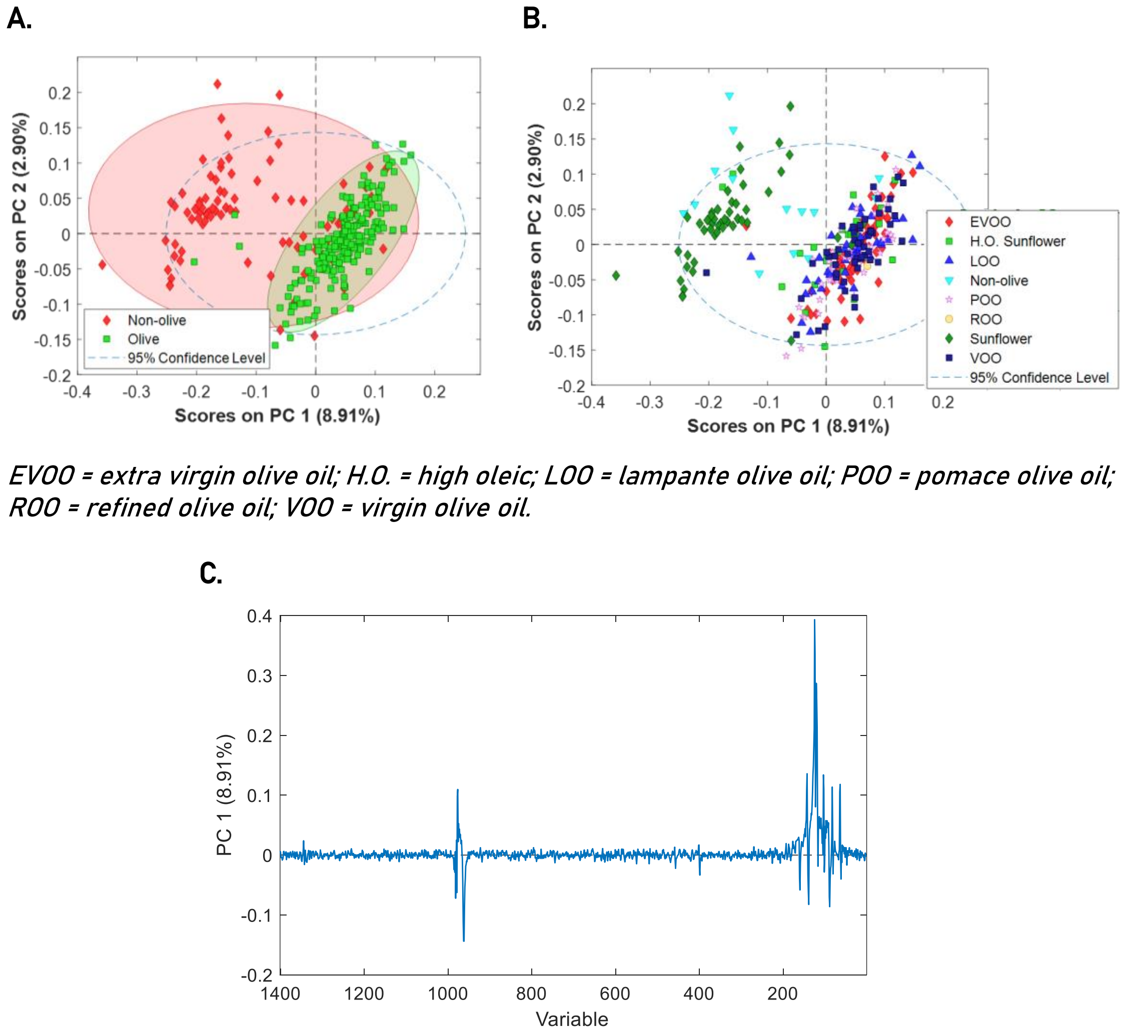
Una vez generado el PCA, se detectaron dos muestras de aceite que quedaron significativamente lejanas al límite de confianza y que fueron descartadas. Se trataba de una muestra categorizada como aceite de oliva virgen y otra como aceite de orujo de oliva. Tras ello, se seleccionaron 4 PCs atendiendo a los eigenvalores, que explicaban una variabilidad acumulada de los datos del 14.31%.

La Figura 3.10 muestra el gráfico de puntuaciones obtenido al representar la PC2 frente a la PC1 con diferentes etiquetas para las muestras, al igual que en el apartado anterior, en los gráficos A y B.

Es reseñable el bajo porcentaje de variabilidad capturada por los PCs, lo cual podría estar justificado por la morfología de la huella instrumental de carbono. En la Figura 3.7B se puede apreciar la cantidad de ruido que constituye la señal, la cual no aporta información de interés para el estudio. La señal de 13C-NMR es mucho más débil con respecto a la de 1H-NMR, es por ello que habitualmente se adquiere con un mayor número de barridos para aumentar la ratio señal/ruido. Sin embargo, los objetivos perseguidos con el proyecto al que pertenecen estos datos, y en concreto con este estudio, se fundamentan en el desarrollo de un método analítico rápido. Es por ello que, a la hora de optimizar las condiciones de adquisición (véase el apartado 4.4 del **Capítulo 4**), se priorizó la adquisición de señales en el menor tiempo posible, a la vez que la obtención de una señal de la máxima calidad informativa.

Esto además queda confirmado cuando se observan los gráficos de cargas (*loadings plot*) para la PC1 en la Figura 3.10C, que es la responsable mayoritaria de la separación entre los grupos mostrados en la Figura 3.10A. En este gráfico se aprecia el peso que tienen los picos característicos de la huella instrumental de aceites vegetales, y la mínima contribución que conlleva en ello el ruido presente en las señales 13C LF-NMR.

A pesar de la baja variabilidad explicada por los PCs seleccionados, los resultados muestran una separación natural entre las muestras que atiende al origen vegetal de los aceites, en línea con los resultados mostrados en el apartado anterior, generados a partir de los datos 1H-NMR. Esto confirma que el uso de una señal analítica como huella instrumental, analizada mediante herramientas quimiométricas, no requiere una máxima resolución espectral. El desafío real es conseguir una señal de la máxima calidad informativa, es decir, que encierre la información necesaria para el objetivo perseguido.



EVOO = extra virgin olive oil; H.O. = high oleic; LOO = lampante olive oil; POO = pomace olive oil; ROO = refined olive oil; VOO = virgin olive oil.

Figura 3.10. Gráfico de puntuaciones de la PC2 frente a la PC1 obtenido del PCA generado a partir de los espectros 13C LF-NMR. Se muestra el mismo gráfico con diferentes etiquetas: para diferenciar entre oliva y no oliva (A), y entre las diversas categorías de aceites vegetales incluidos en el estudio (B). Además, se muestra el gráfico de cargas para la PC1 (C).

Nótese que, en (C), el eje x fue invertido horizontalmente de acuerdo con la representación convencional de señales NMR.

Por otro lado, de igual forma a la que ocurría con los datos 1H LF-NMR, observando la Figura 3.10B donde las muestras aparecen etiquetadas según el origen vegetal y la categoría comercial, la mayoría de los aceites de girasol alto oleico se agrupan con los aceites de oliva.

2.2. Modelado supervisado

2.2.1. Criterios de calidad de aceites de oliva

El principal desafío en el ámbito abordado con este estudio es el de diferenciar la categoría de aceite de oliva, especialmente entre virgen y lampante. Esto viene dado por una serie de criterios de calidad y pureza recogidos en la legislación [**17**,**18**].

Dado que con el análisis exploratorio no se encontraron agrupaciones significativas y suficientes para llevar a cabo esta distinción entre categorías comerciales de aceite de oliva (resultados no mostrados), se examinó el potencial de métodos supervisados.

Para ello, se construyó una matriz de datos incluyendo solo los aceites de oliva de todas las categorías declaradas por el Laboratorio Tello (virgen, virgen extra, lampante y orujo de oliva), haciendo un total de 205 muestras. Tras eliminar los *outliers* detectados en el PCA, el total de muestras fue de 204 y 203, respectivamente para los datos 1H y 13C.

El grado de acidez y el índice de peróxidos son los dos criterios de calidad cuyo análisis resulta complejo. Ambos análisis se basan en métodos analíticos convencionales basados en una valoración volumétrica que debe realizarse por duplicado. Los valores del grado de acidez que delimitan la categoría son: <0.8 % para ser considerado aceite de oliva virgen extra (EVOO), <2.0 % para aceite de oliva virgen (VOO), y considerando el aceite de oliva como lampante (LOO) en aquellos que superen 2.0 % de acidez. Mientras que el límite máximo de índice de peróxidos es de 20 mEq O_2/kg tanto para la categoría EVOO como VOO.

Se decidió seguir una estrategia basada en la clasificación en lugar de la cuantificación, ya que el interés no reside en conocer el valor exacto, sino llevar a cabo un cribado que permita detectar aquellos aceites que de manera

[**17**] Commission Delegated Regulation (EU) 2022/2104 of 29 July 2022 supplementing Regulation (EU) No 1308/2013 of the European Parliament and of the Council as regards marketing standards for olive oil, and repealing Commission Regulation (EEC) No 2568/91 and Commission Implementing Regulation (EU) No 29/2012. Official Journal of the European Union L 284/1, 2022.

[**18**] Commission Implementing Regulation (EU) 2022/2105 of 29 July 2022 laying down rules on conformity checks of marketing standards for olive oils and methods of analysis of the characteristics of olive oil. Official Journal of the European Union L 284/23, 2022.

inequívoca están por debajo de los límites establecidos por la legislación para ser categorizados como EVOO/VOO.

Para ello, los datos de referencia proporcionados por el Laboratorio Tello se utilizaron para generar dos clases para cada uno de los dos criterios, grado de acidez e índice de peróxidos, como muestra la Tabla 3.4.

Tabla 3.4. Criterios establecidos para la generación de grupos/clases de aceites de oliva de acuerdo a dos criterios de calidad y número de muestras de cada grupo.

	Grado de acidez		**Índice de peróxidos**	
	Valor (%)	**Nº muestras**	**Valor (mEq O_2/kg)**	**Nº muestras**
Grupo 1 *	≤ 0.5	180	≤ 10	192
Grupo 2	≥ 0.8	11	> 10	13
Inconcluyentes	0.5 – 0.8	14	N/A	-

(*) El grupo 1 engloba a los aceites de oliva que según estos dos criterios de calidad podría ser considerado EVOO. N/A = no aplica.

Los límites establecidos para el desarrollo de modelos supervisados son más restrictivos que los establecidos por la legislación. De esta forma, se pretende que el modelo sea capaz de discernir las que sin duda alguna están dentro del límite para la categoría EVOO o VOO, y se consideren dudosos aquellos que no lo están, para que posteriormente pasen a ser analizados mediante los métodos analíticos reconocidos por el Consejo Oleícola Internacional (COI).

En primer lugar, se llevó a cabo un análisis exploratorio usando el método *partial least squares* (PLS). La Figura 3.11 muestra los resultados generados a partir de los datos 1H y 13C utilizando como pre-procesado centrado en la media (*mean center*) y autoescalado respectivamente, ya que fueron los que mostraron mejores resultados en este análisis. El número de variables latentes (LVs) seleccionado para cada análisis exploratorio con PLS fue: 5 LVs (acidez/1H), 4 LVs (peróxidos/1H), 3 (acidez/13C) y 3 (peróxidos/13C).

Se observó una separación significativa en ambos criterios de calidad de los aceites a partir de los datos 13C LF-NMR (Figuras 3.11B y 3.11D). Según estos resultados, las muestras con valores de acidez e índice de peróxidos que superaban el límite para ser considerado EVOO/VOO, obtuvieron un valor superior a la unidad en el estadístico T^2-Hotelling. Esto demuestra que es posible realizar un cribado dirigido a evaluar aceites con un grado de acidez <0.5 % y un índice de peróxidos <10 mEq O_2/kg.

Sin embargo, utilizando los datos 1H LF-NMR para el mismo fin, no se obtuvieron resultados relevantes para el objetivo perseguido. Esto podría estar debido a la información química ofrecida por cada una de las dos señales analíticas. Por un lado, el grado de acidez indica la cantidad de ácidos grasos libres presentes en el aceite debida a la descomposición de los triglicéridos, mientras que el índice de peróxidos revela el grado de oxidación, en especial de los ácidos grasos insaturados que forman parte de los triglicéridos. Las pequeñas diferencias a nivel nuclear debidas a estos cambios en la composición de los aceites pueden estar resultando prácticamente invisibles para la señal 1H adquirida, y, sin embargo, sí sean detectables en la señal 13C NMR. Además, los espectros 13C LF-NMR a pesar de mostrar menor señal/ruido, tienen una mayor resolución espectral de los picos que la conforman, pudiendo proveer por tanto de mayor información química para este análisis.

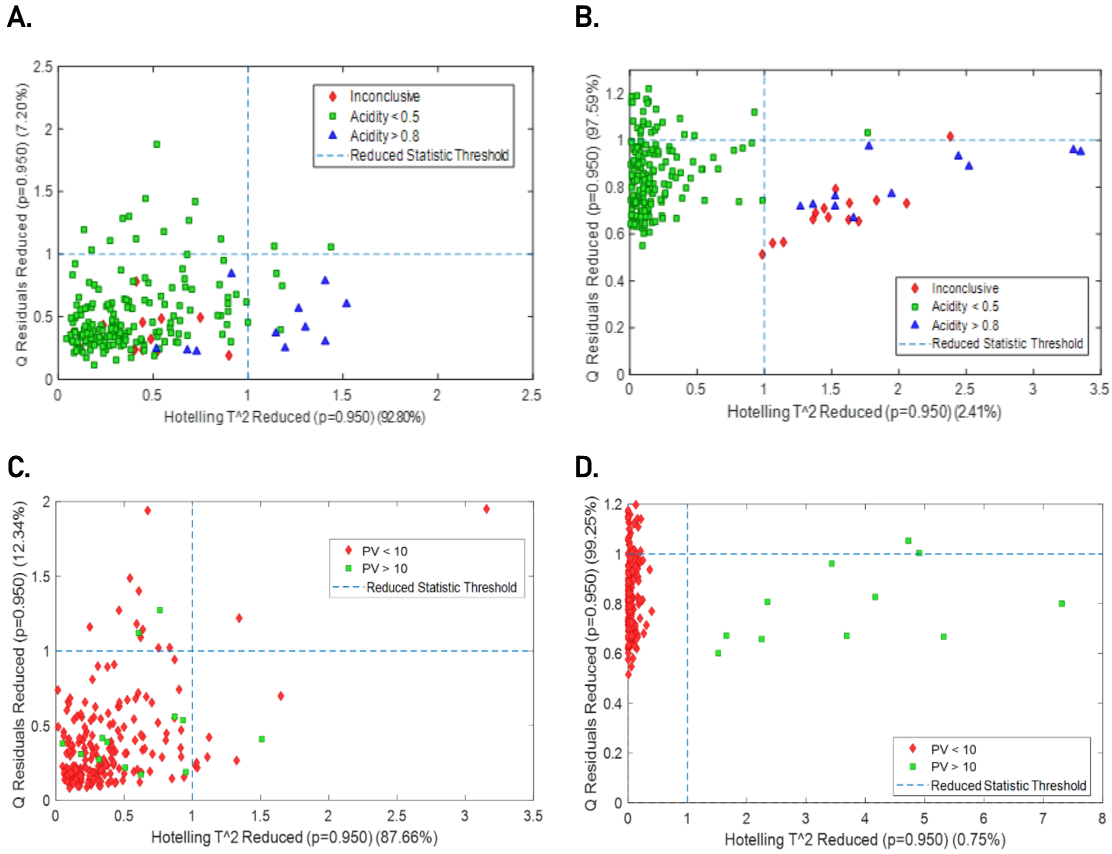


Figura 3.11. Análisis exploratorio con PLS a partir de los datos 1H LF-NMR (A,C) y 13C LF-NMR (B,D) de muestras de aceites de oliva atendiendo a los grupos establecidos para los criterios grado de acidez (A,B) e índice de peróxidos (C,D).

PV = Peroxide value

2.2.2. Proporción de ácidos grasos saturados e insaturados

La proporción de grasas saturadas e insaturadas es un aspecto relevante, debido a la creciente preocupación de los consumidores por el consumo de dichos nutrientes a nivel nutricional y de salud.

Incluyendo todos los aceites vegetales que fueron analizados para este estudio, tanto de oliva como de no oliva, se examinó la posibilidad de predecir la proporción de grasas saturadas a insaturadas presente en cada uno de los aceites, independientemente de la variedad vegetal o categoría comercial. Para ello se utilizó el método PLS y se dividieron los datos en conjunto de entrenamiento y conjunto de validación externa en una proporción aproximada de 75/25 haciendo uso del algoritmo CADEX propuesto por Kennard y Stone [**19**].

La Figura 3.12 ilustra los resultados de los modelos generados a partir de los datos 1H y 13C LF-NMR, donde se muestran a su vez el número de LVs seleccionadas y los diferentes parámetros de calidad calculados: error cuadrático medio de calibración (RMSEC), de validación cruzada (RMSEC) y de predicción (RMSEP), sesgo de calibración, validación cruzada y predicción, y coeficientes de determinación (R^2) de calibración, validación cruzada y de predicción. Las figuras representan los valores predichos por el modelo frente a los valores de referencia proporcionados por el Laboratorio Tello.

Los parámetros de calidad mostraron un buen ajuste en todos los casos, como reflejan los valores del coeficiente de determinación (R^2) para la etapa de calibración o entrenamiento del modelo, los cuales superaron el valor de 0.9 en los modelos generados a partir de datos 13C LF-NMR, y cercanos al 0.9 con los datos 1H LF-NMR.

En la etapa de validación, los resultados de los modelos 1H LF-NMR mostraron una mejor predicción de ambos parámetros (proporción de ácidos grasos saturados e insaturados) con respecto a los 13 LF-NMR. El error de predicción (RMSEP) de estos últimos fue casi tres veces superior, demostrando un posible sobreajuste en el desarrollo del modelo (etapa de entrenamiento).

Los siguientes estudios estarán encaminados a mejorar la etapa de pre-procesado de los espectros 13C LF-NMR para el desarrollo de modelos supervisados a partir de estos datos, de forma que estos no se vean influenciados por la baja señal/ruido que presentan las señales adquiridas.

[**19**]. Kennard, R.W.; Stone, L.A. Computer aided design of experiments. *Technometrics* **1969**, *11*, 137-148. DOI: 10.2307/1266770.

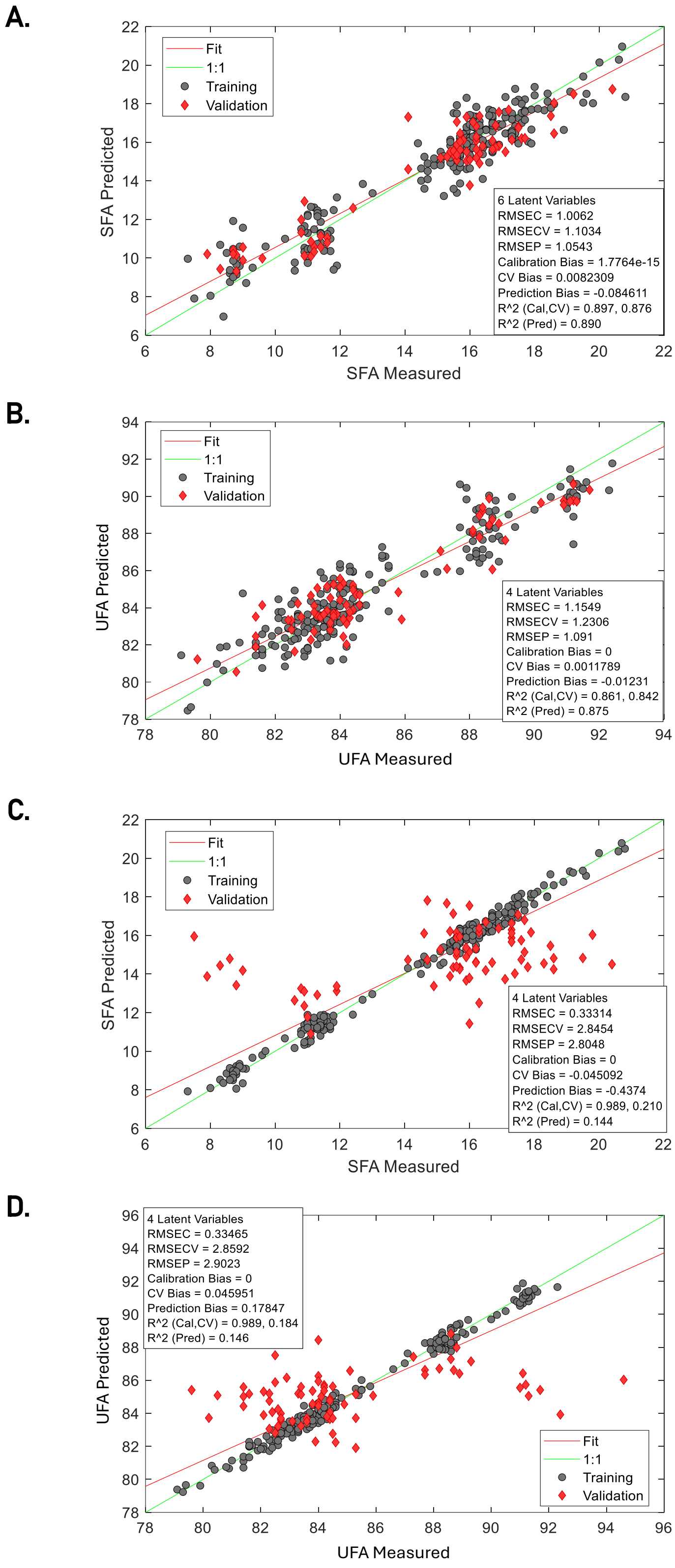


Figura 3.12. Modelos PLS desarrollados para la predicción del contenido en ácidos grasos saturados (A,C) e insaturados (B,D) a partir de datos 1H LF-NMR (A,B) y 13C LF-NMR (C,D) de aceites vegetales.

SFA = saturated fatty acids; UFA = unsaturated fatty acids

3. Conclusiones

En conclusión, los resultados preliminares presentados en este estudio muestran un potencial destacable de la técnica LF-NMR en el desarrollo de un método analítico de cribado para el control de la calidad y autenticidad de aceites de oliva.

La rápida adquisición de las señales 1H y 13C (30 segundos y 2.5 minutos respectivamente), la mínima cantidad de muestra requerida (300 µL) y la posibilidad de desarrollar un método analítico multiparamétrico basado en un enfoque no-dirigido, hacen de esta técnica analítica una candidata apropiada para el desarrollo de un método analítico rápido y poco invasivo, en línea con los principios de la química analítica 'verde'.

En especial, los resultados obtenidos del análisis exploratorio mediante PLS de los datos 13C LF-NMR muestran la plausibilidad de implementar un método de cribado para poder detectar aquellos aceites que cumplen con los requisitos de ser etiquetados como EVOO/VOO en términos de grado de acidez e índice de peróxidos basado en esta técnica. Esto supondría agilizar los análisis rutinarios en laboratorios de control, ya que evitarían analizar estos dos criterios de calidad de un gran número de muestras con la seguridad de que cumplen con los límites recogidos en la legislación. Para este fin es necesario que el modelo quimiométrico desarrollado sea sensible y preciso para la clase objetivo.

De forma complementaria, se obtuvieron resultados satisfactorios en la predicción de la proporción de grasas saturadas e insaturadas en aceites vegetales usando las señales 1H LF-NMR.

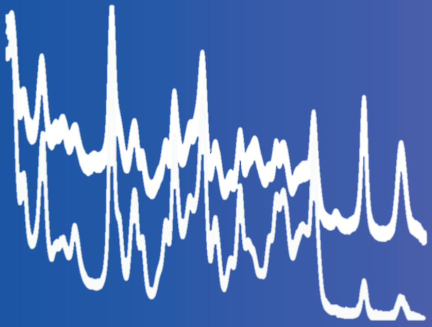
Este estudio continuará progresando y ampliándose para desarrollar modelos predictivos con un mejorado rendimiento analítico, utilizando para ello un mayor número de muestras de aceite, así como explorando otras herramientas quimiométricas, tanto en la etapa de pre-procesado como en la de modelado.

3.5. Contribuciones a congresos

1. A. Arroyo Cerezo, A.M. Jiménez Carvelo, L. Cuadros Rodríguez, J.L. Hidalgo Ruíz, M. Tello Liébana, J.A. Tello Jiménez, J.R. Belmonte Sánchez, A. Garrido Frenich. **Autentificación de aceites de oliva virgen mediante resonancia magnética nuclear de bajo campo (LF-NMR).** [Póster]. *EXPOLIVA XXI Simposio Científico-Técnico. Jaén (España), mayo 2023.*
2. A. Arroyo Cerezo, A.M. Jiménez Carvelo, L. Cuadros Rodríguez, P. Berzaghi. **Study of portable NIR instruments for virgin olive oil quality assurance.** [Oral 10']. *Second SensorFINT International Conference. Berlín (Alemania), junio 2023.*
3. A. Arroyo Cerezo, A.M. Jiménez Carvelo, M. Pellegrino, X. Yang, A.F. Savino, P. Berzaghi. **Development of a portable, low-cost and non-invasive screening of virgin olive oils quality by NIR.** [Póster]. *21st International Conference on Near Infrared Spectroscopy (NIR 2023). Innsbruck (Austria), Agosto 2023.*
4. A. Arroyo Cerezo, A.M. Jiménez Carvelo, L. Cuadros Rodríguez. **The potential of a benchtop LF-NMR equipment for food authentication assessment by a non-targeted approach.** [Póster]. *SensorFINT Final Conference 2024. Córdoba (España), mayo 2024.*
5. A. Arroyo Cerezo, A.M. Jiménez Carvelo, L. Cuadros Rodríguez. **Método de cribado para el análisis de aceites de oliva.** [Oral 5']. *V Congreso / VII Jornadas de Investigadores en Formación Fomentando la Interdisciplinariedad (JIFFI 2024). Granada (España), junio 2024.*

CAPÍTULO 4

Quimiometría y química analítica alimentaria: Aplicaciones innovadoras



4.1. Presentación

La inteligencia artificial está revolucionando la sociedad moderna, y el ámbito científico no es una excepción, habiéndose convertido en una parte indispensable de todas las áreas de la investigación [1]. Sin embargo, la verdadera novedad introducida en los últimos años es el nombre con el que se la conoce, pues ya en el siglo XX nació la quimiometría como ciencia que aplica algoritmos matemáticos y estadísticos, propios de la conocida hoy como inteligencia artificial (IA) y minería de datos (*data mining*), al tratamiento y análisis de datos químicos [2].

Tal y como versa el Capítulo 1, la quimiometría se ha convertido en una parte esencial de la química analítica alimentaria actual, ayudando a abordar los problemas complejos mediante el tratamiento y análisis estadístico y matemático de los datos [2]. Desde sus inicios, ha avanzado a pasos agigantados en su aplicación al diseño de experimentos y la optimización de procesos, la identificación de estructuras y patrones de comportamiento comunes entre los datos y la generación y validación de modelos de aprendizaje automático. La irrupción de la IA ha permitido la inclusión de métodos basados en aprendizaje automático y profundo (*machine learning* y *deep learning*, respectivamente) para el desarrollo de modelos no lineales y más complejos, ampliando así la quimiometría tradicional [3,4].

En esta línea se enmarca el presente capítulo, que recoge dos estudios haciendo uso del análisis de la quimiometría para resolver problemas surgidos en la investigación llevada a cabo durante la tesis doctoral.

Antes de ello, se presenta una publicación en forma de capítulo de libro que aborda una parte fundamental de la quimiometría: el análisis de la similitud. A pesar de las numerosas aplicaciones y utilidades que la involucran, esta estrategia no ha recibido la atención que merece dentro de la quimiometría. Este capítulo de libro pretende sentar las bases que definen dicha estrategia, y asimismo servir de guía sobre los enfoques y la forma de aplicar e interpretar los resultados que arroja.

[1] Tseng, Y.J.; Chuang, P.J.; Appell, M. When machine learning and deep learning come to the big data in food chemistry. *ACS Omega* **2023**, *8,* 15854-15864. DOI: 10.1021/acsomega.2c07722.

[2] Joshi, P.B. Navigating with chemometrics and machine learning in chemistry. *Artif. Intell. Rev.* **2023**, *56*, 9089-9114. DOI: 10.1007/s10462-023-10391-w.

[3] de Carvalho Rocha, W.F.; do Prado, C.B.; Blonder, N. Comparison of chemometric problems in food analysis using non-linear methods. *Molecules* **2020**, *25*, 3025. DOI: 10.3390/molecules25133025.

[4] Rial, R.C. AI in analytical chemistry: Advancements, challenges, and future directions, *Talanta* **2024**, *274*, 125949. DOI: 10.1016/j.talanta.2024.125949.

El manuscrito que se presenta formará parte del libro titulado "Problem-Oriented Analytical Chemistry Driven by Chemometrics" que será publicado en el año 2025 por la editorial Elsevier (ISBN: 9780443221637).

A continuación, el primer estudio presentado se encuentra estrechamente relacionado con el Capítulo 2, ya que surge como consecuencia del uso de la técnica analítica SORS. Dicha técnica permite recoger la contribución espectral de capas subsuperficiales del material medido, resultando en una señal espectral que es una mezcla de contribuciones superficiales y subsuperficiales. Los espectrómetros actuales que aplican esta variante de la espectrometría Raman incorporan en el propio software una etapa de corrección/resolución de estos espectros, diseñada para eliminar la contribución espectral de las capas superficiales. No obstante, este proceso automatizado demostró no ser óptimo dependiendo del material analizado, lo que plantea un desafío que podría ser abordado desde diferentes enfoques: la optimización del proceso de medida, o la optimización del procesado de la señal tras ser adquirida. Dadas las especificaciones técnicas del equipo instrumental bajo uso, la primera estrategia no podía ser abordada, por la imposibilidad de modificar los parámetros instrumentales de interés, a diferencia de una plataforma construida en el laboratorio (*homemade*). Por lo tanto, se planteó la hipótesis de mejorar la corrección de los espectros adquiridos mediante SORS a través de métodos quimiométricos de resolución de señales.

La comprobación de esta hipótesis dio lugar a un estudio realizado en colaboración internacional que derivó en la publicación del siguiente artículo científico en una revista de alto impacto, cuya referencia es:

1. Chemometric enhancement for blind signal resolution from non-invasive spatially offset Raman spectra. *Chemom. Intell. Lab. Syst.* **2023**, *243,* 105027. DOI: 10.1016/j.chemolab.2023.105027.

En el estudio presentado a continuación de este, se hace uso de la metodología de diseño de experimentos (DoE) para optimizar los parámetros instrumentales para la adquisición de señales LF-NMR de la máxima calidad y en el menor tiempo de análisis. Este estudio surge como parte inicial del proyecto de colaboración público-privada NMR-Control (CPP2021-008672). La generación de nuevos métodos analíticos rápidos y poco invasivos siguiendo una estrategia basada en el uso de huellas instrumentales requiere una etapa inicial para la optimización de la fase experimental. Esta necesidad se acrecienta en el caso abordado, dada la novedad de la instrumentación empleada: un equipo de sobremesa basado en la espectrometría de resonancia magnética nuclear de baja frecuencia de campo (LF-NMR). Dicho estudio conlleva dos innovaciones reseñables: (i) el uso de la metodología Taguchi para el diseño experimental y (ii) la aplicación de la teoría de la información como estrategia para evaluar la calidad informativa de una señal analítica.

La ejecución de este trabajo de investigación derivó en la escritura de un manuscrito que ha sido enviado para su publicación en una revista de alto impacto y que se encuentra bajo revisión. El título provisional del citado trabajo es el siguiente:

2. Optimising the acquisition conditions of high information quality low-field NMR signals based on a cutting-edge approach applying information theory and Taguchi's experimental designs - Virgin olive oil as an application example. *Anal. Chim. Acta.* [Bajo revisión].

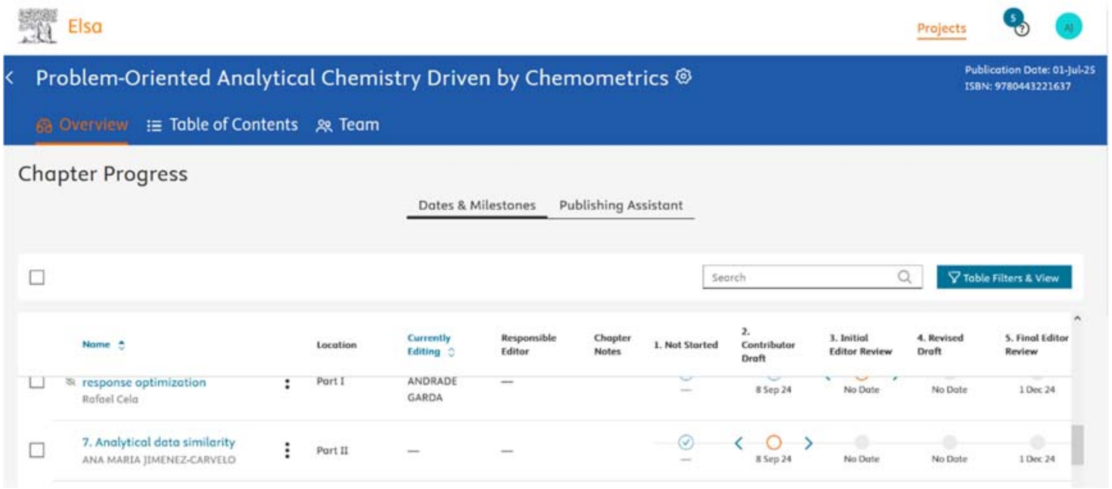
4.2. Capítulo de libro 2

Chapter 7. Analytical data similarity.

Ana M. Jiménez-Carvelo✉, Alejandra Arroyo-Cerezo, Esteban A. Roca-Nasser, Luis Cuadros-Rodríguez

En el libro: Problem-Oriented Analytical Chemistry Driven by Chemometrics, Cuadros-Rodríguez, L., Jiménez-Carvelo, A.M., Andrade-García, J.M., Eds. Elsevier, 2025.

ISBN: 9780443221637.



Problem-Oriented Analytical Chemistry Driven by Chemometrics

Publication Date: 01-Jul-25
ISBN: 9780443221637

Chapter Progress

	Name	Location	Currently Editing	Responsible Editor	Chapter Notes	1. Not Started	2. Contributor Draft	3. Initial Editor Review	4. Revised Draft	5. Final Editor Review
□	response optimization Rafael Cela	Part I	ANDRADE GARDA	—		—	8 Sep 24	No Date	No Date	1 Dec 24
□	7. Analytical data similarity ANA MARIA JIMENEZ-CARVELO	Part II	—	—		✓ —	○ 8 Sep 24	No Date	No Date	1 Dec 24

Capítulo de libro enviado a la editorial Elsevier (29-07-2024)

✉ Corresponding author (e-mail: amariajc@ugr.es)

Abstract

The term "similarity" is difficult to define precisely due to its wide applicability. In data science, similarity analysis is crucial, from basic statistics to complex deep learning analytics. For more than a century, it has been used in various scientific fields, such as ecology, finance, bioinformatics and social sciences, to assess relationships between variables. Modern analytical chemistry makes use of similarity analysis for the interpretation of large datasets and is thus an essential part of data mining and machine learning. However, to date, similarity analysis has rarely been referred to with the importance it deserves within chemometric tools. This chapter lays the fundamental basis for similarity analysis as an indispensable part of analytical chemistry, detailing the different strategies for its application, including exploratory data analysis and pairwise comparison of analytical signals.

Keywords:

Dissimilarity

Similarity indices

Exploratory data analysis

Distance

Correlation

Orientation

Entropy

Contents

1. Introduction

The term similarity poses an outstanding challenge for its precise definition due to its wide variety in contextual applicability. The Oxford English Dictionary (OED) (https://www.oed.com/) defines it as the state or fact of being similar in some way, likeness, resemblance. Hence, similarity analysis could be defined as the study of the degree of likeness or resemblance between two or more material systems concerned.

Particularly in the world of data science, similarity analysis plays a very important role at all levels: from the most basic statistics to the most complex deep learning method [1]. For over a hundred years, similarity analysis has been carried out by means of the so-called association metrics in a multitude of scientific disciplines such as ecology, financial data analysis, bioinformatics, or social sciences to assess the relationship between continuous, ordinal, or categorical variables. With the advancement of data science, similarity analysis has grown as an indispensable part of data mining and machine learning with application in several fields where the analytical chemistry is present [2].

It should be noted that many synonyms are used for similarity: resemblance, correlation, association, relationship, interestingness, comparison, etc. [2]. Furthermore, it is also quite usual to find publications referring indistinctly to their antonyms such us dissimilarity or divergence, among others. This aspect will be dealt with later, although throughout the chapter the term SIMILARITY will be referred to exclusively to avoid a terminological confusion, and sometimes will be referred as DISSIMILARITY because of the terminological duality of both words.

In the field of modern analytical chemistry, the search for meaningful information from vast datasets has led to the evolution of sophisticated mathematical and statistical methodologies. Among these, similarity analysis stands out as a fundamental tool for devising intricate relationships among analytical signals. According to Chapter 2, a two-dimensional analytical signal characterised by two types of features (location and intensity) is essentially a first-order tensor embedded in a vector containing the intensities corresponding to the acquired analytical quantity. Thus, these signals can be described as a set of *n*-tuples $Y = (yi, \ldots, yn)$, such for all $i = 1, \ldots, n$ the elements in Y take values in a set of real values, nonnegative values, binary values, etc [1].

When dealing with a dataset consisting of a series of first-order tensors, each of them being the data profile or the analytical signal of an object, then a two-way array is formed. This contains as many rows as there are objects (or samples) in the dataset and as many columns as there are variables describing each object within the analytical signal. Particularising the definition of similarity to this type of data (data profiles), several applications and ways of applying this similarity analysis can be found. Broadly speaking, these can be divided into: (i) similarity between objects or variables, (ii) similarity between an object and a reference and (iii) similarity between two sets of objects or variables [3]. This classification based on the objective pursued with the similarity analysis can be gathered in another classification based on the mathematical approach to data handling, split into 'comparison of two-by-two objects' and 'comparison of sets of objects'. The first approach encompasses objectives (i) and (ii), for which different similarity metrics (or similarity indices) are often used to address them. While the second approach encompasses objective (iii), and it is widely known in chemometrics as exploratory data analysis, screening or cluster analysis [4].

Similarity analysis is widely used in analytical chemistry field following both aforementioned approaches. However, it is usual to find it as a supporting step for the pursued goal and less common to be the main part of the study, process, or method. For example, as an initial part of the study, similarity analysis is often applied to the analytical profile obtained to ensure proper measurement conditions, adequate or expected measurement, or to assess the stability (or repeatability) of the measurements [5]. Depending on the analytical technique to be used, and the type and condition of the sample, it is possible to have at hand a reference of the target analytical signal to be achieved. This reference may come from a database or from other equipment with better performance than the one to be used for the analysis or the method to be developed. Matching the acquired signal with a database or any other reference can also be applied to analyse the analytical information obtained. That is, as part of the goal of the analysis it is possible to identify, determine or even characterize the sample according to the chemical information collected in the analytical profile through a similarity analysis. Examples of this can be found in different fields of application such as pharmacology [6,7], biochemistry [8] and other chemical industries [9,10].

It is also common to conduct a similarity analysis for the establishment of hypotheses that will then be the starting point for the development of machine learning models. In these cases, the final objective can be very diverse: process monitoring, authentication or quality control, among others. The focus of these studies lies more on the chemical information derived from the analytical signal rather than its quality, as seen in the previous examples. All fields in which analytical chemistry is applied can benefit from this type of similarity analysis applications [11,12,13,14,15]. However, specific examples related to the aspects covered will be detailed throughout the next sections of the chapter.

Hence, it is evident that similarity analysis is a meaningful part of chemometrics that has never been given the limelight it merits. This chapter focuses exclusively on detailing the various strategies that can be followed to perform a similarity analysis and which may be applied in different fields of analytical chemistry. Commonly, the strategy followed in chemometrics involves representing each object or sample as a point in a two- or three-dimensional space, even though the object is described by a set of hundreds or thousands of variables (analytical signal profile). This approach is used to compare sets of samples and is widely recognized as exploratory data analysis. Section 3 of this chapter is entirely devoted to it. Nonetheless, there are many applications where the aim is to precisely determine the similarity between two analytical signals by comparing variable-to-variable. This pertains to assessing the similarity of pairs of signals and will be addressed in section 2 of this chapter.

2. Similarity analysis on pairs of analytical signals

A two-dimension analytical signal can be defined by a combined system of two analytical quantities. As described in Chapter 2, these are (i) the location or position of the signal (spatial, spectral, temporal, etc., this will depend on the coupled analytical technique and measuring device used), and (ii) the intensity of the signal, intended as the magnitude of the energy acquired and transduced by the detector/sensor, in the form it takes, from the analytical response. Consequently, two analytical signals can be dissimilar in one or both quantities.

The position-derived dissimilarities could be caused by distortions, i.e., phenomena of shifting, warping, shortening, or lengthening of the signal. While the intensity-derived dissimilarities between two analytical signals refer to higher/lower analytical responses, different sensibility level, signal noise, etc. These two cases of dissimilarity between two analytical signals could also be interpreted as horizontal and vertical displacements, respectively, considering the analytical signal profile representation. To simply illustrate these cases of dissimilarity, let **A**, **B** and **C** be three intensity vectors of 12 elements, where **A** = [0, 0, 1, 11, 3, 1, 2, 6, 10, 9, 0, 0]; **B** is characterised by one position shifted with respect to **A**, and **C** is calculated by multiplying each element of vector **A** by a non-constant factor which is a linear function of the intensities. Vector **C** could represent an amplified analytical signal under higher sensitivity instrumental conditions compared to the first vector, **A**, but maintaining the signal-to-noise ratio. The analytical profile of these vectors is illustrated in Figure 4.1A and 4.12B.

However, a third kind of dissimilarity source when measuring the similarity of a pair of signals can be found. The shape of the profile may differ on a single peak, band, shoulder, or other specific feature of the instrumental signal. This should not be confused with the second type of dissimilarity source expressed in Figure 4.1B, where the entire profile of the signal is observed to be different, but rather in this third case only one region of the signal differs from the other signal. Figure 4.1C illustrates this third case of dissimilarity between vectors **A** and **D**.

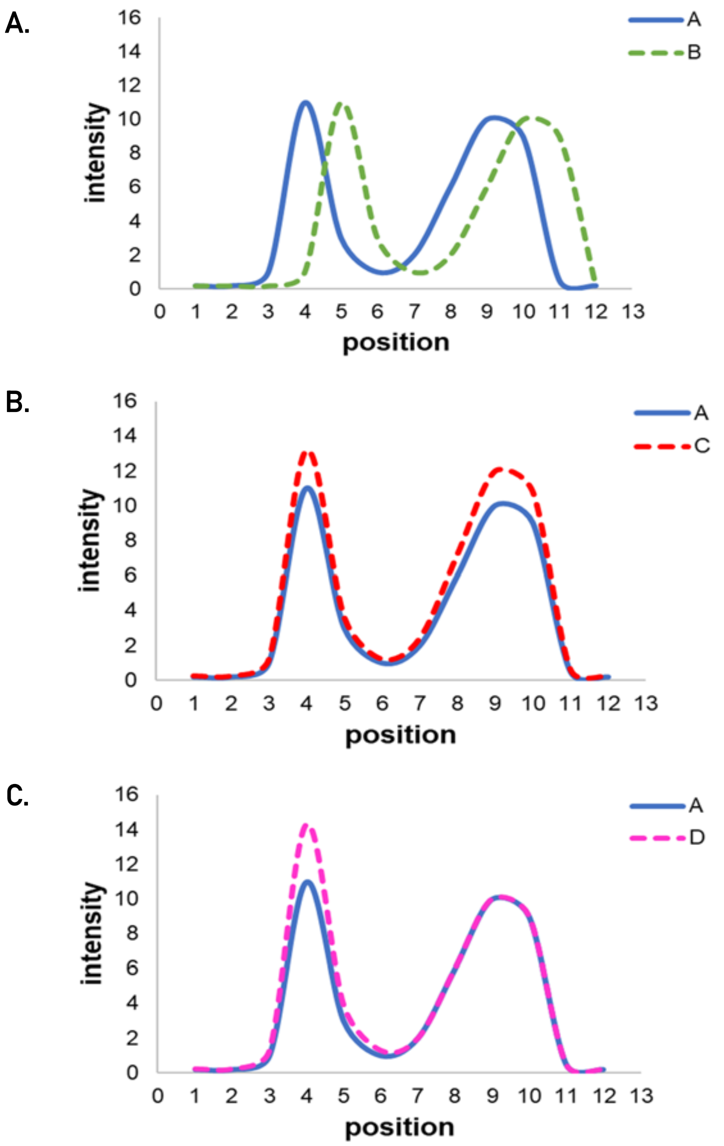


Figure 4.1. Analytical profile of vectors characterised by 12 elements (or variables) length. Each of them represents different kinds of dissimilarity source, see the text for further information.

The type of dissimilarity causing differences between two analytical signals depends on many factors, which characterize the multitude of applications for which similarity analysis can be used. These applications can be focused on the analytical measurement process (i.e., at the instrumental level to assess the quality of the signal acquired, either by matching it with a reference, or to determine the stability of the same measurement under different environmental or temporal conditions, i.e., the repeatability of the measurement [**5,16**]), on the chemical information provided by the signal (to compare how similar are two samples that have characteristics in common [**17**], to authenticate a sample by comparing it with a reference [18], to identify chemical compounds [10], among many others), or on the data treatment performed (to compare with a reference signal after applying pre-processing such as signal alignment [19], or after performing a signal resolution, etc. [20]).

Several strategies have been proposed to measure how close/far are both signals according to their features, thus leading to similarity/dissimilarity analysis. However, the lack of consensus and the scarce importance given over the years to the classification of these strategies is striking. Some authors agree in using the term distance when referring to any measure of dissimilarity, whatever the mathematical form of the function used [3,21]. Others have divided them into two main groups according to the differences measured: those based on spectral amplitude and those based on the shape of the spectrum [22]. While other authors refer broadly to different strategies, but without providing a concise classification [23,24,25].

Therefore, it has been decided here to present the following proposal for the classification of the strategies to be followed to perform a similarity analysis between pairs of analytical signals, namely: approaches based on (i) distance, (ii) spatial orientation, (iii) correlation and (iv) information entropy.

At this point, it should be noted that a dissimilarity metric can be easily transformed into a similarity one. It is even simpler when the calculation takes a normalised value between 0 and 1, since subtracting 1 from one of them (similarity or dissimilarity metric), results in the other one. However, this chapter is focused on the analysis of similarity, understanding that in almost all the applications that are to be applied, what is sought is to know how similar the compared objects are, and therefore the results are expressed in similarity metrics. Even so, many of the strategies presented here focus on measuring dissimilarity. Therefore, throughout the chapter, the terms similarity and dissimilarity will be used, and hints on how to transform one into the other will be provided where appropriate.

As mentioned above, the concept of dissimilarity is often equated with the concept of distance. However, the dissimilarity term is a broader concept. Distance can be defined as the length of the space between two points. It is a number describing how large the space separating two points is [3]. The length of the straight line joining the two points is known as the Euclidean distance [23]. This is the simplest definition of distance used in everyday life in many aspects. Euclidean distance is a very easy concept to understand when considering two points in a two- or three-dimensional space (Cartesian coordinates). But this representation of an analytical signal defined by more than 3 variables such as a chromatogram or a spectrum is not feasible. In these cases, the distance is calculated in pairs between all the variables that comprise the analytical signal.

Nevertheless, there are other functions for calculating distance which do not follow Euclidean geometry. In fact, the Euclidean distance is a part of a group of functions known as the distance-metrics. Note that in mathematics the terms distance and metric could be used as synonymous, but to avoid confusion, the term metric will be avoided in this context and referred to as distance.

To be called distance between two objects ***A*** and ***B***, each represented by a vector that is indirectly a point in n-dimensional space, the measurement must satisfy the axioms:

1) nonnegativity: $d_{AB} \geq 0$
2) reflectivity $d_{AA} = 0$; $d_{BB} = 0$
3) distinguishability: $d_{AB} = 0$ only if $\boldsymbol{A} = \boldsymbol{B}$
4) symmetry: $d_{AB} = d_{BA}$
5) triangularity: $d_{AB} \leq d_{AC} + d_{BC}$

where ***A***, ***B*** and ***C*** are objects in the n-dimensional space, and d refers to the calculated distance [23,26,27].

Axiom 5 is rarely satisfied, and in fact a metric can be called distance when the first four axioms are fulfilled, while the fifth must be fulfilled only to be called metric [3].

Distances could be divided into two groups: unbounded and bounded distances. The former refer to those which after calculation provide values from zero to infinity, while the bounded distances have a maximum value they can reach. It should be noted that for similarity analysis it is highly recommended to use a bounded distance if it is used to compare between several objects; otherwise, the analyst should set a suitable threshold to differentiate between similarity and dissimilarity. However, a simple straightforward function can be applied to transform the unbounded distance into a similarity measure [3]. The simplest way is:

$$\mathrm{s_{AB}} = \frac{1}{1 + \mathrm{d_{AB}}}$$

where s_{AB} is the similarity metric between objects ***A*** and ***B***.

It is also possible to apply a normalisation of a distance-metric in order to get somewhere between zero and an upper bound, e.g., zero to one. This normalisation can be performed by dividing the result by the maximum value that the distance between the two objects can reach.

The Euclidean distance, together with the Manhattan or city-block distance and the Lagrange distance, belong to the family of distances known as Minkowski. Figure 4.2 shows a geometric representation of the three distances for two points in a two-dimensional Cartesian coordinate space. The general formula is the Minkowski formula and is as follows:

$$\mathrm{d_{AB}(Minkowski)} = \left[\sum_{i=1}^{n} |a_i - b_i|^q\right]^{1/q}$$

where *a* and *b* (in lower case) indicate the elements of the vectors ***A*** and ***B***, n is the number of variables coincident with the dimensions of ***A*** and ***B***; q is a parameter assuming a value of 1, 2 and infinity for Manhattan, Euclidean and Lagrange, respectively.

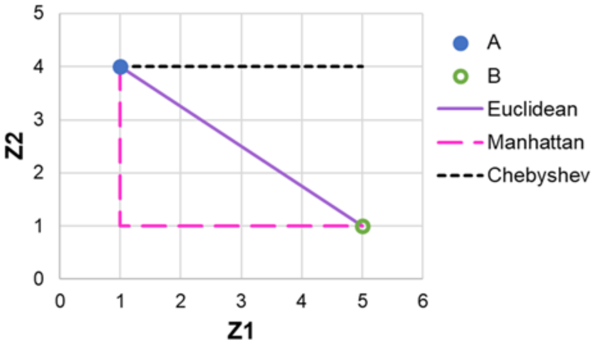


Figure 4.2. Geometric description of the three distances (Euclidean, Manhattan and Lagrange) between two points in a two-dimensional space.

Hence, the calculation of the distance between two analytical signals is one way to estimate the dissimilarity, but it is not the exclusive. Another strategy to determine it is based on calculating the spatial orientation. As with the distance between two points, this calculation becomes simpler to explain in a two-dimensional space. It is based on reporting the angle formed by the two vectors (where each vector is defined as the straight line drawn from the origin of the coordinates system to the point representing the signal). These two signals will be totally similar when the angle is 0°, and totally dissimilar when the angle is 90°. Most widely used strategies to calculate the spatial orientation are the cosine and the angle value itself (expressed in radians), typically calculated as arccosine from cosine. Although it has been given other names in literature, such as spectral angle mapper (SAM) **[20,24,28]**, in the end the applied function is nothing more than calculating the cosine of the angle formed by the two vectors. Some publications refer it to as cosine distance (after subtracting the unit), although the term *distance* should be used cautiously here since it does not strictly measure the distance between two points but rather the spatial orientation between two vectors by calculating the inner product.

Some outstanding advantages of the cosine calculation are that it is bounded between 0 and 1, and it directly measures similarity. The closer the value is to 1, the more similar the objects being compared are, and vice versa: the closer the value is to 0, the more different they are. However, it should be noted that due to the sinusoidal profile of the function, the cosine calculated as a similarity metric is less sensitive to minor differences when the compared objects are very similar. In these situations, it is advisable to use the angle itself instead, which varies more linearly with the angle [29]. Figure 4.3 illustrates the values of cosine and arccosine versus the angle formed by two vectors characterized by two points in a two-dimensional space. The values acquired up to the maximum separation between the two vectors in the first quadrant, i.e. with positive values, are shown, since all analytical signals are characterized by positive intensity values.

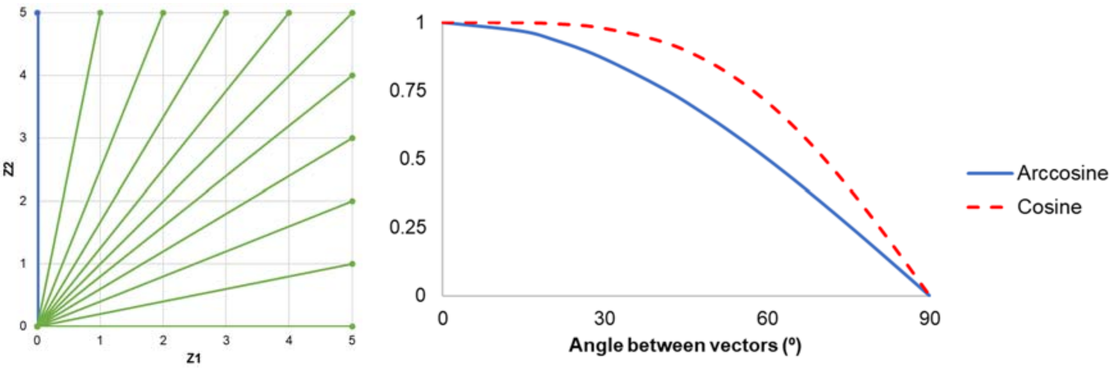


Figure 4.3. Values obtained from cosine and arccosine calculations between two points in a two-dimensional space, depicted at various orientations.

The third group of strategies to calculate the similarity between two vectors is the correlation. Correlation measures the lineal association between variables, but it can be used for similarity analysis in pairs of analytical signals. The Pearson coefficient is the most common function to be applied when calculating correlation, and it is widely known it can adopt values between - 1 and 1, where the negative values show a negative correlation, i.e., an indirect proportion. Therefore, the correlation calculation would be in breach of the first axiom to be called distance, although some publications refer to it as a correlation distance [3]. Correlation reports about the extent of association between the profile of objects, so it does not measure systematic differences between objects. It is therefore a direct measure of similarity, unlike the distances which measure dissimilarity or divergence [3]. To avoid negative values, it is common to use the square of Pearson's coefficient, also known as the coefficient of determination, as an alternative. The correlation metric could be described as calculating the cosine between two normalised vectors. This explains why the result values of both (correlation and cosine) are usually practically identical.

The last group of the classification of strategies for measuring similarity arises from the information theory or mathematical theory of communication, developed by Claude Shannon in the 1940s [30]. The proposal is based on entropy calculation to quantify the information and the amount of uncertainty in a message, so that the information can be transmitted efficiently. This theory has mainly been applied to computer sciences, telecommunications, cryptography, and statistics, among other areas. However, a growing number of applications are to be found in other, very different areas, and one of them is closely related to analytical signals. Based on this theory, several similarity metrics have been developed. They rely on the calculation of the known as Shannon entropy (H_S), defined as:

$$H_S(\mathbf{A}) = -\sum_{i=1}^{n} p_{a,i} \log(p_{a,i}) \ ; \qquad p_{a,i} = \frac{a_i}{\sum_{i=1}^{n} a_i}$$

where $p_{a,i}$ indicates the probability associated with each of the n elements of vector A, and it is calculated by dividing each value by the sum of all, i.e. applying a total sum normalisation (TSN) [31]. Notice that although Shannon proposed the use of binary logarithms (in base 2), it is possible to use other types of logarithms (decimal, natural, …).

From this calculation, Kullback and Leibler introduced a divergence metric, expressed as:

$$KL(\mathbf{A}, \mathbf{B}) = -\sum_{i=1}^{n} p_{a,i} \log\left(\frac{p_{a,i}}{p_{b,i}}\right)$$

so, the higher the value of KL, the greater the divergence, and therefore, the lower the similarity between the two compared objects ***A*** and ***B***. However, KL divergence presents an asymmetry issue. To address this disadvantage, Jeffreys proposed a symmetric version defined as:

$$J(A, B) = KL(A, B) + KL(B, A)$$

Despite this, J has the disadvantage of being an unbounded metric, thus making interpretation more complex as stated previously. As a solution, the Jensen-Shannon divergence (JS) was proposed in 1991:

$$JS(A, B) = H_S(\eta_1 A + \eta_2 B) - \eta_1 H_S(A) - \eta_2 H_S(B)$$

JS divergence is defined for the comparison between probability distributions, it also has the advantage that weights can be assigned to each of the distributions to be compared (η_1 and η_2) [32]. This calculation can be easily transferred to the comparison of analytical signals, using the equation for the calculation of the Shannon entropy (H_S) or the KL divergence shown previously [20].

Besides, other types of metrics have been defined based on information theory as well, such as:

$$s_{AB} = \frac{\sum_{i=1}^{n} \left(p_{a,i} / p_{b,i} \right)}{n}$$

where s_{AB} measures the similarity between the relative abundances, p_i, of two signals ***A*** and ***B***, being *n* the total number of variables of each [33,34]. The calculation of relative abundances applied to analytical signals is simply dividing each analytical response value by the sum of all values, also known as total sum normalisation [31].

Generally, there are no missing values in the analytical signals, i.e., when a spectrum, chromatogram, thermogram, etc., is acquired, every position has an intensity value, even if this is a zero. However, other fields such as biology usually grapple with missing values, e.g. when a characteristic is present or not. This situation extends to analytical chemistry, particularly in the context of metabolic profiling, where it is increasingly common to encounter features whose values are presence or absence. Binary dissimilarity metrics are used for these scenarios [23,35]. In addition, in some analytical techniques it is a common practice to binarize the signals obtained. This is a mathematical transformation of the response values that characterise the signal to binary values, generally 0 and 1, to simplify the complexity of the data or to highlight features of interest. Examples of this practice could include mass spectra, MALDI images, even chromatograms, spectra or in microscopy [36,37,38].

The above-mentioned strategies for measuring similarity variable-to-variable between analytical signal pairs can be used in two ways. In the case of comparing two analytical signals contained in a data vector, i.e., 2D signals: If the average result of all the similarity metric values of each variable is calculated, a scalar is obtained, known as the similarity index. While the vector obtained by comparing two 3D analytical signals is considered a similarity curve or profile. Analytical signals of higher dimensionality can give rise to similarity surfaces or images. The following sections are devoted to similarity indices (2.1) and similarity curves and surfaces (2.2).

2.1. Similarity indices

A similarity index is defined by the IUPAC as a *quantity that describes the equivalence of two objects characterised by multivariate data* [39]. The same definition refers to other terms such as Ward's minimum variance method, Tanimoto similarity index, and city-block (Manhattan), Euclidean and Mahalanobis distances. However, it should be noted not all of them are strictly similarity indices. The Ward's minimum variance method is actually a criterion used in cluster analysis to minimise variance [40]. Tanimoto similarity index is based on the Jaccard coefficient, also known as Soergel distance, when it is used as dissimilarity index. This calculation is most used for binary data, but it is possible to express it in a form that can be applied to continuous data [41]. While the other three mentioned terms are distances (city-block, Euclidean and Mahalanobis), so they can be used to calculate a similarity index.

Therefore, a similarity index is a scalar, i.e., a number representing how similar two objects are. To apply this to the comparison of two analytical signals, each characterised by a first-order tensor, it is necessary to obtain an average of the calculated element-by-element similarity. The calculation of a similarity index to compare between pairs of 2D analytical signals can be expressed in both algebraic and matrix form. The algebraic equation expresses the essence of the similarity calculation by defining the relationship between the signal variables. Let consider the cosine calculation as an example of similarity index to be expressed in algebraic form:

$$\mathrm{COS}\ (\mathbf{A}, \mathbf{B}) = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{a_i} \cdot \mathrm{b_i})}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{a_i^2} \cdot \mathrm{b_i^2}}}$$

where ai and bi are the elements of the two vectors ***A*** and ***B*** of n elements describing two analytical signals, and the i subscript denotes each position of the vector elements.

The matrix algebra facilitates the application of the index calculation in the multivariate framework. As already described, a 2D analytical signal is a vector of intensities, therefore, it can be understood as a single row matrix, where each column is a variable (analytical response) of the signal. Keeping the example of cosine, let's now look at the matrix equation for the calculation of this index:

$$\mathrm{COS}\ (\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \times \mathbf{B}^{\mathrm{T}}}{\sqrt{(\mathbf{A} \times \mathbf{A}^{\mathrm{T}}) \cdot (\mathbf{B} \times \mathbf{B}^{\mathrm{T}})}}$$

where the *T* superscript indicates the transpose of the matrix. Note that the dot "·" symbolises a scalar product, while the cross "×" is referring to a vector (or cross) product. The cross symbolised multiplication of a matrix *(1×n)* by the transpose of another matrix *(n×1)* results in a scalar. Thus, the above equation shows the division of two scalar values, which results in a single value to be used as a similarity index.

To achieve this, one assumption is required: the two signals must have the same number of variables (n). However, discrepancies in the number of elements present in each analytical signal pose a major challenge, especially in some techniques such as chromatography. These variations can distort the similarity index calculation. To mitigate this problem, resampling methods are applied to ensure uniformity of the number of elements in all signals.

A similarity index can be calculated by any of the four strategies described in the previous section: distance, spatial orientation, correlation, and entropy information. The list of similarity indices that can be found in scientific literature is innumerable [3,23,41,42]. Even though not many, there are some publications comparing the outcome of different similarity indices to determine which one(s) is/are the best. However, this is not a decision which can be generalised, since it depends on the scenario to be faced: type of analytical signal, type and length of data, objective sought, etc. Several assumptions need to be considered, with particular emphasis placed on the bounded nature of the index. Typically, scenarios requiring signal pair similarity analysis involve not just two signals, but rather a set of signals seeking comparison either against a single reference or among themselves. Hence, opting for a bounded index is deemed more suitable, which facilitates simpler deductions and conclusions. In turn, this simplifies the decision-making process of a threshold at which the two signals can be considered similar.

Most distances are unbounded, but it is possible to convert an unbounded index into a bounded one. The simplest way is to divide by the maximum value that this index can assume between the given vectors. An example of this is the nearness index (NEAR), which is based on the Euclidean distance normalised and converted to a similarity index. It is calculated by the following algebraic equation:

$$\mathrm{NEAR}(\mathbf{A},\mathbf{B}) = 1 - \mathrm{d}^{\mathrm{N}}(\mathbf{A},\mathbf{B}) = 1 - \sqrt{\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{a_i} - \mathrm{b_i})^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{a_i} + \mathrm{b_i})^2}}$$

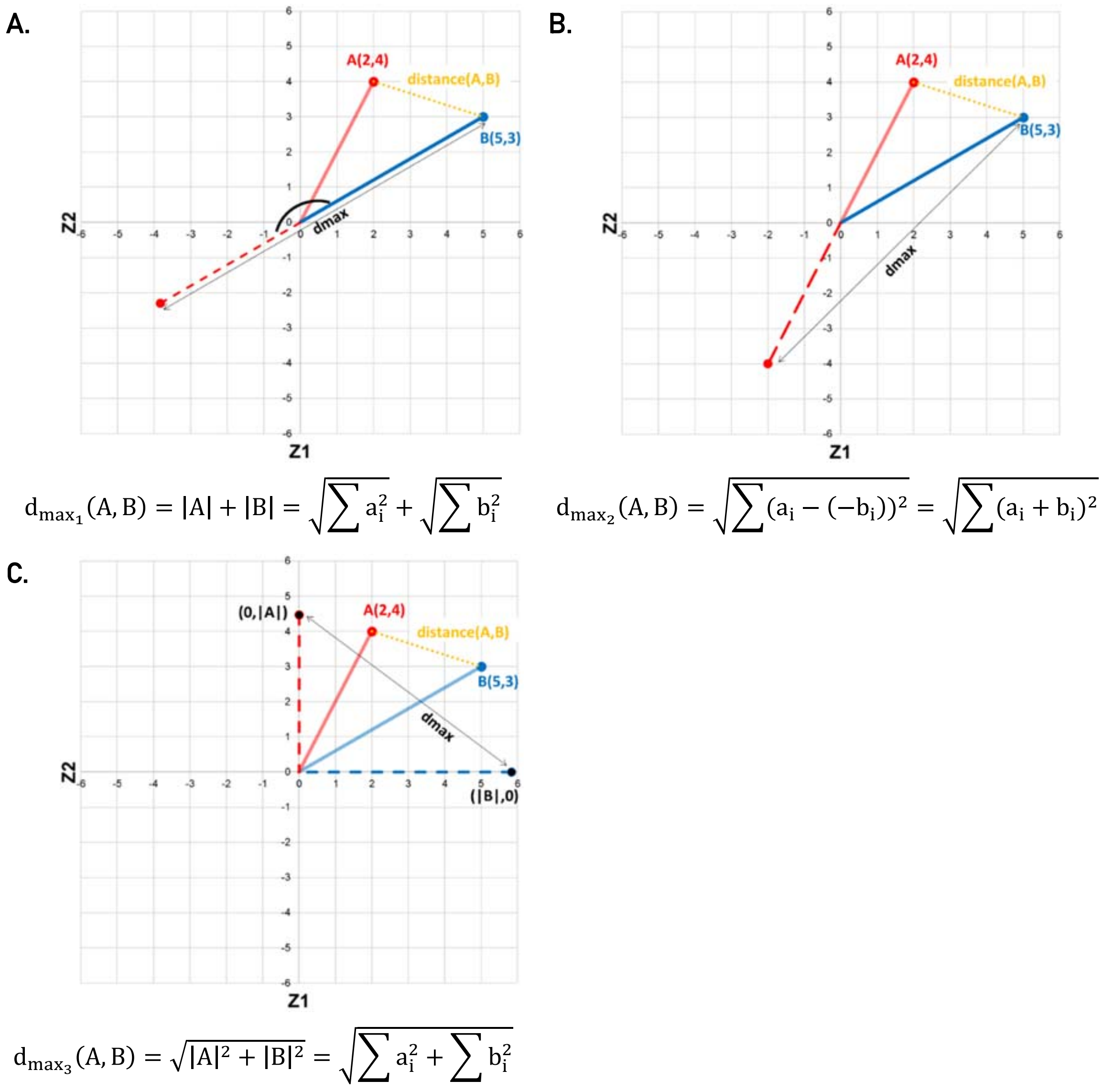
where *d^N* denotes the normalised Euclidean distance between vectors ***A*** and ***B***, i.e. the actual distance divided by the maximum distance. Note that vectors ***A*** and ***B*** each comprise n variables.

The matrix equation of this index would then be:

$$\mathrm{NEAR}\ (\mathbf{A},\mathbf{B}) = 1 - \sqrt{\frac{(\mathbf{A}-\mathbf{B}) \times (\mathbf{A}-\mathbf{B})^{\mathrm{T}}}{(\mathbf{A}+\mathbf{B}) \times (\mathbf{A}+\mathbf{B})^{\mathrm{T}}}}$$

where, as in the previous case, the cross "×" refers to a vector (or cross) product of a matrix *(1×n)* by the transpose of a matrix *(n×1)* resulting in a scalar value.

This similarity index has certain advantages over the use of the classical Euclidean distance, since (i) it is a bounded index and (ii) normalised, i.e., it takes values between 0 and 1, thus facilitating the interpretation of the results [16]. In the equation presented for the nearness index calculation, the maximum distance was taken considering only positive values. That is, if no analytical signal gives rise to a negative response, all its values are in the quadrant of positive numbers. However, if negative values exist in the signal, this must be considered to modify the calculation of the maximum distance. Up to three forms of this distance can be considered, illustrated in Figure 4.4 with an easy example of two signals in a two-dimensional space, A and B, and the Euclidean distance between them.



$$d_{max_1}(A,B) = |A| + |B| = \sqrt{\sum a_i^2} + \sqrt{\sum b_i^2}$$

$$d_{max_2}(A,B) = \sqrt{\sum (a_i - (-b_i))^2} = \sqrt{\sum (a_i + b_i)^2}$$

$$d_{max_3}(A,B) = \sqrt{|A|^2 + |B|^2} = \sqrt{\sum a_i^2 + \sum b_i^2}$$

Figure 4.4. Different ways for expressing and calculating the maximum Euclidean distance between two points in a two-dimensional space.

Figure 4.4A represents the situation where the maximum distance possible between the two signals is assumed to be the one in which both have the same modulus and the same direction, but totally opposite orientations, thus having an angle of 180° (or π radians). On the other hand, in Figure 4.4B the maximum distance is the one where one of the vectors is reflected with respect to the centre of coordinates. Clearly, both options would imply negative values in the analytical signal. While Figure 4.4C represents the maximum distance considering only positive values, that is, when the vectors would be determined by orthogonal directions to each other, forming an angle of 90° (or π/2 radians). The algebraic equation for the calculation of the maximum distance between two vectors is expressed in each figure using the three described ways.

As for the strategy based on spatial orientation, it has already been discussed above and compared between the two most common calculations: cosine and arccosine. Undoubtedly the arccosine used as a similarity index proves to be more sensitive to small differences when the two signals to be compared are very similar. This therefore results in a simpler comparison when faced with a large data set.

Conversely, Pearson's correlation coefficient is taken as the representative of the correlation strategy for the calculation of similarity. It should be known that when the two vectors to be compared are row centred, the results are identical to those obtained by calculating the cosine [23]. For this reason, a priori, if the analytical signals to be compared are very similar to each other, Pearson's correlation coefficient would not be the appropriate similarity index to consider since it is not so sensitive to small differences. However, as stated in the previous section is preferred to use the coefficient of determination, since it takes values between 0 and 1, and the results would be easier to interpret.

At this point, it is worth noting that the use of correlation metrics as a similarity index has been applied and is integrated into the software OpenLab ChemStation from the instrument manufacturer Agilent. They refer to the calculation of the determination coefficient between two spectra to be compared as the "match factor". Years earlier, they also defined a "similarity factor" to report the match of peaks from one chromatogram to another one [43,44].

Finally, as for the fourth strategy considered, the calculation of entropy based on information theory, the Jensen-Shannon index is the one that stands out. Although it has not been widely used in the analytical field so far, it is another strategy that should be taken into account in the search for a similarity index for the comparison of analytical signals. Let's look at the JS algebraic equation to be applied to the comparison of two analytical signals, but expressed as similarity index (S_{JS}):

$$S_{JS}(\mathbf{A},\mathbf{B}) = 1 - \left(\sum_{i=1}^{n} \frac{1}{2} \left[p_{a,i} \cdot \log\left(\frac{p_{a,i}}{p_{m,i}}\right) \right] + \sum_{i=1}^{n} \frac{1}{2} \left[p_{b,i} \cdot \log\left(\frac{p_{b,i}}{p_{m,i}}\right) \right] \right)$$

where $p_{a,i}$ and $p_{b,i}$ correspond to the TSN normalised intensity values of the vectors ***A*** and ***B***, respectively, and both vectors are characterized by a total of n variables each; $p_{m,i}$corresponds to the average between $p_{a,i}$ and $p_{b,i}$. The first part of the equation corresponds to the calculation of the KL divergence between signal ***A*** and the average signal between ***A*** and ***B***, and the second part corresponds to the calculation of the KL divergence between signal ***B*** and the average signal of ***A*** and ***B***. Note that equal weight has been given for signals ***A*** and ***B***, although this can be modified according to the analyst's requirements **[20,32]**. Moreover, it should be noted that for applying this equation, each analytical signal ***A*** and ***B*** must be TSN normalised and given the presence of the logarithm in the equation, no 0 value(s) can be present. This can be easily corrected by adding to the full vector before normalisation a very small value (e.g. 1 x 10^{-8}).

Consider now the matrix form of this calculation:

$$\mathrm{S_{JS}}(\mathbf{A},\mathbf{B}) = 1 - \left(\frac{1}{2}\left[\mathbf{P_A} \times \left(\log\left(\frac{\mathbf{P_A}}{\mathbf{P_M}}\right)\right)^{\mathrm{T}}\right] + \frac{1}{2}\left[\mathbf{P_B} \times \left(\log\left(\frac{\mathbf{P_B}}{\mathbf{P_M}}\right)\right)^{\mathrm{T}}\right]\right)$$

where $\boldsymbol{P_A}$ and $\boldsymbol{P_B}$ are the *(1×n)* TSN normalised compared signal vectors and $\boldsymbol{P_M}$ represents the average between $\boldsymbol{P_A}$ and $\boldsymbol{P_B}$. Once again, the cross "×" refers to a vector (or cross) product of a matrix *(1×n)* by the transpose of a matrix *(n×1)* resulting in a scalar value. Note that this matrix equation represents the mathematical operations to be carried out, but the expression does not follow the rules of matrix algebra. That is, it is not possible to divide matrices, and the inverse of the divisor matrix must be used although some mathematical programs, e.g., MATLAB, can be commanded using an expression similar to that proposed above.

At this point it is worth noting that a key consideration to take into account in the application of similarity indices is the normalisation of the intensity values (the analytical signal). This step should be applied in most applications, with the exception that the goal is to compare differences in these values per se. Different intensity values due to instrumental deviations can considerably affect the similarity index result. Therefore, if the analytical profile of two signals is to be compared in itself, it is highly recommended to normalise the intensities.

Let's now compare several similarity indices following the different strategies described with an example applied to analytical signals. Let there be 3 vectors of 600 variables corresponding to three analytical signals, where one is the reference signal and the other two are experimentally acquired signals, processed by different ways, called experimental signal 1 and experimental signal 2 (see Figure 4.5). These data are part of a broader study that can be consulted at **[29]**.

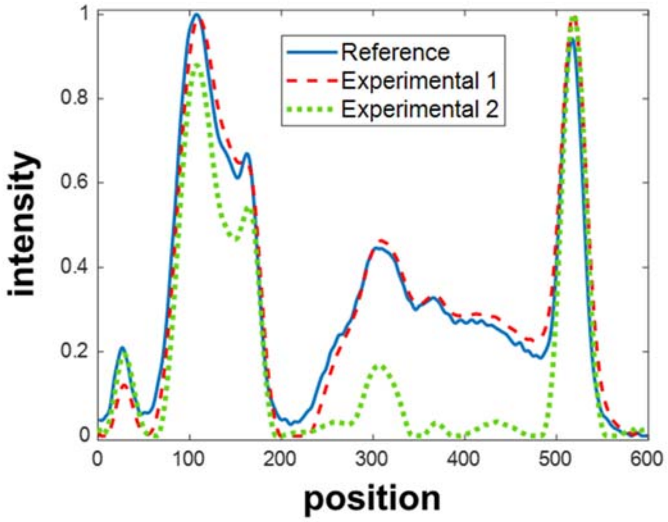


Figure 4.5. Spectra acquired (experimental signal 1 and experimental signal 2) to be compared with a reference spectrum by different similarity indices.

Visually there is no doubt that experimental signal 1 is much more similar to the reference signal than experimental signal 2. To compare both experimental signals with respect to the reference signal, the following similarity indices have been considered: nearness index (NEAR), arccosine (ACOS), coefficient of determination (R^2) and similarity index based on JS divergence (S_{JS}), which correspond respectively to the four strategies for similarity calculation, namely: distance, spatial orientation, correlation, and entropy. Note that the 4 indices take values between 0 and 1, meaning 0 totally different and 1 totally similar. Table 4.1 shows the results of the fourth indices.

Table 4.1. Similarity indices found from comparison of analytical signals showed in Figure 4.4.

	NEAR	ACOS	R^2	S_{JS}
REF *vs* EX1	0.935	0.921	0.968	0.985
REF *vs* EX2	0.716	0.715	0.814	0.865
% difference	22.0	20.6	15.4	12.0

Ref: reference signal, Ex1: experimental signal 1, Ex2: experimental signal 2, NEAR: nearness index, ACOS: arccosine index, R^2: coefficient of determination index, S_{JS}: similarity index based on Jensen-Shannon divergence.

Clearly, as expected, the values obtained from the comparison between the reference signal and experimental signal 2 are lower with respect to the reference signal vs. experimental signal 1. The highest values have been achieved by S_{JS}, indicating that it is less sensitive to the differences between the signals. For a better understanding of the results, the % difference between the index calculated for reference vs experimental 1 signals and reference vs experimental 2 signals have been calculated (see Table 4.1). Based on this, S_{JS} differs the least, followed by the calculated R^2. The latter is not surprising since, as mentioned above, it takes very similar values to the cosine, and the cosine is not very sensitive to small differences when two signals are very similar between them. While the results obtained with NEAR and ACOS indices show a difference of over 20% between the two comparisons, thus indicating that these two indices are more sensitive to minor differences when the signals are very similar and considering this example. Therefore, it is strongly recommended to the reader to consider and calculate different indices for the application of a similarity analysis between pairs of analytical signals, in order to obtain a global and general overview, instead of selecting a single index.

To finish this section, as stated above, similarity indices can only be applied to two signals at a time. However, similarity analysis is usually intended to be performed on a set of signals to evaluate which one is more similar to each other or to a reference, as part of a study. Especially with a large data set, it is an arduous task to analyse the numerical results visually one by one. But then, how to conduct a multiple pairwise comparison? There is a strategy consisting of generating a similarity matrix and working with it to infer conclusions without it being time-consuming. This matrix is characterised by being a square and symmetric matrix, whose dimensions are equal to the number of objects being compared. Furthermore, it should be noted that in certain studies this similarity matrix is the input dataset used for the subsequent application of both supervised and unsupervised multivariate analysis methods, since this way the dimensionality of the input data is considerably reduced [**27**]. The similarity matrix can be constructed by using any of the similarity indices discussed in this section. The most common is by using the correlation coefficient, sometimes named as correlation matrix [**45**].

2.2. Similarity curves (profiles) and similarity surfaces (images)

When dealing with analytical signals of higher dimensions than those discussed so far, i.e. 3D analytical signals or higher (second order data or higher), the calculation of the similarity indices discussed in the previous section does not result in a scalar. In the case of a 3D signals (for instance, an absorption spectrum-chromatogram), a vector, i.e. a similarity profile, is then obtained. It is rational then to assume at this point that the interpretation is not as straightforward as in the case of 2D signals. However, it is possible to represent the vector obtained, thus yielding a similarity curve. This plot provides information on how the similarity between the two compared objects varies throughout the data set. On the horizontal axis, the compared variables are represented, and on the vertical axis, the result of the applied similarity metric is shown. This can be done along the two dimensions characterising the analytical signal. For example, for a signal acquired by gas chromatography coupled to mass-spectrometry (GC-MS), a chromatographic similarity profile or a similarity profile along the mass spectrum can be obtained. These similarity profiles are useful for identifying similarity patterns and discrepancies in the data set.

Figure 4.6 shows an example of the similarity profiles obtained by comparing the 3D analytical signals of olive oil samples acquired by liquid chromatography with a UV-Vis molecular absorption detector. In this example, the NEAR index was used to obtain the similarity curves. The profiles were calculated along the first dimension, that is, for the UV-vis profile along the wavelengths studied.

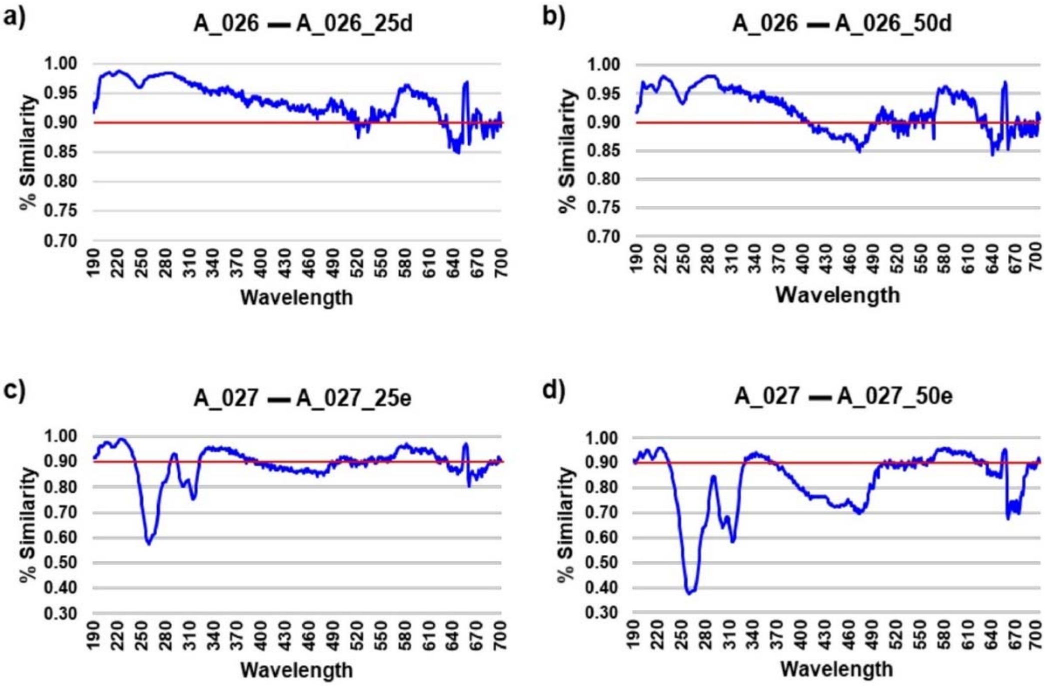


Figure 4.6. Similarity profile plots obtained by comparing several extra virgin olive oil samples measured with liquid chromatography coupled to a UV-Vis molecular absorption detector. Figure reprinted with permission from [**46**].

The matrix form of the equation used is as follows:

$$\mathbf{NEAR}(\mathbf{A},\mathbf{B})_{(\mathrm{n})} = \mathrm{diag}\left\{\mathbf{M}(1) - \left[\sqrt{\frac{(\mathbf{A}-\mathbf{B})\times(\mathbf{A}-\mathbf{B})^{\mathrm{T}}}{(\mathbf{A}+\mathbf{B})\times(\mathbf{A}+\mathbf{B})^{\mathrm{T}}}}\right]\right\}$$

where *M(1)* is an *n×n* square matrix of ones; *A* and *B* are the *m×n* matrices corresponding to the two 3D analytical signals to be compared. Applying this equation results in a vector of *n* elements with values between 0 and 1, which corresponds to the similarity profile along the first dimension of the compared signals [46]. Note that in this equation "diag" denotes the diagonal of the *n×n* matrix obtained from the operation expressed within the braces and T superscript symbolises the transpose of the original matrix.

In addition, it is also possible to obtain the similarity profile along the second dimension. For this, the following equation should be used:

$$\mathbf{NEAR}(\mathbf{A},\mathbf{B})_{(\mathrm{m})} = \mathrm{diag}\left\{\mathbf{M}(1) - \left[\sqrt{\frac{(\mathbf{A}-\mathbf{B})^{\mathrm{T}} \times (\mathbf{A}-\mathbf{B})}{(\mathbf{A}+\mathbf{B})^{\mathrm{T}} \times (\mathbf{A}+\mathbf{B})}}\right]\right\}$$

In this case, *A* and *B* are the same *m×n* matrices; *M(1)* is an *m×m* square matrix of ones, and the obtained NEAR similarity profile is a vector of *m* elements.

The lector should note that the equation represents the mathematical operations to be carried out, but the expression does not follow the rules of matrix algebra. As stated in the previous section (see the text after the matrix equation for the S_{JS} index calculation), it is not possible to divide matrices, and the inverse of the divisor matrix must be used although some mathematical programs, e.g., MATLAB, can be commanded using an expression similar to that proposed above.

From these similarity profiles, and with prior knowledge of the objects to be compared and the chemical information present in the analytical signals, much relevant information about the similarity between the compared signals can be obtained. In the example shown, the authors set a similarity limit of 0.90. This implies that those parts of the similarity profile lying below this limit, indicated by a red line in the figures, are not considered to be similar. Similarity profiles (c) and (d) (see Figure 4.6) depict two regions in which the NEAR index decreases considerably to values even below 0.50. In this case, the authors found that the range between 210 and 270 nm could be related to the composition of unsaturated fatty acids. The signals compared in these figures came from an extra virgin olive oil and an extra virgin olive oil adulterated with olive oil at different concentrations (25% and 50% respectively for (c) and (d)). Therefore, obtaining a similarity profile to compare between these two signals has provided the necessary information that the acquired signal would be able to differentiate and detect such adulteration [46].

It should be noted that any of the similarity indices described in the previous section can be adapted to be calculated on 3D analytical signals to obtain a similarity profile. To this end, the same strategy of calculating on matrices and obtaining the diagonal of the resulting matrix as the similarity profile between the two 3D signals compared should be performed.

There are not many published studies employing the generation of similarity profiles between 3D or higher dimension order analytical signals, and especially not with this term. In contrast to the similarity profile obtained when comparing 3D signals, when dealing with 4D signals (or third order data) a similarity surface or similarity image is then obtained, by means of a data matrix which can be plotted for a visual interpretation of the results. This practice is perhaps more common in the literature than the previous one for 3D signals, or at least the studies are more traceable by specifying the term image, or sometimes the term similarity map [**47,48**]. In this case, each value of the obtained matrix (which could be referred to as a pixel) is a similarity value, calculated by any of the strategies and indices described in the previous section.

An important consideration when working with 3D or higher-dimensional signals is the alignment of the signals, especially in some techniques such as chromatography or nuclear magnetic resonance spectrometry, where small deviations can occur. This is less common in other optical spectrometric techniques. Note that this aspect is of equal importance for 2D analytical signals, however it is a more complex and less known step to implement for higher dimensional signals. A misalignment of the signals to be compared can lead to erroneous similarity results [**47**]. This is why applying alignment techniques as part of the pre-processing could be crucial before applying a similarity analysis. Another solution to this misalignment is the proposed "tile-based" analysis. It was also given the name "DotMap" to the algorithm developed by Sinha et al. [**49**]. In this case, similarity is not analysed on a value-by-value basis (pixel-by-pixel could be said in the case of 4D analytical signals resulting in a similarity surface), but a relatively small region is chosen, yet large enough to avoid excluding comparable information from one signal to another in case they are misaligned. In this case, the average value of the selected region is calculated for both analytical signals to be compared, and the calculation of the respective similarity index is applied [**47**].

Lastly, it should be noted that there is another common strategy to study similarities between pairs of analytical signals by means of a similarity surface. However, this is not applied to 3D or higher dimensions data, but to 2D analytical signals. It is known as 2D correlation spectroscopy (2D-COS), because is widely used for spectrometric data. The output of 2D-COS can be plotted as a contour or mesh plot, thus, to identify similarities and differences between two spectra [**50,51**]. Although its main application is on a unique object of which two spectra were acquired, it can also be applied to study similarities between analytical signals of two objects, so it could be considered another strategy to perform a similarity analysis.

3. Similarity analysis on analytical signals sets

When studying analytical signal sets consisting of a considerable number of samples, the similarity analysis may be approach with a specific set of tools, different of the one described in the previous section. While it may be possible to apply conventional similarity indices to these types of datasets, its interpretation becomes more and more challenging as the number of objects to be compared increases. Therefore, it is important that studied data can be condensed while preserving enough information to easily separate into groups with similar characteristics, without overwhelming potential interpretations. The objective is to simplify the data while preserving its most meaningful information.

This section is devoted to studying the methods used to test the similarity of datasets based on multivariate mathematical approaches. The techniques to be laid out in this section can be described as multivariate exploratory data analysis (MEDA) methods, which are tools that provide a summary of the main characteristics of data and often use graphical representations to facilitate the interpretation of the results [39]. These approaches can be distance-based methods (data clustering or grouping) and covariance-based methods (data decomposition into latent factors). MEDA can be a useful tool to detect natural tendencies in the studied data, and it will serve as a precursor to supervised methods. The latter being focused on building a model from known classes which can be applied to new data to assign a class to unknown objects [39].

It should be noted that no method is one-size-fits-all, i.e., different datasets may require different approaches. The analyst must keep in mind that some methods can be better at describing patterns in one dataset and others not as much, which can be challenging. For this reason, it is common to apply several multivariate methods to the same dataset to find which can describe the emerging patterns better. On the other hand, it is critical to highlight the role an analyst has when working with MEDA. Since no information is provided to the algorithms on how the data may separate, it is the task of the analyst to interpret the results and correctly assign patterns with all known complementary information about the data.

Finally, it is important to mention that the goal of this section is not to provide deep mathematical and statistical explanations on how each method works, but to lay out examples of when these tools might be useful. For further details on how these methods work, the reader can refer to specialised literature provided throughout this section.

3.1. Cluster analysis

Cluster analysis is a statistical multivariate method that seeks to organise a given dataset into distinct clusters. The goal is to find an optimal grouping for which the observations or objects within each cluster are similar, but the clusters are dissimilar to each other [52]. Therefore, members within a cluster share similar characteristics, but are unrelated to members of other clusters. Once the cluster analysis is done, natural groupings in the data can be studied graphically by plotting the observations. To find out how closely related samples are, some type of similarity analysis should be done. In the context of cluster analysis, there are multiple ways to measure similarity between samples, the most common way, as counterintuitive as it may seem, is to measure dissimilarity by using distance. As discussed in previous sections, there are multiple distance metrics that can be used; the most common in cluster analysis being Euclidean and Manhattan distances. These can be applied to the data to create a similarity (or dissimilarity if distances are used) matrix [23,53], which is one of the building blocks of the clustering algorithms that will be discussed in this section.

There are two main types of clustering methods: hierarchical clustering and non-hierarchical clustering; their main difference being the approach. Hierarchical clustering does not require a predetermined number of clusters. Therefore, after applying the algorithm, it is the user who decides the number of clusters based on the results. On the other hand, non-hierarchical clustering requires a pre-established number of clusters to run. Even though most of the time the number of clusters are given by the user, some non-hierarchical clustering algorithms will find the optimal number of clusters beforehand.

Hierarchical clustering can be divided into two methods: agglomerative and divisive. Agglomerative clustering algorithms, also called "bottom up", start by making each item its own cluster. Then, clusters are merged until one cluster remains. Divisive algorithms or "top-down" work the other way around. All items start as members of a single cluster and, as the calculations advance, these are split until every item is its own cluster [54]. The end goal of hierarchical clustering is to create a dendrogram, which is a tree diagram used to illustrate the arrangement of clusters [39]. Figure 4.7 shows the hierarchy of clusters formed during the clustering process. The height of the nodes indicates the similarity of its branches. Items more similar to each other are combined at lower heights, whereas items that are less similar are combined higher up [54]. Clusters are identified by drawing horizontal lines (cut-off heights) in the dendrogram. The number of clusters is determined by the intersections between these cut-off heights and vertical lines. For instance, setting a cut-off height (illustrated in red in Figure 4.7) will result in the selection of two clusters.

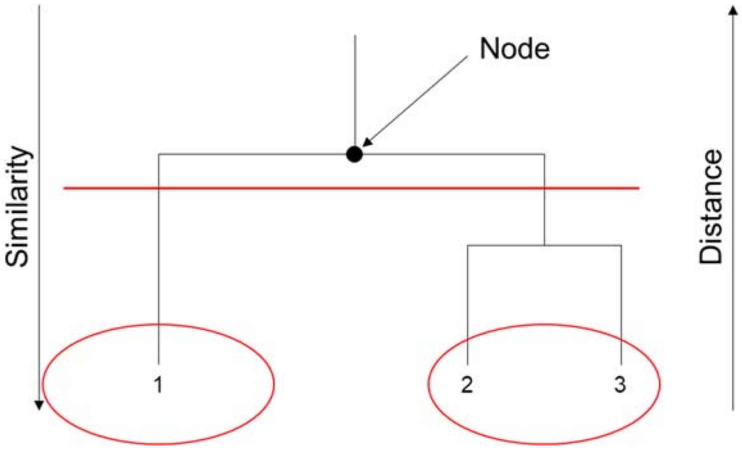


Figure 4.7. Graphical display of a dendrogram as an output example of cluster analysis.

Non-hierarchical clustering, on the other hand, will partition the members of the dataset into a specific number of pre-determined clusters. Non-hierarchical clustering algorithms assign each member to only one cluster (although some methods, like fuzzy clustering, assign the probability of belonging to a cluster) based on its distance to the centroid of the cluster. The algorithm works by iterating until the number of pre-assigned clusters is reached, maximising intra-cluster similarity and minimising inter-cluster similarity [52]. Figure 4.8 shows how clusters could be formed around centroids depending on the number of clusters pre-assigned by the user, like it would be seen by using a K-means clustering algorithm.

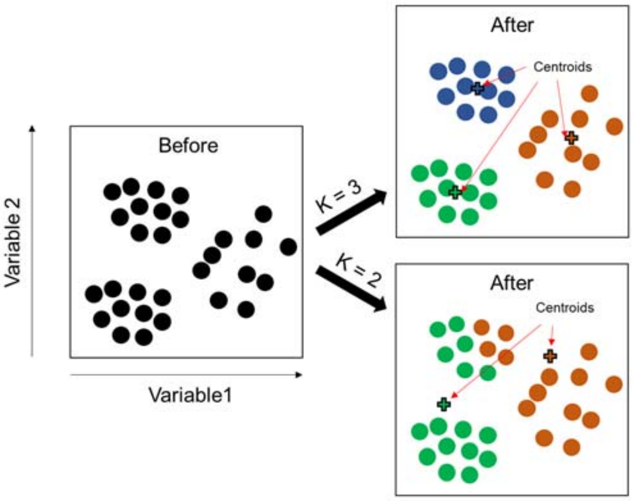


Figure 4.8. K-means cluster analysis output depiction.

The selection of the appropriate type of clustering method will come down to factors such as number of samples, the goal of the analysis, and knowledge on the number of clusters. For example, if the number of samples is too large, non-hierarchical clustering may be better because of its computational simplicity. Moreover, when sample size grows larger, it becomes more challenging to display the data using dendrograms. However, if the number of clusters is unknown *a priori*, hierarchical clustering may be beneficial. Regardless, choosing one, other, or both, will most likely come down to trial and error.

Finally, to demonstrate the utility of cluster analysis, it is convenient to discuss possible scenarios of what could be encountered. For illustrative purposes, a data set consisting of 20 samples (simulated chromatograms) was *ad-hoc* created using MATLAB R2022b. These were labelled from 1 to 20 and divided into 4 distinct groups based on differences introduced regarding peak heights, shown on Figure 4.9. The objective is to determine if cluster analysis is capable of grouping similar samples together. It should be noted that these chromatograms have no chemical basis and were created exclusively to display the MEDA methods and their differences.

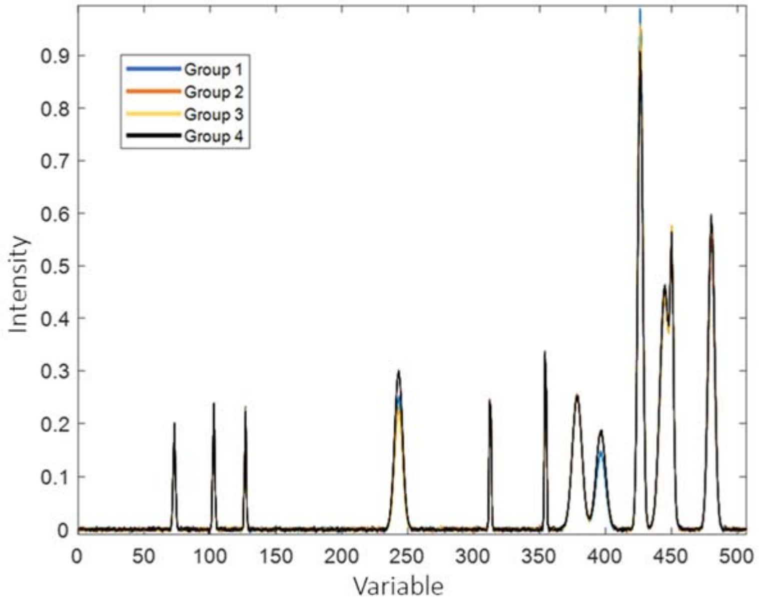


Figure 4.9. Simulated chromatograms using MATLAB R2022b software.

The chromatograms are visually very similarity, which makes grouping them a difficult task if no similarity tool is used. Figure 4.10 shows the outcome of the hierarchical cluster analysis applied to the areas of the chromatographic peaks.

As expected, 4 clusters consisting of 5 samples each were obtained. Furthermore, a great deal of information could be extracted from the obtained results. The dendrogram splits at a high distance for every group, which means that there is very low inter-cluster similarity. In addition, the nodes of each sample within the same cluster appear at very short distances, which signify high intra-cluster similarity. Finally, results also show that groups 2 and 3 are the most similar due to the short distance of the last node that separates them.

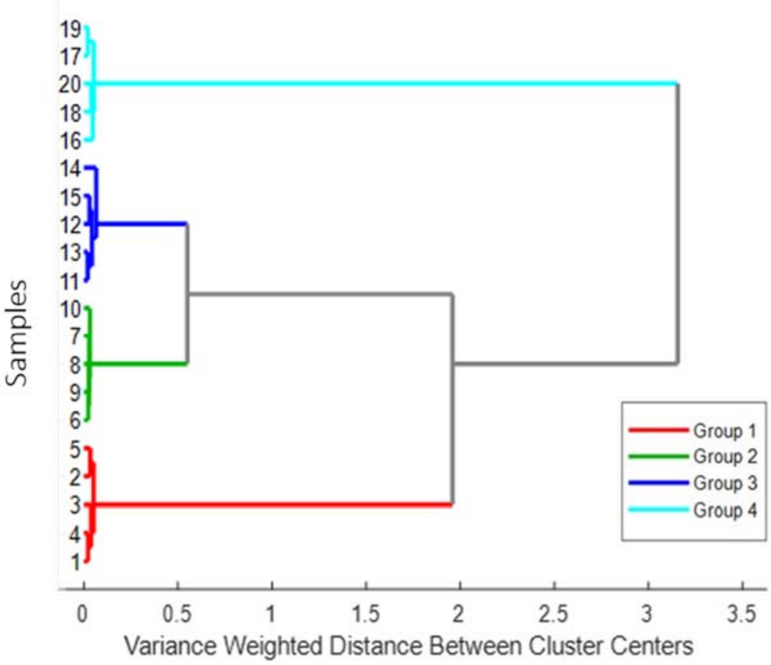


Figure 4.10. Hierarchical cluster analysis applied to selected peaks of the studied chromatograms. Results generated via PLS_Toolbox 9.3.1.

3.2. Factor analysis

Factor analysis (FA) is described by the IUPAC as *the matrix decomposition of a data matrix into the product of a score matrix and the transpose of the loadings matrix* [39]. Formally, this definition could be expressed as follows:

$$\mathbf{Y} = \mathbf{S} \times \mathbf{L}^{\mathrm{T}}$$

Where *Y* is the original data matrix, and *S* and *L* are the scores and loadings matrices, respectively [23,55]. Note that if *Y* is a *n×m* matrix, which embeds the original data, organized in *n* rows (objects or samples) and *m* variables, *S* and *L* are matrices *n×k* and *m×k*, where *k* indicates the number of factors; this number is chosen by the user (typically less than 10).

It is based on the principal of dimensionality reduction: reducing a fairly large number of variables into a small number of latent factors/variables that explain the observed correlations among variables. This method arises from the fact that objects tend to have unmeasurable (or hard to measure directly) characteristics related to measurable ones. For example, intelligence quotient (IQ) scores, which are the standard for measuring human intelligence, could be considered a latent factor that describes the performance on solving specific tests, which would be the dependent variables. By summarising these variables into one single score, which could be called a latent variable, the data have been simplified and its interpretation made far easier. All in all, latent factors will summarise a large number of dependent variables, therefore samples with similar factor scores will group together and be more similar than those with different scores.

The first step when performing a FA is to decide the method for factor extraction. The most common one is through Principal Component Method [52,53], which is useful when capturing as much variability as possible is desired, rather than understanding the underlying latent factors. This method assigns the maximum variance to the first factor, assigns the second most amount of variance to the second factor, and so on. This way each factor explains a unique percentage of variance. Another consideration when performing FA is the type of rotation used. Two types can be considered: orthogonal and oblique, being its goal to find a new frame of reference in which the factors are more interpretable [52]. It could be compared to finding new fits to a series of points, different rotations would correspond to different fitting methods and will yield different outcomes, and although no one is better than the other in every case, the most used is Varimax rotation (orthogonal).

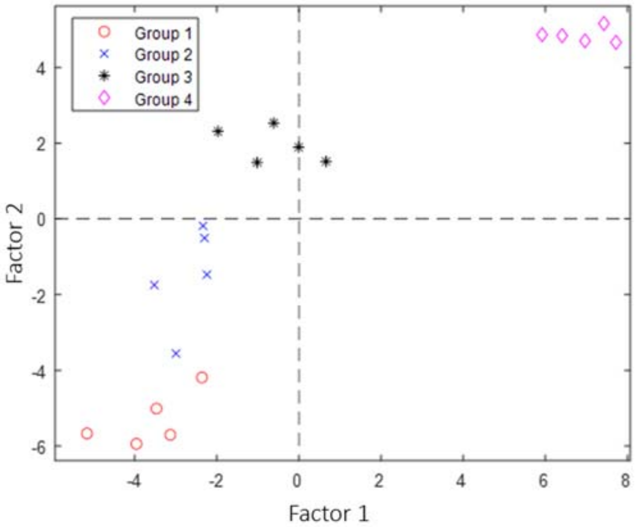


Figure 4.11. Factor Analysis applied to the selected chromatographic peaks using STATGRAPHICS Centurion software. Scores plot of factor 2 vs factor 1 shown.

To detail the applicability of FA as similarity analysis, take the same example laid out in the previous section. Once the FA is performed on the chromatographic data, the results can be seen in Figure 4.11.

Some tendencies in the separation of the data can be clearly observed. Factor 1 is exceptional at separating samples of group 4 from the others, while factor 2 clearly separates all the groups. In this case it should be noted that groups 2 and 3 are the most similar, which is not surprising, as cluster analysis on the same data (Section 3.1) resulted in high similarity between those two groups. Although only factor scores results are shown in this figure, FA outputs factor loadings, which provide more information of the analysis. These explain the influence of dependent variables on each factor.

3.3. Principal components analysis (PCA): loadings and scores

PCA is the most used chemometric method for MEDA. Its goal is to reduce the dimensionality of a dataset into a new coordinate system known as the principal components. This new coordinate system retains as much variability explaining the original data as possible and makes the interpretation of the data easier for the user. Finding the values of the principal components involve solving an eigenvalue/eigenvector problem or, alternatively, the solution can be obtained via singular value decomposition of the data matrix [**56**]. Furthermore, it should be noted that the same matrix decomposition described in section 3.2 for FA is applicable to PCA, where in this case instead of number of factors, it is referred to as principal components. The goal is also to find natural groupings in the data. Samples that are most similar will find themselves close together when graphing a PCA scores plot.

Although PCA is related and sometimes (mistakenly) confused with FA, they are quite different in nature. PCA's objective is to explain a large part of the variance of the variables, while FA seeks to account for correlations (covariance) between the original variables. Furthermore, PCA might be useful if the goal of the analysis is to simply reduce the variables down into a linear combination of smaller components, while FA might be more appropriate if it is believed that there is some hidden construct defining the relationship among variables [**52**].

Firstly, it is important to understand the concept of a principal component (PC). Even though the word component is used, it does not allude to a single compound or dependent variable but rather to a source of variance within the dataset. The first PC is a linear combination explaining the maximum variance in the original data, the second PC captures the second most variance in a direction orthogonal to the first PC, and so on [**52**]. Orthogonality is the most important feature of PCA, which translates into PCs being uncorrelated and each one capturing an independent source of variability.

Scores and loadings provide the user with information on how samples and variables are related to each other [57]. The scores plot will be the most helpful tool when looking for similarity between samples within a dataset, as it will serve as a visual display of how closely they are. Moreover, this enhances the trends in the grouping of samples and, the closer they are in space, the more similar they are according to the observed PC. In contrast, the PCA loadings give information about the contribution of the original variables on the PCs. For analytical signals containing a large number of variables (e.g. IR spectra or chromatograms), this means that loading plots will describe which regions of the original variables are responsible for the score values of a certain PC.

Let's take again the wine data to run an example of PCA to study similarities in a dataset (see Figure 4.9). When cluster analysis was applied, samples grouped with closely related members. PCA allows finding another view on the possible groupings of similar samples (see Figure 4.12). For a better interpretation of the results, labels were included.

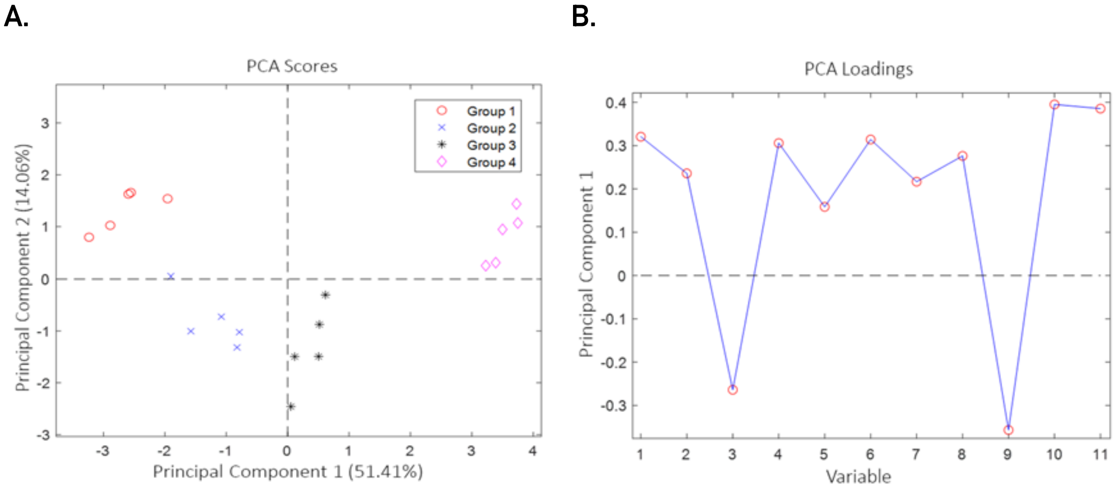


Figure 4.12. Results of applying PCA to the selected chromatographic peaks using STATGRAPHICS Centurion software. Scores plot of PC2 vs PC1 (**A**) and loading plot of PC1 (**B**).

PC1 explains 51.4% of the variability, while PC2 explains 14.1%. By plotting only PC1 and PC2 a large amount of information can be acquired while reducing the dimensionality of the dataset by a great deal. The results show some data behaviour trends, yielding 4 distinguishable groups (see Figure 4.12A). What these results show, essentially, is that the PCs shown (PC1 vs PC2) are excellent at describing the natural grouping of samples. It is common to plot other combinations of PCs in order to find other types of natural grouping in the dataset. Furthermore, it is worth noting that regardless of PCA and FA having similar objectives, each method attributed different importance to the characteristics of the samples, therefore yielding particular scores.

Attending to the PCA loadings, high intensity in the peaks means that those chromatographic peaks have a higher contribution to the studied PC (see Figure 4.12B). Positive values indicate a positive association between the variables and the represented PC and vice versa. In this case, the following interpretation could be made: an increase in the intensity of the peaks 10 and 11 of a sample will result in an increase of the score in PC1 of that sample (and vice-versa). This should be then studied to identify which chemical information is provided on that region of the chromatogram.

3.4. Multivariate analysis of variance (MANOVA)

MANOVA is a statistical technique that extends from ANOVA to situations where there are two or more dependent variables [**58**]. It tests if a series of sample sets differ from each other by examining multiple dependent variables simultaneously and the interactions between them. Fundamentally, MANOVA involves the comparison of means across independent groups, based on the means of two or more dependent variables, providing insight into whether the groups demonstrate statistically significant differences. Although one could argue that the alternative to MANOVA would be to conduct ANOVA for each dependent variable, this approach is far from convenient. Using this approach (i) could increase the likelihood of finding statistical differences when there are none (i.e., committing a type 1 error) and (ii) would not determine whether independent variables are related to combinations of dependent variables [**58**].

To conduct a MANOVA analysis, statistic programs usually provide several alternative tests [**58**]. These are based on accepting or rejecting that no statistical difference is present between the means of the dependent variables. Therefore, accepting this conclusion means there is no sufficient evidence to support a significant difference between the independent groups based on the comparison of means of the dependent variables. In principle, the dependent variables should be normally distributed, present homogeneity of covariances and absence of multicollinearity, although non-severe deviations from these requirements do not invalidate the results, at least in a first approximation.

In short, regarding similarity analysis, significant values obtained from the mentioned methods indicate dissimilarity between one or more sample sets. Conversely, non-significant values suggest statistical similarity among the groups or not enough significant evidence to conclude the groups are different. MANOVA, unlike some of the other methods studied, presents limitations when applied to analytical signals containing more measured variables than observations [**59,60**]. Therefore, using this method for full chromatograms, spectra, or similar analytical signals, is not feasible and becomes impractical.

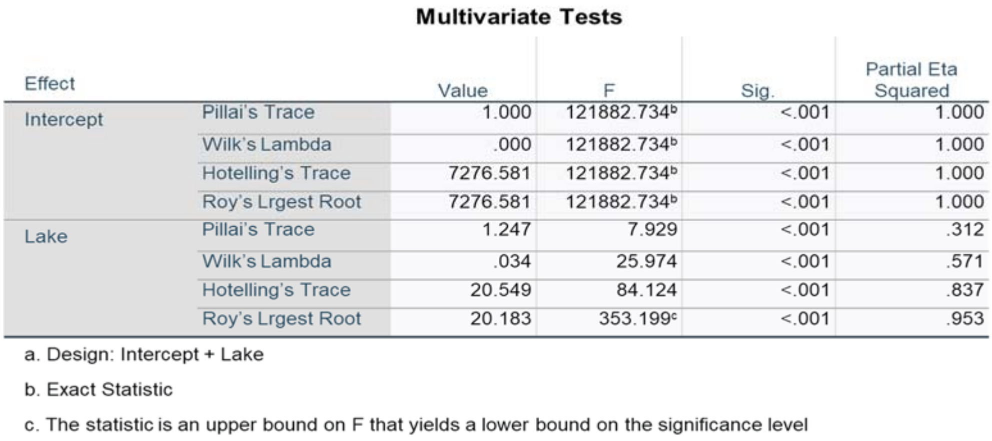
This would be feasible in the case of a targeted approach where, for example, only the retention times of the compounds of interest are used and not the whole chromatogram [61]. For this reason, the following example will deviate from the ones mentioned previously.

Table 4.2 shows the results an environmental study where various parameters of water samples from different lakes were measured. The objective is to determine if these findings are different in any of the locations, so MANOVA will allow simultaneous comparison of each sample. 15 samples were taken in each location to account for any possible deviation in the measurements in water that could arise from sampling location within the same lake.

Table 4.2. Data collected on natural waters from different lakes.

	No samples	pH	Dissolved oxygen (mg/L)	Nitrate content (mg/L)	Turbidity (NTU)
Lake 1	15	6.44	7.45	1.14	23
Lake 2	15	7.72	7.92	1.18	34
Lake 3	15	6.48	7.94	1.19	24
Lake 4	15	6.49	7.81	1.19	23
Lake 5	15	6.45	7.82	1.16	23

Although some differences can be seen in one of the lakes just by looking at the data, no conclusion can be drawn until proper analysis is performed. After MANOVA is applied by means of specific statistical software, a table of results is provided, as the one shown in Figure 4.13.



Multivariate Tests

Effect		Value	F	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	1.000	121882.734[b]	<.001	1.000
	Wilk's Lambda	.000	121882.734[b]	<.001	1.000
	Hotelling's Trace	7276.581	121882.734[b]	<.001	1.000
	Roy's Lrgest Root	7276.581	121882.734[b]	<.001	1.000
Lake	Pillai's Trace	1.247	7.929	<.001	.312
	Wilk's Lambda	.034	25.974	<.001	.571
	Hotelling's Trace	20.549	84.124	<.001	.837
	Roy's Lrgest Root	20.183	353.199[c]	<.001	.953

a. Design: Intercept + Lake

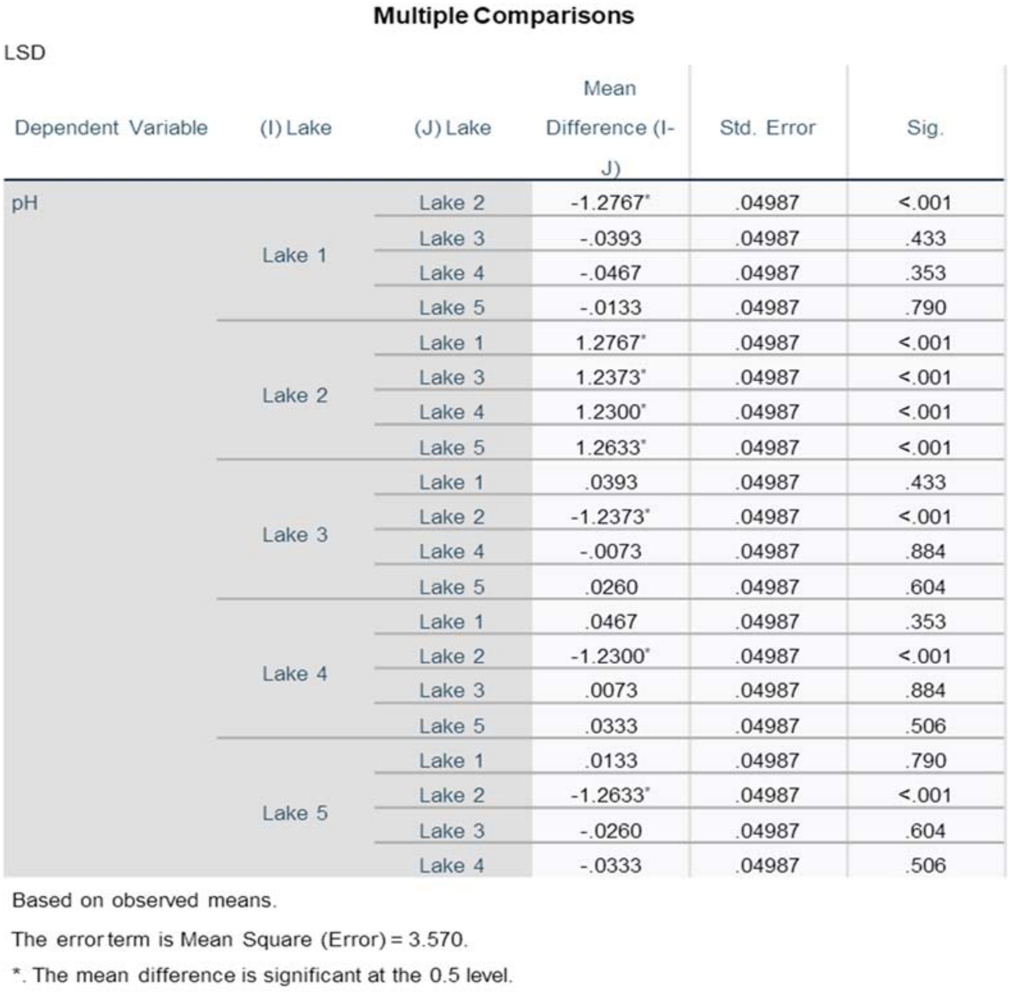
b. Exact Statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level

Figure 4.13. Results of MANOVA applied to the data of the lakes presented in Table 4.2. Note that this output was obtained by using IBM SPSS software, version 28.0.

In this case all the tests conclude that one or more of the Lakes are different from the rest.

It should be noted that MANOVA does not provide information about where the dissimilarity between groups is present. Thus, post-hoc testing should be performed to understand where the differences between the observations lie [58]. Therefore, with the results obtained from MANOVA, a least significant difference (LSD) test can be performed as a post-hoc test, obtaining the results shown in Figure 4.14. This post-hoc test provides insights on which sample groups appear to be dissimilar. In this case, when groups are compared, it is evident that Lake 2 is to blame. The significance value of the pH data collected from Lake 2 is below 0.05 when compared to every other lake, indicating statistically significant differences.



Multiple Comparisons

LSD

Dependent Variable	(I) Lake	(J) Lake	Mean Difference (I-J)	Std. Error	Sig.
pH	Lake 1	Lake 2	-1.2767*	.04987	<.001
		Lake 3	-.0393	.04987	.433
		Lake 4	-.0467	.04987	.353
		Lake 5	-.0133	.04987	.790
	Lake 2	Lake 1	1.2767*	.04987	<.001
		Lake 3	1.2373*	.04987	<.001
		Lake 4	1.2300*	.04987	<.001
		Lake 5	1.2633*	.04987	<.001
	Lake 3	Lake 1	.0393	.04987	.433
		Lake 2	-1.2373*	.04987	<.001
		Lake 4	-.0073	.04987	.884
		Lake 5	.0260	.04987	.604
	Lake 4	Lake 1	.0467	.04987	.353
		Lake 2	-1.2300*	.04987	<.001
		Lake 3	.0073	.04987	.884
		Lake 5	.0333	.04987	.506
	Lake 5	Lake 1	.0133	.04987	.790
		Lake 2	-1.2633*	.04987	<.001
		Lake 3	-.0260	.04987	.604
		Lake 4	-.0333	.04987	.506

Based on observed means.

The error term is Mean Square (Error) = 3.570.

*. The mean difference is significant at the 0.5 level.

Figure 4.14. Results of applying least significant difference (LSD) test to the data of the lakes presented in Table 4.2. Note that this output was obtained by using IBM SPSS software, version 28.0.

3.5. Coupling ANOVA and PCA

Coupling ANOVA and PCA takes the best of statistics and chemometrics to develop a powerful tool for the analysis of datasets with large number of variables. This strategy addresses the disadvantage of MANOVA, which cannot be applied when the number of observations is smaller than the number of variables. Moreover, as for MANOVA, the absence of multicollinearity is required. This is rarely satisfied in continuous analytical signals, such as spectra or chromatograms, where a peak or band is formed by a set of variables being all correlated [**62**].

In ANOVA-PCA, the data matrix is firstly decomposed in effects matrices and a residual matrix, containing the residuals non-explained by the model. Then, the PCA is applied to each of the effects matrices combined with the residual matrix. The results show the importances of each of the factors comparing with the residual error. In short, ANOVA looks for the significance of each factor under study, while ANOVA-PCA explores not only the significance, but also the relationship between the factors by studying the groupings [**63**]. There is another alternative in which the residual matrix is not included before applying PCA, known as ANOVA-simultaneous components analysis (ASCA). ASCA has demonstrated greater discriminatory power between factors, although it runs the risk of over-fitting dissimilarities [**63**]. Both approaches are very useful in the context of Design of Experiments (DoE) to visualise variability not explained by the experimental factors and the similarities in a dimension-reduced space.

Now that the basis of PCA and ANOVA coupling have been roughly laid out, let's delve into an example illustrating how applying ASCA to a dataset may look and what can be expected from the results. Consider a dataset of NIR spectra of corn samples from different locations and harvesting periods, aimed at evaluating the presence of mycotoxins. While the conventional chemometric approach would involve applying PCA, challenges arise when spectra differences caused by mycotoxin presence are overshadowed by variations in the spectra caused by other variables. Here is where ASCA comes into play.

The output obtained after applying ASCA is shown in Table 4.3, which details the effect of each factor on the total variance and the PCs that best explain each factor. It is worth noting that in many cases the factor of interest contributes relatively less to the overall variability compared to others. However, as a first step, spectra of the samples and a matrix containing all the relevant factors must be provided.

Table 4.3. Results obtained from ASCA analysis applied to NIR spectra of corn samples with different characteristics.

Factor	PCs	Effect
Mycotoxin presence	3	15.4
Location	3	12.3
Harvesting Period	2	20.4
Residuals	-	51.9

The PCs of each factor can then be plotted, like in any PCA, to visualise the variability of the PCs corresponding to each effect and look for patterns. In this study, the "mycotoxin presence" effect should ideally groups samples based on the presence of mycotoxins, as the other effects (location and harvesting period) are theoretically neglected. This way, ASCA effectively captures the variability of the most meaningful factor for analysis, significantly mitigating the overshadowing effects of other variables.

4. Summary and conclusions

This chapter has provided an overview on similarity analysis, its applications, and how to use common exploratory analysis methods as similarity analysis tools. First, the term similarity was defined as the study of the degree of likeness or resemblance between two objects (in the case of this chapter, between analytical signals or analytical features), and its importance to chemometrics was emphasised. The definition of a two-dimensional analytical signal was clarified and the two analytical magnitudes that compose it were specified as its position and intensity. Because pairs of analytical signals can vary in both of these dimensions, dissimilarity metrics can be calculated by analysing these differences. In addition, a third dissimilarity source was introduced, one where two analytical signals are identical except for a single peak, band or region.

The lack of consensus over the classification of the strategies used to report how far/close signals are was brought to notice. For this reason, a proposed classification of strategies to perform similarity analysis between pairs of analytical signals was laid out. This classification is based on the measuring approach used: (i) distance, (ii) spatial orientation, (iii) correlation, and (iv) entropy. Several examples of similarity indices based on the different strategies were presented. Furthermore, challenges of calculating similarity indices on analytical signals of more than two dimensions were specified, providing insights on how these problems could be overcome.

Similarity analysis of signal sets was approached through the different strategies which are part of the so-called multivariate exploratory data analysis. These tools are a common way to compare data sets containing a large number of analytical signals. These strategies can be classified according to the mathematical approaches into distance-based and covariance-based methods. Along with detailed explanations on how these exploratory methods work, analytical applications were laid out. Furthermore, examples on applying said methods were shown, proving how useful these could be to find natural groupings contained in real data sets. The importance of the decisions making by the analyst was also discussed, emphasising that the success of exploratory analysis is highly related to its interpretation.

Finally, it should be noted that similarity analysis, although not given the credit it deserves to date, is a crucial part of chemometrics. The applications of similarity analysis, from between pairs of analytical signals to sets of hundreds of signals, allow analysts to easily reach conclusions on the degree of relatedness of their data. It is hoped that by stating the importance of similarity analysis, and by extending theory to practical applications, it can finally be seen as the indispensable tool for analytical chemistry, and more specifically in chemometrics.

Acknowledgements

AMJC acknowledges the Grant (RYC2021-031993-I) funded by MICIU/AEI/10.13039/501100011033 and "European Union NextGeneration EU/PRTR". **AAC** gratefully acknowledges the Spanish Ministry of Science, Innovation and Universities for a pre-doctoral fellowship FPU (FPU20/04711, Formación del Profesorado Universitario, -*Training of University Lecturer*-).

References

[1] I. Batyrshin, Towards a general theory of similarity and association measures: Similarity, dissimilarity and correlation functions, J. Intell. Fuzzy Syst. 36 (2019) 2977-3004.

Doi: 10.3233/JIFS-181503.

[2] I. Batyrshin, Data science: Similarity, dissimilarity and correlation functions, in: G.S. Osivop, A.I. Panov, K.S. Yakovlev (Eds.), Artificial Intelligence, Springer, Cham, Switzerland, 2019, pp. 13-28.

[3 R. Todeschini, D. Ballabio, V. Consonni, distances and similarity measures in chemometrics and chemoinformatics, Encyclopedia of Analytical Chemistry - Online, John Wiley & Sons, Hoboken, NJ, US, 2020.

Doi: 10.1002/9780470027318.a9438.pub2.

[4] G. Alaerts, J. Van Erps, S. Pieters, M. Dumarey, A.M. van Nederkassel, M. Goodarzi, J. Smeyers-Verbeke, Y. Vander Heyden, Similarity analyses of chromatographic fingerprints as tools for identification and quality control of green tea, J. Chromatogr. B 910 (2012) 61-70.

Doi: 10.1016/j.jchromb.2012.04.031.

[5] J. Xue, Z. Yang, L. Han, L. Chen, Study of the influence of NIRS acquisition parameters on the spectral repeatability for on-line measurement of crop straw fuel properties, Fuel 117 (2014) 1027-1033.

Doi: 10.1016/j.fuel.2013.10.017.

[6] Y. Tang, Y. Liang, K.T. Fang, Data mining in chemometrics: Sub-structures learning via peak combinations searching in mass spectra, J. Data Sci. 1 (2003) 481-496.

[7] Y. Liu, Q. Meng, R. Chen, J. Wang, S. Jiang, Y. Hu, A new method to evaluate the similarity of chromatographic fingerprints: weighted Pearson product-moment correlation coefficient, J. Chromatogr. Sci. 42 (2004) 545-550.

Doi: 10.1093/chromsci/42.10.545.

[8] B.M. Teska, C. Li, B.C. Winn, K.K. Arthur, Y. Jiang, J.P. Gabrielson, Comparison of quantitative spectral similarity analysis methods for protein higher-order structure confirmation, Anal. Biochem. 434 (2013) 153-165.

Doi: 10.1016/j.ab.2012.11.018.

[9] J.B. Loudermilk, D.S. Himmelsbach, F.E. Barton, J.A. De Haseth, Novel search algorithms for a mid-infrared spectral library of cotton contaminants, Appl. Spectrosc. 62 (2008) 661-670.

[10] A. Thong, N. Basri, W. Chew, Comparison of untargeted gas chromatography-mass spectrometry analysis algorithms with implications to the interpretation and putative identification of volatile aroma compositions, J. Chromatogr. A 1713 (2024) 464519.

Doi: 10.1016/j.chroma.2023.464519.

[11] T. Matsushita, J.J. Zhao, N. Igura, M. Shimoda, Authentication of commercial spices based on the similarities between gas chromatographic fingerprints, J. Sci. Food Agric. 98 (2018) 2989-3000.

Doi: 10.1002/jsfa.8797.

[12] X. He, J. Li, W. Zhao, R. Liu, L. Zhang, X. Kong, Chemical fingerprint analysis for quality control and identification of Ziyang green tea by HPLC, Food Chem. 171 (2015) 405-411.
Doi: 10.1016/j.foodchem.2014.09.026.

[13] D. Granato, P. Putnik, D.B. Kovačević, J. Sousa Santos, V. Calado, R.S. Rocha, A. Gomes Da Cruz, B. Jarvis, O. Ye Rodionova, A. Pomerantsev, Trends in chemometrics: Food authentication, microbiology, and effects of processing, Compr. Rev. Food Sci. Food Saf. 17(2018) 663-677.
Doi: 10.1111/1541-4337.12341.

[14] M. Bovens, B. Ahrens, I. Alberink, A. Nordgaard, T. Salonen, S. Huhtala, Chemometrics in forensic chemistry – Part I: implications to the forensic workflow, Forensic Sci. Int. 301 (2019) 82-90.
Doi: 10.1016/j.forsciint.2019.05.030.

[15] H. Shin, S. Oh, D. Kang, Y. Choi, Protein quantification and imaging by surface-enhanced Raman spectroscopy and similarity analysis, Adv. Sci. 7 (2020) 1903638.
Doi: 10.1002/advs.201903638.

[16] R. Pérez-Robles, N. Navas, S. Medina-Rodríguez, L. Cuadros-Rodríguez, Method for the comparison of complex matrix assisted laser desorption ionization-time of flight mass spectra. Stability of therapeutical monoclonal antibodies, Chemom. Intell. Lab. Syst. 170 (2017) 58-69.
Doi: 10.1016/j.chemolab.2017.09.008.

[17] A. Arroyo-Cerezo, A.M. Jiménez-Carvelo, A. González-Casado, I. Ruisánchez, L. Cuadros-Rodríguez, The potential of the spatially offset Raman spectroscopy (SORS) for implementing rapid and non-invasive in-situ authentication methods of plastic-packaged commodity foods – Application to sliced cheeses, Food Control 146 (2023) 109522.
Doi: 10.1016/j.foodcont.2022.109522.

[18] B. Lavine, J. Almirall, C. Muehlethaler, C. Neumann, J. Workman Jr., Criteria for comparing infrared spectra – A review of the forensic and analytical chemistry literature, Forensic Chem. 18 (2020) 100224.
Doi: 10.1016/j.forc.2020.100224.

[19] T.G. Bloemberga, J. Gerretzena, A. Lunshofa, R. Wehrensc, L.M.C. Buydensa, Warping methods for spectroscopic and chromatographic signal alignment: A tutorial, Anal. Chim. Acta 781 (2013) 14-32.
Doi: 10.1016/j.aca.2013.03.048.

[20] Z. Liu, M. Huang, Q. Zhu, J. Qin, M.S. Kim, A packaged food internal Raman signal separation method based on spatially offset Raman spectroscopy combined with FastICA, Spectrochim Acta A Mol Biomol Spectrosc. 275 (2022) 121154.
Doi: 10.1016/j.saa.2022.121154.

[21] D.A. Sheena, W.F.C. Rocha, K.A. Lippa, D.W. Bearden, A scoring metric for multivariate data for reproducibility analysis using chemometric methods, Chemom. Intell. Lab. Syst. 162 (2017) 10-20.
Doi: 10.1016/j.chemolab.2016.12.010.

[22] R., Rao, B. Zbell, Normalized spectral similarity score (NS3) as an efficient spectral library searching method for hyperspectral image classification, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 4 (2011) 226-240.
Doi: 10.1109/JSTARS.2010.2086435.

[23] R.G. Brereton, Exploratory data analysis, in: R.G. Brereton (Ed.), Chemometrics for Pattern Recognition, John Wiley & Sons, Chichester, U.K., 2009, pp. 47-106.

[24] R. Zeng, J.P. Zhang, K. Cai, W.C. Gao, W.J. Pan, C.Y. Jiang, P.Y. Zhang, B.W. Wu, C.H. Wang, X.Y. Jin, D.C. Li, How similar is "similar", or what is the best measure of soil spectral and physiochemical similarity?, PLoS ONE 16 (2021) e0247028.
Doi: 10.1371/journal.pone.0247028.

[25] D.J. Degnan, J.E. Flores, E.R. Brayfindley, V.L. Paurus, B.M. Webb-Robertson, C.S. Clendinen, L.M. Bramer, Characterizing families of spectral similarity scores and their use cases for gas chromatography–mass spectrometry small molecule identification, Metabolites 13 (2023) 1101.
Doi: 10.3390/metabo13101101.

[26] J.E. Gentle, Vectors and vector spaces, in: J.E. Gentle (Ed.), Matrix Algebra. Theory, Computations, and Applications in Statistics, Springer, New York, USA, 2007, pp. 9-36.

[27] P. Zerzucha, B. Walczak, Concept of (dis)similarity in data analysis, Trends Anal. Chem. 38 (2012) 116-128.
Doi: 10.1016/j.trac.2012.05.005.

[28] L. Ramirez-Lopez, T. Behrens, K. Schmidt, R.A. Viscarra Rossel, J.A.M. Demattê, T. Scholten, Distance and similarity-search metrics for use with soil vis–NIR spectra, Geoderma 199 (2013) 43-53.
Doi: 10.1016/j.geoderma.2012.08.035.

[29] A. Arroyo-Cerezo, M. Medina-García, L. Cuadros-Rodríguez, D.N. Rutledge, A.M. Jiménez-Carvelo, Chemometric enhancement for blind signal resolution from non-invasive spatially offset Raman spectra, Chemom. Intell. Lab. Syst. 243 (2023) 105027.
Doi: 10.1016/j.chemolab.2023.105027.

[30] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379-423.
Doi: 10.1002/j.1538-7305.1948.tb01338.x.

[31] J. Walach, P. Filzmoser, K. Hron, Data normalization and scaling: Consequences for the analysis in omics sciences, in: J. Jaumot, C. Bedia, R. Tauler (Eds.), Data Analysis for Omics Sciences: Methods and Applications, in: D. Barceló (Ed.), Comprehensive Analytical Chemistry, vol. 82, Elsevier, Amsterdam, 2018, pp. 165-196.

[32] R. Joshi, S. Kumar, A dissimilarity measure based on Jensen Shannon divergence measure, Int. J. Gen. Syst. 48 (2019) 280-301.
Doi: 10.1080/03081079.2018.1552685.

[33] B. Ceccanti, G. Masciandaro, C. Macci, Pyrolysis-gas chromatography to evaluate the organic matter quality of a mulched soil, Soil Till. Res. 97 (2007) 71-78.

Doi: 10.1016/j.still.2007.08.011.

[34] M.P. Rueda, F. Comino, V. Aranda, A. Domínguez-Vidal, M.J. Ayora-Cañada, Analytical pyrolysis (Py-GC-MS) for the assessment of olive mill pomace composting efficiency and the effects of compost thermal treatment, J. Anal. Appl. Pyrol. 168 (2022) 105711.
Doi: 10.1016/j.jaap.2022.105711.

[35] S.H. Cha, S. Yoon, C.C. Tappert, Enhancing binary feature vector similarity measures, CSIS Technical Reports 18 (2005).

[36] T. Bien, K. Koerfer, J. Schwenzfeier, K. Dreisewerd, J. Soltwisch, Mass spectrometry imaging to explore molecular heterogeneity in cell culture, Proc. Natl. Acad. Sci. 119 (2022), e2114365119.
Doi: 10.1073/pnas.2114365119.

[37] Q. Zhou, L. Chen, C. Peng, Full-field vibration intensity chromatographic analysis based on vision measurement for rotating machinery, in Advances in Guidance, Navigation and Control: Proceedings of 2020 International Conference on Guidance, Navigation and Control, ICGNC 2020 (October 23–25, 2020), Springer: Tianjin, China, 5505-5514.
Doi: 10.1007/978-981-15-8155-7_455.

[38] F. Hollaus, S. Brenner, R. Sablatnig, CNN based binarization of multispectral document images, in: International Conference on Document Analysis and Recognition ICDAR (September 2019) 533-538.
Doi: 10.1109/ICDAR.2019.00091.

[39] D.B. Hibbert, Vocabulary of concepts and terms in chemometrics (IUPAC Recommendations 2016), Pure & Appl Chem 88 (2016) 407-443.
Doi: 10.1515/pac -2015-0605.

[40] F. Murtagh, P. Legendre, Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?, J. Classif. 31 (2014) 274-295.
Doi: 10.1007/s00357-014-9161-z.

[41] D. Bajusz, A. Rácz, K. Héberger, Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?, J. Cheminf. 7 (2015) 1-13.
Doi: 10.1186/s13321-015-0069-3.

[42] S.S. Choi, S.H. Cha, C.C. Tappert, A survey of binary similarity and distance measures. J. Syst. Cybern. Informatics 8 (2010) 43-48.

[43] M. Stahl, Peak purity analysis in HPLC and CE using diode-array technology, Agilent Technologies 2003.

[44] Agilent Technologies, Understanding your spectra module, Agilent OpenLAB CDS ChemStation Edition, Ed. 01/13, 2013.

[45] F.A. Franchina, D. Zanella, E. Lazzari, P.H. Stefanuto, J.F. Focant, Investigating aroma diversity combining purge-and-trap, comprehensive two-dimensional gas chromatography, and mass spectrometry, J. Sep. Sci. 43 (2020) 1790-1799.
Doi: 10.1002/jssc.201900902.

[46] F. Ortega-Gavilán, A.M. Jiménez-Carvelo, L. Cuadros-Rodríguez, M.G. Bagur-González, The chromatographic similarity profile – An innovative methodology to detect fraudulent blends of virgin olive oils, J. Chromatogr. A 1679 (2022) 463378. Doi: 10.1016/j.chroma.2022.463378.

[47] A. Barcaru, G. Vivó-Truyols, Use of bayesian statistics for pairwise comparison of megavariate data sets: Extracting meaningful differences between GCxGC-MS chromatograms using Jensen–Shannon divergence, Anal. Chem. 88 (2016) 2096-2104.
Doi: 10.1021/acs.analchem.5b03506.

[48] Y. Jiang, S. Wang, H. Qin, B. Li, Q. Li, Similarity quantification of 3D surface topography measurements, Measurement 186 (2021) 110207.
Doi: 10.1016/j.measurement.2021.110207.

[49] A.E. Sinha, J.L. Hope, B.J. Prazen, E.J. Nilsson, R.M. Jack, R.E. Synovec, Algorithm for locating analytes of interest based on mass spectral similarity in GC × GC–TOF-MS data: analysis of metabolites in human infant urine, J. Chromatogr. A 1058 (2004) 209-215.
Doi: 10.1016/j.chroma.2004.08.064.

[50] I. Noda, Recent developments in two-dimensional (2D) correlation spectroscopy, Chin Chem. Lett. 26 (2015) 167-172.
Doi: 10.1016/j.cclet.2014.10.006.

[51] Y. Park, S. Jin, I. Noda, Y.M Jung, Continuing progress in the field of two-dimensional correlation spectroscopy (2D-COS), part I. Yesterday and today, Spectrochim. Acta A Mol. Biomol. Spectrosc. 281 (2022) 121573.
Doi: 10.1016/j.saa.2022.121573.

[52] A.C. Rencher, Methods of Multivariate Analysis. 2nd Ed., John Wiley & Sons, Inc., 2002.
Doi: 10.1002/0471271357.

[53] W.K. Härdle, L. Simar, Applied Multivariate Statistical Analysis, 5th Ed., Springer, Cham, Switzerland, 2019.

[54] A.J. Izenman, Modern Multivariate Statistical Techniques. Regression, Classification, and manifold Learning, Springer, New York, USA, 2008.

[55] K. Héberger, Chemoinformatics-multivariate mathematical-statistical methods for data evaluation, in: K. Vékey, A. Telekes, A. Vertes (Ed.), Medical Applications of Mass Spectrometry, Elsevier, Amsterdam, Netherlands, 2008, pp. 141-169.

[56] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments. Philos. Transact. A Math. Phys. Eng. Sci. 374 (2016) 20150202.
Doi: 10.1098/rsta.2015.0202.

[57] K. Kumar, Principal component analysis: Most favourite tool in chemometrics, Resonance 22 (2019 747-759.
Doi: 10.1007/s12045-017-0523-9

[58] R. Warne, A primer on multivariate analysis of variance (MANOVA) for behavioural scientists, Pract. Assess. Res. Eval. 19 (2014) 17.
Doi: 10.7275/sm63-7h70

[59] C. Bertinetto, J. Engel, J. Jansen, ANOVA simultaneous component analysis: A tutorial review, Anal. Chim. Acta: X 6 (2020) 100061.
Doi: 10.1016/j.acax.2020.100061.

[60] G. Zwanenberg, H.C. Hoefsloot, J.A. Westerhius, J.J. Jansen, A. K. Smilde, ANOVA-principal component analysis and ANOVA-simultaneous component analysis: a comparison, J. Chemom. 25 (2011) 561-567.
Doi: 10.1002/cem.1400.

[61] L. Stáhle, S. Wold, Multivariate analysis of variance (MANOVA), Chemom. Intell. Lab. Syst. 9 (1990) 127-141.
Doi: 10.1016/0169-7439(90)80094-M.

[62] P.D.B. Harrington, N.E. Vieira, J. Espinoza, J.K. Nien, R. Romero, A.L. Yergey, Analysis of variance–principal component analysis: A soft tool for proteomic discovery, Anal. Chim. Acta 544 (2005) 118-127.
Doi: 10.1016/j.aca.2005.02.042.

[63] G. Zwanenburg, H.C. Hoefsloot, J.A. Westerhuis, J.J Jansen, A.K. Smilde, ANOVA–principal component analysis and ANOVA–simultaneous component analysis: a comparison, J. Chemom. 25 (2011) 561-567.
Doi: 10.1002/cem.1400.

4.3. Artículo científico 5

Chemometric enhancement for blind signal resolution from non-invasive spatially offset Raman spectra.

Publicado en 2023 en la revista Chemometrics and Intelligent Laboratory Systems 243, 105027.

DOI: 10.1016/j.chemolab.2023.105027.

Chemometrics and Intelligent Laboratory Systems 243 (2023) 105027

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Chemometric enhancement for blind signal resolution from non-invasive spatially offset Raman spectra

Alejandra Arroyo-Cerezo [a,b,*,1], Miriam Medina-García [a,*,1], Luis Cuadros-Rodríguez [a,b], Douglas N. Rutledge [c,d], Ana M. Jiménez-Carvelo [a]

[a] Department of Analytical Chemistry, University of Granada, C/ Fuentenueva s/n, 18071, Granada, Spain
[b] Biohealth Research Institute (ibs.GRANADA), University of Granada, Granada, Spain
[c] Faculté de Pharmacie, Université Paris-Saclay, 91400, Orsay, France
[d] Muséum National d'Histoire Naturelle, 75005, Paris, France

[*] Corresponding author. Department of Analytical Chemistry, University of Granada, C/ Fuentenueva s/n, 18071, Granada, Spain.
E-mail addresses: arroyoc@ugr.es (A. Arroyo-Cerezo), miriammedina@ugr.es (M. Medina-García).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.chemolab.2023.105027
Received 11 September 2023; Received in revised form 24 October 2023; Accepted 1 November 2023
Available online 7 November 2023
0169-7439/© 2023 Elsevier B.V. All rights reserved.

Highlights:

- Mixed signals of substances measured noninvasively through container were resolved.
- More accurate spectra compared to the equipment software resolution were obtained.
- Potential of MCR and ICA methods in SORS resolution was demonstrated.
- Similarity indexes based on different mathematical principals were applied.

Keywords:

Mixed Raman spectra

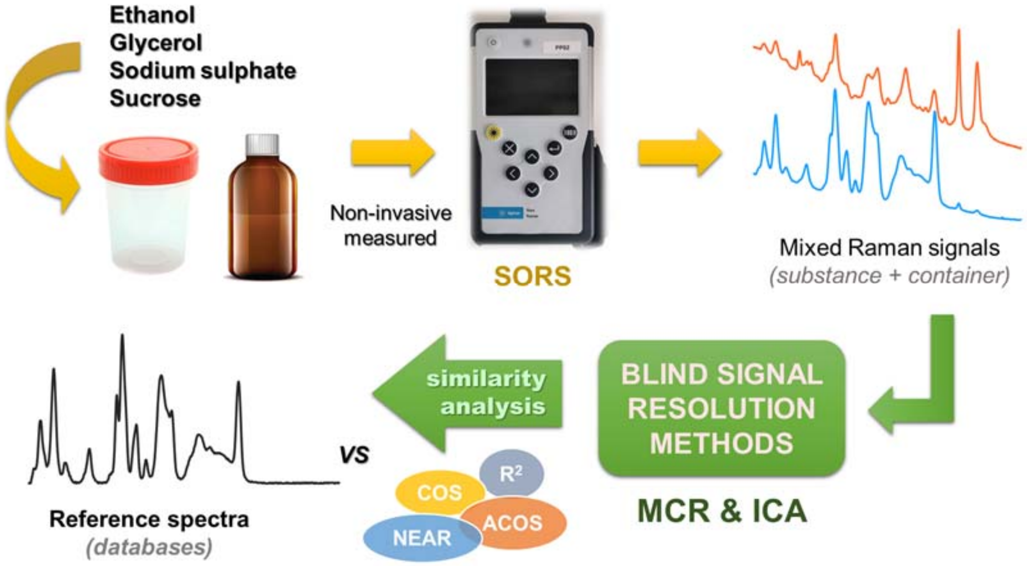
Spectra separation

Independent components analysis

Multivariate curve resolution

Similarity analysis

Graphical abstract



Abstract

Spatially offset Raman spectroscopy (SORS) is a promising spectroscopic technique that enables the collection of Raman signals from deeper layers of materials compared to conventional Raman spectroscopy. SORS equipment allows acquiring Raman spectra of substances through packaging, making it a non-invasive analytical technique. Note that the acquired spectrum is always the result of mixing two contributions: (1) the spectrum of the container material, and (2) the spectrum of the substance inside the container. Nowadays, SORS equipment are supported by software capable of removing the surface contribution from the recorded spectral data. However, the optimal extraction of this contribution is not achieved in all cases. This study explored the potential of two chemometric methods, Multivariate Curve Resolution (MCR) and Independent Components Analysis (ICA), for resolving the mixed spectra acquired by Vaya Raman equipment (Agilent) of four standard substances (sucrose, anhydrous sodium sulphate, ethanol 96% and glycerol) analyzed through containers of two different plastic materials (polypropylene and polyethylene terephthalate). The two resulting resolved spectra (by MCR and by ICA) and the one resolved by the equipment software were compared by similarity analysis with the Raman spectra of the standard substances available in recognized databases intended as target spectra. Similarity analysis was performed by calculating four similarity indexes, namely: cosine, arccosine, coefficient of determination and nearness index. Both MCR and ICA methods successfully extracted pure Raman spectra of the test substances more similar to the standard substances than the spectra resolved by the equipment software. These results highlighted the potential of both methods for resolving complex mixed signals, such as those acquired through the SORS technique, thereby enhancing its utility.

1. Introduction

Spatially offset Raman spectroscopy (SORS) is an innovative spectroscopic mode characterized by collecting the Raman signal of the measured material at a point shifted with respect to the laser incidence point in contrast to conventional Raman spectroscopy, where the signal is collected at the incident point of the beam. This has allowed to focus attention on the development of non-invasive analytical methods based on SORS in different fields such as medical, pharmaceutical, food, among others [1]. Thanks to the offset when performing a SORS measurement, it is possible to retrieve the Raman signal from deep layers of the measured material, since the laser can strike deeper into non-reflective materials. Some of the most relevant and recent applications using SORS are the quantification of active ingredients in hand sanitizers measured through original container [2], the identification of changes in red blood cell concentrates within standard plastic transfusion bags [3], the detection of drugs in containers [4] or the characterization and quality assessment of packaged food [5].

Since the birth of the technique by Matousek in 2005 [6], homemade modified conventional Raman spectrometers have been used to perform SORS measurements. However, it was commercialized shortly thereafter, and SORS-based portable and handheld equipment are now available, such as those supplied by Agilent Technologies (Santa Clara, CA, USA) [7].

The technology developed to perform SORS measurements is based on the acquisition of two independent signals in a single measurement. The first signal acquired corresponds to the Raman spectrum collected at the same point of incidence of the laser (hereafter referred to as 'zero spectrum'). This is equivalent to a conventional Raman measurement, resulting in a higher contribution from the surface layers of the measured material. The second signal is collected at a point shifted from the point of incidence of the laser ('offset spectrum' from now on). In this case, the collected signal contains a higher contribution from the deep layers of the measured material, as the laser travels deeper allowing the scattered light to be collected through the surface layers. The depth reached by the laser will depend on the offset applied [8]. However, the offset-collected signal must again pass back through the surface layers to be collected, and therefore the offset spectrum also contains a contribution from these layers, although in smaller proportion.

Thus, to obtain the spectrum of the measured material inside the container without any contribution from it, it is necessary to optimize the measurement or post-measurement processing [9]. At this point, it should be noted that the post-measurement extraction of the pure spectrum from the interior could be performed mainly in two ways: scaled subtraction or by multivariate statistical methods [10]. In this context, some software implemented in commercial SORS equipment typically perform a scaled subtraction of the zero spectrum from the offset spectrum, followed by other spectral corrections such as 0-1 intensity scaling (normalization). However, when the material to be measured is more chemically complex or the composition and thickness of the containers are not of the same material as the one being evaluated, the automation of the spectral resolution process falls short, and the results may not be optimal. This could be the situation in the internal quality control (IQC) of finished products to be marketed, in the pharmaceutical or food industry, where the materials measured are more complex and there is a greater variety of containers. As stated by Mosca et al. [1], no specific software for SORS spectra resolution is currently available. Most studies using SORS are focused on optimizing the distance (offset) at which the measurements should be performed [8,11] or the appropriate thickness of the surface layers [12,13]. But when the offset spectrum is acquired, the Raman signal must pass through the surface layers, so its contribution will always be present in this spectrum. In addition, the optimal offset is difficult to determine [9]. Therefore, the data processing stage after acquisition should be the focal point to optimize the SORS results and obtain the pure Raman spectrum of the subsurface layers.

Spectral data acquired by SORS should be considered as a mixture of signals. Some studies have investigated the use of chemometric methods such as Principal Components Analysis (PCA) or Non-negative Matrix Factorization (NMF) to separate Raman signals of each of the chemical compounds in the measured material [14]. Multivariate Curve Resolution (MCR) [15] is another widely used method to separate Raman signals from specific chemical species [16]. Other methods such as Band Target Entropy Minimization (BTEM) [17,18] or Independent Components Analysis (ICA) [19] have been used to this purpose. However, these studies aimed to identify, quantify, or detect specific chemical compounds from spectra acquired by conventional Raman. There are few studies focused on optimizing the processing and resolution of the signal collected by SORS. Only Liu et al. [9,20] used FastICA to achieve pure spectra of inner layers from the spectral data by SORS measurements, and Churchwell et al. [21] to obtain pure Raman spectra of bones measured by SORS through the skin via BTEM.

In this context, SORS measurements could be considered within the framework of the so-called 'cocktail party problem' of a signal unmixing, which is solved by Blind Signal Separation (BSS) methods [22]. As part of these methods are the aforementioned methods such as ICA [23,24] and MCR [14], since both decompose the mixed input signals into components that should correspond to the original independent signals. Both methods have often been compared although they should not always be used for the same purpose, it depends on the input data and applied algorithm [25]. When dealing with the problem of blind extraction of source signals, a crucial step is to assure that the results are as expected, that is, that the original pure signals were successfully extracted.

In this sense, evaluating the similarity between the signal obtained and the target signal would be the most objective and optimal way to do so. Over time, many similarity indexes have been used to compare two spectra with each other, since by a single numeric parameter value it is possible to compare two signals, embedded in either data vectors consisting of hundreds or thousands of elementary data [26]. Some indexes are based on the correlation or the normalized covariance between data such as the so-called Similarity Index (SI) which is calculated from the coefficient of correlation (r) or the coefficient of determination (R2), both for the whole spectral range and for selected regions of interest [27,28]. Other indexes are based on the angle formed by the two data vectors such as the Spectral Angle Mapper (SAM) calculated from the cosine or arccosine of that angle [25,29,30]. Furthermore, several indexes based on the

Euclidean distance between the two vectors have also been developed [31], such as the Nearness Index (NEAR) proposed by researchers of our group [32,33]. And some studies have used the Spectral Information Divergence (SID), which is based on information entropy but is usually applied to pixel comparisons in hyperspectral images [34,35]. In addition, the use of error metrics, such as the Root Mean Square (RMS), also called the quadratic mean, has also been proposed [36]. In short, there is a wide range of criteria available, so it is difficult to have a single similarity index, since they respond to different parameters [37]. There is so far no consensus on the most suitable index to be used in similarity analysis, and many studies propose the joint use of several indexes to get an overall similarity result instead of a single index [38] or even using hybrids such as SAM-SID [39,40].

The goal of the present study was to evaluate the application of two chemometric methods (ICA and MCR) for the resolution of signals acquired by SORS in order to obtain the pure Raman spectrum of a substance measured through a container. The main purpose was testing if the resulting spectrum would be acceptably similar to that acquired by a direct Raman measurement of the substance (i.e., not through the container), and thus demonstrate this technique could be used in parallel to the conventional technique for ICQ of complex manufactured products, such as food, so that a new model development is not required. For this end, four recognized substances (ethanol, glycerol, sodium sulphate anhydrous and sucrose), whose pure Raman signals are fully characterized, and two containers of different materials (polypropylene and polyethylene terephthalate) were used. The results obtained with ICA and MCR were evaluated by similarity analysis with the target Raman spectra and compared with the final spectrum provided by the software of the SORS equipment used. For this, four similarity indexes were used based on (i) the coefficient of determination, (ii) the cosine of the angle formed by both data vectors, (iii) the angle itself, and (iv) the Euclidean distance between the two spectral data vectors. For a better comparison and visualization of the results, the index values have been 0-1 normalized, where in all cases the value 0 denotes maximum dissimilarity and 1 shows full sameness.

2. Material and methods

2.1. Trials: test substances and containers

Four standard substances were selected for this study in order to test the reliability of the extraction of the corresponding Raman spectra, namely: glycerol 99% provided by Panreac Quimica SLU (Barcelona, Spain), ethanol 96% by VWR (Darmstadt, Germany), anhydrous sodium sulphate by Panreac Quimica SLU (Barcelona, Spain) and sucrose by Sigma-Aldrich (Missouri, USA).

Two containers of different plastic materials were used for this study: colourless translucent polypropylene (PP) and transparent brown polyethylene terephthalate (PET). Both materials are characterized by a well-established Raman spectrum of sufficient intensity. This resulted in a total of eight trials to be measured by SORS.

2.2. Data acquisition

SORS measurements were performed using a portable Vaya Raman instrument (Agilent, Technologies, Santa Clara, USA). This spectrometer collects the Raman signal in the range 350-2000 cm^{-1}, with spectral resolution of 12-20 cm^{-1}, after excitation with a laser at 830 nm with 450 mW of power. Each substance was measured 10 times through each of the two containers. In addition, the empty containers (PP and PET) were measured 10 times to provide additional spectral information for the blind signal resolution stage. Each measurement was performed in less than 2 minutes in dark conditions so that ambient light could not affect the Raman spectra acquired by the instrument.

The spectrometer provides three files per measurement: (i) a zero spectrum (Raman signal collected without offset), (ii) an offset spectrum (Raman signal collected with 7 mm offset) and (iii) a final spectrum (a baseline corrected spectrum 0-1 normalized after scaled subtraction performed by the equipment software). The latter will be referred to as the final-SORS from now on. Figure 4.15 shows an example of the three spectra (zero, offset and final-SORS) obtained from a single measurement. Note that both the zero spectrum and offset spectrum acquired per trial contain contributions from both the substance and the material container.

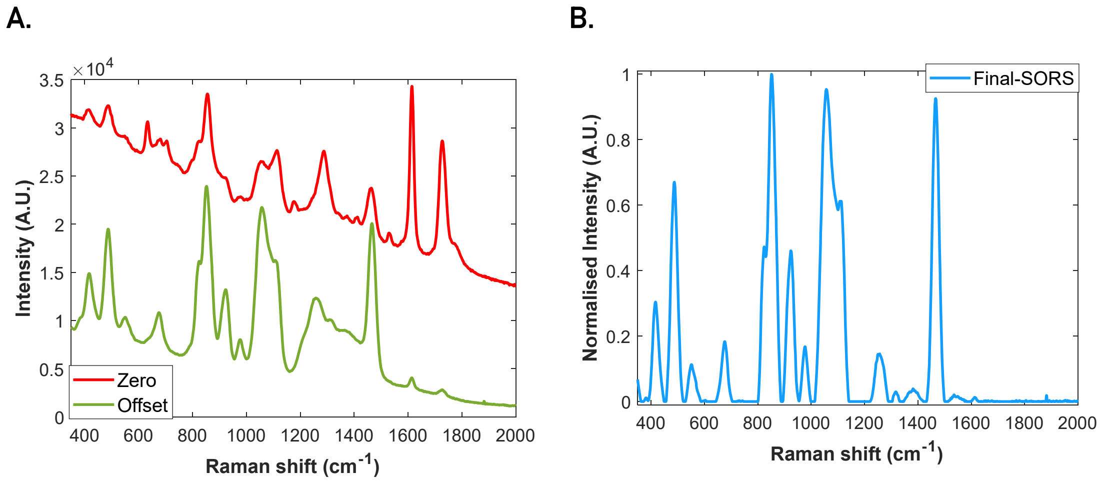


Figure 4.15. Raman spectra acquired of glycerol packaged in polyethylene terephthalate (PET) from a single SORS measurement: (**A**) zero and offset unprocessed spectra and (**B**) final-SORS spectrum resolved and processed (baseline and intensity normalization) by the equipment software.

2.2.1. Reference Raman spectra

In order to have recognized reference of the pure Raman spectra of the standard substances under study and to be able to carry out a subsequent valid comparison for the evaluation of the results, the Raman spectra of the four test substances were extracted from free databases of spectral libraries. The glycerol and ethanol spectra were retrieved from the Raman Spectral Database of KnowItAll [41], the anhydrous sodium sulphate spectrum from NICODOM Ltd Raman Spectral Database [42] and the sucrose spectrum from Specarb Raman Spectra of carbohydrates database [43]. From now on these spectra will be called target spectra.

2.3. Data processing

The zero and offset spectra were used to process and resolve the pure Raman spectrum of the substance inside the container. For this purpose, the files were exported in 'comma separated value' (CSV) format and converted to MATLAB format (ver R2022a, Mathworks, MA, USA), thus each spectrum was in a data vector consisting of 1651 elementary data points with positive intensity values. These data vectors were used to create a matrix for each trial consisting of 30 spectra: 10 zero and 10 offset spectra of the test substance measured through the corresponding container, plus 10 zero spectra of the empty container. The use of the zero spectra of the empty container was necessary to provide information on the spectral contribution of the container (PP or PET), which was also implicit in the measurements of the substance inside the containers, both in the zero and offset spectra (although to a lesser contribution in the latter). Only baseline correction was performed as pre-processing by two methods: weighted least squares (WLS) [44] and Whittaker filter [45].

2.3.1. Blind signal resolution

Two chemometric methods were used for the resolution of the SORS mixed signals in order to extract the pure Raman spectrum of the four test substances: Independent Components Analysis (ICA) and Multivariate Curve Resolution (MCR). Note that the 30 spectra × 1651 variables matrix of each trial were the input data signals for both methods. Additionally, a Raman spectrum of the empty container (average of the 10 replicates) was imposed when applying both ICA and MCR methods, so that one of the extracted components corresponded to the spectrum of the container, and so another component could match the spectrum of the material inside. In this way, the purpose of extracting the spectrum of the inside substance as pure as possible was achieved. Notice the decision to use the spectrum of empty container instead the spectrum of the substance inside, arises from the aim of the study: to test the usability of SORS for final product IQC in food industries. In that situation, having a single reference spectrum of a food product is difficult (e.g., there is no single spectrum characterizing all olive oils, since no two olive oils are the same), but the container is already well known, and the spectrum of the empty container is easily available in that company. A more detailed explanation of how this empty container spectrum was imposed for each method is given below for each method.

For each matrix (one per trial) and each chemometric tool (ICA and MCR), three models were developed from the dataset (i) without pre-processing, (ii) baseline corrected applying WLS, and (iii) baseline corrected applying Whittaker filter. The results were then evaluated by performing a similarity analysis to choose the optimal pre-processing method for the acquired data.

Independent Component Analysis (ICA)

ICA was performed using the Joint Approximate Diagonalization of Eigenmatrices (JADE) algorithm [**46**], running in the MATLAB environment. The optimal number of Independent Components (ICs) for each case was determined using the Random_IC_by_Blocks method [**47**] and the Durbin-Watson method [**46**]. Random_ICA_by_blocks randomly distribute the rows of the X-matrix into 2 blocks of approximately equal sizes. For each of these predefined blocks, ICA models were computed, with from 1 to Fmax ICs. In each case, the correlation between the extracted ICs and the empty container spectrum was calculated in order to find the number of ICs in which one of them achieves the maximum correlation with the given empty container spectrum. When too many ICs are extracted, the correlations decrease. To avoid the possibility of a bias being introduced by a particular distribution of the rows into the blocks, the whole procedure is repeated k times resulting in different sets of blocks, producing a broader perspective for the selection of the optimal number of ICs. The Durbin-Watson method calculates a signal to noise ratio for each row in the residual matrix after extraction of each successive IC. When too many ICs have been extracted, the Durbin-Watson limit criterion increases.

Each trial matrix was decomposed into two characteristic matrices consisting of vectors: one containing the source signals (ICs) present in the mixed input signal, and another containing the proportions of the extracted signals (comparable to PCA scores) [**22**]. The first matrix obtained from decomposition was intended here to contain at least the two pure Raman signals for each trial: (i) the test substance and (ii) the material container.

Multivariate Curve Resolution (MCR)

The PLS_Toolbox software (ver 9.1, Eigenvector Research Inc. MA, USA) was used under MATLAB environment to perform the MCR analyses. Two components were manually selected for all cases, based on our starting hypothesis: the number of output signals should be equal to the number of spectral components present: (i) the substance inside and (ii) the container material. An iterative resolution method was applied, i.e., the resolution process was optimized considering the Raman spectrum of the empty container as a spectral contrast constraint for each MCR developed [14], assigning 100% to one component and 0% to the other component. In addition to this, no further constraints were applied beyond the default ones (non-negativity constraint on both contributions and spectra).

2.3.2. Similarity analysis

Four different similarity metrics have been explored in this study. The similarity analysis between two spectra was based on the evaluation of the covariance (coefficient of determination) between the two signal vectors, the difference in directions in the n-dimensional orthogonal space defined by the number of components of the vectors (cosine of the angle and the angle itself), and the Euclidean distance (nearness index) between the vectors concerned. Moreover, all similarity indexes values were 0-1 normalized in order to facilitate comparison of the values. A short summary of each metric is presented below, showing both algebraic and matrix equations for ease of use. Note that the matrix equations are formulated for one-row matrices (vectors). In addition, the dot "·" is used to symbolize a scalar product, while the cross "×" is indicating a vector product.

Cosine (COS) and arccosine (ACOS)

The similarity between two vectors could be expressed by the cosine of the angle (θ) they form. This is a widely used metric to compare spectra and assess their similarity, e.g., if the cosθ value is 1, θ is zero and the two spectral vectors have the same orientation. The estimation of the COS index, i.e., the cosine between two vectors, $\mathbf{Y_A}$ and $\mathbf{Y_B}$, could be calculated using the equations (1) [32]:

Algebraic equation
$$\mathrm{COS}\ (\mathbf{Y_A}, \mathbf{Y_B}) = \frac{\sum(y_{A_i} \cdot y_{B_i})}{\sqrt{\sum {y_{A_i}}^2 \cdot \sum {y_{B_i}}^2}} \tag{1}$$

Matrix equation
$$\mathrm{COS}\ (\mathbf{Y_A}, \mathbf{Y_B}) = \frac{\mathbf{Y_A} \times \mathbf{Y_B}^{\mathrm{T}}}{\sqrt{(\mathbf{Y_A} \times \mathbf{Y_A}^{\mathrm{T}}) \cdot (\mathbf{Y_B} \times \mathbf{Y_B}^{\mathrm{T}})}}$$

The T superscript denotes the transposed matrix.

Due to the intrinsic sinusoidal form of the cosine function, it is insensitive to minor differences between similar signals since it takes values very close to 1. For this reason, it was also decided to calculate the arccosine index.

The arccosine is the mathematical function that calculates the angle formed by two vectors from the value given by the cosine. In other words, it is the inverse function of the cosine. Obviously, this function varies linearly with the angle and is therefore more responsive or sensitive than the cosine function when the signals are very similar, i.e., when the angle formed by them is close to 0.

This ACOS index could be estimated from COS value according to equation (2):

Algebraic equation

$$\mathrm{ACOS}^{\mathrm{N}}\,(\mathbf{Y_A},\mathbf{Y_B}) \;=\; \theta^{\mathrm{N}}\,(\mathbf{Y_A},\mathbf{Y_B}) \;=\; 1-\frac{1}{\theta_{\max}}\cdot \arccos\,(\mathrm{COS}\,\theta) \qquad (2)$$

To normalize the arccosine function, it has been divided by the maximum angle (θ_{max}) possible for the vectors in the domain of positive intensity values, in this case, by π/2 rad (or 90°).

Note that both the COS and ACOS indexes only address the orientation of the two vectors in space, so that changes in the global intensity of the signal, i.e., the corresponding moduli (of one or both), do not result in a change in the calculated value.

Coefficient of determination (R^2)

The coefficient of determination is a measure of similarity commonly derived from the variance analysis used to compare two datasets. It is calculated as the square of the Pearson correlation coefficient between two datasets. The coefficient of determination reports the ratio between the square of the covariance and the products of the variances of the elements of both vectors. As in the previous case, the R^2 index is not affected by the signal intensities, i.e., the lengths of the vectors.

Therefore, the R^2 value between two vectors, denoted by $\mathbf{Y_A}$ and $\mathbf{Y_B}$, respectively, could be calculated using the equations (3) **[32]**:

Algebraic equation

$$\mathrm{R}^2\,(\mathbf{Y_A},\mathbf{Y_B}) = \frac{\sum\left((y_{A_i}-\bar{y}_A)\cdot(y_{B_i}-\bar{y}_B)\right)^2}{\sum\left(y_{A_i}-\bar{y}_A\right)^2\cdot\sum\left(y_{B_i}-\bar{y}_B\right)^2} \qquad (3)$$

Matrix equation

$$\mathrm{R}^2\,(\mathbf{Y_A},\mathbf{Y_B}) \;=\; \frac{[\,(\mathbf{Y_A}-\mathbf{\bar{Y}_A})\times(\mathbf{Y_B}-\mathbf{\bar{Y}_B})^{\mathrm{T}}\,]^2}{[(\mathbf{Y_A}-\mathbf{\bar{Y}_A})\times(\mathbf{Y_A}-\mathbf{\bar{Y}_A})^{\mathrm{T}}]\cdot[(\mathbf{Y_B}-\mathbf{\bar{Y}_B})\times(\mathbf{Y_B}-\mathbf{\bar{Y}_B})^{\mathrm{T}}]}$$

Nearness index (NEAR)

The Nearness index, previously defined **[32]**, measures the Euclidean distance between the points that define the extremes of the signal vectors in the orthogonal n dimensional space. Equations (4) and (6) define this index for the vectors $\mathbf{Y_A}$ and $\mathbf{Y_B}$ in algebraic and matrix forms respectively, where d^N

symbolizes the normalized distance and is calculated as the Euclidean distance divided by the maximum distance (d_{max}). The maximum distance corresponds to the situation in which the difference between moduli of the vectors is maximal, and the bounded angle is maximal (90°, i.e., each vector lies on one of the axes of the n-dimensional space), interpreted as follows for the d_{max} calculation, considering a hypothetical situation in which: the module of one vector is reduced to 0, and the module of the other vector concentrates the moduli of the two original vectors, by their sum. Thus, the angle formed by both is implicitly the maximum. Equations (5) and (7) show the calculation of d^N.

Algebraic equations

$$\mathrm{NEAR}(\mathbf{Y_A},\mathbf{Y_B}) = 1 - \mathrm{d}^N(\mathbf{Y_A},\mathbf{Y_B}) = 1 - \sqrt{\frac{\sum(y_{A_i} - y_{B_i})^2}{\sum(y_{A_i} + y_{B_i})^2}} \quad (4)$$

$$\mathrm{d}^N(\mathbf{Y_A},\mathbf{Y_B}) = \frac{\mathrm{d}(\mathbf{Y_A},\mathbf{Y_B})}{\mathrm{d}_{max}(\mathbf{Y_A},\mathbf{Y_B})} = \sqrt{\frac{\sum(y_{A_i} - y_{B_i})^2}{\sum(y_{A_i} + y_{B_i})^2}} \quad (5)$$

Matrix equations

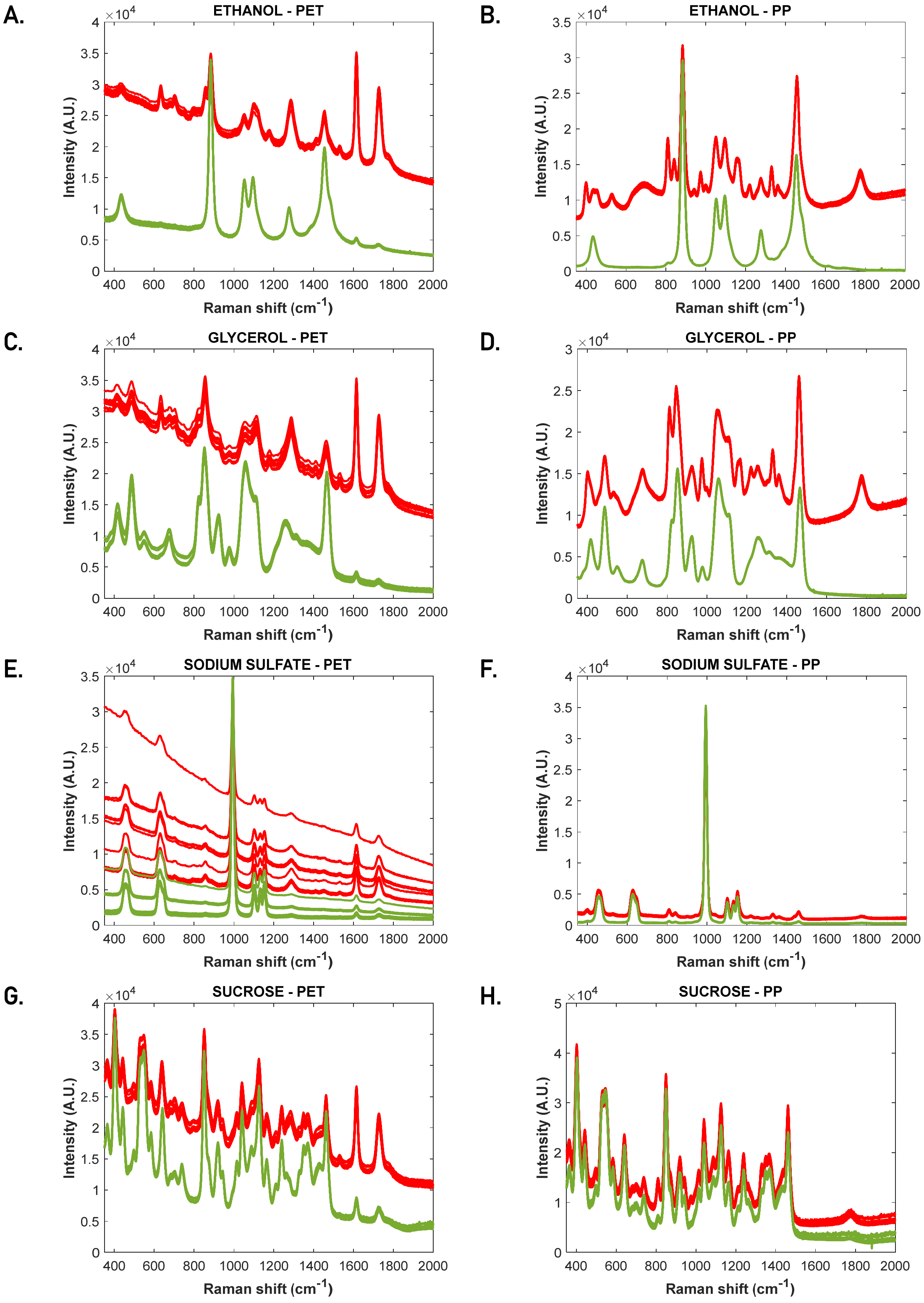
$$\mathrm{NEAR}(\mathbf{Y_A},\mathbf{Y_B}) = 1 - \sqrt{\frac{\sum(\mathbf{Y_A} - \mathbf{Y_B}) \times (\mathbf{Y_A} - \mathbf{Y_B})^T}{\sum(\mathbf{Y_A} + \mathbf{Y_B}) \times (\mathbf{Y_A} + \mathbf{Y_B})^T}} \quad (6)$$

$$\mathrm{d}^N(\mathbf{Y_A},\mathbf{Y_B}) = \frac{\mathrm{d}(\mathbf{Y_A},\mathbf{Y_B})}{\mathrm{d}_{max}(\mathbf{Y_A},\mathbf{Y_B})} = \sqrt{\frac{\sum(\mathbf{Y_A} - \mathbf{Y_B}) \times (\mathbf{Y_A} - \mathbf{Y_B})^T}{\sum(\mathbf{Y_A} + \mathbf{Y_B}) \times (\mathbf{Y_A} + \mathbf{Y_B})^T}} \quad (7)$$

3. Results and discussion

3.1. Acquired Raman spectra

The acquired Raman spectra (zero and offset) from the eight trials are shown in Figure 4.16, in which the characteristic bands of each container material (PET and PP) as well as of the four test substances concerned are visible. Regarding the substances under study, the typical bands at around 884, 1052, 1096 and 1454 cm^{-1} of ethanol can be seen in Figure 4.16A and 4.16B. Note that the characteristic bands of higher intensity in the Raman spectrum of ethanol are usually found in the interval 2800-3000 cm^{-1}, out of the range of the spectra acquired in this study [**48**]. The distinctive Raman peaks of glycerol are present at 415, 488, 852, 1057 and 1466 cm^{-1} [**49**], but this spectrum also shows a characteristic region between 1200 and 1400 cm^{-1}. The symmetric sulphate stretch peak of sodium sulphate is located at 994 cm^{-1} (Figure 4.16E and 4.16F) [**50**]. Lastly, the characteristic bands of the sucrose Raman spectrum in the 350-1500 cm^{-1} interval can be seen in Figure 4.16G and 4.2H, at 403, 551, 851, or 1127 cm^{-1} [**51**].


A.
ETHANOL - PET
×10⁴
Intensity (A.U.)
Raman shift (cm⁻¹)
0
0.5
1
1.5
2
2.5
3
3.5
4
400
600
800
1000
1200
1400
1600
1800
2000
B.
ETHANOL - PP
×10⁴
Intensity (A.U.)
Raman shift (cm⁻¹)
0
0.5
1
1.5
2
2.5
3
3.5
C.
GLYCEROL - PET
×10⁴
Intensity (A.U.)
Raman shift (cm⁻¹)
D.
GLYCEROL - PP
×10⁴
Intensity (A.U.)
Raman shift (cm⁻¹)
E.
SODIUM SULFATE - PET
×10⁴
Intensity (A.U.)
Raman shift (cm⁻¹)
F.
SODIUM SULFATE - PP
×10⁴
Intensity (A.U.)
Raman shift (cm⁻¹)
G.
SUCROSE - PET
×10⁴
Intensity (A.U.)
Raman shift (cm⁻¹)
H.
SUCROSE - PP
×10⁴
Intensity (A.U.)
Raman shift (cm⁻¹)
5

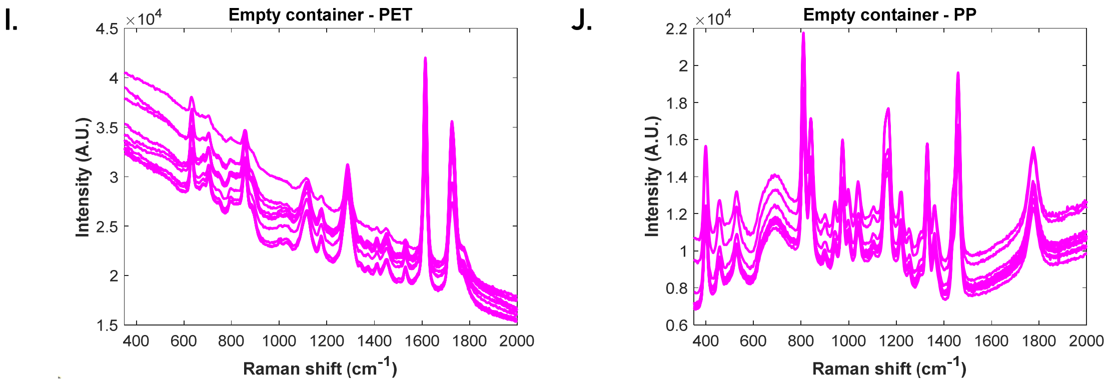


Figure 4.16. Unprocessed Raman spectra acquired with spatial offset (green lines) and without offset (red lines) of (**A**,**B**) ethanol, (**C**,**D**) glycerol, (**E**,**F**) sodium sulphate and (**G**,**H**) sucrose measured through polyethylene terephthalate (PET) and polypropylene (PP); zero Raman spectra of empty containers: PET (**I**) and PP (**J**).

In addition, Figures 4.16I and 4.16J show the Raman spectra acquired of both empty PET and PP containers, respectively, measured with zero offset. The main two distinctive peaks of PET at 1614 and 1726 cm^{-1} [52] are observed in both zero and offset spectra of all trials where this container was used (Figure 4.16A, C, E and G). This confirmed that surface contribution is still present in the Raman spectra even if acquired with offset, as previously mentioned. Particularly noteworthy are the spectra acquired for sodium sulphate. On the one hand, great variability is observed among the 10 repetitions of this material measured through PET (Figure 4.16E), while the opposite occurs when measuring through PP (Figure 4.16F). This variability is greater in the spectra acquired without offset than with offset (red lines). This could be related to the physical state of the sodium sulphate, which is generating a higher scattering of the laser light. On the other hand, it is worth noting that in all PET spectra background fluorescence seems to be present, which could also be caused by the brown colour of the container, whereas PP was colourless. According to Olds et al. [53], although SORS technique is intended to supress fluorescence contribution, the offset spectrum could still capture some fluorescence from surface. This is even more noticeable in the spectra of the empty PET container (Figure 4.16I).

Although the main typical Raman peaks of PP are outside the spectral range of our equipment (between 2800 and 3000 cm^{-1}), other related peaks in the 800-1500 cm^{-1} interval and around 500 cm^{-1} are present, e.g., some pairs of double peaks such as 810 and 844 or 1330 and 1361 cm^{-1} [54]. The highest PP peak in this spectral range is visible at 1460 cm-1 but overlapped with peaks of all the substances except sodium sulphate (Figure 4.16F), although its intensity here is weaker due to the high sulphate signal.

The final-SORS spectra obtained with the instrument (once resolved by the equipment software) are shown in supplementary material (Figure 4.S1). The same figure also shows the target spectra of the corresponding standard substances from spectral databases. The signal processing performed by the software seemed to work well in some cases, such as ethanol (Figure 4.S1A and 4.S1B), but in other situations it was not optimal, for example in the case of sodium sulphate through PP (Figure 4.S1F). Other examples where the extraction of the spectra was not performed correctly were the final-SORS glycerol spectra acquired through both PET and PP (Figure 4.S1C and 4.S1D). Compared to the target spectrum of the substance, it is visible that peaks present in region 1200-1400 cm^{-1} are lost. This could be due to the processing method used for baseline correction, which may not be optimal for all cases.

For these reasons, this study aimed at verifying the optimal way to extract the signals acquired by the SORS technique and to obtain the pure Raman spectrum of the measured substance, regardless of the nature of the container material.

3.2. Signal resolution

The ICA and MCR methods are generally applied to identify or quantify the proportion of a particular substance in complex mixtures such as chemicals, pharmaceuticals or food products [55] or also to separate overlapping or noisy signals [18]. But here both chemometric methods were expected to be useful to resolve the mixture of signals resulting from performing measurements with the SORS technique and to extract the two pure source signals.

The goal of the ICA algorithm is to extract one or more individual component signals from a raw signal that is a mixture of overlapping source signals, without the need for prior knowledge of these components. In this study, it was expected to resolve the signal mixture by obtaining two pure components: the container (PET or PP) and the substance within. The signal mixture matrix containing the spectral data was used to develop ICA models in three ways, as explained in section 2.3: without pre-processing, baseline corrected by WLS method and baseline corrected by Whittaker filter method. The optimal number of ICs, determined by the Random_ICA_by_Blocks and Durbin-Watson methods, was 2 in most cases, except for two trials where it was 3 ICs. However, this third independent component consisted of a mixture of artifacts: baseline, dispersion and some negative contributions and we have not been able to find a scientific explanation for this selection, which appears to be arbitrarily performed by the software. In fact, only the first two ICs are significant for the study and whether or not to include the third one does not influence the results.

Although the main purpose was to extract the pure spectrum of the test substance, this was not used to impose to have the maximum correlation with one IC because in the potential future application it will not always be possible to have one unique target spectrum of the products inside, as explained above. However, under the product IQC scenario, it will always be possible to have the pure spectrum of the container by performing a Raman measurement of the empty container. Note that the use of this instrument for in line production control would not be suitable or would not provide advantages, since the line must be stopped to perform the measurement, which, although fast, needs at least one minute. Therefore, its use is focused on the finished product level in a non-invasive way.

An example of the source signals obtained by the ICA model after baseline correction of the data by WLS in the case of glycerol in a PET container is given in Figure 4.17A, together with the spectrum provided of the empty container (PET) after WLS correction and the target spectrum of glycerol. Two components were enough to resolve the mixture of signals and obtain the two expected source spectra: IC1 corresponded to the target spectrum of glycerol, and IC2 to the Raman spectrum of PET. If compared with the result after the processing of the spectrometer itself (Figure 4.S1C) just visually the results obtained from ICA of the glycerol pure spectrum were closer to the target spectrum than this final-SORS spectrum. For instance, this was noticeable by considering the spectral region between 1200 and 1400 cm^{-1}, which was lost when the spectral processing was performed by the equipment software.

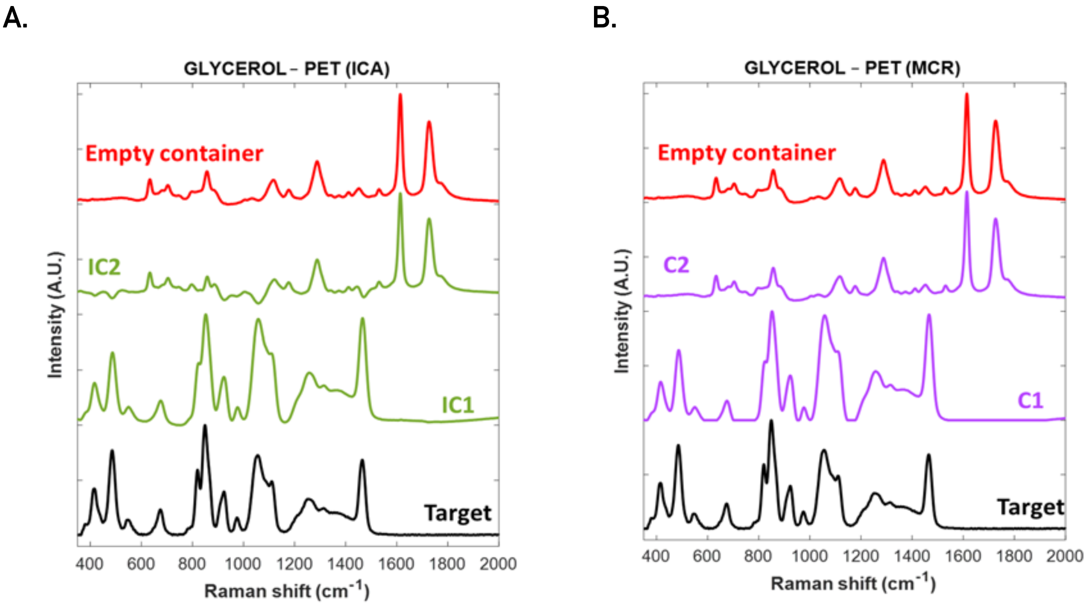


Figure 4.17. Components resulting from applying **(A)** ICA and **(B)** MCR considering glycerol SORS spectra, measured trough polyethylene terephthalate container, as input data (after baseline correction by WLS method) for blind signal resolution.

Note: IC1 and IC2 are the independent components obtained as results from ICA, while C1 and C2 are the components obtained as results from MCR.

MCR is another widely used method to solve the signal mixing problem. This method works by applying a bilinear model in its simplest form to extract the qualitative profiles of the individual sources that constitute the mixed signal, and the proportion profiles of each of the sources [14]. Again, for this study, the aim was to obtain two components, one corresponding to the Raman spectrum of the substance inside and the other to that of the container material (PET or PP).

Firstly, MCR was used without applying any spectra contrast constraints. Since the results were satisfactory but not optimal, it was attempted to improve them by adding the Raman spectrum of the empty container as a constraint, as was imposed for the development of ICA models. The use of constraints in MCR significantly improves the results since it allows enhancing the resolution of the chemical profiles when the given condition is known [14]. It is intended that this model could serve as a real application to extract the pure spectrum of the material inside when the container is the same, for example in case of a company at the IQC level of finished product. Therefore, the spectrum of the empty container was imposed as a hard spectral constraint so that one of the two components (output data) corresponded to the spectral signal of the container, and the other to the substance inside. For comparison, Figure 4.17B shows the results of MCR developed using the same input data (glycerol measured through PET) and pre-processing method (WLS baseline correction) as the ICA shown above. As can be seen, the first component corresponded to the pure spectrum of glycerol that could be correctly resolved and separated from the characteristic Raman signal of PET (component 2).

At first glance, there were no significant differences to determine which of the two methods, ICA or MCR, works better for the resolution of the acquired mixed SORS signals. For this reason, an analysis of the similarity between the obtained output signals (the extracted Raman spectra of the substances under study) and the target spectra was performed. This was conducted not only to determine if one of the methods is preferable for blind signal resolution, but also to establish which pre-processing method is more suitable for solving the problem posed here (none, baseline correction by WLS or by Whittaker filter).

3.3. Performance assessment

Four similarity indexes (COS, ACOS, R^2 and NEAR) were calculated to assess the performance of the results. It should be noted that the index values calculated are easy to interpret, as they all assess similarity by taking values between 0 and 1, where 0 means totally different and 1 means identical. For example, in the case of the ACOS presented, the arccosine is commonly calculated without being normalized, named Spectral Angle Mapper (SAM) [9,19], which makes it difficult to interpret and it is necessary to select a limit beyond which the result would be acceptable, considering the maximum and minimum values this index can assume. Therefore, for this study, the four calculated indexes were normalized to range between 0 and 1, in order to facilitate the comparison of each of the resolved signals for each trial using the three MCR models, the three ICA models and the final-SORS spectrum provided by the equipment software. All these resolved signals were compared with the corresponding target spectrum. The results of the calculated indexes are shown in Table 4.4. For each resolution method, the pre-processing with which the best results were obtained are indicated in bold type.

In most cases, the lowest similarity values were for the final-SORS spectra. Particularly noteworthy are the low values obtained for the final-SORS spectrum of sodium sulphate measured through the PP container. This was to be expected, given the spectrum obtained as discussed above, which can be seen in the supplementary material (Figure 4.S1F). However, very low similarity values were also found in the results obtained when applying ICA to the unprocessed glycerol and ethanol signals acquired through the PP container. This result is because the resulting resolved signal included negative spectral contributions, probably from the container signal, and as a consequence the similarity to the target spectrum tended to be poor. This fact only reinforces the need to apply a baseline correction to the Raman spectra. In fact, the highest similarity indexes were obtained when WLS pre-processed method was applied for baseline correction, regardless of the resolved signal method (MCR or ICA), the test substance studied, or the container material used. This was probably because the baseline correction with this method is smoother than with the Whittaker filter.

There were only three exceptions where the best result was not obtained using WLS filtered data. Two of these occurred when MCR was used to resolve the sodium sulphate spectra measured through either the PET or the PP container. In this case, spectra baseline corrected by the Whittaker filter gave the best solution. This was also the case for MCR applied to ethanol through PP, where the optimal result was for the non-processed data. However, none of the similarity indexes differed by more than 0.02, which is negligible. Also note that in some other cases, the results for the non-processed data and the baseline corrected WLS data were very similar, namely: MCR applied to PP-glycerol and to PP-sucrose data, and ICA applied to PP-sodium sulphate data. Nevertheless, the differences were just 0.001, which is negligible. Therefore, although the two methods used are not based on the same assumptions [24], for the purpose intended here they have proved to be equally useful and render practically the same satisfactory results. Accordingly, it could be concluded that the most suitable method for pre-processing the Raman spectra acquired by SORS in order to obtain the pure Raman spectrum of the substance inside packaging is the baseline correction by WLS.

Table 4.4. Values of similarity indexes by comparing Raman spectra resolved by MCR and ICA and final-SORS spectrum vs the respective target spectrum for each of the eight trials.

Trial		Pre-processing	COS	ACOS	R^2	NEAR
PET / Ethanol	MCR	*None*	0.953	0.804	0.880	0.838
		Whittaker	0.883	0.689	0.769	0.739
		WLS	**0.953**	**0.804**	**0.896**	**0.845**
	ICA	*None*	0.876	0.680	0.873	0.682
		Whittaker	0.906	0.722	0.770	0.778
		WLS	**0.953**	**0.805**	**0.895**	**0.832**
	Final-SORS		0.826	0.619	0.683	0.669
PP / Ethanol	MCR	*None*	**0.954**	**0.806**	**0.890**	**0.845**
		Whittaker	0.886	0.693	0.773	0.741
		WLS	0.944	0.786	0.879	0.830
	ICA	*None*	0.387	0.253	0.441	0.154
		Whittaker	0.829	0.622	0.690	0.681
		WLS	**0.955**	**0.807**	**0.882**	**0.845**
	Final-SORS		0.822	0.614	0.678	0.664
PET / Glycerol	MCR	*None*	0.964	0.828	0.912	0.792
		Whittaker	0.948	0.793	0.873	0.834
		WLS	**0.978**	**0.866**	**0.933**	**0.861**
	ICA	*None*	0.922	0.747	0.923	0.692
		Whittaker	0.959	0.817	0.876	**0.850**
		WLS	**0.978**	**0.865**	**0.934**	0.844
	Final-SORS		0.936	0.771	0.860	0.816
PP / Glycerol	MCR	*None*	**0.981**	**0.877**	**0.944**	0.852
		Whittaker	0.943	0.784	0.861	0.829
		WLS	**0.981**	0.874	0.943	**0.879**
	ICA	*None*	0.373	0.244	0.584	0.250
		Whittaker	0.870	0.671	0.725	0.689
		WLS	**0.981**	**0.874**	**0.942**	**0.856**
	Final-SORS		0.935	0.769	0.860	0.816
PET / Sodium sulphate	MCR	*None*	0.759	0.548	0.596	0.508
		Whittaker	**0.816**	**0.607**	**0.654**	**0.588**
		WLS	0.811	0.602	0.646	0.582
	ICA	*None*	0.633	0.436	0.610	0.370
		Whittaker	0.789	0.579	**0.652**	0.550
		WLS	**0.809**	**0.600**	0.643	**0.577**
	Final-SORS		0.613	0.420	0.354	0.396

PP / Sodium sulphate	MCR	*None*	0.808	0.599	0.641	0.576
		Whittaker	**0.815**	**0.606**	**0.652**	**0.588**
		WLS	0.808	0.599	0.641	0.578
	ICA	*None*	**0.803**	**0.594**	0.640	**0.567**
		Whittaker	0.781	0.570	**0.649**	0.539
		WLS	0.802	0.593	0.639	0.566
	Final-SORS		0.359	0.234	0.099	0.208
PET / Sucrose	MCR	*None*	0.972	0.849	0.904	0.794
		Whittaker	0.942	0.781	0.845	0.819
		WLS	**0.975**	**0.858**	**0.918**	**0.828**
	ICA	*None*	0.939	0.776	0.906	0.692
		Whittaker	0.953	0.804	0.843	**0.827**
		WLS	**0.974**	**0.854**	**0.915**	0.798
	Final-SORS		0.927	0.756	0.820	0.806
PP / Sucrose	MCR	*None*	0.963	0.826	0.874	0.780
		Whittaker	0.943	0.783	0.841	**0.820**
		WLS	**0.963**	**0.826**	**0.881**	0.796
	ICA	*None*	0.944	0.787	0.848	0.720
		Whittaker	0.885	0.691	0.783	0.644
		WLS	**0.964**	**0.828**	**0.882**	**0.761**
	Final-SORS		0.860	0.659	0.658	0.725

***PET**: polyethylene terephthalate; **PP**: polypropylene; **MCR**: Multivariate Curve Resolution; **ICA**: Independent Components Analysis; **Final-SORS**: resolved spectrum by the software of the Spatially Offset Raman Spectroscopy instrument; **Whittaker**: baseline correction by Whittaker filter; **WLS**: baseline correction by weighted least squares; **COS**: cosine; **ACOS**: arccosine; **R^2**: coefficient of determination; **NEAR**: nearness index.*

The resolved spectra that resulted in the best values for the performance metrics (highest similarity indexes values) using both MCR and ICA for each of the eight trials, marked in bold type in Table 4.18, are shown in Figure 4.18, together with the respective target spectrum and final-SORS spectrum. As for ethanol measured through PET (Figure 4.18A), despite the best result being obtained with ICA, it was not possible to remove all PET contribution from the extracted ethanol spectrum, whereas it was feasible using MCR. Although this was visually remarkable, this was only evident when examining the NEAR values, while all other calculated indexes yielded identical values for the MCR and ICA results.

When ethanol was measured through the PP container, there were no major differences between the spectra resolved using ICA and MCR, as the values of the calculated indexes were practically the same. As for the glycerol spectra, the results were certainly better when performing the blind signal resolution manually compared to the ones achieved by the equipment software resolution, as mentioned in the previous section. Part of the information in the region between 1200 and 1400 cm^{-1} of the glycerol spectrum was lost in the final-SORS spectrum, while ICA and MCR resulted in a pure spectrum very close to the target spectrum (Figure 4.18B and 4.18F). In fact, the highest value of all similarity indexes calculated was from the MCR applied to the PP-glycerol data. For this trial, again all indexes yielded similar results for MCR and ICA, except the NEAR index, showing that it is more sensitive to small differences. This will be further discussed below.

The pure Raman spectrum of sodium sulphate was successfully resolved by ICA and MCR for both data sets, i.e., measured through PET and PP. The improvement with respect to the resolution by the equipment software was very significant in this case as can be seen in Figures 4.18C and 4.18G, especially acquired through PP, where the NEAR for the final-SORS spectrum was 0.208 and the R^2 was 0.099, reflecting a very low similarity with the target spectrum.

Finally, the most complex case in this study was for the Raman spectra of sucrose, as can be seen in Figures 4.18D and 4.18H. Nevertheless, the similarity analysis showed very good results, with values greater than 0.95 for COS and greater than 0.80 for NEAR except in the case of ICA for the PP-sucrose data. As in the case of the glycerol, some spectral information was lost in the final-SORS spectrum of sucrose measured through PP in the region between 1200 and 1400 cm^{-1}, which did not occur when applying MCR and ICA. In addition, the most intense peak of the final-SORS spectrum of sucrose through PET did not match with the target spectrum (851 cm^{-1} instead of 403 cm^{-1}).

The baseline correction carried out by the equipment to obtain the final-SORS spectrum was stronger than using WLS, e.g., note the loss of the rounded shape of the baseline around 800 cm^{-1} found in the ICA-resolved spectrum (Figure 4.18H). Therefore, based on these results, manual processing of the spectral data is more appropriate to obtain pure Raman spectra. In general, both ICA and MCR performed equally well in the blind signal resolution. The biggest difference found between them was in the NEAR calculated from the Raman data of PP sucrose, but even then MCR was only 0.06 higher than ICA. This contrasts with the study by Liu et al. [19] where they obtained better results by using ICA than MCR, although it is worth noting the difference in the data handled, as they used line-scan Raman imaging for data acquisition.

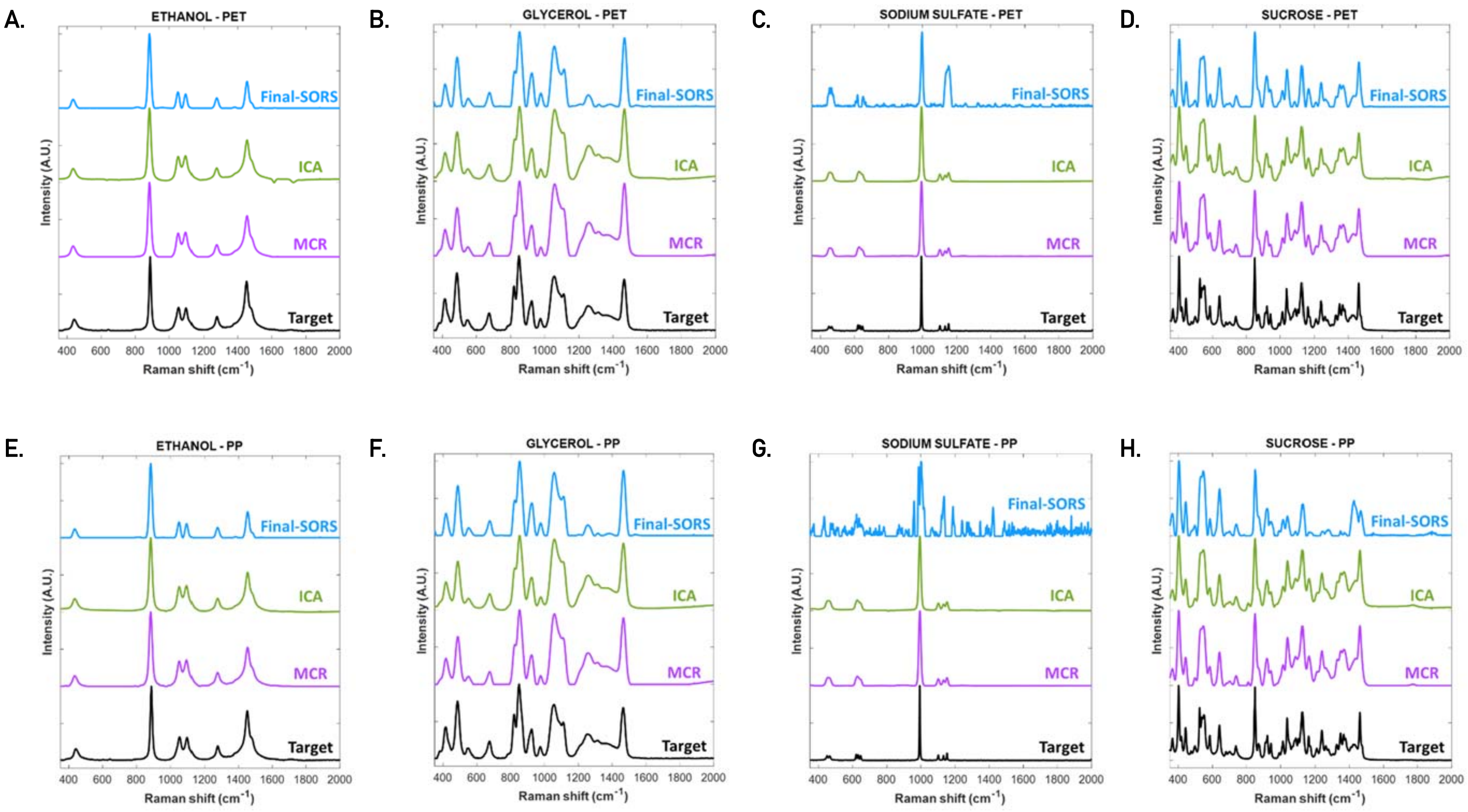


Figure 4.18. ICA and MCR resolved spectra together with the respective final-SORS spectra (processed by the equipment software) and target spectra of (**A,E**) ethanol, (**B,F**) glycerol, (**C,G**) sodium sulphate and (**D,H**) sucrose measured through polyethylene terephthalate (PET) and polypropylene (PP).

3.4. Comprehensive evaluation of similarity indexes

When analyzing the information provided by each similarity index, the COS index gave the highest values. This makes sense since the studied signals were very similar and COS is rather insensitive to minor changes and for this reason, the COS index qualified the resolved signals as practically equal to the target spectra [31]. Conversely, the ACOS values were slightly lower since this index is more responsive to small differences between similar signals. R^2 usually presented intermediate values between the COS and ACOS indexes, and NEAR index gave the lowest values, suggesting that it is more sensitive to small differences than the other indexes.

At this point, it was decided to further analyze the four indexes and study in-depth the differences, by plotting the corresponding values, taken in pairs, and performing a lineal regression to compare the results provided by them in cases where the pattern of behaviour follows a linear trend. In principle, performing an ordinary least-squares (OLS) regression might seem sufficient for this purpose. However, note that the OLS regression works by minimizing the residual only in the vertical direction (on the Y-axis) and it should be only used when the experimental error of the values of one of the variables (to be plotted on the X-axis) is negligible compared to that of the other variable. However, the two variables to be regressed (similarity indexes) come from the same experimental data, and there is no reason to consider that the precision associated with the calculation of the values for each of the indexes is very different. In such cases, applying orthogonal-distance least-squares (ODLS) regression is the most appropriate approach. The main advantage of this type of regression is that the fitted linear equation is independent of which variables are assigned to the Y and X axes. Consequently, an ODLS regression was conducted using Solver, an Excel add-in program that applies an iterative method to perform the regression [56].

As a starting point, it should be pointed that NEAR values have previously been shown to vary quasi-linearly along the similarity grade [31] and similar behaviour is expected for ACOS. In addition, it is obvious that the cosine of an angle neither follows a linear dependence on the angle considered, which is accentuated at low values of the angle (see Figure 4.S2A). This results in the COS being very insensitive to small differences between the two sets of values under study. The other indexes, R^2 and NEAR, also present this non-linear trend with respect to COS (Figures 4.S2B and 4.S2C), thus explaining the highest values reached with this index. Consequently, for none of these three cases (ACOS vs COS, R^2 vs COS and NEAR vs COS) has a straight line been fitted by regression.

While the two ODLS regressions performed using ACOS, i.e., NEAR vs ACOS, and R^2 vs ACOS, shown in Figure 4.19, render interesting results. The first striking feature is the large scattering of R^2 values for low similarities (values below 0.6), which disqualifies this index for assessing similarity in this domain.

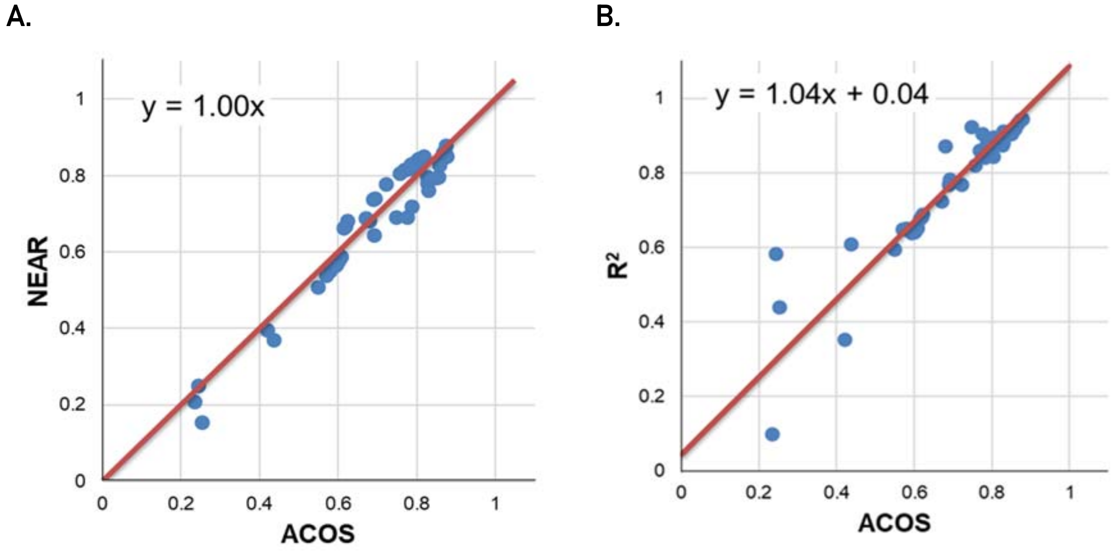


Figure 4.19. Orthogonal-distance least-squares (ODLS) regression plots of (**A**) NEAR vs ACOS, and (**B**) R^2 vs ACOS (see the text for a more detailed explanation).

The ODLS regression of NEAR vs ACOS (see Figure 4.19A), resulted a straight line of slope 1 and intercept zero. This indicates both indexes yield the same similarity or dissimilarity between the two data sets being compared. Thus, any of them, ACOS and NEAR, could be employed when studying the similarity between a resolved Raman spectrum and the reference spectrum of the substance. However, before extending this statement to any case it is necessary to take into account a particular consideration. Note that for their calculation, the spectra to be compared (i.e., resolved spectra) were 0-1 normalized in intensity. This was done because the collected reference spectra used to compare with were already normalized. Moreover, the goal pursued was to achieve the same spectral profile, since the intensity may also be conditioned by the technical specifications of the equipment used. However, NEAR considers the difference between the vectors (location and intensity of the spectrum peaks) in both spatial orientation and module, while the other indexes only consider orientation. It would therefore be more sensitive to small differences, but this is beyond the scope of this study since here the intensity of the profiles compared were normalized, hence there is no difference in module, and the dissimilarity or similarity between them only depends on the spatial orientation.

In another instance, where the spectra were not normalized, we would only observe this behaviour when plotting R^2 vs ACOS (or vice-versa). The fitted ODLS regression equation of R^2 vs ACOS showed the estimated values of the intercept and slope to be practically 0 and 1, respectively (Figure 4.19B). Since both indexes only assess differences in orientation (angle), it is not surprising that the values returned are equivalent. Note as expected that plotting R^2 vs NEAR (Figure 4.S2D) is practically identical to R^2 vs ACOS, since ACOS and NEAR have been shown to behave the same.

4. Conclusions

This study describes two data processing methodologies applied to the resolution of mixed Raman spectra acquired in non-invasive SORS mode through containers, in order to obtain the spectrum of the substance inside. The raw acquired signal is a mixture of the characteristic Raman spectra of the packaging material and the substance inside. The software commanding the equipment itself already provides a resolved spectrum, which however, sometimes does not give the optimal result. Nevertheless, the application of two chemometric methods for blind signal resolution, MCR and ICA, enabled the extraction of the pure Raman spectra of the four standard substances under test (ethanol, glycerol, sodium sulphate and sucrose), measured through polypropylene and polyethylene terephthalate containers. The resulting resolved spectra were compared with the Raman spectra of the four substances available in recognized databases, as well as the spectra provided by the equipment software. A considerable improvement in results was obtained when ICA and MCR were applied, in relation to those given by the equipment, particularly in the case of sodium sulphate and glycerol. For the comparison of spectra, similarity analysis tools were applied, based on the use of four indexes suitable for this purpose.

The careful study of the values found for the different indexes has shown that any of them captures the information of these comparisons even if the results were slightly different, for this particular study, since the intensity of the spectra was 0-1 normalize. This study demonstrates that both MCR and ICA are useful methods for the resolution of mixed Raman spectra obtained by SORS, highlighting the potential use of the technique for ICQ of finished manufactured products, followed by proper data treatment.

Acknowledgments

AAC gratefully acknowledges the Spanish Ministry of Universities for a pre-doctoral fellowship FPU (FPU20/04711, Formación del Profesorado Universitario). In addition, **AMJC** acknowledges the Grant (RYC2021-031993-I) funded by MCIN/AEI/501100011033 and "European Union NextGenerationEU/PRTR".

References

[1] S. Mosca, C. Conti, N. Stone, P. Matousek, Spatially offset Raman spectroscopy, Nat. Rev. Dis. Primers, 1 (2021) e21. https://doi.org/10.1038/s43586-021-00019-0.

[2] N. Gupta, J.D. Rodriguez, H. Yilmaz, Through-container quantitative analysis of hand sanitizers using spatially offset Raman spectroscopy, Commun. Chem. 4 (2021) e126. https://doi.org/10.1038/s42004-021-00563-6.

[3] M.Z. Vardaki, H.G. Schulze, K. Serrano, M.W. Blades, D.V. Devine, R.F. Turner, Assessing the quality of stored red blood cells using handheld Spatially Offset Raman spectroscopy with multisource correlation analysis, Spectrochim. Acta A Mol. Biomol. Spectrosc. 276 (2022) e121220. https://doi.org/10.1016/j.saa.2022.121220.

[4] C. Zhang, M. Huang, L. Kong, Two-step spatially offset Raman spectroscopy technique for rapid and non-invasive detection of drugs in containers—simulation and experiment. Laser Phys. Lett. 18 (2021) e125601. https://doi.org/10.1088/1612-202X/ac2eeb.

[5] A. Arroyo-Cerezo, A.M. Jiménez-Carvelo, A. González-Casado, I. Ruisánchez, L. Cuadros-Rodríguez, The potential of the spatially offset Raman spectroscopy (SORS) for implementing rapid and non-invasive in-situ authentication methods of plastic-packaged commodity foods–Application to sliced cheeses, Food Control 146 (2023) e109522. https://doi.org/10.1016/j.foodcont.2022.109522

[6] P. Matousek, I.P. Clark, E.R. Draper, M.D. Morris, A.E. Goodship, N. Everall, M. Towrie, W.F. Finney, A.W. Parker, Subsurface probing in diffusely scattering media using spatially offset Raman spectroscopy, Appl. Spectrosc, 59 (2005) 393-400. https://doi.org/10.1366/0003702053641450.

[7] A. Arroyo-Cerezo, A.M. Jimenez-Carvelo, A. González-Casado, A. Koidis, L. Cuadros-Rodríguez, Deep (offset) non-invasive Raman spectroscopy for the evaluation of food and beverages–A review, LWT 149 (2021) e111822. https://doi.org/10.1016/j.lwt.2021.111822.

[8] S. Mosca, P. Dey, M. Salimi, B. Gardner, F. Palombo, N. Stone, P. Matousek, Spatially offset Raman spectroscopy – How deep?, Anal. Chem. 93 (2021) 6755-6762. https://doi.org/10.1021/acs.analchem.1c00490.

[9] Z. Liu, M. Huang, Q. Zhu, J. Qin, M.S. Kim, A packaged food internal Raman signal separation method based on spatially offset Raman spectroscopy combined with FastICA, Spectrochim. Acta A Mol. Biomol. Spectrosc. 275 (2022) e121154. https://doi.org/10.1016/j.saa.2022.121154.

[10] M. Bloomfield, P.W. Loeffen, P. Matousek, Detection of concealed substances in sealed opaque plastic and coloured glass containers using SORS, in: C. Lewis, D. Burgess, R. Zamboni, F. Kajzar, E.M. Heckman (Eds.), Optics and Photonics for Counterterrorism and Crime Fighting VI and Optical Materials in Defence Systems Technology VII, SPIE, 2010, Vol. 7838, pp. 51-65.

[11] K. Chao, S. Dhakal, J. Qin, Y. Peng, W.F. Schmidt, M.S. Kim, D.E. Chan, A spatially offset Raman spectroscopy method for non-destructive detection of gelatin-encapsulated powders, Sensors 17 (2017) e618. https://doi.org/10.3390/s17030618.

[12] S. Lohumi, H. Lee, M.S. Kim, J. Qin, B.K. Cho, Through-packaging analysis of butter adulteration using line-scan spatially offset Raman spectroscopy, Analyt. Bioanal. Chem. 410 (2018) 5663-5673. https://doi.org/10.1007/s00216-018-1189-1.

[13] J. Qin, M.S. Kim, W.F. Schmidt, B.K. Cho, Y. Peng, K. Chao, A line-scan hyperspectral Raman system for spatially offset Raman spectroscopy, J. Raman Spectrosc. 47 (2015) 437-443. https://doi.org/10.1002/jrs.4825.

[14] M. Szymańska-Chargot, P.M. Pieczywek, M. Chylińska, A. Zdunek, Hyperspectral image analysis of Raman maps of plant cell walls for blind spectra characterization by nonnegative matrix factorization algorithm, Chemom. Intell. Lab. Syst. 151 (2016) 136-145. https://doi.org/10.1016/j.chemolab.2015.12.015.

[15] A. de Juan, R. Tauler, Multivariate Curve Resolution: 50 years addressing the mixture analysis problem – A review, Anal. Chim. Acta 1145 (2021) 59-78. https://doi.org/10.1016/j.aca.2020.10.051.

[16] M. Tian, C.L. Morais, H. Shen, W. Pang, L. Xu, Q. Huang, F.L. Martin, Direct identification and visualisation of real-world contaminating microplastics using Raman spectral mapping with multivariate curve resolution-alternating least squares, J. Hazard. Mater. 422 (2022) e126892. https://doi.org/10.1016/j.jhazmat.2021.126892.

[17] E. Widjaja, N. Crane, T.C. Chen, M.D. Morris, M. Ignelzi Jr, B.R. McCreadie, Band-target entropy minimization (BTEM) applied to hyperspectral Raman image data. Appl, Spectrosc. 57 (2003) 1353-1362. https://doi.org/10.1366/000370203322554509.

[18] E. Widjaja, R.K.H. Seah, Application of Raman microscopy and band-target entropy minimization to identify minor components in model pharmaceutical tablets, J. Pharm. Biomed. Anal. 46 (2008) 274-281. https://doi.org/10.1016/j.jpba.2007.09.023.

[19] J. Yao, H. Su, Z. Yao, Blind source separation of coexisting background in Raman spectra, Spectrochim. Acta A Mol. Biomol. Spectrosc. 238 (2020) e118417. https://doi.org/10.1016/j.saa.2020.118417.

[20] Z. Liu, M. Huang, Q. Zhu, J. Qin, M.S. Kim, Evaluating performance of SORS-based subsurface signal separation methods using statistical replication Monte Carlo simulation, Spectrochim. Acta A Mol. Biomol. Spectrosc. 293 (2023) 122520. https://doi.org/10.1016/j.saa.2023.122520.

[21] J.H. Churchwell, K. Sowoidnich, O. Chan, A.E. Goodship, A.W. Parker, P. Matousek, Adaptive band target entropy minimization: Optimization for the decomposition of spatially offset Raman spectra of bone, J. Raman Spectrosc., 51 (2020) 66-78. https://doi.org/10.1002/jrs.5749.

[22] X. Shi, X., Blind Signal Processing: Theory and Practice, Springer, Shanghai, 2011.

[23] Y.B., Monakhova, D.N. Rutledge, Independent components analysis (ICA) at the "cocktail-party" in analytical chemistry, Talanta 208 (2020) e120451. https://doi.org/10.1016/j.talanta.2019.120451.

[24] A. Tharwat, Independent component analysis: An introduction, Appl. Comput. Inform. 17 (2021) 222-249. https://doi.org/10.1016/j.aci.2018.08.006.

[25] D.J.R. Bouveresse, D.N. Rutledge, Independent components analysis: theory and applications, in: C. Ruckebusch (Ed.), Resolving Spectral Mixtures with Applications from Ultrafast Time-Resolved Spectroscopy to Super-Resolution Imaging, Elsevier, Netherlands, 2016, pp. 225-277.

[26] M. Fang, W. Ju, W. Zhan, T. Cheng, F. Qiu, J. Wang, A new spectral similarity water index for the estimation of leaf water content from hyperspectral data of leaves, Remote Sens. Environ. 196 (2017) 13-27. https://doi.org/10.1016/j.rse.2017.04.029.

[27] J.A. Ramírez-Rincón, M. Palencia, E.M. Combatt, Separation of optical properties for multicomponent samples and determination of spectral similarity indices based on FEDS0 algorithm, Mater. Today Commun. 33 (2022) e104528. https://doi.org/10.1016/j.mtcomm.2022.104528.

[28] D. Cozzolino, D. Bureš, L.C. Hoffman, Evaluating the Use of a Similarity Index (SI) Combined with near Infrared (NIR) Spectroscopy as Method in Meat Species Authenticity, Foods 12 (2023) e182. https://doi.org/10.3390/foods12010182.

[29] F.A. Kruse, A.B. Lefkoff, J.W. Boardman, K.B. Heidebrecht, A.T. Shapiro, P.J. Barloon, A.F.H. Goetz, The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data, Remote Sens. Environ. 44 (1993) 145-163. https://doi.org/10.1016/0034-4257(93)90013-N.

[30] J.A. Meima, D. Rammlmair, Investigation of compositional variations in chromitite ore with imaging Laser Induced Breakdown Spectroscopy and Spectral Angle Mapper classification algorithm, Chem. Geol. 532 (2020) e119376. https://doi.org/10.1016/j.chemgeo.2019.119376.

[31] J. Sun, L. Yu, J. Li, G. Ding, Similarity analysis on spectrum state evolutions, in: Q. Liang, X. Liu, Z. Na, W. Wang, J. Mu, B. Zhang (Eds.), Communications, Signal Processing, and Systems: Proceedings of the 2018 CSPS Volume II: Signal Processing, Springer, Singapur, 2020, pp. 502-510.

[32] R. Pérez Robles, N. Navas, S. Medina Rodríguez, L. Cuadros Rodríguez, Method for the comparison of complex matrix assisted laser desorption ionization-time of flight mass spectra. Stability of therapeutical monoclonal antibodies, Chemometr. Intell. Lab. Syst. 170 (2017) 58-67. https://doi.org/10.1016/j.chemolab.2017.09.008.

[33] L. Valverde-Som, C. Ruiz-Samblás, F.P. Rodríguez-García, L. Cuadros-Rodríguez, Multivariate approaches for stability control of the olive oil reference materials for sensory analysis – Part I: Framework and fundamentals, J. Sci. Food Agric. 98 (2018) 4237–4244. https://doi.org/10.1002/jsfa.8948.

[34] C.I. Chang, An information-theoretic approach to spectral variability, similarity, and discrimination for hyperspectral image analysis, IEEE Trans. Inf. 46 (2000) 1927-1932. https://doi.org/10.1109/18.857802.

[35] J. Farifteh, F. Van Der Meer, E.J.M. Carranza, Similarity measures for spectral discrimination of salt-affected soils, Int. J. Remote Sens. 28 (2007) 5273-5293. https://doi.org/10.1080/01431160701227604.

[36] J. Xue, Z. Yang, L. Han, L. Chen, Study of the influence of NIRS acquisition parameters on the spectral repeatability for on-line measurement of crop straw fuel properties, Fuel 117 (2014) 1027-1033. https://doi.org/10.1016/j.fuel.2013.10.017.

[37] R. Zeng, J.P. Zhang, K. Cai, W.C. Gao, W.J. Pan, C.Y. Jiang, P.Y. Zhang, B.W. Wu, C.H. Wang, X.Y. Jin, D.C. Li, How similar is "similar," or what is the best measure of soil spectral and physiochemical similarity?, Plos one 16 (2021) e0247028. https://doi.org/10.1371/journal.pone.0247028.

[38] Y. Bi, S. Li, L. Zhang, Y. Li, W. He, J. Tie, F. Liao, X. Hao, Y. Tian, L. Tang, J. Wu, H. Wang, Q. Xu, Quality evaluation of flue-cured tobacco by near infrared spectroscopy and spectral similarity method. Spectrochim. Acta A Mol. Biomol. Spectrosc. 215 (2019) 398-404. https://doi.org/10.1016/j.saa.2019.01.094.

[39] M. Naresh Kumar, M.V.R. Seshasai, K.S. Vara Prasad, V. Kamala, K.V. Ramana, R.S. Dwivedi, P.S. Roy, A new hybrid spectral similarity measure for discrimination among Vigna species, . Int. J. Remote Sens. 32 (2011) 4041-4053. https://doi.org/10.1080/01431161.2010.484431.

[40] B. Melit Devassy, S. George, P. Nussbaum, T. Tessamma, Classification of forensic hyperspectral paper data using hybrid spectral similarity algorithms, J. Chemom. 36 (2022) e3387. https://doi.org/10.1002/cem.3387.

[41] J. Wiley & Sons, Inc., KnowItAll Raman Spectral Database, 2023. https://sciencesolutions.wiley.com/solutions/technique/raman/ (accessed 7 February 2023).

[42] NICODOM Ltd., Raman Spectra Libraries, Raman Spectra Databases, 2021. http://www.raman-spectra.com/ (accessed 31 January 2023).

[43] Engelsen S.B., Database on Raman spectra of carbohydrates. http://www.models.life.ku.dk/~specarb/specarb.html (accessed 10 January 2023).

[44] Eigenvector Research, Inc. "Advanced Preprocessing: Noise, Offset, and Baseline Filtering", Eigenvector Research Wiki. https://wiki.eigenvector.com/index.php?title=Advanced_Preprocessing:_Noise,_Offset,_and_Baseline_Filtering.

[45] E.T. Whittaker, On a new method of graduation, Proc. Edinb. Math. Soc. 41 (1922) 63-75. https://doi.org/10.1017/S0013091500077853.

[46] D.N. Rutledge, D.J.R. Bouveresse, Independent Components Analysis with the JADE algorithm, Trends Anal. Chem. 50 (2013) 22-32. https://doi.org/10.1016/j.trac.2013.03.013.

[47] D.J.R. Bouveresse, A. Moya-González, F. Ammari, D.N. Rutledge, Two novel methods for the determination of the number of components in independent

components analysis models. Chemom. Intell. Lab. Syst. 112 (2012) 24-32. https://doi.org/10.1016/j.chemolab.2011.12.005.

[48] F. Li, Z. Men, S. Li, S. Wang, Z. Li, C. Sun, Study of hydrogen bonding in ethanol-water binary solutions by Raman spectroscopy, Spectrochim. Acta A Mol. Biomol. Spectrosc. 189 (2018) 621-624. https://doi.org/10.1016/j.saa.2017.08.077.

[49] C.M. Gryniewicz-Ruzicka, S. Arzhantsev, L.N. Pelster, B.J. Westenberger, L.F. Buhse, J.F. Kauffman, Multivariate calibration and instrument standardization for the rapid detection of diethylene glycol in glycerin by Raman spectroscopy, Appl. Spectrosc. 65 (2011), 334-341. https://doi.org/10.1366/10-05976.

[50] J. Qiu, X. Li, X. Qi, X., Raman spectroscopic investigation of sulphates using mosaic grating spatial heterodyne raman spectrometer, IEEE Photonics J. 11 (2019) 1-12. https://doi.org/ 10.1109/JPHOT.2019.2939222.

[51] K. Ilaslan, I.H. Boyaci, A. Topcu, Rapid analysis of glucose, fructose and sucrose contents of commercial soft drinks using Raman spectroscopy, Food Control 48 (2015) 56-61. https://doi.org/10.1016/j.foodcont.2014.01.001.

[52] E. Rebollar, S. Pérez, M. Hernández, C. Domingo, M. Martín, T.A. Ezquerra, J.P. García-Ruiz, M. Castillejo, Physicochemical modifications accompanying UV laser induced surface structures on poly (ethylene terephthalate) and their effect on adhesion of mesenchymal cells, Phys. Chem. Chem. Phys. 16 (2014) 17551-17559. https://doi.org/10.1039/C4CP02434F.

[53] W.J. Olds, E. Jaatinen, P. Fredericks, B. Cletus, H. Panayiotou, E.L. Izake, Spatially offset Raman spectroscopy (SORS) for the analysis and detection of packaged pharmaceuticals and concealed drugs, Forensic Sci. Int. 212 (2011) 69-77. https://doi.org/10.1016/j.forsciint.2011.05.016.

[54] S. Zhao, M. Danley, J.E. Ward, D. Li, T.J. Mincer, An approach for extraction, characterization and quantitation of microplastic in natural marine snow using Raman microscopy, Anal. Methods 9 (2017) 1470-1478. https://doi.org/10.1039/C6AY02302A.

[55] S.J. Mazivila, J.L. Santos, A review on multivariate curve resolution applied to spectroscopic and chromatographic data acquired during the real-time monitoring of evolving multi-component processes: From process analytical chemistry (PAC) to process analytical technology (PAT), TrAC, Trends Anal. Chem. 157 (2022) 116698. https://doi.org/10.1016/j.trac.2022.116698.

[56] M. Delgado-Aguilar, L. Valverde-Som, L. Cuadros-Rodríguez, Solver, an Excel application to solve the difficulty in applying different univariate linear regression methods, Chemometr. Intell. Lab. Syst. 178 (2018) 39-46. https://doi.org/10.1016/j.chemolab.2018.04.018.

SUPPLEMENTARY INFORMATION *(Artículo científico 5)*

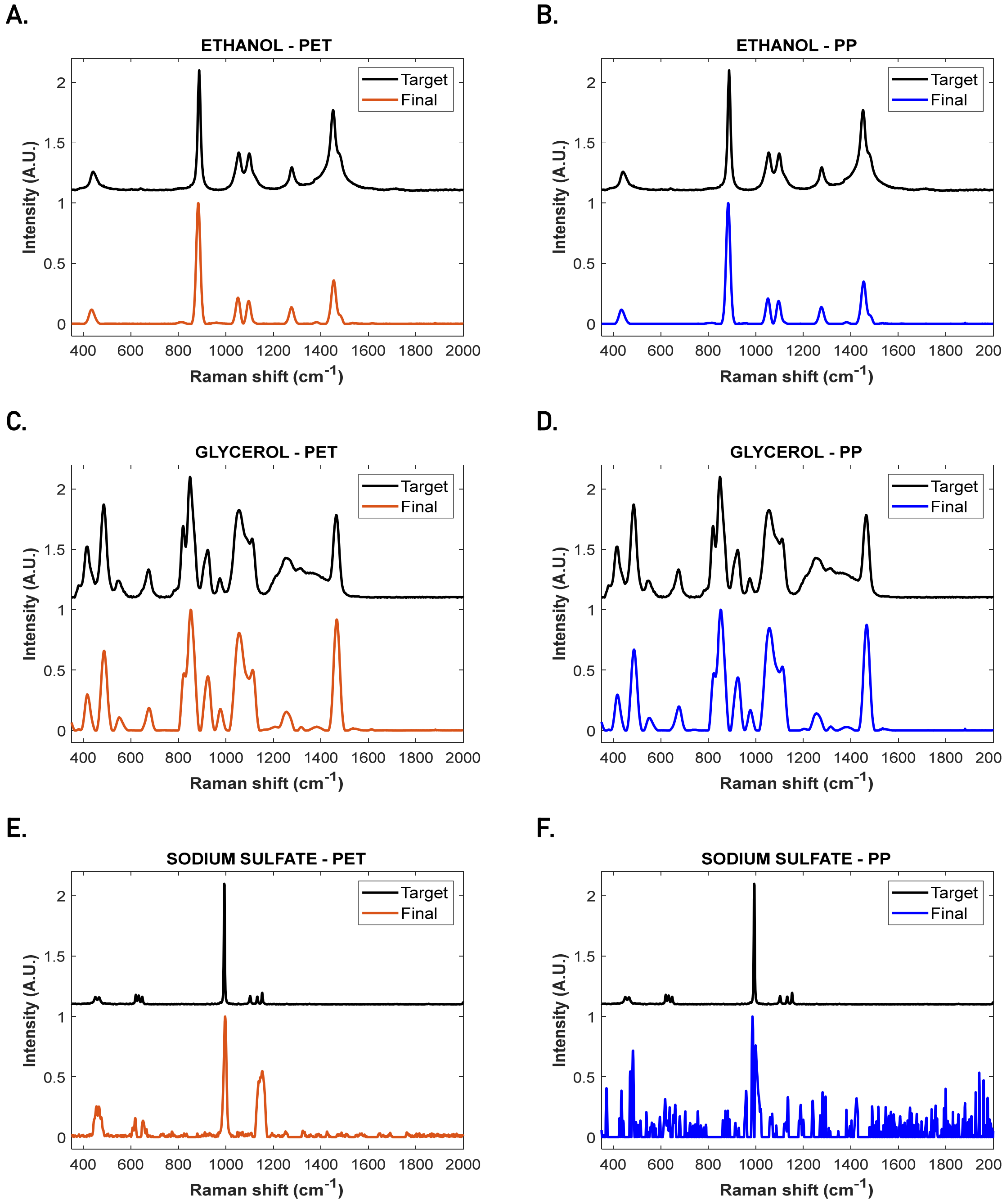


Figure 4.S1. Final-SORS spectra (processed by the equipment software) of (A,B) ethanol (C,D), glycerol (E,F) sodium sulphate and (G,H) sucrose measured through polyethylene terephthalate (PET) and polypropylene (PP), and target spectra of the test substances.

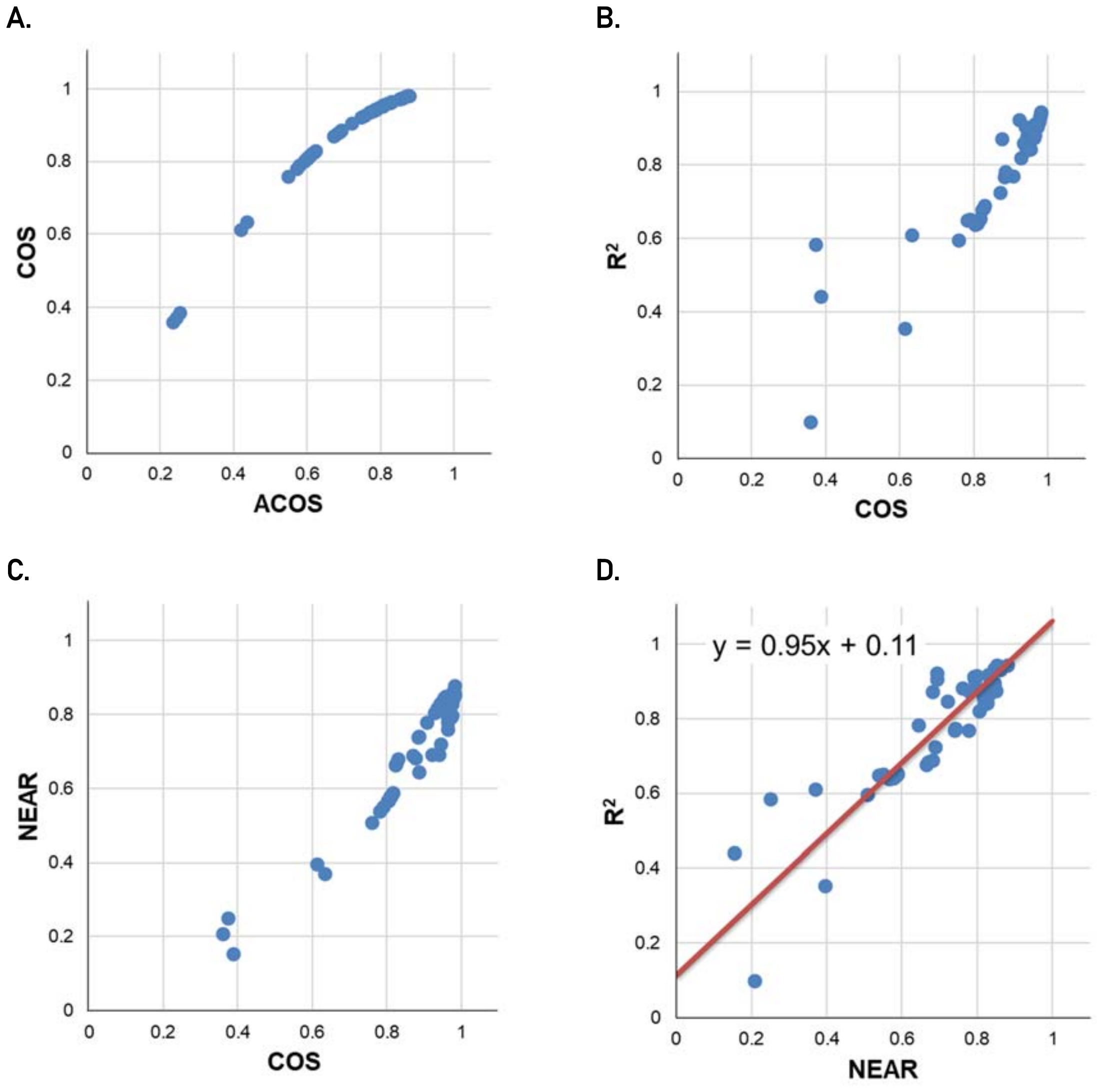


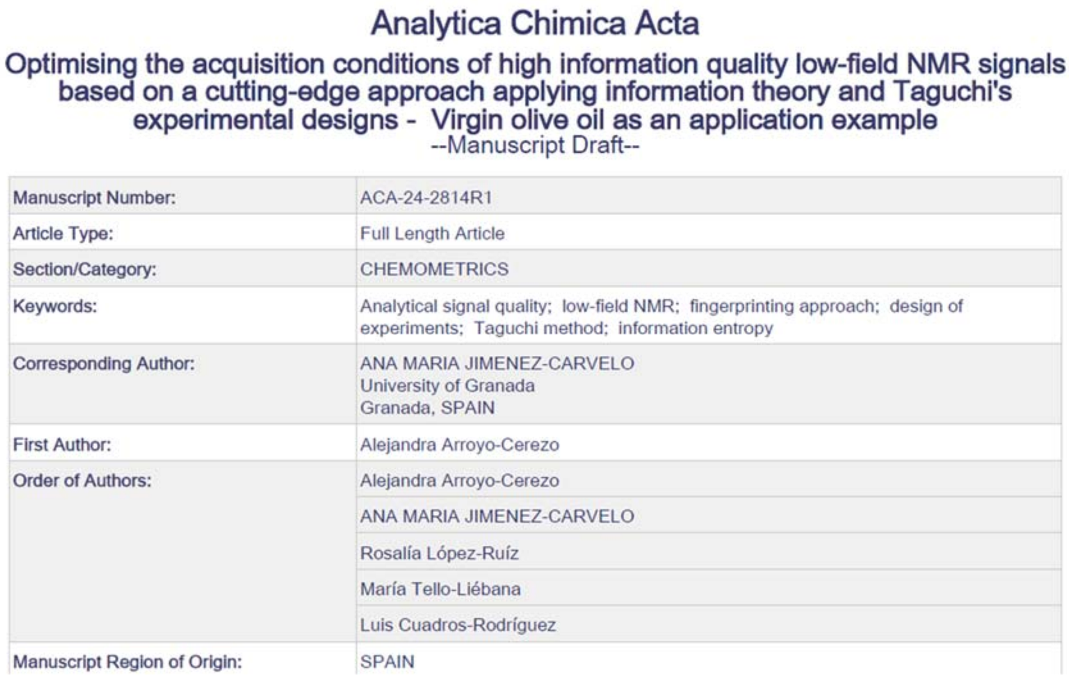
Figure 4.S2. Final-SORS spectra (processed by the equipment software) of (**A,B**) ethanol (**C,D**), glycerol (**E,F**) sodium sulphate and (**G,H**) sucrose measured through polyethylene terephthalate (PET) and polypropylene (PP), and target spectra of the test substances.

4.4. Artículo científico 6

Optimising the acquisition conditions of high information quality low-field NMR signals based on a cutting-edge approach applying information theory and Taguchi's experimental designs – Virgin olive oil as an application example.

Alejandra Arroyo-Cerezo ✉, Ana M. Jiménez-Carvelo ✉, Rosalía López-Ruíz, María Tello-Liébana, Luis Cuadros-Rodríguez.

Enviado a la revista Analytica Chimica Acta por primera vez en agosto 2024.

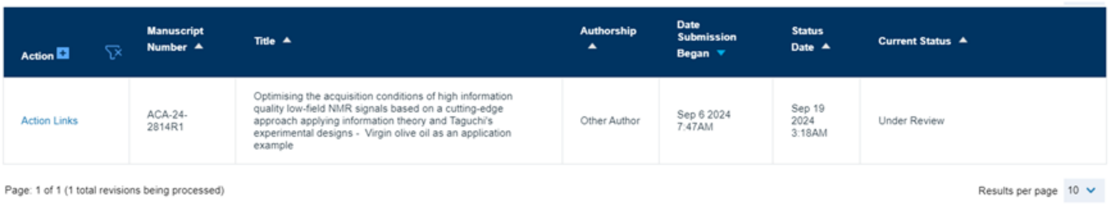


Analytica Chimica Acta

Optimising the acquisition conditions of high information quality low-field NMR signals based on a cutting-edge approach applying information theory and Taguchi's experimental designs - Virgin olive oil as an application example

--Manuscript Draft--

Manuscript Number:	ACA-24-2814R1
Article Type:	Full Length Article
Section/Category:	CHEMOMETRICS
Keywords:	Analytical signal quality; low-field NMR; fingerprinting approach; design of experiments; Taguchi method; information entropy
Corresponding Author:	ANA MARIA JIMENEZ-CARVELO University of Granada Granada, SPAIN
First Author:	Alejandra Arroyo-Cerezo
Order of Authors:	Alejandra Arroyo-Cerezo
	ANA MARIA JIMENEZ-CARVELO
	Rosalía López-Ruíz
	María Tello-Liébana
	Luis Cuadros-Rodríguez
Manuscript Region of Origin:	SPAIN



Action	Manuscript Number	Title	Authorship	Date Submission Began	Status Date	Current Status
Action Links	ACA-24-2814R1	Optimising the acquisition conditions of high information quality low-field NMR signals based on a cutting-edge approach applying information theory and Taguchi's experimental designs - Virgin olive oil as an application example	Other Author	Sep 6 2024 7:47AM	Sep 19 2024 3:18AM	Under Review

Page: 1 of 1 (1 total revisions being processed)

✉ Corresponding author (e-mail: arroyoc@ugr.es)

✉ Corresponding author (e-mail: amariajc@ugr.es)

Highlights:

- Instrument settings for ^{1}H and ^{13}C LF-NMR signals acquisition were optimised.
- Taguchi experimental designs were performed to optimise a robust system.
- A novel proposal to a priori assess the information quality of an analytical signal is presented.
- Information theory was applied to select the most informative fingerprint.

Keywords:

Analytical signal quality

Low-field NMR

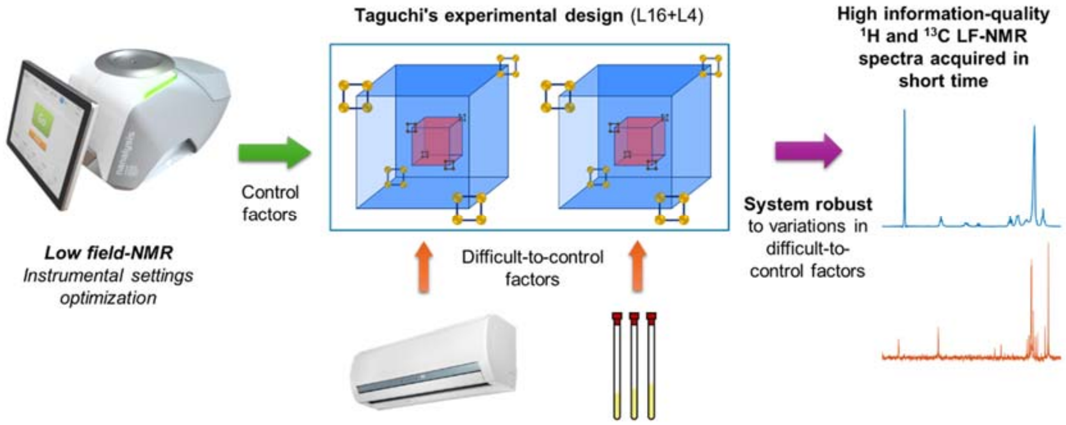
Fingerprinting approach

Design of experiments

Taguchi method

Information entropy

Graphical abstract


Taguchi's experimental design (L16+L4)
High information-quality ^{1}H and ^{13}C LF-NMR spectra acquired in short time
Control factors
System robust to variations in difficult-to-control factors
Low field-NMR Instrumental settings optimization
Difficult-to-control factors

Abstract

Background:

Developing a new spectrometric analytical method based on a fingerprinting approach requires optimization of the experimental stage, particularly with novel instruments like benchtop low-field NMR spectrometers. To ensure high-quality LF-NMR spectra before developing the multivariate model, an experimental design to optimize instrument conditions is essential. However, difficult-to-control factors may be critical for optimization. Taguchi methodology addresses these factors to obtain a system robust to variation. This study uses the Taguchi methodology to optimize instrument settings for acquiring high-quality ^{1}H and ^{13}C LF-NMR signals in a short time from virgin olive oil (VOO).

Results:

Two experimental trials (for ^{1}H and ^{13}C signals, respectively) were carried out and analysed to find an optimal and robust combination of instrument settings against changes in two difficult-to-control factors: ambient temperature and small deviations of the NMR tube volume (700 ± 50 µL). The responses to be optimised, run time and spectral information quality, were analysed separately and jointly, as some factors showed opposite behaviour in the effect on the responses. Multiple response analysis based on suitable desirability functions yielded a combination of factors resulting in desirability values above 0.8 for ^{1}H LF-NMR signals and almost 1.0 for ^{13}C LF-NMR signals.

In addition, a novel approach to assess the information quality of an analytical signal was proposed, addressing a major challenge in analytical chemistry. By applying information theory and calculating information entropy, this approach demonstrated its potential for selecting the highest quality (i.e. most informative) analytical signals.

Significance:

The acquisition instrument conditions of LF-NMR were successfully optimised using Taguchi methodology to acquire highly informative ^{1}H and ^{13}C spectra in a minimum run time. The importance lies in the future development of non-targeted analytical applications for VOO quality control. In addition, the innovative use of information entropy to a priori assess the signal quality represents a significant advance and proposes a solution to a long-standing challenge in analytical chemistry.

1. Introduction

Spectrometric techniques are currently the leading candidates for developing new rapid, less expensive, environmentally friendly, non-invasive and/or non-destructive analytical methods based on a fingerprinting approach for food quality control. Among them, nuclear magnetic resonance spectrometry (NMR) stands out as a powerful technique in terms of analytical performance. The information provided by these spectra is more complete than that of other spectrometric techniques [1]. NMR enables the detection, characterisation and quantitation of substances present in the material even at low concentrations. In the field of food analysis, it has shown great potential for structural characterisation of substances, and for authentication purposes in a variety of foods such as beef, coffee, cheese, fish, honey, vegetable oils, spices, tomato and several beverages (beer, wine, juices, milk) [2,3].

Two types of NMR spectrometers can be distinguished according to the applied field frequency: high frequency and high-resolution spectrometers (> 250 MHz); and low frequency and high-resolution spectrometers (40-100 MHz), known as HF-NMR and LF-NMR respectively. The first, HF-NMR, is the one conventionally used in research studies and is the one providing the great advantages mentioned above in terms of analytical performance. However, it entails certain drawbacks and is the main reason for the lack of application in food industries. These are (i) the high economic investment involved, (ii) the requirements of large spaces and specialised infrastructure, (iii) the maintenance required and (iv) the experienced technical personnel demanded. The main reason of these disadvantages lies in the cryogenically cooled superconducting magnets they incorporate to generate the high magnetic field. In contrast, LF-NMR spectrometers use permanent magnets, usually made of rare earth elements, which are more stable and do not require costly maintenance. Moreover, the use of these magnets allows for the development of more compact and accessible instruments (benchtop) and provides a higher and more homogeneous magnetic field compared NMR relaxometers (20-40 MHz), the pioneering low-field NMR instruments, which are limited to determine relaxation or diffusion parameters due to the low resolution and the poor magnet homogeneity [4,5].

At the beginning of the present decade, technological advances made it possible to launch benchtop LF-NMR spectrometers improving the sensitivity and therefore the resolution of the data acquired [6]. NMR spectral data provide a wealth of chemical information of the measured material. It is one of the most powerful techniques for food analysis [4]. However, due to the compositional complexity of foods, even using HF-NMR it is not possible to separate all compounds present in a food product unless a targeted approach to specific compounds of interest is applied at the data analysis stage [7]. Whereas, when the aim is to study the material in its entire compositional set, e.g., for food authentication purposes, a non-targeted approach should be applied in these cases [8]. The application of this approach involves considering the NMR spectrum as a non-specific instrumental fingerprint of the measured material that contains all the useful information characterising it. By appropriate data processing and analysis using chemometrics it is possible to develop multi-parametric analytical methods, in line with the principles of green analytical chemistry (GAC) [9].

Developing a new analytical method complying with the premises of (i) being fast and (ii) applying a non-targeted approach, undoubtedly requires the optimisation of the experimental stage, particularly the acquisition conditions of fingerprints. This need becomes even more critical when using novel instruments such as LF-NMR benchtop spectrometers, and particularly in the field of food analysis, where there is a certain lack of experience in the use of NMR [2]. For this situation, performing an experimental design to optimise the acquisition instrument conditions would be a convenient solution to ensure that the acquired LF-NMR spectral data are sufficiently meaningful for the proposed objective prior to analytical method development. Design of experiments (DoE) is widely applied in analytical chemistry for processes and reactions optimisation [10]. Using this field of chemometrics, it is possible to study and understand how and how much certain factors affect the analytical response(s). Besides, with a proper DoE it is possible to minimise the number of runs to optimise the process, generating a design space where there is a higher probability to find the optimal values of the factors affecting the process. Thus, DoE should be considered part of the GAC [11].

Numerous types of methodologies can be followed to perform a DoE, being the most common and usual screening designs based of 2-level factorial designs, among others. However, there is a methodology not widely explored in analytical chemistry that is named as its developer: Taguchi designs [10]. Through its methodology, Taguchi considers the existence of factors which are difficult-to-control, referred to as noise factors. Aware of this, Taguchi methodology seeks to generate a robust system and to optimise the response(s) by minimising the variability of the concerned responses due to slight modifications in the nominal value of difficult-to-control factors. The objective of this methodology is to optimise the system for implementation in an off-line total quality control. For that purpose, it is necessary being able to control the difficult-to-control factors while the experiment is being carried out in the laboratory [10,12].

In addition to the aforementioned points, there is a challenge to be addressed in the field of analytical chemistry: determining which analytical signal is most informative. This challenge could be overcome by applying information theory. The information theory was born in the 1940s with the aim of evaluate the quantity of information and the uncertainly in a message [13]. More information transmitted is associated with increased knowledge and reduced uncertainty. The greater the uncertainty, the less the useful information [14]. Following this principle, Shannon proposed the calculation of entropy as a metric of the amount of information, considering that the lower its value, the more information is present in the message. This theory, Shannon's entropy calculation and all its successive extensions, modifications and new proposals, are widely used in disciplines such as computer sciences, telecommunications, cryptography, and statistics, and in some experimental disciplines, particular in ecology.

Analytical chemistry, often referred to as the science of chemical information, was no stranger to the advent of information theory, and at the end of the 20th century there was a growing interest that culminated in the publication of a valuable review [15] and specialised textbook [16], both authored by Eckschlager and Danzer, the pioneers in the application of this subject in analytical chemistry. However, the proposals, clearly ahead of its time, did not have an impact on the analytical community and interest waned, with the result that publications devoted to information theory have been residual throughout the 21st century. Among them, perhaps a chapter in a new handbook focusing on the metrological foundations of analytical chemistry deserves to be highlighted [17]. In this regard, as far as the evaluation of analytical signals is concerned, the application of information theory has been underused in the past. Some examples can be found for use as a similarity index between two signals [18,19,20]. Even specifically for the NMR spectrometry technique, the use of entropy has been proposed as a measure of the spectrum information but applied to the comparison (again, similarity analysis) [21], or as an alternative to the use of the Fourier transform for the reconstruction of spectra [14], which is known as Maximum Entropy Method [22]. Thus, none of them applied this metric as a way of assessing the quality of an analytical signal, which in the end is merely knowing how much information the signal provides.

In this regard, the present study aims to optimise the LF-NMR acquisition instrument settings employing Taguchi methodology to obtain high informative one-dimensional ^{1}H and ^{13}C LF-NMR spectra in the lowest possible run time. For this purpose, the case of study was the virgin olive oil (VOO), since the further purpose is to develop qualitative and quantitative multivariate analytical methods for non-targeted applications using fingerprinting methodology, focused on the quality and authenticity control of VOO by using LF-NMR. Additionally, a proposal based on information entropy is presented to a priori evaluate the information quality of an analytical signal, i.e. before the development of the analytical multivariate model/method from instrumental fingerprints. It should be noted that this approach could be applied to different food matrices following the same steps that those presented in this study.

2. Measuring the spectral information: the entropy as a suitable information metric

When comparing two different signals acquired with the same analytical technique, it can be an easy task to decide visually which one has higher information quality. Consider, for example, two NMR spectra of the same sample, but acquired with different field (low and high). Undoubtedly, the spectrum acquired with HF-NMR will have a higher quality, as it provides more chemical information, simply because the signals are better resolved and therefore there is less overlap than in an LF-NMR spectrum (see Figure 4.S3A in supplementary material). However, this becomes a more difficult task if similar signals are compared, such as two signals acquired with the same instrument where only some instrument settings were changed. In such a situation, it would not be possible to decide visually the best analytical signal quality, as can be seen in Figure 4.S3B (supplementary material). All in all, it is a matter of answering the question: which analytical signal provides more information? Therefore, the solution to this challenge could only be to find a way for measuring the amount of useful information available in an analytical signal.

In the field of analytical chemistry, when applying the fingerprinting approach, the common practice is to develop a multivariable model from the obtained analytical signals and then evaluate whether the model is fit-for-purpose. If the model is not valid for its purpose, it is due to the fact that the analytical signal is not sufficiently informative about the concerned target feature in the material under study (i.e., it is not a high-informative signal). Consequently, a different acquisition mode or type of signal must be considered, and the model must be redeveloped and revalidated. Ultimately, this is a tedious and time-consuming process. Therefore, the present study proposes to evaluate *a priori* the information quality of the signal and address this issue. This approach is applied for the first time to optimise an analytical signal.

2.1. Assessment of the information quality of the spectra: entropy

The information theory was adopted in the present study. Based on this theory, Shannon and Rényi entropies were proposed [13,23], and the calculation was adapted in this study to apply them to continuous signals. Thus, the information metrics were calculated following equations (1) and (2).

$$\mathrm{H_S}(\mathbf{Y}) = -\sum_{\mathrm{i}} \mathrm{p_i} \cdot \log_2 \mathrm{p_i} \tag{1}$$

$$\mathrm{H_R}(\mathbf{Y}) = -\log_2\left(\sum_{\mathrm{i}} \mathrm{p_i}^2\right) \tag{2}$$

where H_S and H_R are the Shannon and Rényi entropies, respectively, of the analytical signal embedded in the vector of signal intensities (***Y***) constituted for n elements (or variables), each one symbolised by y_i, and p_i symbolises each intensity value constituting the vector ***Y*** after scaling by total sum normalisation (TSN) [**24**], i.e., p_i values are calculated as follows:

$$p_i = \frac{y_i}{\sum_i y_i} \tag{3}$$

Both strategies to calculate the information entropy were used in this study to compare the results. The values found for each signal (spectrum) obtained will be used as the response variable in the optimisation process. Note that the objective will be to minimise the entropy of the signal, which is equivalent to concluding that the signal has less uncertainty, and therefore contains more useful information.

In order to make it easier for the reader understanding the usefulness of entropy as a metric of Information, let us show a basic example using a very simple signal which is shown in Figure 4.20, where H_S and H_R calculations were applied (equations 1 and 2). Note that in the top row of the figure the raw signals are shown and in the bottom row the respective signal after scaling by TSN. The most informative of the three signals is (c), which shows the lowest entropy values, both H_S and H_R. Signal (a) implies that the instrument responds continuously, regardless of the features or properties of the material under measurement. It is the one that shows the highest entropy, and therefore the lowest information quality. This could be considered as the maximum entropy for a signal of n = 5 variables which does not provide useful information. While signal (b) provides information but lower than (c), and therefore results in an in-between entropy value.

The same calculation was applied to the NMR signals shown in Figure 4.S3A, acquired with low and high field. The HS values were 12.06 and 10.62, and the HR values were 10.88 and 9.66, respectively for low and high field. With this it was proven that effectively for the signal with higher quality (HF-NMR) a lower entropy value is obtained, also according with Belton [**21**].

3. Material and methods

3.1. Chemicals and samples

One marketed sample of virgin olive oil (VOO) was the material chosen to prepare the test portions and carried out the spectra acquisitions.

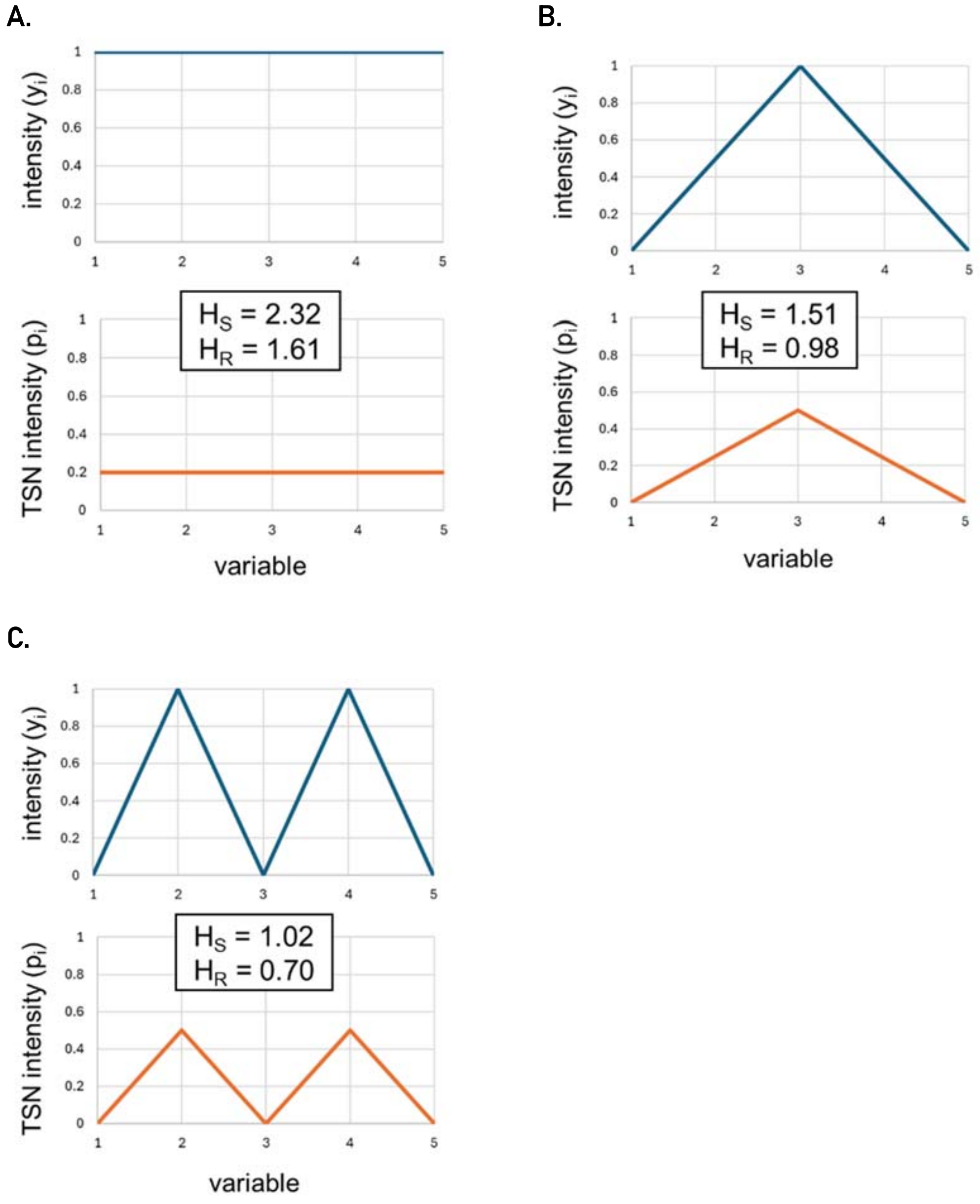


Figure 4.20. Calculation of the Shannon information entropy (H_S) and Rényi information entropy (H_R) of three simple signals (n = 5).
Note that the figures below, i.e., signals represented in orange, correspond to the signals above (in blue) after applying the total sum normalisation.

Non-deuterated bromoform 98% provided by Panreac Quimica SLU (Barcelona, Spain) was used as the internal standard (IS). It should be noted that non-deuterated bromoform was used as IS instead of the common deuterated chloroform for several reasons. On the one hand, in order to develop a more economical analytical method based on LF-NMR, non-deuterated solvents are preferred. On the other hand, bromoform is much less volatile than chloroform, so the material system to be measured becomes more stable than using chloroform. Another important issue when selecting a suitable IS for use in NMR is that it should provide a unique and very well-defined signal, and outside the spectral region due to the concerned material in study, in this case VOO. This had to be satisfied in the present study for both ^{1}H and ^{13}C NMR, and the non-deuterated bromoform met all these requirements, thus it was selected as the IS.

3.2. Experimental design: Taguchi methodology

Two experimental trials were performed to optimise the acquisition instrument conditions for ^{1}H LF-NMR and for ^{13}C LF-NMR spectra acquisition. Among all the instrument settings, 5 were selected as control factors for both trials due to their relevance in the responses to be optimised. These were: number of scans, scan delay, pulse angle, number of points and pre-scans (the last is known as dummy scans in NMR terminology, however, to avoid confusion with dummy factors in DoE, this term was avoided and replaced by pre-scans). The scan delay is also known as relaxation time or recovery delay. Note that, the number of points setting should not be confused with the number of variables contained in each acquired spectrum.

Taguchi methodology was chosen because two critical yet difficult-to-control factors during routine analysis were identified. In addition, as stated in the introduction, this methodology allows to generate a robust testing procedure to small variations on difficult-to-control factors and minimising at the same time the number of experimental runs to be performed, thus providing simplicity when performing an optimising study. On the one hand, the LF-NMR spectrometer to be used in the study is highly sensitive to temperature changes, and the allowable room temperature range is narrow, between 20 and 25 °C. Maintaining an accurate temperature in a research laboratory is already a difficult task. But it is even more difficult in a routine control laboratory, due to the continuous input and output of analysts and the availability of other instrumentation inside the room. On the other hand, the volumes to be introduced into standard NMR tubes are very small, usually around 700 µL. Slight deviations in the sample volumes introduced into the tube results in variations in the height of the liquid inside the tube. This could affect the analytical signal intensity, since the magnetic field must be adjusted with a particular tube liquid height before starting the measurements. Therefore, the possibility of generating a system robust to variations in these two factors was explored.

Hence, the difficult-to-control factors for both designs were the room temperature and the volume introduced into the tube. The experimental designs were performed at two levels. Table 4.5 shows the levels for each of the control factors and difficult-to-control factors for both trials (^{1}H and ^{13}C).

Table 4.5. Levels adopted by each one of the control factors and difficult-to-control factors for the optimisation of ^{1}H and ^{13}C LF-NMR spectra acquisition settings applying the Taguchi methodology.

		^{1}H LF-NMR DoE		^{13}C LF-NMR DoE	
Factors	Level	Min	Max	Min	Max
Control	No scans	2	20	30	300
	Scan delay (s)	2	20	2	20
	Pulse angle (°)	30	90	30	90
	No points	1024	16384	1024	16384
	Pre-scans	0	2	0	2
Difficult-to-control	Room temp (°C)	20	25	20	25
	Sample vol (mL)	650	750	650	750

No: number; temp: temperature; vol: volume

Taguchi methodology divides the experimental layout into two arrays: an internal array for control factors and an external array for difficult-to-control factors. The combination of both gives rise to the crossed array [10]. Due to the number of control factors proposed for the experimental designs, an L16 design was chosen for the internal array and L4 for the external array following Taguchi's terminology. These Taguchi designs correspond to 2^{5-1} and 2^{2} factorial designs, respectively. This experimental design is outlined in Figure 4.21. The reader may refer to [12,25] for further information on this terminology and the designs proposed by Taguchi.

The analysis of the response(s) in the Taguchi methodology is performed only on the internal array, while the external array is the one providing robustness (i.e., stability of the response against small changes in the process value of the difficult-to-control factors). That is, the external array consists of 4 replicates for each of the 16 runs of the internal array, and the combination of both internal and external arrays results in the 64 runs. For each replicate run only the levels of the difficult-to-control factors change. Table 4.6 shows the experimental layout for a Taguchi L16+L4 design from coded levels of the factors (1 and 2) and two spectral responses: run time, t, and information entropy, H; this last used as a metric of the useful information in the spectrum.

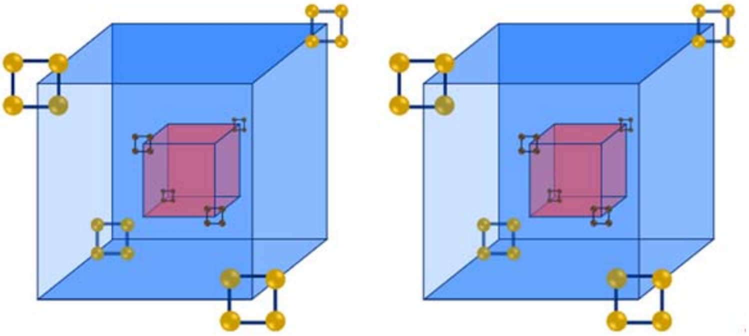


Figure 4.21. Taguchi L16+L4 design layout, which corresponds to a 25-1 + 22 factorial designs, for the internal and external arrays respectively. Note that each cube is constructed for three factors, e.g., A, B and C. The inner and outer cubes denote the levels of the fourth factor (D), while each set of two cubes, left and right, is intended to identify the two levels of the fifth factor (E).

The 4 responses from each replicate run are converted into a single response value, referred to in this study as robust response (RR). Several ways of calculating this value were proposed by Taguchi, being the signal-to-noise ratio (SNR) the best known, since it is the one that seeks robustness by minimising the variability caused by the difficult-to-control factors in the system [**25**].

Two robust responses were considered. Firstly, the variability defined by the 4 replicates from each run of the 16 internal array trials was calculated from the standard deviation (RR_S). This robust response is aimed at verifying if the difficult-to-control factors affect to the pseudo-repeatability of each internal array run. The following step involves the search for the optimal acquisition instrument conditions, being in this case to minimise the run time and information entropy. For this purpose, the Taguchi's signal-to-noise ratio (RR_{SNR}) (smaller the better) was used. Equations (4) and (5) show the calculation of each of these two ways of estimating the RR values for one response.

$$\mathbf{RR_S} = \frac{\sum_i (y_i - \bar{y})^2}{n-1} \tag{4}$$

$$\mathbf{RR_{SNR}} = -10 \cdot \log\left(\frac{\sum_i {y_i}^2}{n}\right) \tag{5}$$

where y_i is the value of the response of the i-th replicate of each run, $\bar{y}$ is the mean of the replicates per run, and n is the total number of replicates per run.

Table 4.6. Experimental layout of the Taguchi design applied to LF-NMR signals acquisition

Run	Control factors					*Difficult-to-control factors*	Responses (Y)									
	A	B	C	D	E	*Temperature*	1		2		1		2		Robust responses (RR_y)	
						Volume	1		1		2		2			
1	1	1	1	1	1		$t_{1,1}$	$H_{1,1}$	$t_{1,2}$	$H_{1,2}$	$t_{1,3}$	$H_{1,3}$	$t_{1,4}$	$H_{1,4}$	RR_{t1}	RR_{H1}
2	1	1	1	2	2		$t_{2,1}$	$H_{2,1}$	$t_{2,2}$	$H_{2,2}$	$t_{2,3}$	$H_{2,3}$	$t_{2,4}$	$H_{2,4}$	RR_{t2}	RR_{H2}
3	1	1	2	1	2		$t_{3,1}$	$H_{3,1}$	$t_{3,2}$	$H_{3,2}$	$t_{3,3}$	$H_{3,3}$	$t_{3,4}$	$H_{3,4}$	RR_{t3}	RR_{H3}
4	1	1	2	2	1		$t_{4,1}$	$H_{4,1}$	$t_{4,2}$	$H_{4,2}$	$t_{4,3}$	$H_{4,3}$	$t_{4,4}$	$H_{4,4}$	RR_{t4}	RR_{H4}
5	1	2	1	1	2		$t_{5,1}$	$H_{5,1}$	$t_{5,2}$	$H_{5,2}$	$t_{5,3}$	$H_{5,3}$	$t_{5,4}$	$H_{5,4}$	RR_{t5}	RR_{H5}
6	1	2	1	2	1		$t_{6,1}$	$H_{6,1}$	$t_{6,2}$	$H_{6,2}$	$t_{6,3}$	$H_{6,3}$	$t_{6,4}$	$H_{6,4}$	RR_{t6}	RR_{H6}
7	1	2	2	1	1		$t_{7,1}$	$H_{7,1}$	$t_{7,2}$	$H_{7,2}$	$t_{7,3}$	$H_{7,3}$	$t_{7,4}$	$H_{7,4}$	RR_{t7}	RR_{H7}
8	1	2	2	2	2		$t_{8,1}$	$H_{8,1}$	$t_{8,2}$	$H_{8,2}$	$t_{8,3}$	$H_{8,3}$	$t_{8,4}$	$H_{8,4}$	RR_{t8}	RR_{H8}
9	2	1	1	1	2		$t_{9,1}$	$H_{9,1}$	$t_{9,2}$	$H_{9,2}$	$t_{9,3}$	$H_{9,3}$	$t_{9,4}$	$H_{9,4}$	RR_{t9}	RR_{H9}
10	2	1	1	2	1		$t_{10,1}$	$H_{10,1}$	$t_{10,2}$	$H_{10,2}$	$t_{10,3}$	$H_{10,3}$	$t_{10,4}$	$H_{10,4}$	RR_{t10}	RR_{H10}
11	2	1	2	1	1		$t_{11,1}$	$H_{11,1}$	$t_{11,2}$	$H_{11,2}$	$t_{11,3}$	$H_{11,3}$	$t_{11,4}$	$H_{11,4}$	RR_{t11}	RR_{H11}
12	2	1	2	2	2		$t_{12,1}$	$H_{12,1}$	$t_{12,2}$	$H_{12,2}$	$t_{12,3}$	$H_{12,3}$	$t_{12,4}$	$H_{12,4}$	RR_{t12}	RR_{H12}
13	2	2	1	1	1		$t_{13,1}$	$H_{13,1}$	$t_{13,2}$	$H_{13,2}$	$t_{13,3}$	$H_{13,3}$	$t_{13,4}$	$H_{13,4}$	RR_{t13}	RR_{H13}
14	2	2	1	2	2		$t_{14,1}$	$H_{14,1}$	$t_{14,2}$	$H_{14,2}$	$t_{14,3}$	$H_{14,3}$	$t_{14,4}$	$H_{14,4}$	RR_{t14}	RR_{H14}
15	2	2	2	1	2		$t_{15,1}$	$H_{15,1}$	$t_{15,2}$	$H_{15,2}$	$t_{15,3}$	$H_{15,3}$	$t_{15,4}$	$H_{15,4}$	RR_{t15}	RR_{H15}
16	2	2	2	2	1		$t_{16,1}$	$H_{16,1}$	$t_{16,2}$	$H_{16,2}$	$t_{16,3}$	$H_{16,3}$	$t_{16,4}$	$H_{16,4}$	RR_{t16}	RR_{H16}

Control factors are: A = number of scans, B = scan delay, C = pulse angle, D = pre-scans, E = number of points.
Responses are: t = run time; H = information entropy; RR = robust response

3.3. LF-NMR measurements

3.3.1. Sample preparation

An aliquot of the VOO sample was mixed with non-deuterated bromoform in a 1:2 ratio (VOO:bromoform) within a 1.5 mL cylindrical plastic vial with a conical bottom (an Eppendorf tube). Since the volume to be introduced into the NMR tube is one of the difficult-to-control factors in this study, two 5-mm standard NMR tubes were carefully filled with 650 µL and 750 µL, respectively, of the prepared VOO:bromoform solution.

3.3.2. LF-NMR spectra acquisition

LF-NMR spectra were acquired using a Nanalysis Benchtop 100PRO NMR Spectrometer (Nanalysis Corp., Calgary, Canada), which was equipped with an autosampler with a capacity of 25 tubes, operating at 100 MHz ^{1}H-frequency and 25.6 MHz ^{13}C-frequency.

The runs to be carried out for each experimental trial were divided in two groups depending on the temperature to be set in the laboratory: 20 °C and 25 °C, in order to have enough time to temper the room to the target temperature without disturbing the instrument stability.

For the H-NMR spectra, the acquisition instrument settings that were not critical to be included as control factors in the experimental trials were set to: 20 ppm spectral width, 6 ppm spectral center and automatic receiver gain. While for the C-NMR, the instrument settings were set to 240 ppm spectral width, 100 ppm spectral center, automatic receiver gain and decoupling mode active for ^{1}H nucleus.

3.4. Data processing

Spectral data were exported in .dx format from the instrument, processed in MestReNova (v14.2.0-26256, Mestrelab Research S.L., Santiago de Compostela, Spain) in order to carry out the phase and baseline correction, followed of a zero-filling to reach an spectral size of 2048 variables for all the spectra. Then, these were exported as .csv file. The information entropy was estimated with an in-house function programmed in MATLAB (R2022a version, Mathworks Inc., Natick, MA, USA). Statgraphics Centurion XVIII software package (version 18.1.12, Statgraphics Technologies, Inc., Virginia, USA) was used for the data treatment and interpretation of the obtained responses from each experimental trial.

As stated in 3.2 section, the considered responses were run time and information entropy. The first one is provided by the instrument and was directly introduced in Statgraphics. While the information entropy, as stated, has been estimated from the application of information theory by applying equations (1) and (2) presented in section 2.1.

4. Results and discussion

4.1. Acquired ^{1}H and ^{13}C LF-NMR spectra

An example of the acquired ^{1}H and ^{13}C LF-NMR spectra is shown in Figure 4.22. The highest peak in both spectra corresponds to the peak of the non-deuterated bromoform (IS), while the rest of the peaks correspond to the VOO fingerprint. The IS signal is found at approximately 6.9 ppm in the ^{1}H NMR spectrum and approximately 11 ppm in the ^{13}C NMR spectrum. It should be noted that non-deuterated bromoform is usually stabilised in ethanol. Therefore, it is possible to see the ethanol quadruplet signal in the ^{1}H NMR spectrum at approximately 3.8 ppm. However, as this signature will be present in all acquired spectra, this will not be significant when chemometrics is applied to develop a multivariate model of classification or quantification from instrumental fingerprints.

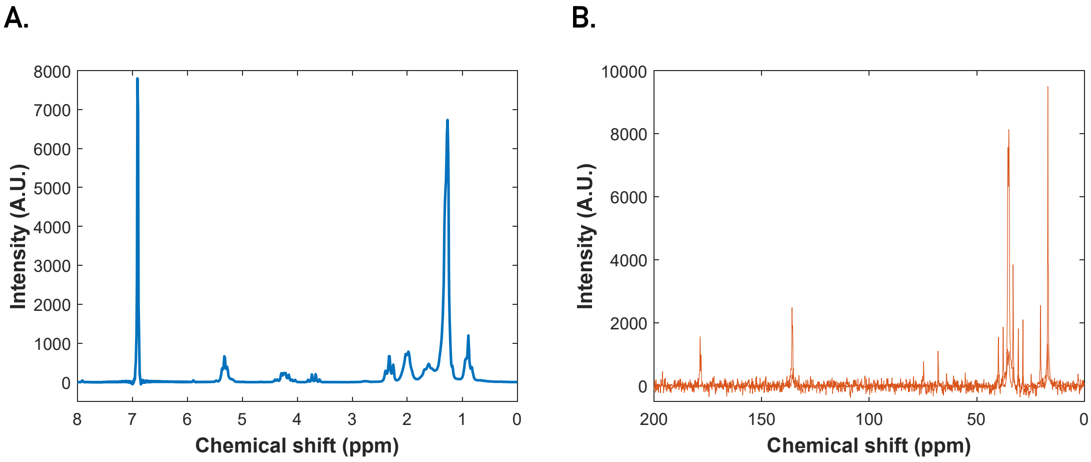


Figure 4.22. Acquired (A) ^{1}H and (B) ^{13}C LF-NMR spectra of the VOO:bromoform test solutions.

The main characteristic peaks of the VOO can be seen in the ^{1}H LF-NMR spectrum (Figure 4.22A), such as the highest one at around 1.2 ppm related to acyl chains. Despite the lower resolution compared to the conventional HF-NMR spectra, it is possible to assign small signals to other compounds, such as the peak around 0.85 ppm to fatty acids present in olive oil except linolenic, peaks between 2 and 2.8 ppm to unsaturated fatty acids such us linolenic and linoleic, or the small peaks around 4.2 and 5.15 ppm usually attributed to glycerol skeleton from triacylglycerols [26,27].

Regarding the VOO ^{13}C LF-NMR spectrum, all the characteristic peaks usually found in conventional HF-NMR, regarding those reported in the literature, can be observed in Figure 4.22B. The first isolated peak appearing around almost 180 ppm, and those peaks at 70 and 75 ppm can be attributed to triacylglycerols, while those between 130 and 140 ppm to unsaturated fatty acids. The rest of the peaks located between 10 and 40 ppm use to be assigned to carbons forming acyl chains [**26**,**28**].

4.2. Optimisation of acquisition instrument settings for ^{1}H-NMR spectra

Upon completing the experimental part of the study, 64 ^{1}H LF-NMR spectra were acquired. Figure 4.S4A (supplementary material) shows 16 raw spectra, which were acquired with the different combinations of the 5 control factors shown in Table 4.6. As explained, these 16 runs were performed in quadruplicate by changing the levels of the two difficult-to-control factors, room temperature and tube volume. It is worth mentioning the results of runs 4 and 11 where shifted spectra are observed in comparison with the others (see Figure 4.23, where runs 3 and 12 are also shown for visual comparison). Probably, bromoform did not have enough time to relax after the pulse and the signal collected from the IS had a lower intensity than the highest peak of the VOO. This led the instrument to recognise the latter signal as that of bromoform, completely shifting the entire spectrum, since the highest peak of VOO is usually at about 1.5 ppm and the instrument placed it at 6.9 ppm, which is the usual chemical shift of bromoform. This occurred in all four replicates of both runs. Since the purpose is to obtain high informative signals, using bromoform as the IS (hence as the highest peak of the spectrum), these spectra were considered to be of very low quality for the intended purpose.

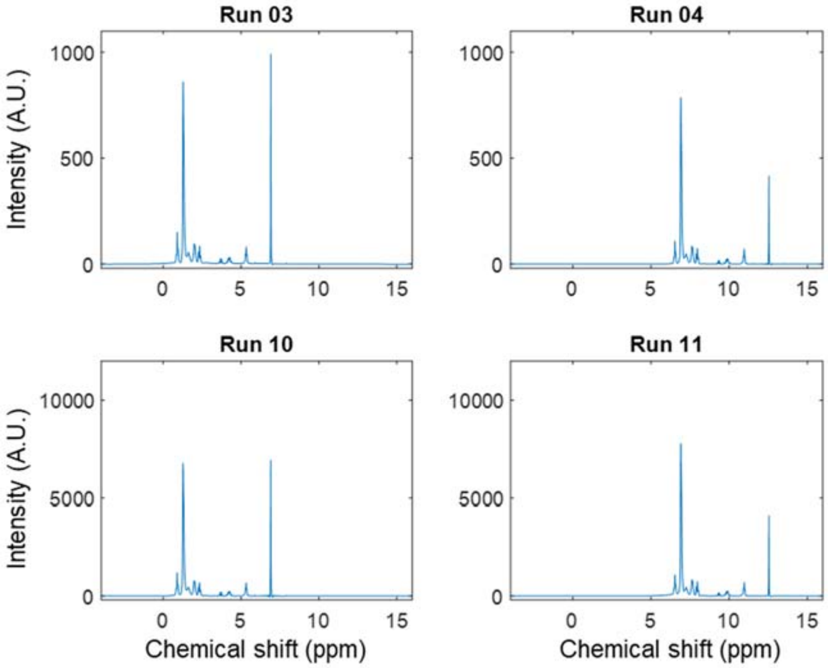


Figure 4.23. ^{1}H LF-NMR acquired spectra from runs 3, 4, 10 and 11 (see the combination of control factors in Table 4.6). *Note that spectra from 4 and 11 are shifted with respect to the others.*

As some spectra were acquired with 1024 points and others with 16384 points, they were processed in MestReNova so that all spectra had the same number of variables (2048). This was necessary to evaluate the amount of information present by calculating the information entropy, since a lower or higher number of variables affects this result [**21**]. However, this is already part of one of the factors to be controlled (number of points) and it would be redundant to study this variable as part of the response and that is why the number of variables of all the spectra has been equalised.

After signal processing, the spectra were trimmed to include only the instrumental fingerprint of the VOO for the calculation of the information entropy. Finally, signals of 513 variables in the range 0.5 - 5.5 ppm were obtained. After that, the spectra were TSN scaled, and the information entropies were calculated. The values of the H_S and H_R are shown in Table 4.S1 (supplementary material). The maximum information entropy was assigned to the low-quality spectra (runs 4 and 11), calculated as the logarithm of the number of variables [**15**]. Meanwhile, for the response run time, the minimum value was 10.3 seconds, and the maximum value was 11 minutes and 20 seconds.

Next, the variability (RR_S) and SNR (RR_{SNR}) robust responses were studied individually. A 2nd grade polynomial model was fitted which included the interaction and linear terms, and after excluding the non-significant interactions (P-value < 0.05), the Pareto charts shown in Figure 4.S5 (supplementary material) were obtained.

If only the RR_S robust responses are examined, all five control factors seem to produce variability in the Shannon entropy response. On the contrary, only number of points and number of scans factors had significant effect on the Rényi entropy, together with the same interactions that were significant for Shannon entropy. However, only two factors (number of scans and pulse angle) have such a significant effect over run time response.

On the other hand, analysing the results from RR_{SNR}, it was observed that the control factors generated opposite behaviours in the responses. That is, for example, as expected, increasing the scan delay and the number of points increased the run times and decreased the information entropies, resulting in an increase of signal information quality. Note that by using the RR_{SNR}, the optimisation goal (minimising, in this case) is already implicit. Therefore, the higher the value of RR_{SNR}, the closer the response (time or entropy) is to the minimum. This means that the optimal value of each factor will be opposite for each response. This made it necessary to perform a second step: a multiple response analysis based on the desirability function [**29**], to find the optimal value of each factor that satisfies the optimality of both responses.

The RR_{SNR} Pareto charts for the information entropy response show that the most significant effect is caused by the interactions rather than by the individual factors. But even more striking is that one of the interactions includes the pre-scans factor whose effect was not found to be significant. This uncommon outcome in experimental design analysis could be justified by looking at the results of 2 of the 16 runs. These are those already discussed above (see the spectra of runs 4 and 11 in Figure 4.23), which were given the maximum information entropy value (Table 4.S1 in supplementary material). This reveals that a combination of 2 seconds scan delay, 90° pulse angle and 1024 number of points is not adequate to provide ^{1}H LF-NMR VOO sufficiently informative signals for the intended purpose.

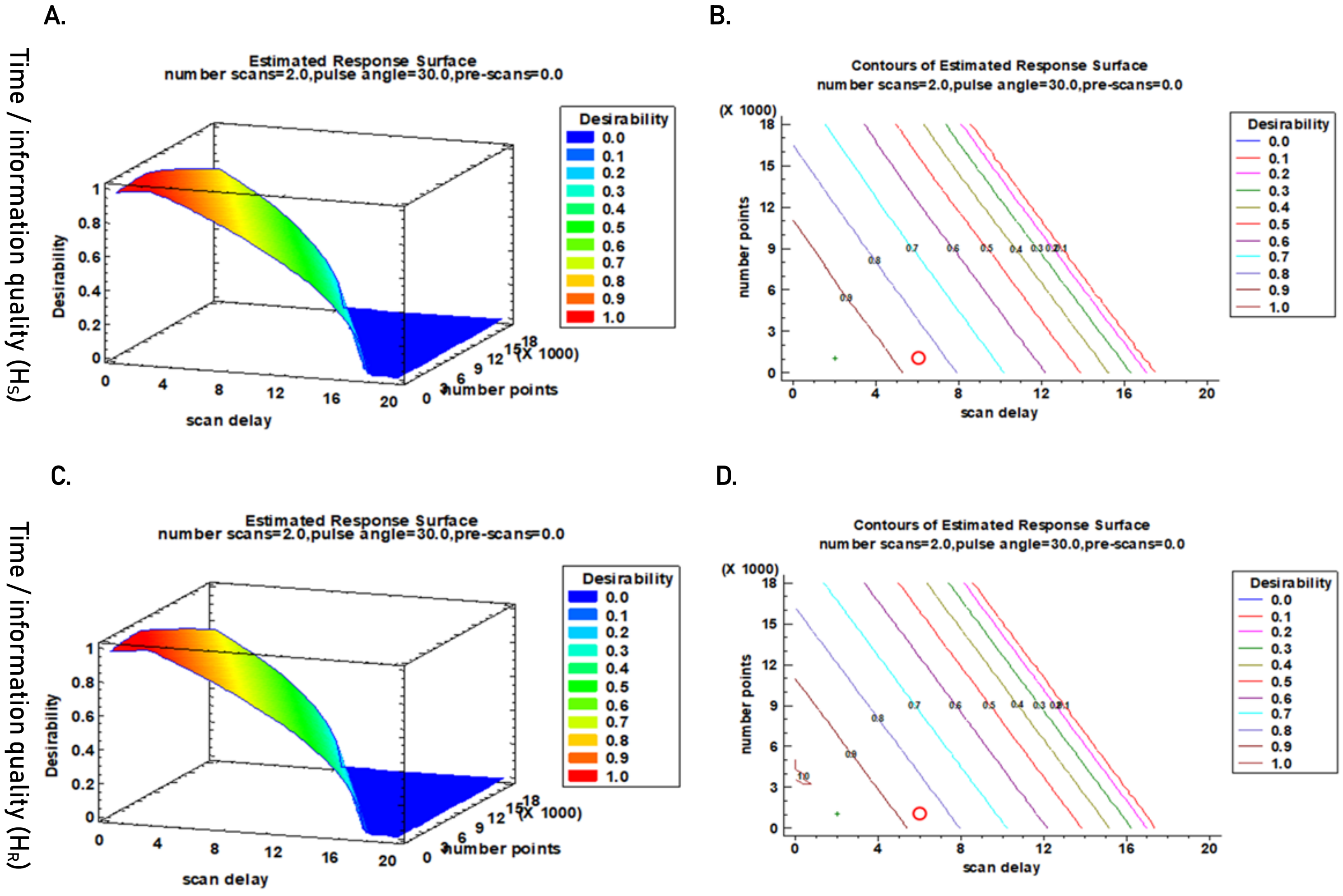
Two of the five control factors had a significant effect on both responses (run time and information entropy): the scan delay and the number of points. While the other three control factors only showed a significant effect on one of the two responses. This led the multiple response analysis to focus on optimising the control factors that significantly affected both responses (scan delay and number of points). The rest of the factors were set to the optimum level obtained to optimise the response where it was found to be a significant factor. These were as follows: pulse angle: 30°, pre-scans: 0, and number of scans: 2.

The estimated time-H_S and time-H_R response surfaces were obtained (see Figures 4.10A and 4.10C, respectively). There is a sharp drop in the desirability function and a plane where the desirability is 0 as the value of both control factors increases. This is mainly because the increase in both factors leads to a drastic increase in the run time of the ^{1}H LF-NMR spectra. However, this change is not as abrupt on the information entropy. Nevertheless, there is a factor region where the desirability function approaches the optimal value (1).

By plotting the surface contours, the model identifies the minimum levels of both factors as the optimal point (see Figures 4.10B and 4.10D): specifically, 2 seconds of scan delay and 1024 points. This configuration maintains the desirability above 0.9 when using H_S as the signal information metric, and nearly at 1.0 when using H_R. However, it is known that bromoform (IS) requires sufficient time to relax after receiving the pulse. The authors demonstrated that this condition affects the peak intensity, as 2 seconds was an insufficient scan delay time. This is evident in 2 of the 16 spectra shown in Figure 4.23, as previously mentioned. Therefore, it was decided to increase the scan delay value to 6 seconds. This is shown in Figure 4.24B and 4.10D with a red circle, where the desirability remains above 0.8.

Therefore, finally the optimal acquisition settings for ^{1}H LF-NMR spectra were: 2 scans, 6 seconds scan delay, 30° pulse angle, 1024 number of points and 0 pre-scans. This yielded a run time of 15 seconds per spectrum.

Figure 4.24. Response surfaces estimated from multiple response analysis (desirability function) for ^{1}H LF-NMR signal acquisition optimisation for run time and information entropies: (A) Shannon entropy and (C) Rényi entropy responses, and their respective contours (B) and (D).



4.3. Optimisation of acquisition instrument settings for ^{13}C-NMR spectra

The 64 acquired ^{13}C LF-NMR spectra (shown in Figure 4.S4B) were also processed in MestReNova to ensure consistence in the number of variables (2048). After that, the spectra were trimmed to include only the instrumental fingerprint of the VOO, thus removing the IS signal. This resulted in spectra in the 18 - 180 ppm range contained in vectors of 1348 variables. As in the previous section for ^{1}H LF-NMR, the ^{13}C LF-NMR spectra were TSN scaled and the information entropies, H_S and H_R, were calculated. They are shown in Table 4.S2 (supplementary material). For these set of spectra, the minimum run time was approximately 2.5 minutes, and the maximum was more than 2 hours (129 minutes).

The same data analysis sequence as in the previous section was conducted. First, the influence of the control factors on the two responses was studied separately using RR_S and RR_{SNR} again assuming a 2nd-degree polynomial model, consisting only of the linear terms and the 2nd-order interactions. After excluding non-significant interactions, the Pareto charts shown in Figure 4.S6 (supplementary material) were obtained.

Regarding RR_S study, all the factors affected the between-replicate variability of the run time. In the case of the information entropy response, different factors were significant when comparing H_S and H_R. Specifically, pulse angle, number of scans, and scan delay were significant for H_S, while, surprisingly, only the interactions showed a significant effect on between-replicate variability for H_R. Note that to be considered significant effect the P-value should be below 0.05. However, it can be considered doubtful up to a P-value limit below 0.20. In the RR_S results for the study of information entropy using Rényi's entropy, the scan delay, pulse angle and pre-scans factors were found to be below 0.20 (specifically the P-value was 0.17, 0.08 and 0.16 respectively). Therefore, they are within the limit of doubtfulness and could be considered significant.

The data analysis from RR_{SNR} yielded similar results to the previous ^{1}H LF-NMR trial. That is, opposite behaviours were observed in some control factors on the responses, such as the number of scans and the scan delay. In addition, the effect of the number of scans was significant in both cases, which led again to perform the next step: multiple response analysis, to find the optimal point that satisfies the optimisation of both responses.

It should be noted that for the pulse angle, the effect was opposite to the one found for ^{1}H LF-NMR trial. The result indicates that the optimum level of this instrument setting is the minimum, 30°, for the acquisition of ^{1}H LF-NMR signals and the maximum, 90°, for ^{13}C LF-NMR. This is in some agreement with the scientific literature, as although it is common to use a 90° pulse for ^{1}H NMR spectra acquisition, occasionally a 30° pulse has been applied [1,30].

Next, it was decided to study simultaneously the factors of scan delay and number of scans. The other factors were set to the optimum value obtained for the response where their effect was significant: pulse angle at 90°, pre-scans at 0 and number of points at 1024. Figure 4.25 shows the results of the multiple response analysis, both the estimated response surfaces and the contours.

The observed results are very similar to the one found for ^{1}H LF-NMR trial. The sharp drop of the desirability function is observed as the values of the two factors concerned increase. The experimental region where the desirability reaches unity is smaller in this case, and much larger the area where the desirability is 0. The optimum point predicted by the model (see Figure 4.25B and 4.11D) is at the minimum level studied for both control factors, i.e., 30 scans and 2 seconds of scan delay. At this point, the desirability remains well above 0.9, being 1.0 in the study of signal quality by H_S (Figure 4.25B) and 0.99 by H_R (Figure 4.25D). On this occasion, unlike in the previous section, the decision made was to accept the optimum point offered by the model and not to increase the scan delay. This is because the ^{13}C LF-NMR signal of the IS is not affected by a lower scan delay.

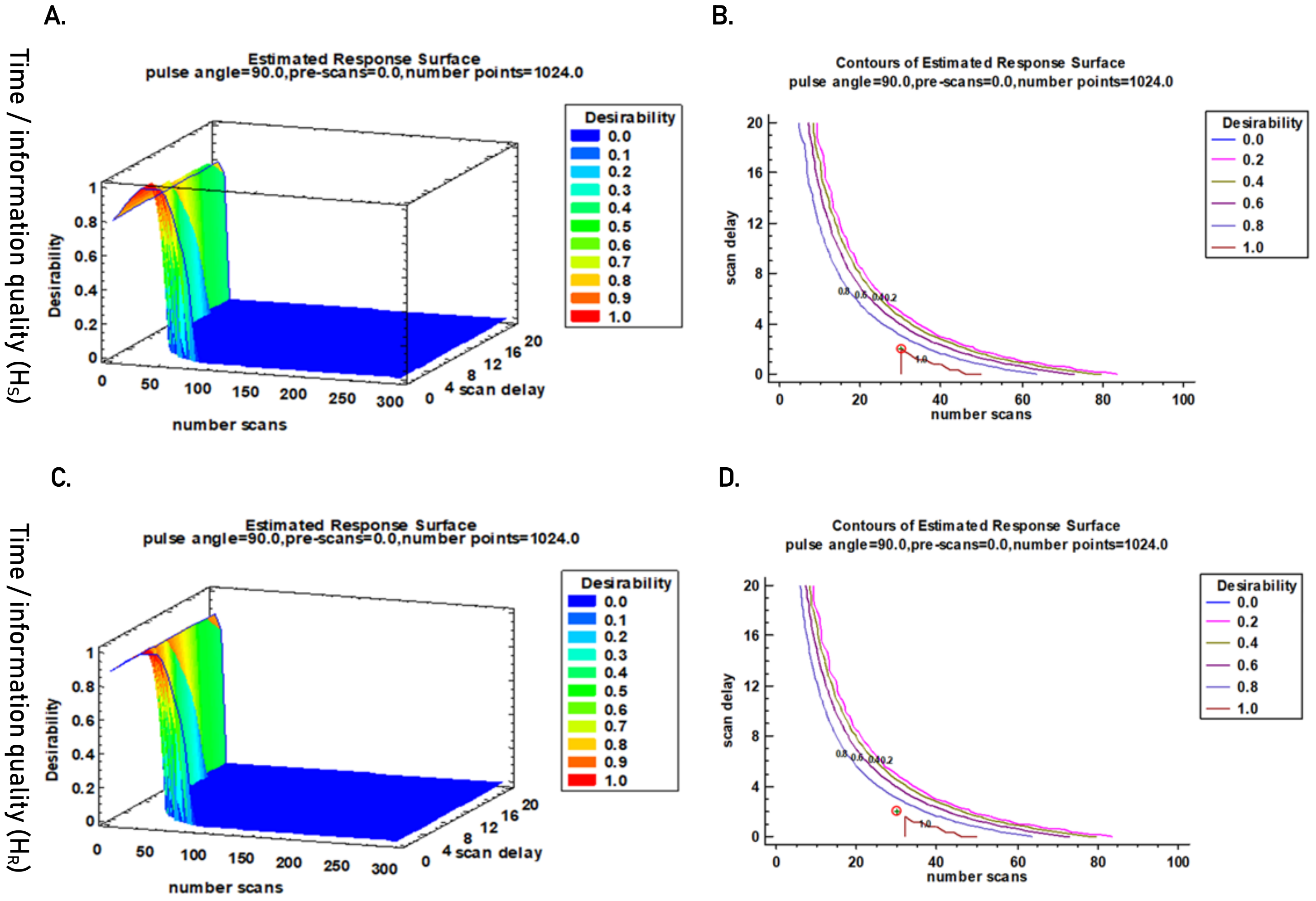
Therefore, the resulting values after optimisation of the instrument conditions for ^{13}C LF-NMR signal acquisition were: 30 scans, 2 seconds of scan delay, 90° pulse angle, 1024 number of points and 0 pre-scans. This combination of instrument settings resulted in a run time per spectrum of approximately 2.5 minutes.

4.4. Evaluation of the difficult-to-control factors influence

Finally, the effect of difficult-to-control factors on responses was independently examined and the experimental design was analysed applying the crossed array option which is provided by the software. To this end, it is not necessary to calculate the robust responses values, but all the runs of the internal and external matrix, i.e., all 64 runs, are jointly studied. After excluding non-significant interactions, the Pareto charts were obtained (Figure 4.S7, supplementary material).

Regarding the ^{1}H LF-NMR trial, it is observed that the volume to be introduced into the tube had a significant effect (P-value < 0.05) on the run time response, while the room temperature was not significant, although it lies in the doubtfulness interval since the P-value for this factor was 0.14. In contrast, temperature did have a significant effect on the information entropy response by H_S (P-value < 0.05) and very close to the limit of significance for H_R (P-value = 0.06). While the tube volume factor was not significant for this response (P-value > 0.2).

Figure 4.25. Response surfaces estimated from multiple response analysis (desirability function) for ^{13}C LF-NMR signal acquisition optimisation for run time and information entropies: (A) Shannon entropy and (C) Rényi entropy responses, and their respective contours (B) and (D).



Different findings were observed for the optimisation of ^{13}C LF-NMR signals. In this case, neither of the two difficult-to-control factors caused a significant or noticeable effect on the run time response. This indicates that slight fluctuations in room temperature and tube volume are not as critical for the acquisition of ^{13}C LF-NMR signals (P-values were 0.55 and 0.71, respectively). In contrast, these two difficult-to-control factors had a significant effect on the information quality estimated through the H_R calculation (P-value < 0.05). However, only the volume showed a significant effect for information quality through H_S (P-value < 0.05), while for room temperature, the P-value was > 0.2. The results from both trials confirm the initial hypothesis that the two difficult-to-control factors could be critical for the acquisition of high-informative LF-NMR signals.

In addition, another noteworthy conclusion is the difference seen in the latter stage between the H_S and H_R results. Generally, throughout this study, no other relevant differences were observed between both, and similar results and conclusions were drawn. In line with other sciences where Rényi entropy is applied, certainly using the latter it is possible to see differences related to the probability of an event occurring [23]. In fact, the variability found in the present study between all H_S with respect to all H_R calculated for the same signals, expressed in terms of relative standard deviation, was 9.7 % and 17.9 % in H_S and H_R, respectively, for ^{1}H LF-NMR signals, and 4.3 % and 10.1 % for ^{13}C NMR signals (as shown in supplementary material, Tables 4.S1 and 4.S2).

5. Final remarks

This study describes how the acquisition instrument conditions of LF-NMR signals, both ^{1}H and ^{13}C, can be determined to ensure the highest informational quality in the shortest possible run time. By using the Taguchi methodology, the acquisition settings have been optimised to obtain a robust combination when faced with variations in two factors considered critical but difficult-to-control. Therefore, within the ranges studied for room temperature and small deviations of volume in the NMR-standard tube not exceeding 22.5 ± 2.5 °C and 700 ± 50 µL respectively, it is possible to effectively acquire ^{1}H and ^{13}C LF-NMR high informative spectra of virgin olive oils. Thus, these can be used as instrumental fingerprints in the development of new rapid and non-destructive analytical methods based on a fingerprinting approach.

In addition, a secondary objective was to propose a way to *a priori* assess the fit-for-purpose of an analytical signal, an unresolved challenge in current analytical chemistry that would save considerable experimental time when developing new non-targeted analytical methods based on instrumental fingerprints. The proposal based on information theory and the calculation of the information entropy, after applying it in the present study, has proved to be of great potential for the desired purpose: to quantify the information of the analytical signal, and ultimately select which analytical signal is likely to lead to better qualitative or quantitative prediction models.

Funding

This work was supported by the MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR (public-private collaboration project CPP2021-008672).

Acknowledgments

AAC gratefully acknowledges the Spanish Ministry of Sciences, Innovation and Universities for a pre-doctoral fellowship FPU (FPU20/04711, Formación del Profesorado Universitario -Training of University Lecturer-). In addition, **AMJC** acknowledges the Grant (RYC2021-031993-I) funded by MICIU/AEI/10.13039/501100011033 and the European Union NextGenerationEU/PRTR. **RLR** acknowledges to the Andalusia Ministry of Economic Transformation, Industry, Knowledge and Universities for financial support from "Ayudas para Captación, Incorporación y Movilidad de Capital Humano de I + D + i (PAIDI 2020)".

References

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

[2] E. Hatzakis, Nuclear magnetic resonance (NMR) spectroscopy in food science: A comprehensive review, Compr. Rev. Food Sci. Food Saf. 18 (2019) 189-220. https://doi.org/10.1111/1541-4337.12408.

[3] A.P. Sobolev, F. Thomas, J. Donarski, C. Ingallina, S. Circi, F.C. Marincola, D. Capitani, L. Mannina, Use of NMR applications to tackle future food fraud issues, Trends Food Sci. Technol. 91 (2019) 347-353. https://doi.org/10.1016/j.tifs.2019.07.035.

[4] H.Y. Yu, S. Myoung, S. Ahn, Recent applications of benchtop nuclear magnetic resonance spectroscopy, Magnetochemistry 7 (2021) 121. https://doi.org/10.3390/magnetochemistry7090121.

[5] J. Giberson, J. Scicluna, N. Legge, J. Longstaffe, Developments in benchtop NMR spectroscopy 2015–2020, in G.A. Webb (Ed.), Annual Reports on NMR Spectroscopy, Vol. 102, Academic Press, USA, 2021, pp. 153-246.

[6] D. Galvan, L.M. de Aguiar, E. Bona, F. Marini, M.H.M. Killner, Successful combination of benchtop nuclear magnetic resonance spectroscopy and chemometric tools: A review, Anal. Chim. Acta 1273 (2023) 341495. https://doi.org/10.1016/j.aca.2023.341495.

[7] T. Head, R.T. Giebelhaus, S.L. Nam, A.P. de la Mata, J.J. Harynuk, P.R. Shipley, Discriminating extra virgin olive oils from common edible oils: Comparable performance of PLS-DA models trained on low-field and high-field 1H NMR data, Phytochem. Anal. 2024 (2024) 1-8. https://doi.org/10.1002/pca.3348.

[8] G.M. Kamal, J. Uddin, M.S. Tahir, M. Khalid, S. Ahmad, A.I. Hussain, Nuclear Magnetic Resonance Spectroscopy in Food Analysis, in M.S. Khan, M.S. Rahman (Eds.), Techniques to Measure Food Safety and Quality: Microbial, Chemnical, and Sensory, Springer, Switzerland, 2021, pp. 137-168.

[9] A. Gałuszka, Z. Migaszewski, J. Namieśnik, The 12 principles of green analytical chemistry and the SIGNIFICANCE mnemonic of green analytical practices, Trends Anal. Chem. 50 (2013) 78-84. https://doi.org/10.1016/j.trac.2013.04.010.

[10] M. Cavazzuti, Design of experiments, in M. Cavazzuti (Ed.), Optimization methods: from theory to design scientific and technological aspects in mechanics, Springer, New York, 2013, pp. 13-42.

[11] K. Kalinowska, M. Bystrzanowska, M. Tobiszewski, Chemometrics approaches to green analytical chemistry procedure development, Curr. Opin. Green Sustain. Chem. 30 (2021) 100498. https://doi.org/10.1016/j.cogsc.2021.100498.

[12] R.K. Roy, A Primer on the Taguchi Method, 2nd ed, Society of Manufacturing Engineers, Southfield, 2010.

[13] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.

[14] K.M. Wright, Maximum entropy methods in NMR data processing, in: D.N. Rutledge (Ed.), Signal treatment and signal analysis in NMR, in: B.G.M. Vandeginste, S.C. Rutan (Eds.) Data handling in science and technology, vol 18, Elsevier, Amsterdam, 1996, pp. 25-43.

[15] K. Eckschlager, V. Štěpánek, K. Danzer, A review of information theory in analytical chemometrics, J. Chemom. 4 (1990) 195-216. https://doi.org/10.1002/cem.1180040303.

[16] K. Eckschlager, K. Danzer, Information Theory in Analytical Chemistry, John Wiley & Sons, New York, 1994.

[17] K. Danzer, Analytical Chemistry – Theoretical and Metrological Fundamentals, Springer-Verlag Berlin Heidelberg, 2007, ch. 9, pp. 265-282.

[18] C. García, T. Hernández, F. Costa, B. Ceccanti, G. Masciandaro, M. Calcinai, Evaluation of the organic matter composition of raw and composted municipal wastes, Soil Sci. Plant Nutr. 39 (1993) 99-108. https://doi.org/10.1080/00380768.1993.10416979.

[19] Z. Liu, M. Huang, Q. Zhu, J. Qin, M.S. Kim, A packaged food internal Raman signal separation method based on spatially offset Raman spectroscopy combined with FastICA, Spectrochim. Acta A Mol. Biomol. 275 (2022) 121154. https://doi.org/10.1016/j.saa.2022.121154.

[20] M.P. Rueda, F. Comino, V. Aranda, A. Domínguez-Vidal, M.J. Ayora-Cañada, Analytical pyrolysis (Py-GC-MS) for the assessment of olive mill pomace composting efficiency and the effects of compost thermal treatment, J. Anal. Appl. Pyrol. 168 (2022) 105711. https://doi.org/10.1016/j.jaap.2022.105711.

[21] P.S. Belton, How much information is there in an NMR measurement? in: I.A. Farhat, P.S. Belton, G.A. Webb (Eds.), Magnetic Resonance in Food Science. From Molecules to Man, RSC Publishing, Dorset, 2007, pp. 177-183.

[22] J.C. Hoch, M.W. Maciejewski, M. Mobli, A.D. Schuyler, A.S. Stern, Nonuniform sampling and maximum entropy reconstruction in multidimensional NMR, Acc. Chem. Res. 47 (2014) 708-717. https://doi.org/10.1021/ar400244v.

[23] A. Rényi, On measures of entropy and information, in Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, vol 1: contributions to the theory of statistics vol 4, University of California Press, 1961 January, pp. 547-562.

[24] J. Walach, P. Filzmoser, K. Hron, Data normalization and scaling: Consequences for the analysis in omics sciences, in: J. Jaumot, C. Bedia, R. Tauler (Eds.), Data analysis for omics sciences: Methods and applications, in: D. Barceló (Ed.), Comprehensive Analytical Chemistry, vol 82, Elsevier, Amsterdam, 2018, pp. 165-196.

[25] Y. Wu, A. Wu, Taguchi methods for robust design, American Society of Mechanical Engineers, New York, 2000.

[26] R. Sacchi, F. Addeo, L. Paolillo, 1H and 13C NMR of virgin olive oil. An overview, Magn. Reson. Chem. 35 (1997) S133-S145. https://doi.org/10.1002/(SICI)1097-458X(199712)35:13<S133::AID-OMR213>3.0.CO;2-K.

[27] R.M. Alonso-Salces, L.A. Berrueta, B. Quintanilla-Casas, S. Vichi, A. Tres, M.I. Collado, C. Asensio-Regalado, G.E. Viacava, A.A. Poliero, E. Valli, A. Bendini, T. Gallina Toschi, J.M. Martínez-Rivas, W. Moreda, B. Gallo, Stepwise strategy based on 1H-NMR fingerprinting in combination with chemometrics to determine the content of vegetable oils in olive oil mixtures, Food Chem. 366 (2022) 130588. https://doi.org/10.1016/j.foodchem.2021.130588.

[28] S. Guyader, F. Thomas, V. Portaluri, E. Jamin, S. Akoka, V. Silvestre, G. Remaud, Authentication of edible fats and oils by non-targeted 13C INEPT NMR spectroscopy, Food Control 91 (2018) 216-224. https://doi.org/10.1016/j.foodcont.2018.03.046.

[29] L. Vera Candioti, M.M. De Zan, M.S. Cámara, H.C. Goicoechea, Experimental design and multiple response optimization. Using the desirability function in analytical methods development, Talanta 124 (2014) 123-138. https://doi.org/10.1016/j.talanta.2014.01.034.

[30] M. Defernez, E. Wren, A.D. Watson, Y. Gunning, I.J. Colquhoun, G. Le Gall, D. Williamson, E.K. Kemsley, Low-field 1H NMR spectroscopy for distinguishing between arabica and robusta ground roast coffees, Food Chem. 216 (2017) 106-113. http://dx.doi.org/10.1016/j.foodchem.2016.08.028.

SUPPLEMENTARY INFORMATION *(Artículo científico 6)*

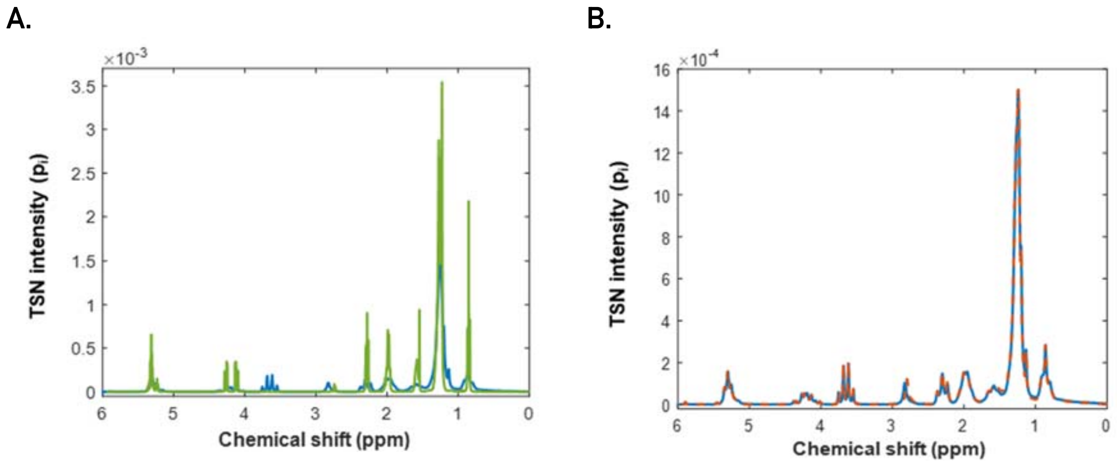


Figure 4.S3. NMR normalised signals (^{1}H spectra) of the same sample acquired: (A) by different instrument, namely: 100 MHz LF-NMR (blue line) and 400 MHz HF-NMR (green line), and (B) by different acquisition settings and same instrument (100 MHz LF-NMR).

TSN intensity = total sum normalised intensity

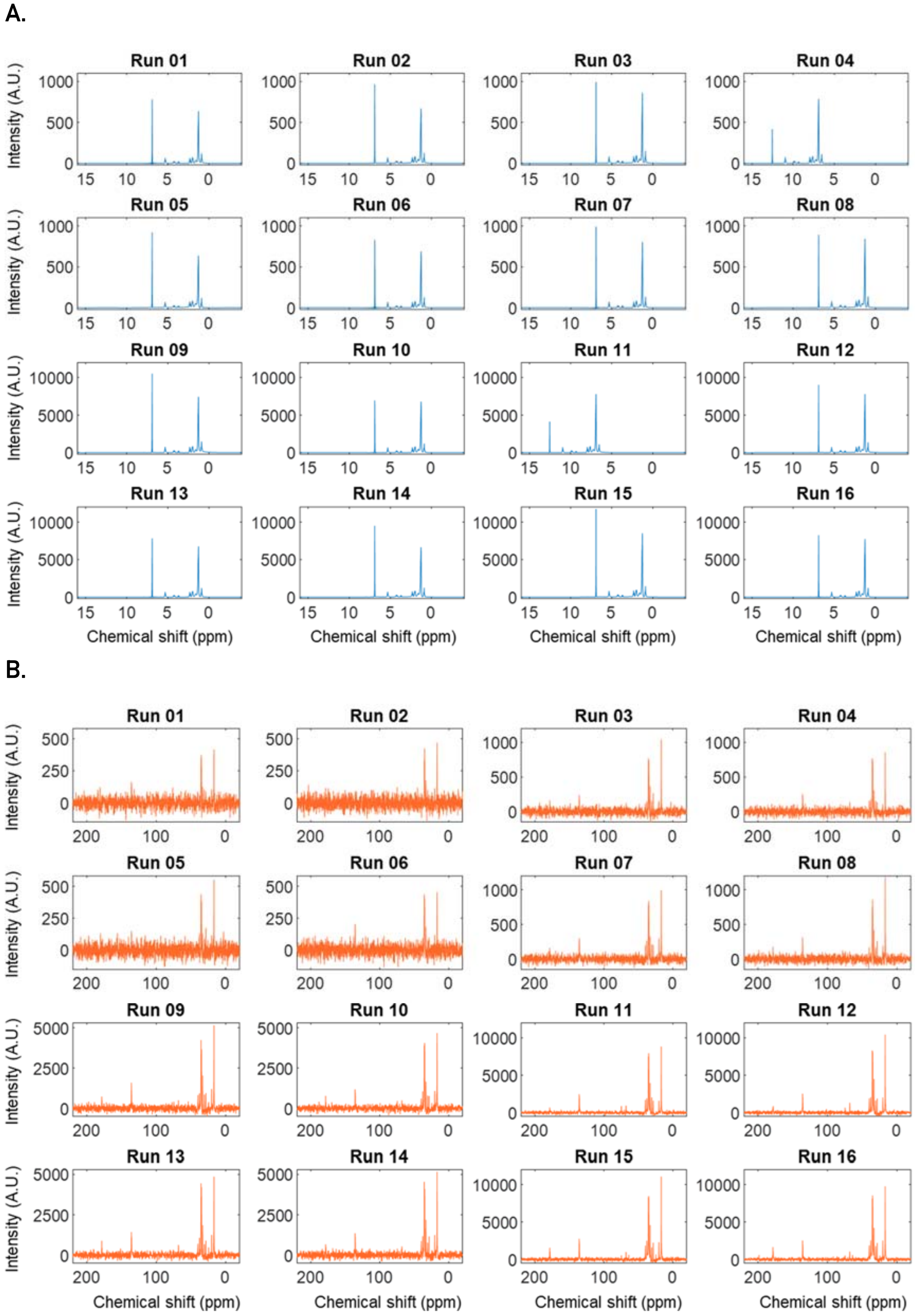


Figure 4.S4. (A) ^{1}H LF-NMR and (B) ^{13}C LF-NMR acquired spectra after applying the 16 combinations of the different control factors included in the experimental trial (see Table 4.6 in the manuscript).

Table 4.S1. Values resulting from the calculation of the information entropy (both Shannon, H_S, and Rényi, H_R) from the 64 ^{1}H LF-NMR spectra.

Run	Information quality of ^{1}H LF-NMR spectra							
	Shannon entropy (H_S)				Rényi entropy (H_R)			
	Rep 1	Rep 2	Rep 3	Rep 4	Rep 1	Rep 2	Rep 3	Rep 4
1	6.76	6.71	6.76	6.78	5.59	5.60	5.59	5.63
2	7.20	7.24	7.26	7.27	5.93	5.97	5.98	5.99
3	7.12	7.10	7.20	7.21	5.78	5.76	5.84	5.85
4	9.00	9.00	9.00	9.00	9.00	9.00	9.00	9.00
5	7.29	7.23	7.29	7.30	6.03	5.94	6.00	6.03
6	6.79	6.78	6.79	6.77	5.62	5.60	5.62	5.62
7	6.68	6.67	6.68	6.66	5.47	5.45	5.47	5.46
8	7.15	7.12	6.96	7.26	5.81	5.77	5.66	5.90
9	6.90	7.23	7.30	7.28	5.73	5.94	6.02	5.99
10	6.82	6.78	6.82	6.78	5.66	5.63	5.66	5.62
11	9.00	9.00	9.00	9.00	9.00	9.00	9.00	9.00
12	7.05	7.11	7.24	7.00	5.72	5.76	5.88	5.66
13	6.79	6.77	6.79	6.77	5.64	5.63	5.64	5.60
14	7.24	7.23	7.29	7.27	5.97	5.95	6.00	5.97
15	7.16	6.94	7.15	7.15	5.82	5.64	5.80	5.77
16	6.70	6.68	6.70	6.64	5.51	5.47	5.51	5.42
RSD	*9.7 %*				*17.9 %*			

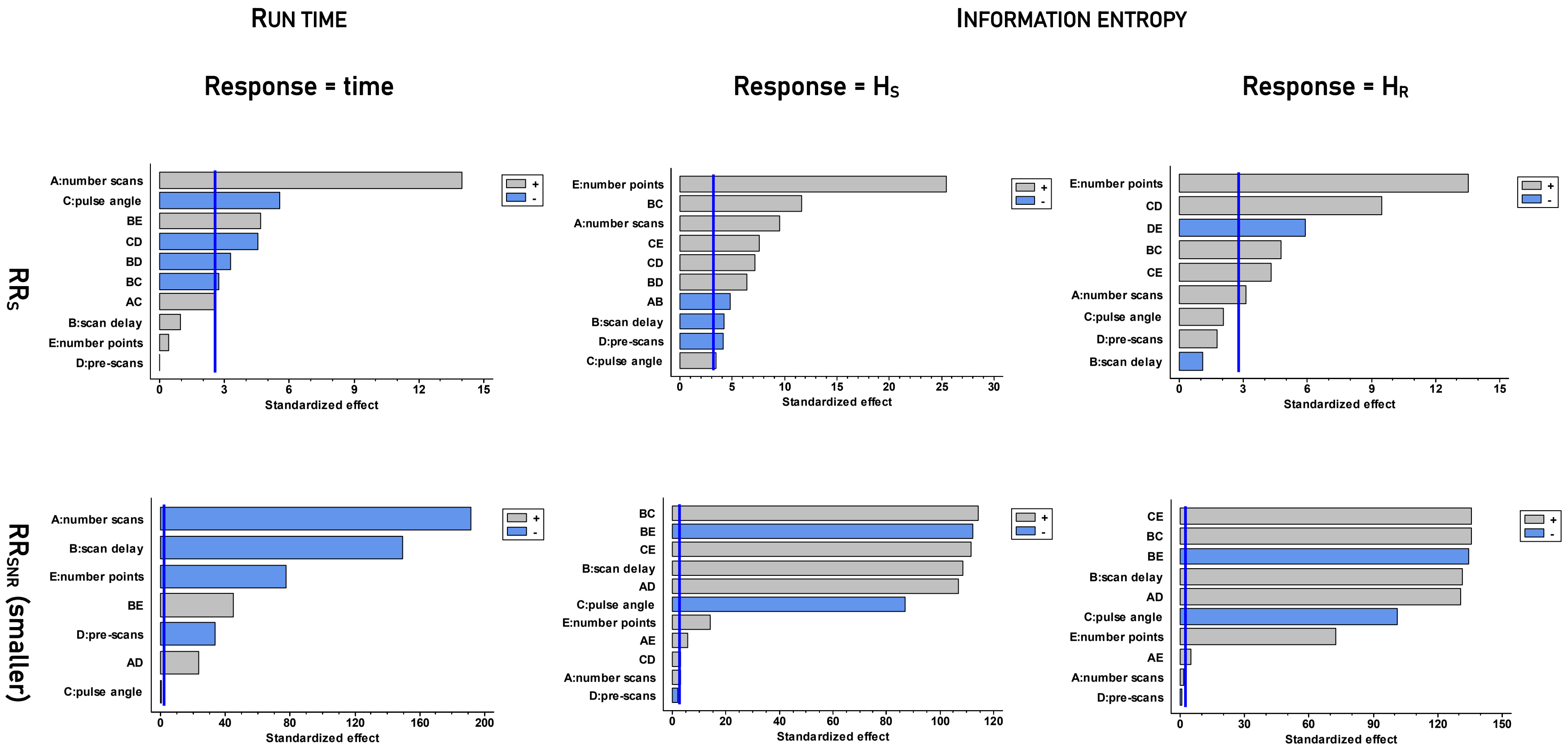
Rep 1 - Rep 4 = the four replicates from the external array; RSD = relative standard deviation

Table 4.S2. Values resulting from the calculation of the information entropy (both Shannon, H_S, and Rényi, H_R) from the 64 ^{13}C LF-NMR spectra.

Run	Information quality of ^{13}C LF-NMR spectra							
	Shannon entropy (H_S)				Rényi entropy (H_R)			
	Rep 1	Rep 2	Rep 3	Rep 4	Rep 1	Rep 2	Rep 3	Rep 4
1	9.78	9.81	9.80	9.77	8.97	8.96	9.01	8.99
2	9.95	9.92	9.94	9.93	9.51	9.42	9.50	9.46
3	9.75	9.73	9.78	9.70	8.80	8.74	8.93	8.70
4	9.74	9.69	9.77	9.68	8.78	8.65	8.84	8.63
5	9.91	9.90	9.91	9.91	9.42	9.35	9.41	9.36
6	9.92	9.90	9.90	9.91	9.39	9.36	9.34	9.42
7	9.64	9.70	9.69	9.63	8.58	8.68	8.67	8.56
8	9.72	9.62	9.71	9.70	8.74	8.58	8.72	8.73
9	9.54	9.38	9.49	9.44	8.24	7.94	8.17	8.04
10	9.39	9.34	9.39	9.40	7.95	7.91	7.98	7.95
11	8.83	8.80	8.92	8.77	7.04	7.04	7.14	7.12
12	8.86	8.84	8.91	8.88	7.11	7.14	7.21	7.15
13	9.35	9.38	9.44	9.33	7.88	7.94	8.05	7.84
14	9.43	9.40	9.43	9.41	8.03	8.00	8.08	8.00
15	8.85	8.80	8.83	8.73	7.15	7.08	7.13	7.10
16	8.77	8.75	8.83	8.76	7.07	7.02	7.10	7.09
RSD	*4.3 %*				*10.1 %*			

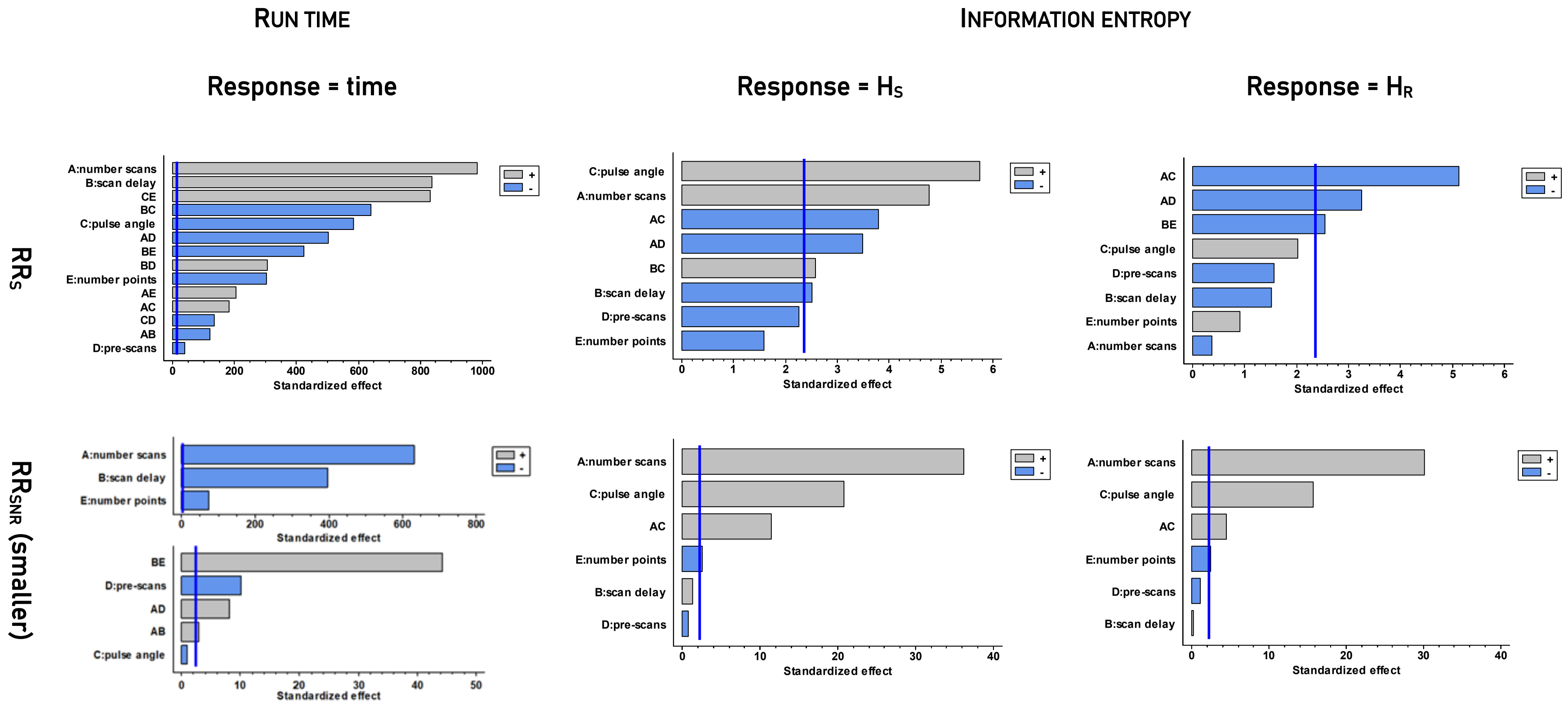
Rep 1 - Rep 4 = the four replicates from the external array; RSD = relative standard deviation

Figure 4.S5. Pareto charts obtained from the analysis of the responses individually (run time and information quality) of the design for the optimisation of the ^{1}H LF-NMR signal acquisition. Results are shown both by calculating robust responses RR_S and RR_{SNR} (see manuscript text for more information). The blue vertical line indicates the significance threshold (P-value < 0.05) on the responses).



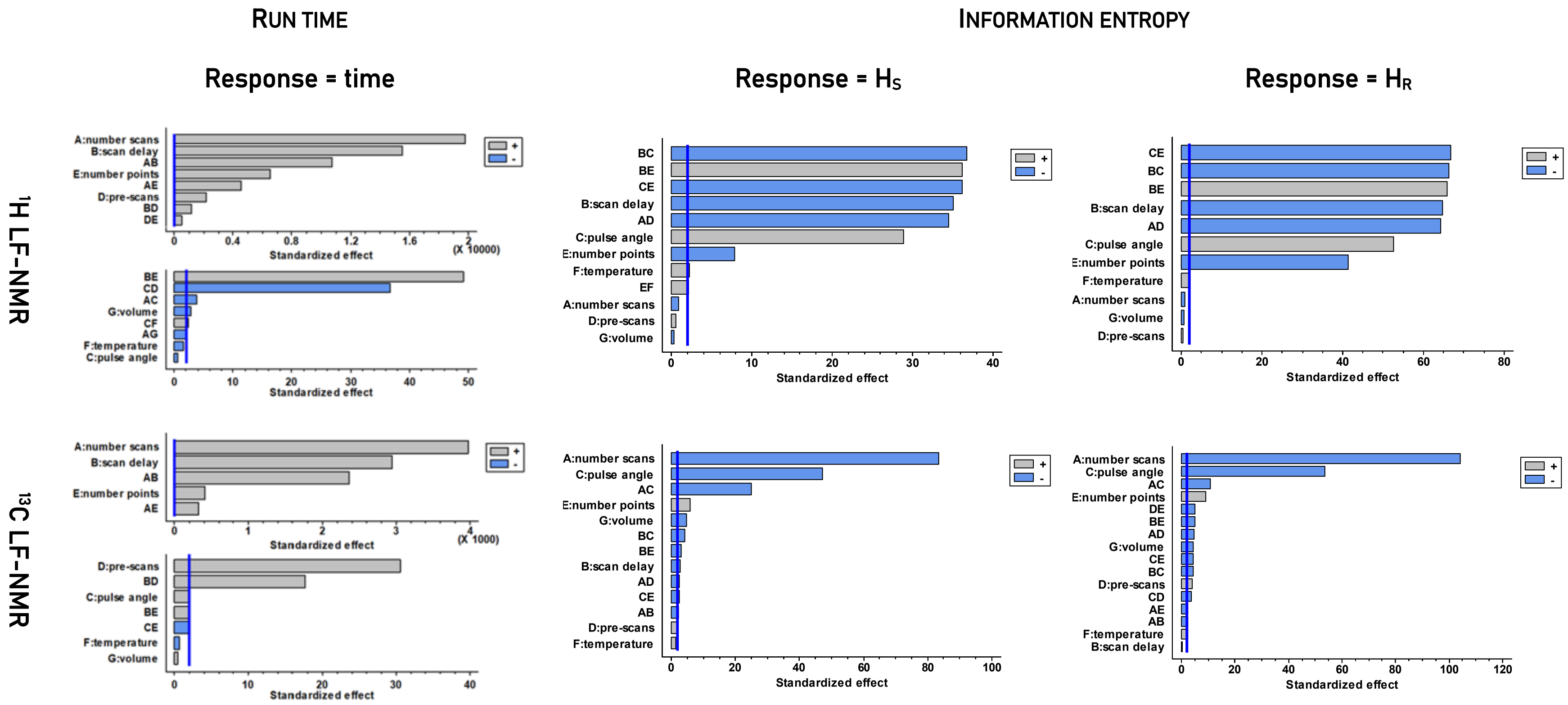
H_S = Shannon entropy; H_R = Rényi entropy; RR = robust response

Figure 4.S6. Pareto charts obtained from the analysis of each individual response (run time and information quality) of the design for the optimisation of the ^{13}C LF-NMR signal acquisition. Results are shown both by calculating RR_S and RR_{SNR} (see manuscript text for more information). The blue vertical line indicates the significance threshold (P-value < 0.05) on the response.



H_S = Shannon entropy; H_R = Rényi entropy; RR = robust response.
Note that the figure on the lower left (corresponding to the study of the time response by RRSNR) has been divided into two according to the magnitude of the standardised effect of the factors or interactions on the robust response. This has been done to improve the visualisation of those significant factors. The significance threshold (P-value < 0.05) is indicated by the vertical blue line.

Figure 4.S7. Pareto charts obtained from the analysis of the responses individually (run time and information quality) of the crossed array, i.e. taking into account the difficult-to-control factors as control factors, for the optimisation of the ^{1}H (first row) and ^{13}C LF-NMR (second row) signals acquisition. The blue vertical line indicates the significance threshold (P-value < 0.05) on the response.



H_S = Shannon entropy; H_R = Rényi entropy.

Note that the figures on the left (corresponding to the study of the time response) have been divided into two according to the magnitude of the standardised effect of the factors or interactions on the response. This has been done to improve the visualisation of those significant factors. The significance threshold (P-value < 0.05) is indicated by the vertical blue line.

4.5. Contribuciones a congresos

1. A. Arroyo Cerezo, A.M. Jiménez Carvelo, M. Medina García, L. Cuadros Rodríguez. **Study of the correction of spectral data obtained by SORS using chemometric blind signal separation (BSS).** [Oral 15']. *XI Colloquium Chemometricum Mediterraneum (CCM 2023). Padua (Italia), junio 2023.*
2. A. Arroyo Cerezo, Á. Fernández Crespo, E.A. Roca Nasser, A.M. Jiménez Carvelo, L. Cuadros Rodríguez. **Optimización de parámetros de adquisición de espectros LF-NMR altamente informativos para su utilización como huellas instrumentales no específicas en el desarrollo de métodos analíticos multivariable.** [Oral 15']. *XXIV Reunión de la Sociedad Española de Química Analítica (SEQA 2024). Zaragoza (España), julio 2024.*

CAPÍTULO 5

Discusión integrada

5.1. Métodos analíticos ‘verdes’ para el análisis de alimentos

El desarrollo sostenible es una prioridad global actualmente, como refleja la Agenda 2030 para el desarrollo sostenible adoptada por casi 200 países en el año 2015. En línea con estos objetivos, existe una eminente necesidad de desarrollar métodos analíticos sostenibles y respetuosos con el medio ambiente para el control de la calidad y autenticidad de alimentos que puedan ser transferidos e implementados en la industria alimentaria y laboratorios de control de la conformidad, tanto oficiales como privados.

En consonancia con lo expuesto, esta tesis doctoral se ha centrado en la aplicación de la metodología de huellas instrumentales como estrategia clave en el desarrollo de nuevos métodos analíticos sostenibles, destinados a la evaluación de la calidad y autenticidad de alimentos.

En definitiva, tal y como se ha manifestado a lo largo de los capítulos anteriores y con especial énfasis en el **Capítulo 1**, la presente tesis doctoral se ha basado en el enfoque verde de la química analítica alimentaria. Uno de los 12 principios de la química analítica verde (GAC) es el desarrollo de métodos automatizados y miniaturizados. Por ello, la instrumentación empleada para ejecutar los estudios de investigación se caracteriza por ser reducida en tamaño con respecto a los equipos instrumentales usados de forma convencional, como reflejan el **Capítulo 2** y el **Capítulo 3**. Asimismo, se ha priorizado el uso de técnicas analíticas no destructivas y/o no invasivas, con el objetivo de reducir o incluso eliminar el uso de disolventes y reactivos químicos.

Al abordar la calidad y autenticidad de alimentos, destaca el reto de detectar adulteraciones y/o falsificaciones. Uno de los alimentos con mayor incidencia en casos de fraude es el aceite de oliva. Además, su categoría de mayor calidad, el aceite de oliva virgen extra, es un producto que se encuentra sometido a una rigurosa normativa que requiere la aplicación de diversos métodos analíticos para la evaluación de la conformidad del etiquetado. Por esta razón, el aceite de oliva ha sido el protagonista del **Capítulo 2** de la presente tesis doctoral.

El desarrollo de métodos analíticos con un enfoque no dirigido lleva implícito el uso indispensable de la inteligencia artificial, fundamento de la minería de datos y el aprendizaje automático. La quimiometría ya ha proporcionado las herramientas necesarias para solventar problemas surgidos durante el desarrollo de la tesis y que han dado forma al **Capítulo 4**. Y ello sin dejar atrás su papel fundamental en el tándem formado junto a la espectrometría para realizar los estudios recogidos en los capítulos anteriores a este.

El objetivo final de estas tesis es desarrollar métodos analíticos multivariable basados en técnicas rápidas y poco invasivas, que puedan servir como soporte en forma de métodos de cribado para los convencionalmente utilizados, o incluso puedan llegar a sustituirlos. Con este fin, se han explorado distintas técnicas analíticas espectrométricas y se han empleado una variedad de herramientas quimiométricas. A continuación, se presenta una discusión integrada, dividida en tres aspectos clave, considerados como los más relevantes de la investigación llevada a cabo.

5.2. Miniaturización instrumental

A lo largo de los capítulos anteriores que constituyen la tesis, se ha destacado la existencia de los 12 principios de la GAC, siendo uno de ellos la miniaturización. Asimismo, se prioriza el uso de una técnica analítica directa y la realización de medidas *in situ*. El uso de equipos instrumentales miniaturizados, que a veces implican su portabilidad y por tanto la posibilidad de realizar medidas *in situ*, a su vez proporciona una serie de ventajas que de forma indirecta permiten cumplir con la mayoría del resto de principios de la GAC. Esto implica un volumen mínimo requerido de muestra, reducción del uso de energía, reducción o incluso supresión de la generación de residuos analíticos, o el aumento de la seguridad del analista [**1**].

Cabe destacar que el término miniaturización en química analítica implica una reducción de escala del proceso analítico completo, es decir, no solo implica a la parte instrumental sino también el resto de las etapas: preparación de muestras, separación, detección, etc. [**2**]. Sin embargo, en este apartado se contempla en exclusividad la miniaturización instrumental aplicada a la química analítica alimentaria (por tanto, la etapa de detección), dado que ésta ya implica la reducción en las otras etapas mencionadas del proceso analítico.

La miniaturización instrumental es relativamente reciente, y aun requiere de un mayor desarrollo e investigación para su aplicación como revela la literatura científica reciente [**3**,**4**,**5**]. A pesar de ello, en la última década el aumento en el mercado de equipos miniaturizados es notable, así como su aplicación en

[1] Gałuszka, A.; Migaszewski, Z.; Namieśnik, J. The 12 principles of green analytical chemistry and the SIGNIFICANCE mnemonic of green analytical practices. *Trends Anal. Chem.* **2013**, *50*, 78-84. DOI: 10.1016/j.trac.2013.04.010.

[2] Agrawal, A.; Keçili, R.; Ghorbani-Bidkorbeh, F.; Hussain, C.M. Green miniaturized technologies in analytical and bioanalytical chemistry. *Trends Anal. Chem.* **2021**, *143*, 116383. DOI: 10.1016/j.trac.2021.116383.

[3] Rodríguez-Saona, L.; et al. Miniaturization of optical sensors and their potential for high-throughput screening of foods. *Curr. Opin. Food Sci.* **2020**, *31*, 136-150. DOI: 10.1016/j.cofs.2020.04.008.

[4] Ünlüer, Ö.B.; Ghorbani-Bidkorbeh, F.; Keçili, R.; Hussain, C.M. Future of the modern age of analytical chemistry: Nanominiaturization. In *Handbook on Miniaturization in Analytical Chemistry*, Hussain, C.M., Ed.; Elsevier: Amsterdam, Netherlands, 2020; pp. 277-296.

[5] Mejía-Carmona, K.; Maciel; E.V.S., Lanças, F.M. Miniaturized liquid chromatography applied to the analysis of residues and contaminants in food: A review. *Electrophoresis* **2020**, *41*, 1680-1693. DOI: 10.1002/elps.202000019.

investigación. Las técnicas espectrométricas son las más susceptibles de ser miniaturizadas a nivel instrumental. Se puede distinguir entre equipos de sobremesa (*benchtop*), de mano (*handheld*) y portátiles (*portable*).

En esta línea, en el **Capítulo 2** se ha descrito el uso de un equipo de mano y portátil, para la adquisición de señales espectrométricas Raman (véase la Figura 5.1). Este equipo conlleva otras ventajas reseñables en armonía con la GAC, ya que se fundamenta en una variante de la espectrometría Raman. Se trata de la espectrometría Raman con compensación espacial (SORS) que, por sus especificaciones técnicas, permite llevar a cabo medidas no invasivas a través de capas superficiales del material bajo estudio. El desarrollo de esta variante condujo a la existencia en el mercado de equipos portátiles, cumpliendo así con prácticamente todos los principios del enfoque verde de la química analítica. El uso de estos equipos supone que el método analítico a desarrollar sea: no destructivo y no invasivo, ausente del uso de disolventes y demás agentes químicos, seguro para el analista y el medioambiente, y eficiente en términos energéticos y económicos.

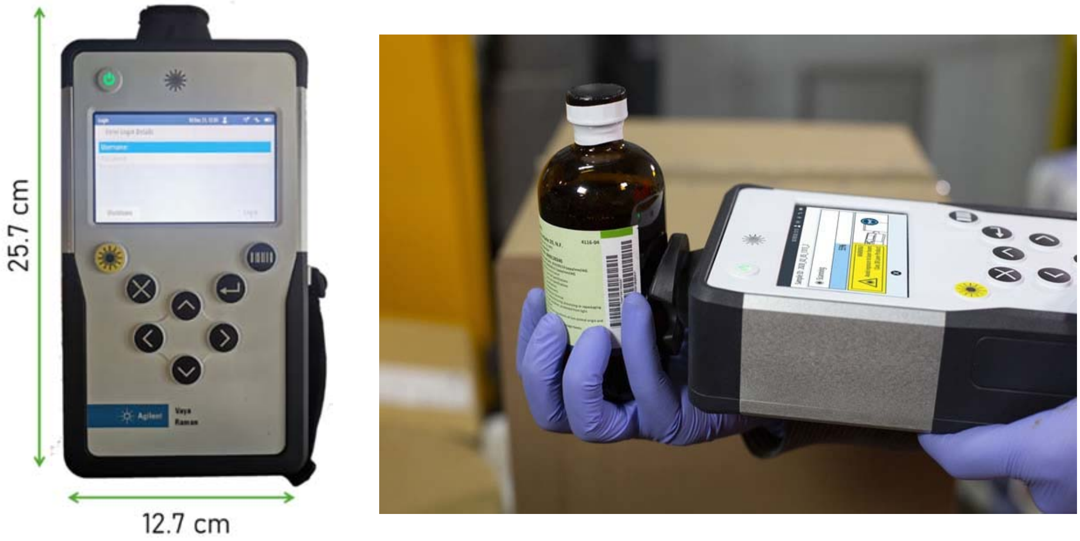


Figura 5.1. Espectrómetro portátil basado en la técnica SORS que ha sido empleado en los estudios presentados en forma de publicaciones científicas en el **Capítulo 2**.

El uso de esta técnica para el análisis de alimentos era, hasta la fecha, escaso. La señal Raman adquirida proporciona una gran cantidad de información química del material medido, tratándose de una compleja huella instrumental que debe ser tratada y analizada siguiendo un enfoque no dirigido. La Figura 5.2. ilustra a modo de ejemplo la señal Raman de dos productos alimenticios, margarina y queso en lonchas, que fueron adquiridas de forma no invasiva a través del envase original.

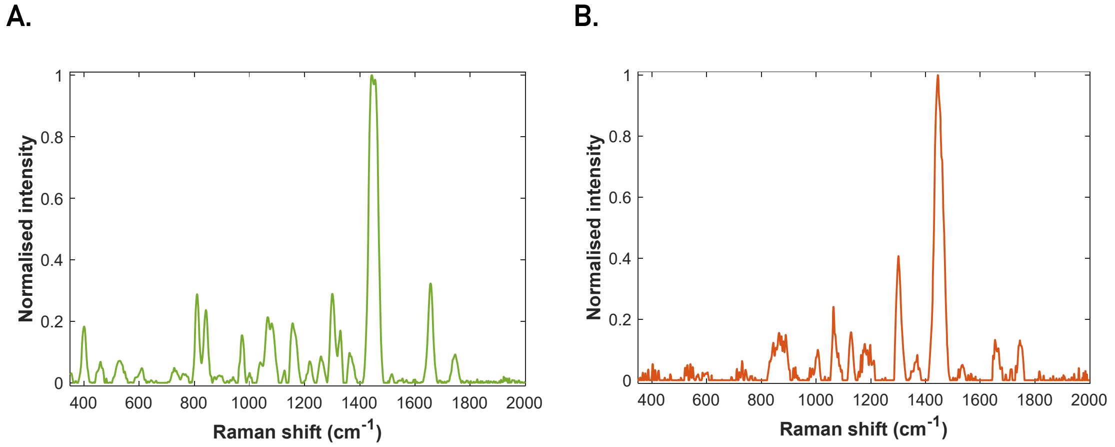


Figura 5.2. Señal Raman de una muestra de margarina (A) y de queso en lonchas (B) adquiridas de forma no invasiva a través de su envase original con SORS.

Ambas señales son parte de los dos estudios presentados en forma de publicación científica en el citado capítulo, siendo la matriz de ensayo margarina y queso en lonchas, respectivamente en los artículos científicos 1 y 2. La aplicación del tándem SORS-quimiometría ha demostrado tener un gran potencial para la autentificación de alimentos de forma rápida y no invasiva.

La elección de estos dos productos alimenticios como objeto de estudio estuvo motivada por su idoneidad para el uso de la técnica SORS, ya que ambos se comercializan contenidos en un envase de plástico, incoloro o coloreado, un material permeable al láser empleado para la adquisición de los espectros.

La margarina es un producto que se encuentra regulado en Europa actualmente por la normativa 1308/2013 [6]. La denominación de las diferentes categorías que pueden encontrarse en los lineales de los supermercados viene definida por la cantidad de grasa y por el origen de los ingredientes utilizados. Sin embargo, en otros países esto puede diferir, incluso dentro del mismo continente pueden encontrarse países en los que se emplean ingredientes diferentes.

En este primer estudio se investigó la conformidad del etiquetado de un conjunto de 62 muestras de materia grasa para untar de diferentes orígenes geográficos y con diverso porcentaje de grasa mediante la aplicación del tándem SORS-quimiometría. Se desarrollaron modelos de aprendizaje automático para evaluar la conformidad del etiquetado atendiendo a tres aspectos principales:

[6] Regulation (EU) No 1308/2013 of the European Parliament and of the Council of 17 December 2013 establishing a common organization of the markets in agricultural products, Official Journal of the European Union, 2013 (consolidated version on 2017).

(i) Clasificación de las muestras según su origen geográfico

(ii) Clasificación de las muestras según ingredientes de interés

(iii) Cuantificación de la cantidad de grasa presente en las muestras

La etapa de tratamiento y análisis de los datos se discutirá en profundidad en el apartado 5.4. del presente capítulo.

Cabe destacar que se llevó a cabo el estudio de forma complementaria con un espectrómetro Raman convencional, en el marco de una colaboración internacional, con el fin de comparar los resultados. A la hora de evaluar el rendimiento analítico proporcionado por un equipo instrumental miniaturizado, es esencial realizar a la vez esta comparación con el equipo convencional ya que en muchas ocasiones la miniaturización conlleva una disminución de la calidad de la señal adquirida. Es por ello que es necesario examinar la bondad y adecuación de la señal adquirida, y de los resultados, al objetivo perseguido.

Esta comprobación dio lugar a resultados incluso mejores cuando se desarrollaron modelos quimiométricos con los datos adquiridos por SORS respecto a los adquiridos con el espectrómetro Raman convencional. Este hecho parece estar motivado por la contribución de la auto-fluorescencia a los espectros Raman, uno de los principales desafíos cuando se trabaja con esta técnica. Sin embargo, la modalidad SORS está diseñada en gran parte para abordar este desafío, siendo capaz de suprimir dicha contribución de la fluorescencia [**7**].

Por otro lado, la segunda publicación aborda el estudio de conformidad del etiquetado de quesos en lonchas fabricados a partir de leche de diferente origen animal: queso, vaca y oveja. Aquellos quesos elaborados con leche de cabra o de oveja habitualmente suponen un coste superior que los de vaca y los de mezcla de leches de diferentes orígenes, un hecho que hace susceptible de fraude a los quesos con mayor valor económico en el mercado.

En la actualidad, la autentificación de este producto se lleva a cabo mediante el análisis de la caseína, realizando su caracterización por electroforesis, en modo isoelectroenfoque [**8**]. El principal inconveniente de este análisis es la complejidad del procedimiento a seguir. Esto pone de manifiesto la necesidad de tener al alcance un método analítico más rápido, sencillo y económico de

[7] Matousek, P.; Parker, A.W. Non-invasive probing of pharmaceutical capsules using transmission Raman spectroscopy. *J. Raman Spectrosc.* **2007**, *38*, 563-567. DOI: 10.1002/jrs.1688.

[8] Regulation (EU) No 2018/150 of 30 January 2018 amending Implementing Regulation (EU) 2016/1240 as regards methods for the analysis and quality evaluation of milk and milk products eligible for public intervention and aid for private storage, Official Journal of the European Union, 2018.

aplicar para evaluar la conformidad del etiquetado y la ausencia de fraude en este producto.

El estudio llevado a cabo para evaluar el potencial del espectrómetro portátil SORS, y en combinación con quimiometría, tuvo como objetivo el desarrollo de un método analítico de cribado para la autentificación del origen animal de muestras de queso en lonchas. Para ello se aplicó un enfoque no dirigido a los datos espectrales de un total de 80 muestras. Se siguió la estrategia de *one input-class* (1iC) para el desarrollo de modelos de clasificación, tras realizar un estudio de similitud para evaluar previamente la existencia de diferencias significativas entre las señales de las muestras con diferente origen animal. De forma complementaria, se desarrollaron modelos de cuantificación del contenido graso y proteico de las muestras de queso incluidas en el estudio. En el apartado 5.4 se discutirán en profundidad los resultados obtenidos tras aplicar diversas herramientas quimiométricas.

En términos generales, ambos estudios arrojaron unos resultados que confirman el potencial de la técnica SORS en combinación con la quimiometría para el desarrollo de métodos analíticos de cribado para la autentificación de productos lácteos. Destaca la gran ventaja que proporciona el uso de un equipo portátil, ya que una vez desarrollado el método podría aplicarse para realizar análisis de producto terminado *in situ* directamente en los lineales del supermercado.

El tercer estudio que forma parte de la tesis doctoral y se encuentra recogido en forma de publicación científica en el **Capítulo 3** también hace uso de instrumentación miniaturizada. En esta ocasión, se emplearon dos espectrómetros portátiles basados en la técnica espectrométrica de infrarrojo cercano (NIR). Se trata de una de las espectrometrías más aplicadas en el análisis de alimentos a nivel de investigación, y a su vez de la que mayor variedad comercial existe de equipos miniaturizados [**3**,**9**].

Como parte de una estancia de investigación realizada en la Universidad de Padua (Italia), se exploró el potencial de estos dos equipos mostrados en la Figura 5.3 para el desarrollo de un método analítico multivariable de cribado para la evaluación de la calidad de aceites de oliva.

Las 195 muestras de aceites vegetales utilizadas para este estudio fueron primeramente analizadas por un laboratorio oficial, adscrito al Gobierno de Italia, siguiendo los métodos reconocidos por el Consejo Oleícola Internacional (COI). Estos resultados fueron empleados como valores de referencia en el desarrollo de modelos de aprendizaje automático para la caracterización de los

[9] Grabska, J.; Beć, K.B.; Huck, C.W. Portability of miniaturized food analytical systems 4.0. In *Food Industry 4.0*; Hassoun, A., Ed.; Academic Press: Massachusetts, USA, 2024; pp. 189-231.

aceites a partir de los datos espectrales NIR. A su vez, se adquirieron los espectros NIR con un espectrómetro convencional con el objetivo de comparar los resultados, al igual que en el primer estudio elaborado con la técnica SORS. En la Figura 5.4 se muestra un ejemplo de las tres señales espectrales adquiridas con los diferentes equipos.

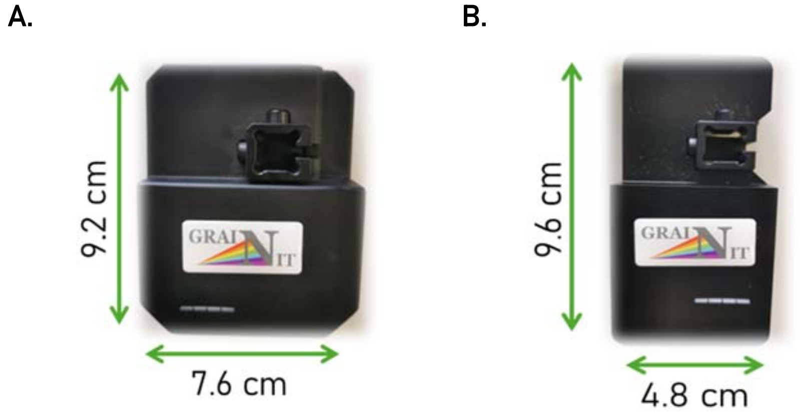


Figura 5.3. Espectrómetros NIR portátiles usados en el estudio presentado en forma de publicación científica en el **Capítulo 3**.

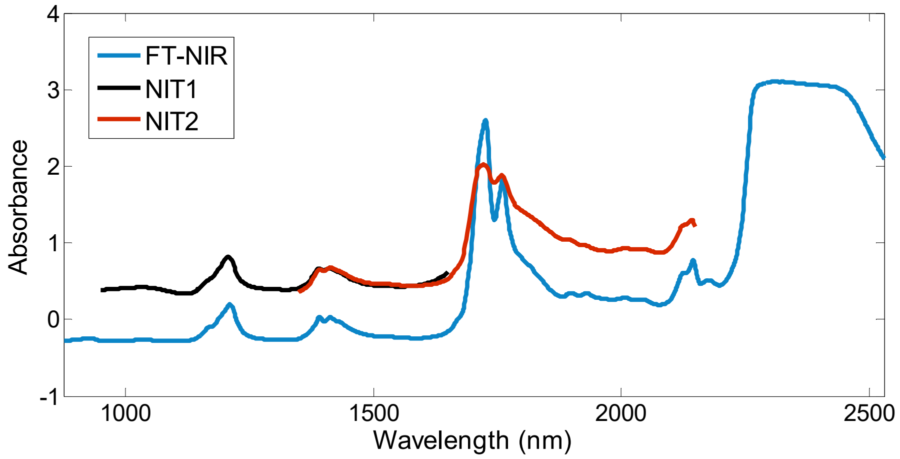


Figura 5.4. Espectros NIR adquiridos con los dos equipos portátiles miniaturizados (NIT1 y NIT2) y un equipo convencional de sobremesa (FT-NIR).

Conforme a lo esperado, se obtuvieron mejores resultados con los datos adquiridos con el espectrómetro NIR convencional de sobremesa. No obstante, la calidad de los modelos desarrollados con datos de los equipos miniaturizados arrojó resultados satisfactorios y prometedores. Las ventajas que supondrían el uso de un instrumento portátil como el empleado para la industria o el laboratorio que desee implementarlo son destacables. La aplicación de este método de cribado desarrollado podría suponer un ahorro significativo tanto en lo económico como en tiempos de análisis, aliviando así la carga de trabajo del laboratorio correspondiente.

En esta misma línea, y como parte de un proyecto de colaboración público-privada (CPP2021-008672), se enmarca el estudio presentado a continuación del anterior en el **Capítulo 3**, que pretende ser materializado en forma de publicación científica una vez llegue a término. La miniaturización instrumental también es protagonista en este estudio. Se trata de la técnica analítica de resonancia magnética nuclear de bajo campo (LF-NMR) aplicada mediante el uso de un equipo de sobremesa (*benchtop*) como el mostrado en la Figura 5.5, que ofrece amplias ventajas con respecto a su formato convencional en términos de requerimientos de espacio y mantenimiento, así como en lo económico.

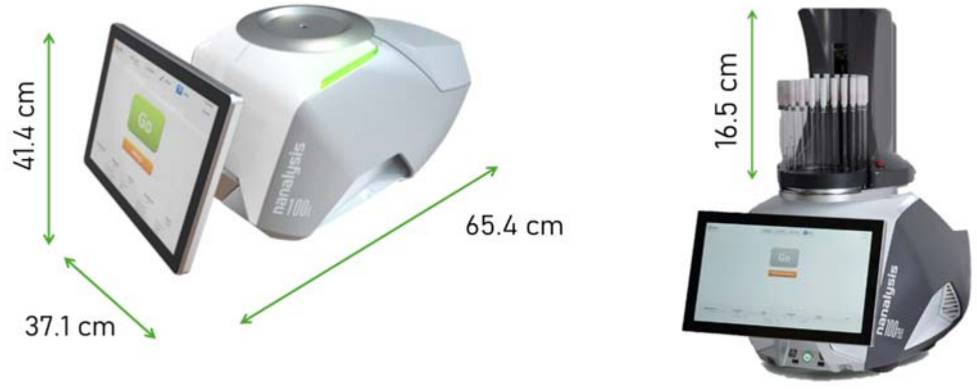


Figura 5.5. Espectrómetro LF-NMR de sobremesa, mostrando el automuestreador (derecha), usado en el estudio 1 presentado en el **Capítulo 3**.

Los resultados preliminares obtenidos a partir de las señales 1H y 13C LF-NMR de más de 200 muestras de aceites vegetales evidencian el potencial de la técnica para el desarrollo de métodos analíticos de cribado en lo que respecta al análisis de la calidad de aceites de oliva. Esta aplicación continuará siendo objeto de estudio tras finalizar la tesis doctoral.

Los cuatro estudios aludidos ponen de manifiesto el potencial uso de equipos miniaturizados para el desarrollo de métodos analíticos de cribado en el control de la calidad y autenticidad de alimentos, cumpliendo así con los principios de la GAC. No cabe duda de que es necesario seguir avanzando y ampliar el trabajo realizado para que estos métodos puedan ser transferibles e implementados en la industria, como se persigue con el citado proyecto que se encuentra en pleno desarrollo. Aún así, los resultados detallados demuestran lo prometedor de las técnicas analíticas empleadas, que auguran ser unas candidatas óptimas para la futura complementación y/o sustitución de los métodos analíticos actualmente reconocidos como oficiales en el control alimentario.

5.3. Aceite de oliva como matriz de ensayo

El aceite de oliva es uno de los alimentos con mayor relevancia en España, como uno de los principales productores y consumidores a nivel mundial y por su relevancia cultural, económica y gastronómica. Es por ello que asegurar la calidad de este producto es crucial, con especial interés en las categorías de mayor calidad: virgen y virgen extra.

El análisis de este producto es el protagonista del **Capítulo 3** de la presente tesis, debido a la importancia justificada en el párrafo anterior, y sumado a la dilatada experiencia en el análisis de esta matriz alimentaria del grupo de investigación en el que se ha desarrollado la tesis.

La normativa vigente en la actualidad que regula la calidad del aceite de oliva detalla los parámetros de pureza y calidad que deben ser analizados en el control oficial, y para ello aplica los métodos analíticos reconocidos por el COI. Tal y como se detalla en la presentación del citado capitulo, ello implica el uso de diversas técnicas analíticas y otras determinaciones llevadas a cabo mediante métodos clásicos no instrumentales [**10,11**]. Realizar estos análisis que quedan lejos de seguir los principios de la GAC, supone además largos tiempos, que conlleva a una alta carga de trabajo en laboratorios de control rutinario y de control oficial de aceites de oliva.

La tendencia actual hacia el desarrollo de métodos analíticos 'verdes' orientados a evaluar la calidad de los alimentos es un hecho destacado en el transcurso de este trabajo. En el capítulo de libro recogido en el **Capítulo 1** se pone de manifiesto la existencia de diversas estrategias para evaluar la sostenibilidad de estos nuevos métodos. De esta forma es posible estimar el grado de cumplimiento con los principios de la GAC, y comparar la mejora con respecto a los métodos convencionales para el mismo fin, o bien entre diferentes nuevos métodos que persigan ser más respetuosos con el medio ambiente.

Estas estrategias se basan en la aplicación de diversos algoritmos que dan como resultado valores numéricos para examinar la sostenibilidad de los

10. Commission Delegated Regulation (EU) 2022/2104 of 29 July 2022 supplementing Regulation (EU) No 1308/2013 of the European Parliament and of the Council as regards marketing standards for olive oil, and repealing Commission Regulation (EEC) No 2568/91 and Commission Implementing Regulation (EU) No 29/2012. Official Journal of the European Union L 284/1, 2022.

11. Commission Implementing Regulation (EU) 2022/2105 of 29 July 2022 laying down rules on conformity checks of marketing standards for olive oils and methods of analysis of the characteristics of olive oil. Official Journal of the European Union L 284/23, 2022.

nuevos métodos o procedimientos analíticos desarrollados. Sin embargo, hasta la fecha se basan en una evaluación *ex-post*, es decir, evalúan la sostenibilidad del método una vez desarrollado. Por el contrario, la posibilidad de realizar una evaluación *ex-ante* para ejecutar la toma de decisiones de forma previa al desarrollo del método ofrecería grandes ventajas.

Además, ninguna de las propuestas estaba dirigida a la evaluación de métodos no destructivos y/o no invasivos para el análisis de la calidad de los alimentos. De estas carencias surge la propuesta recogida en el apartado 3.2 de la presente tesis. Se centra en la evaluación de la 'blancura' de un método analítico (*analytical method whiteness*) espectrométrico no destructivo antes de ser desarrollado. Un método analítico blanco sería aquel que cumple con todos los requisitos de tres aspectos a evaluar representados por tres colores: (i) rojo para el rendimiento analítico, (ii) verde para la sostenibilidad del método y (iii) azul para la productividad y eficiencia práctica. Esto forma parte de una metodología ya desarrollada por otros autores y conocida como RGB, por los tres colores que la caracterizan [**12**].

Esta propuesta se publicó acompañada de una plantilla para llevar a cabo la evaluación *ex-ante*, que a su vez fue testada para la comparación de tres métodos analíticos con el mismo objetivo: el control de la calidad y/o autenticidad de aceites de oliva. La diferencia entre los tres métodos radicaba en la técnica empleada: espectrometrías NIR, Raman (modalidad SORS) y LF-NMR, mediante el uso de los instrumentos miniaturizados que se han empleado a lo largo de la investigación de esta tesis.

La evaluación de la 'blancura' del método va más allá de los principios de la GAC, que incide principalmente en el 'verde', ya que tiene en cuenta otros aspectos de gran importancia en el desarrollo de nuevos procedimientos analíticos como es el rendimiento y la productividad. A pesar de que el rendimiento analítico no pueda ser evaluado antes del desarrollo del método, se sugiere establecer los requisitos mínimos según el objetivo perseguido, y en caso de resultar satisfactorio en los otros dos aspectos (sostenibilidad y productividad), evaluar de nuevo el método de forma *ex-post* para verificar que se alcanzan los requisitos mínimos de rendimiento formulados por el analista. Los resultados de la ejemplificación del uso de esta propuesta para evaluar la 'blancura' de los tres métodos analíticos mostraron la suficiente objetividad para realizar dicha evaluación. Todos obtuvieron una puntuación superior al 80%.

Las dos publicaciones científicas que siguen a la anterior en el **Capítulo 3** dan respuesta a la necesidad de tener a disposición un método analítico de cribado

[12] Nowak, P.M.; Wietecha-Posłuszny, R.; Pawliszyn, J. White Analytical Chemistry: An approach to reconcile the principles of Green Analytical Chemistry and functionality. *Trends Anal. Chem.* **2021**, *138*, 116223. DOI: 10.1016/j.trac.2021.116223.

que pueda de alguna manera agilizar la carga de trabajo de los laboratorios dedicados al control de la calidad y autenticidad del aceite de oliva. En ambos se plantea el uso de dos de las tres técnicas espectrométricas evaluadas en la publicación anterior, y en combinación con herramientas quimiométricas para desarrollar un método analítico rápido, poco invasivo y siguiendo la metodología de huellas instrumentales.

Por un lado, la espectrometría NIR ha sido ampliamente puesta en valor por su potencial para el análisis de aceites de oliva en múltiples publicaciones científicas [13,14]. La principal novedad aportada con el primer estudio del **Capítulo 3** reside en el uso de instrumentos miniaturizados portátiles y de bajo coste. Los modelos cuantitativos desarrollados a partir de los datos espectrales permitieron detectar tres muestras de aceite de oliva declaradas como categoría virgen extra, que sin embargo no cumplían los límites de algunos de los parámetros para ser etiquetados como tal. La Figura 5.7 ilustra un ejemplo de ello, donde se observa la detección de estos tres aceites como anómalos mediante un análisis exploratorio de los datos adquiridos con ambos instrumentos. Asimismo, fue posible predecir aquellos parámetros que se encontraban fuera de rango para que estas tres muestras puedan ser consideradas como virgen extra, mediante los modelos de cuantificación desarrollados.

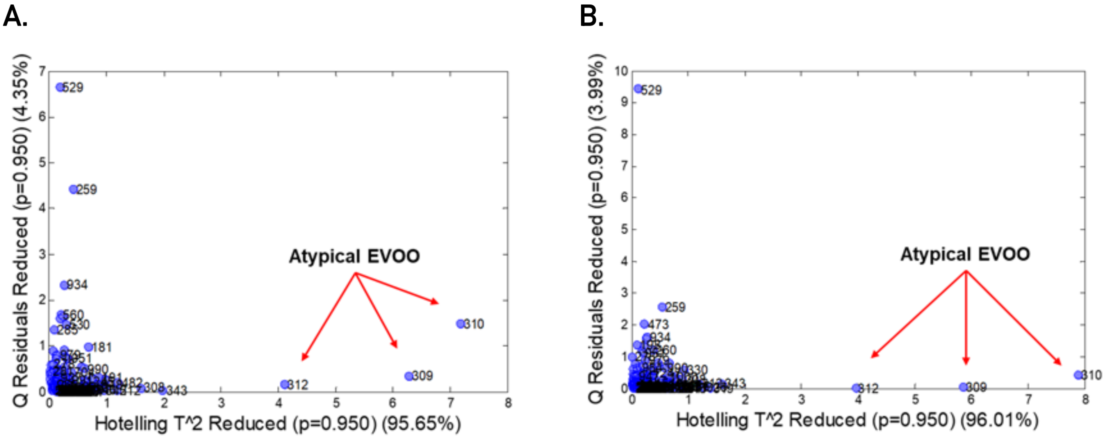


Figura 5.7. Ejemplo de la detección de tres muestras anómalas etiquetadas como aceite de oliva virgen extra a través de un análisis exploratorio realizado con los datos espectrales NIR adquiridos con los dos equipos mostrados en la Figura 5.3.

[13] Ordoudi, S.A.; Strani, L.; Cocchi, M. Toward the non-targeted detection of adulterated virgin olive oil with edible oils via FTIR spectroscopy & chemometrics: research methodology trends, gaps and future perspectives. *Molecules* **2023**, *28*, 337. DOI: 10.3390/molecules28010337.

[14] Cappelli, A.; et al. Applying spectroscopies, imaging analyses, and other non-destructive techniques to olives and extra virgin olive oil: a systematic review of current knowledge and future applications. *Agriculture* **2024**, *14*, 1160. DOI: 10.3390/agriculture14071160.

Cabe destacar que, de los cuatro criterios de calidad considerados en legislación, la acidez no se encontraba fuera de rango en estas tres muestras. Por otro lado, en el índice de peróxidos sí resultaron estar fuera de rango, pero también lo estuvieron muestras correctamente etiquetadas como virgen extra. Este criterio es un indicador clave del estado de oxidación, y por lo tanto está íntimamente relacionado con la calidad del aceite de oliva. Sin embargo, es muy inestable en el tiempo, y el análisis debe llevarse a cabo de forma rigurosa. Los resultados de este estudio y de acuerdo con la literatura científica **[15]** demuestran que con la huella instrumental analizada con un enfoque no dirigido es posible llevar a cabo un adecuado control de la calidad de los aceites de oliva y la detección de posibles muestras anómalas mediante un rápido análisis y sin la necesidad de efectuar múltiples análisis con un enfoque dirigido a obtener el valor de un único parámetro que, sin embargo, puede no proveer de una visión integral sobre las características y el estado de la muestra.

Por tanto, podría decirse que la implementación en la industria oleícola de métodos analíticos de cribado basados en un enfoque no dirigido y en línea con los principios de la GAC es cada vez más tangible, aunque es necesario continuar el trabajo abordando las siguientes etapas, es decir, la transferencia y la implementación de los métodos desarrollados.

Por otro lado, la espectrometría LF-NMR es también una técnica de reciente implantación y que está mostrando resultados prometedores para este fin. En comparación con la espectrometría NIR, la LF-NMR está menos explorada en este ámbito ya que los espectrómetros de bajo campo son relativamente recientes en el mercado, mientras que los convencionales de alto campo suponen diversos inconvenientes para ser aplicados en la industria alimentaria. Estos aspectos se encuentran recogidos en forma de manuscrito en el **Capítulo 4**, concretamente en el apartado 4.4. Se trata del último artículo enviado para su publicación en una revista científica que forma parte de la presente tesis. Tiene como objetivo la optimización de las condiciones instrumentales para la adquisición de señales LF-NMR de aceites de oliva de alta calidad informativa. Este estudio surge al comienzo de la investigación que forma parte del proyecto de colaboración público-privada CPP2021-008672, donde se puso de manifiesto la necesidad de buscar los parámetros instrumentales óptimos para el uso del equipo, dada su novedad en el mercado y los objetivos perseguidos con su aplicación.

15. García Martín, J.F. Potential of near-infrared spectroscopy for the determination of olive oil quality. *Sensors*, **2020**, *22*, 2831. DOI: 10.3390/s22082831.

Tras obtener las condiciones instrumentales óptimas para la adquisición de señales LF-NMR en el menor tiempo posible de análisis y de la máxima calidad, se llevó a cabo el estudio presentado en el apartado 3.4 del **Capítulo 3**. Éste que bien podría considerarse como estudio preliminar, arrojó resultados que incitaban al optimismo, y que continuarán ampliándose para la consecución del objetivo final del proyecto: la implementación de la LF-NMR en laboratorios de control rutinario de aceites de oliva y productos de otros sectores industriales.

Esta técnica, respecto a la espectrometría NIR, supone una mayor inversión económica inicial. No obstante, la bibliografía científica demuestra que la espectrometría NMR proporciona mejores resultados en términos de rendimiento analítico [**16**]. Es deseable tener a disposición una variedad de alternativas de métodos analíticos con el mismo fin, que provean de diferentes opciones adaptables a las necesidades de la industria o laboratorio interesado en su implementación.

En definitiva, la implementación de un método analítico 'blanco' como herramienta de cribado para analizar la calidad y autenticidad del aceite de oliva en laboratorios de control rutinario es algo factible y cada vez más cercano a convertirse en una realidad.

5.4. Minería de datos y aprendizaje automático (quimiometría)

Los métodos de minería de datos y de aprendizaje automático aplicados a datos químicos, tradicionalmente denominados como métodos quimiométricos, conforman un conjunto de herramientas que se ha convertido en un objeto indispensable de la química analítica alimentaria actual, y en concreto para la ejecución de la presente tesis doctoral.

Todos los estudios realizados han requerido de alguna forma la aplicación de herramientas quimiométricas. Por un lado, en el **Capítulo 2** y el **Capítulo 3** se ha hecho uso de diversos métodos de análisis supervisado y no supervisado para el desarrollo de modelos de aprendizaje, al igual que en el estudio recogido en forma de publicación científica en el apartado 4.3 (**Capítulo 4**). Mientras que, por otro lado, el último estudio presentado en el **Capítulo 4** hace uso de herramientas de optimización de procesos. Además, este mismo capítulo recoge una propuesta en forma de capítulo de libro que examina el estado del arte y sienta las bases de una parte indispensable de la quimiometría, que es el análisis de la similitud, hasta ahora poco reconocida como independiente.

[16] McDowell, D.; et al. Low vs high field 1H NMR spectroscopy for the detection of adulteration of cold pressed rapeseed oil with refined oils. *LWT-Food Sci. Technol.* **2019**, *111*, 490-499. DOI: 10.1016/j.lwt.2019.05.065.

5.4.1. Aplicación de métodos no supervisados

Agrupamiento

Se trata de la estrategia más conocida de los métodos no supervisados o análisis exploratorio. Su aplicación como paso previo para llevar a cabo un modelado supervisado es fundamental. El método usado por excelencia es el análisis de componentes principales (PCA). En esta tesis se ha aplicado fundamentalmente para el establecimiento de hipótesis de partida de forma previa al modelado supervisado, así como para la detección de valores atípicos (*outliers*). Su uso permitió reconocer el agrupamiento natural de las muestras de margarina (apartado 2.2 del **Capítulo 2**) atendiendo al origen geográfico de fabricación, lo que llevó posteriormente al desarrollo de modelos supervisados.

Por otro lado, para el primer estudio del **Capítulo 3** se empleó PCA en la detección valores atípicos a partir de los datos de referencia proporcionados por el laboratorio. Este paso es necesario cuando se trabaja con datos de referencia para el desarrollo de modelos de predicción, para prevenir el sobreajuste del modelo, evitar distorsiones en la interpretación de los resultados y generar un modelo robusto. Este análisis dio como resultado la detección de tres muestras con valores atípicos que no fueron incluidas en el posterior desarrollo de los modelos de predicción. Esta detección se llevó a cabo evaluando los valores residuales Q (*Q-residuals*), que indican la discrepancia entre los datos observados (valores de referencia) y la variabilidad capturada por los componentes principales del PCA generado. Este análisis permitió también detectar tres muestras atípicas, en el sentido en que diferían significativamente del resto, como se observa en la Figura 5.7 ya comentada en el apartado anterior. Al contrario que los *outliers*, estas muestras fueron detectadas por altos valores del estadístico T^2-Hotelling, que pone de manifiesto la diferencia con respecto al promedio, es decir, muestra aquellos puntos que se alejan de lo esperado.

PCA también fue el punto de partida del último estudio recogido en el **Capítulo 3** para realizar un análisis exploratorio de los datos adquiridos mediante LF-NMR. A través de este análisis, se comprobó que estas señales analíticas encierran información química suficiente para separar entre muestras de aceites de oliva de otros aceites vegetales no oliva en base a su composición. No fue posible observar un agrupamiento natural en base a la categoría comercial de los aceites de oliva. Sin embargo, este propósito será perseguido mediante el empleo de métodos supervisados.

Resolución de mezclas

Los métodos de resolución están dirigidos a extraer las señales y componentes puros que conforman la señal analítica de una matriz compleja. Esta estrategia ha sido aplicada para mejorar el tratamiento de los datos posterior a la adquisición de medidas SORS, como queda recogido en el **Capítulo 4**. Dado que la señal adquirida es una mezcla de la contribución de las capas superficiales y subsuperficiales, es necesario llevar a cabo un tratamiento de la señal analítica para obtener la señal resuelta del material objeto de estudio. A pesar de que el software del equipo instrumental empleado incluía un proceso automatizado para realizar esta acción, en ocasiones el resultado no era el esperado.

Esto llevó a la ejecución del citado estudio, donde se hizo uso de dos métodos quimométricos: el método de resolución de curvas multivariable o *multivariate curve resolution* (MCR) y el análisis de componentes independientes o *independent component analysis* (ICA). Los resultados obtenidos mostraron una mejora considerable con respecto a la resolución llevada a cabo por el software integrado en el equipo empleado en todos los experimentos realizados, tal y como muestra la Figura 5.8 a modo de ejemplo.

No se vieron diferencias significativas en los resultados entre ambos métodos usados, ICA y MCR, lo que indica que ambos métodos son útiles para el objetivo perseguido. Estos modelos podrían ser implementados como parte del software para la resolución de señales SORS, proporcionando un avance significativo en el procesado tras la adquisición de estas señales espectrales de muestras complejas.

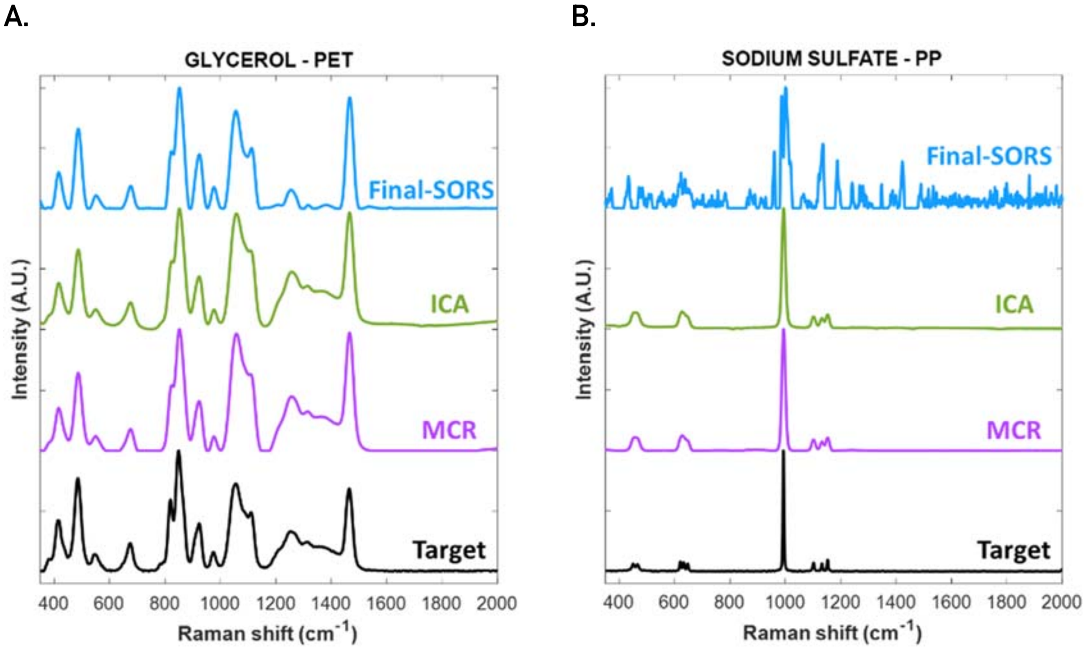


Figura 5.8. Resultados obtenidos tras aplicar dos métodos quimiométricos de resolución de señales: ICA y MCR, en comparación con la señal objetivo (target) y la señal resuelta por el propio software del equipo empleado (Final-SORS).

Análisis de la similitud

La comparación entre señales analíticas es una estrategia ampliamente utilizada en quimiometría, que hasta la fecha no parece haber recibido toda la atención y valor que merece en el ámbito quimiométrico.

Sin embargo, en la presente tesis se ha utilizado ampliamente esta estrategia, como ya venía siendo parte del grupo de investigación de pertenencia. Antes de surgir la posibilidad para sentar las bases sobre las que descansa el análisis de similitud de señales analíticas y ser materializadas en forma de capítulo de libro, ya en estudios anteriores se hizo uso del mismo.

Ejemplo de ello es el segundo estudio presentado en forma de publicación científica y que queda recogido en el apartado 2.3 del **Capítulo 2**. Se llevó a cabo un análisis de similitud de los espectros SORS de las muestras de queso en lonchas de diferentes orígenes animales, con el objetivo principal de explorar la idoneidad de estas señales para el objetivo perseguido: el desarrollo de un método analítico capaz de diferenciar el origen animal con fines de autentificación.

Para este fin se hizo uso de un índice de similitud previamente desarrollado por el grupo de investigación en el que se ejecutó la tesis. Recibe el nombre de *nearness index (NEAR)* y se fundamenta en el cálculo de la distancia euclídea normalizada entre dos vectores [**17**]. Se analizó la similitud global entre los espectros de muestras con el mismo origen animal, y entre las muestras con distinto origen animal. Los resultados se resumen en la Tabla 5.1, donde se resaltan aquellos valores que pertenecen a la comparación de muestras de queso fabricado con leche del mismo origen animal.

Tabla 5.1. Resultados de la comparación por pares de muestras de queso en lonchas de diferente origen animal mediante el cálculo del índice NEAR a partir de espectros SORS.

	Vaca	Cabra	Oveja
Vaca	**0.868**		
Cabra	0.445	**0.876**	
Oveja	0.291	0.743	**0.844**

17. Pérez-Robles, R.; Navas, N.; Medina-Rodríguez, S.; Cuadros-Rodríguez, L. Method for the comparison of complex matrix assisted laser desorption ionization-time of flight mass spectra. Stability of therapeutical monoclonal antibodies. *Chemom. Intell. Lab. Syst.* **2017**, *170*, 58-67. DOI: 10.1016/j.chemolab.2017.09.008.

Estos resultados demostraron que las señales SORS adquiridas encerraban la información relevante y necesaria para el objetivo perseguido, dando lugar así a la hipótesis de partida que llevó al siguiente paso de generación de un modelo de clasificación supervisado.

Por otra parte, los resultados del estudio para optimizar la resolución de señales SORS mediante métodos quimiométricos (ICA y MCR) se evaluaron a través de un análisis de similitud de las señales resultantes (véase apartado 4.3 del **Capítulo 4**). En este caso, además de hacer uso del índice NEAR, se empleó también el cálculo del coseno (índice COS), del ángulo (arco coseno) (índice ACOS) y del coeficiente de determinación (índice R^2) entre dos vectores a comparar, cada uno de ellos representativo de un espectro que contiene el mismo número de elementos o variables. Esta evaluación se convirtió en una forma sencilla y fácilmente interpretable de los resultados.

Además, el estudio se completó con un profundo análisis de la información proporcionada por cada uno de los cuatro índices empleados. Con ello, se observó que el índice COS y R^2 son los menos sensibles a pequeñas diferencias cuando se trata de comparar señales muy similares entre ellas. Mientras que el índice ACOS, seguido del índice NEAR resultaron ser más sensibles a pequeños cambios, y además muestran un comportamiento bastante lineal, a diferencia de los dos anteriores.

A partir de este estudio, fue cuando tuvo lugar el comienzo de la escritura del capítulo de libro 2 presentado en el **Capítulo 4** de esta tesis. Esto vino motivado por el interés suscitado a raíz de este análisis de los índices de similitud, junto a la escasa bibliografía encontrada en la que sustentar los resultados obtenidos.

Al mismo tiempo, esta aportación dio a conocer el uso poco conocido de la teoría de la información en el ámbito de la química analítica. Gracias a ello, se pudo hacer frente a un desafío encontrado durante el desarrollo del último estudio que forma parte del **Capítulo 4**, dirigido a optimizar las condiciones experimentales para la adquisición de señales LF-NMR. El reto hallado es el de evaluar la calidad informativa de una señal analítica de forma previa al desarrollo de un nuevo método.

Ello condujo a la formulación de una propuesta para evaluar la calidad de una señal analítica en términos de cantidad de información basada en el cálculo de la entropía de la información, formulada por Shannon **[18]**, y posteriormente generalizada por Rényi **[19]**. Este estudio ha supuesto una propuesta innovadora

[18] Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379-423. DOI: 10.1002/j.1538-7305.1948.tb01338.

[19] Rényi, A. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol 1: contributions to the theory of statistics vol 4, University of California Press, 1961 January, pp. 547-562.

para la resolución de este desafío, que será ampliada tras finalizar la tesis como parte de la investigación posdoctoral.

5.4.2. Aplicación de métodos supervisados

El uso de métodos supervisados para la investigación durante este doctorado ha sido imprescindible para la generación de modelos de aprendizaje automático, tanto para fines de clasificación como de cuantificación. Prueba de ello son los estudios recogidos a lo largo del **Capítulo 2** y **Capítulo 3**.

Aplicaciones cualitativas (métodos analíticos de clasificación)

Comenzando con el estudio que tiene por objetivo la autentificación no invasiva de margarinas mediante la técnica SORS, para la clasificación según el origen geográfico de fabricación se siguió la estrategia ilustrada en forma de diagrama en la Figura 5.9. En un primer paso, se trató de discriminar las muestras según su procedencia geográfica por continentes: Europa o Marruecos. Aquellas muestras clasificadas con origen Europa se utilizaron para construir un modelo que diferenciase entre muestras de España y no España, a continuación, las muestras procedentes de Francia y para terminar Reino Unido (véase **Capítulo 2**, apartado 2.2).

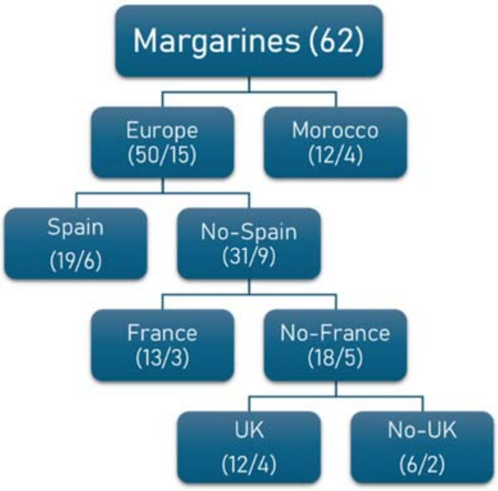


Figura 5.9. Diagrama de clasificación de las muestras de margarina según su origen geográfico. Entre paréntesis se indica el número total de muestras perteneciente a cada clase seguido del número de muestras incluidas para el conjunto de validación.

Los resultados de esta cadena de clasificaciones binarias, también denominada 'árbol de clasificación', dieron lugar a valores de sensibilidad y especificidad superiores a 0.8 en la mayoría de los casos cuando se estudiaron las métricas de calidad de los modelos generados. Además, destaca la calidad superior en los modelos generados a partir de los datos SORS con respecto a los datos adquiridos mediante la técnica Raman convencional, como ya se ha puesto de manifiesto anteriormente.

De igual forma, los modelos de clasificación desarrollados con los mismos datos según la presencia de ingredientes relevantes en su composición obtuvieron resultados satisfactorios, con especial mención a aquellos modelos para detectar la presencia de fitoesteroles o de aceite de oliva (valores de precisión superiores a 0.8 en todos los casos).

Además, para el desarrollo de estos modelos se compararon los resultados ofrecidos por tres métodos distintos de clasificación: (i) modelado flexible e independiente por analogía de clases (SIMCA), (ii) análisis discriminante mediante regresión parcial de mínimos cuadrados (PLS-DA), y (iii) sistemas de aprendizaje automático mediante vectores de soporte (SVM). En todos los casos, con SIMCA se obtuvieron las mejores métricas de calidad. Cabe destacar que se empleó la estrategia de una clase de entrada, conocida como *one class classification* (OC) o *one input-class* (1iC.) De acuerdo con la opinión de algunos investigadores, esta debería ser la estrategia de elección cuando el objetivo es la autentificación de productos alimenticios, donde la clase de entrada sea el alimento 'genuino' que tenga la característica especial que la haga vulnerable a posibles fraudes o adulteraciones [**20**].

Se siguió esta misma estrategia y método quimiométrico (1iC-SIMCA) en el siguiente estudio para la autentificación no invasiva de quesos en lonchas mediante la técnica SORS (véase **Capítulo 2**, apartado 2.3). Para ello, la clase objetivo estaba compuesta por los quesos fabricados con leche de vaca. Para la etapa de validación externa del modelo se emplearon 3 conjuntos de validación diferentes. Los resultados fueron satisfactorios, ya que se alcanzaron valores de precisión muy cercanos al 90%.

Aplicaciones cuantitativas (métodos analíticos de cuantificación)

Ambos estudios (margarinas y quesos) se ampliaron con el desarrollo de modelos de cuantificación de los macronutrientes de interés, concretamente del contenido graso en margarinas, y el contenido graso y proteico en quesos. El método quimiométrico empleado para ello fue PLS. La Figura 5.10 resume los resultados de predicción de los respectivos conjuntos de validación con los modelos desarrollados.

Como ya se ha comentado anteriormente, estos resultados demostraron el gran potencial de la técnica SORS para la autentificación rápida y no invasiva de productos lácteos como son la margarina y el queso en lonchas.

20. Rodionova, O.; Titova, A.V.; Pomerantsev, A.L. Discriminant analysis is an inappropriate method of authentication. *Trends Anal. Chem.* **2016**, *78*, 17-22. DOI: 10.1016/j.trac.2016.01.010.

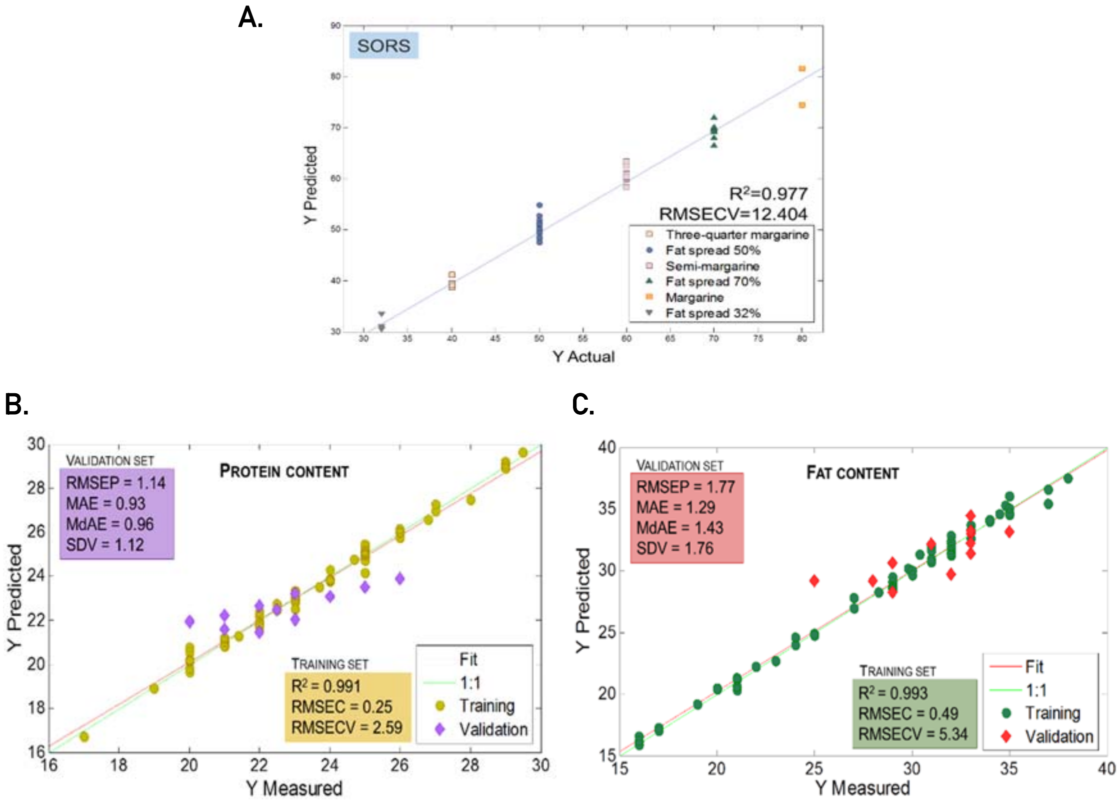


Figura 5.10. Resultados de los modelos de cuantificación del contenido graso en margarinas (A) y de contenido proteico (B) y graso (C) en quesos en lonchas generados a partir de datos adquiridos de forma no invasiva mediante SORS.

PLS también fue el método de elección en el desarrollo de modelos de predicción para la cuantificación de parámetros de interés en la calidad de aceites de oliva a partir de datos NIR (aparatado 3.3 del **Capítulo 3**). Los parámetros objeto de estudio de estos modelos de cuantificación fueron los siguientes criterios de calidad y pureza de aceites de oliva: grado de acidez, índice de peróxidos, K232, además de los ácidos palmítico, palmitoleico, oleico y linoleico. Los mejores resultados predictivos se obtuvieron para los ácidos grasos. Prueba de ello son los parámetros de calidad de los modelos desarrollados (véase la Tabla 3.2 incluida en el citado apartado).

Una novedad introducida en la evaluación de la calidad de los mismos es el parámetro de exactitud (*accuracy*) o fiabilidad (*reliability*) calculado como el cociente entre el error estándar de predicción (SEP) del modelo quimiométrico y el error estándar del laboratorio (SEL) en la obtención de valores de referencia. Cuanto menor sea este parámetro, mayor es la fiabilidad del modelo desarrollado. Así, los resultados del conjunto de predicción de ácido palmítico y oleico mostraron valores <1.5, lo que de acuerdo con bibliografía se considera una precisión excelente [21]. Mientras que la predicción del grado de acidez y del parámetro K232 resultaron tener una precisión buena (valor <3).

Llegado este punto, cabe destacar la importancia de tener a disposición unos valores de referencia veraces, para evitar cometer errores en el desarrollo de los modelos de aprendizaje automático. Asimismo, de cara a la transferencia e implantación de métodos analíticos multivariable con un enfoque basado en el uso de huellas instrumentales inespecíficas, es de gran importancia llevar a cabo una etapa posterior a la validación externa del modelo con la inclusión de muestras "desconocidas", que sirvan como conjunto de predicción.

5.4.3. Optimización de procesos (DoE)

Finalmente, se ha hecho uso de la metodología diseño de experimentos para optimizar las condiciones experimentales en la adquisición de señales LF-NMR de aceites vegetales de la máxima calidad y en el menor tiempo posible.

Se hizo uso de una metodología poco aplicada en el ámbito de la química analítica, que, sin embargo, puede ofrecer ventajas relevantes. En ciertas ocasiones existen factores que afectan al proceso a ser optimizado, pero cuyo control es difícil de asumir. Este fue el caso encontrado al enfrentar el reto de optimizar la adquisición de señales LF-NMR, ya que se encontraron dos factores críticos y a la vez difíciles de controlar como son la temperatura de la habitación y el volumen de muestra a introducir en el tubo de medida.

Ante estas situaciones, existe una metodología conocida por el nombre de su desarrollador: Taguchi, que persigue optimizar un sistema haciéndolo robusto antes posibles variaciones en esos factores difíciles de controlar dentro de un rango específico.

La aplicación de esta metodología al último estudio presentado en el **Capítulo 4** condujo a conseguir el objetivo perseguido, ya que se generaron dos combinaciones de parámetros instrumentales, para la adquisición de señales 1H y 13C, robustas ante variaciones en los dos factores considerados de difícil control dentro del rango de estudio. Para ello se hizo uso del análisis de respuesta múltiple, basado en la función de deseabilidad, dado que se consideraron dos respuestas para su optimización: el mínimo tiempo de análisis y la máxima calidad informativa, que claramente tienen objetivos contrapuestos por lo que las condiciones que mejoraban una de ellas empeoraban la otra, y viceversa.

Además, se corroboró que los factores considerados de difícil control afectaban a la variabilidad de ambas respuestas y en ambos experimentos (1H y 13C).

[21] Shenk, J.; Westerhaus, M. Calibration the ISI way. In *Near Infrared Spectroscopy: The Future Waves;* Davies, AMC.; Ed.; NIR Publications: Montreal, Canada, 1996; pp. 198-202.

Para finalizar, cabe resaltar que la estancia internacional realizada en la Universidad de Copenhague permitió adquirir y reforzar conocimientos en lo que respecta a la aplicación de diversas herramientas quimiométricas para el tratamiento y análisis de datos químicos complejos. A pesar de no haberse materializado en unos resultados reseñables, supuso grandes aportaciones a la formación predoctoral, así como el establecimiento de una colaboración internacional que espera ser continuada en investigaciones futuras.

Estudios complementarios y otras tareas

Paralelamente al desarrollo de la presente tesis doctoral, se ha colaborado en diversos estudios y actividades, que se encuentran directa o indirectamente relacionados con el tema principal abordado en esta tesis, y que quedan recogidos a continuación.

1. Publicaciones científicas

1. Cuadros-Rodríguez, L., Ortega-Gavilán, F., Martín-Torres, S., Arroyo-Cerezo, A., Jiménez Carvelo, A.M. Chromatographic fingerprinting and food identity/quality: Potentials and challenges. *J. Agric. Food Chem.* **2021**, *69*, 14428-14434. DOI: 10.1021/acs.jafc.1c05584.
2. Yang, X., and Arroyo-Cerezo, A., Berzaghi, P., Magrin, L. Comparative Near-Infrared (NIR) spectroscopy calibrations performances of dried and undried forage on dry and wet matter bases. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2024**, *316*, 124287. DOI: 10.1016/j.saa.2024.124287.
3. Medina-García, M. Roca-Nasser, E.A., Martínez-Domingo, M.A., Valero, E.M., Arroyo-Cerezo, A., Cuadros-Rodríguez, L., Jiménez-Carvelo, A.M. Towards the establishment of a green and sustainable analytical methodology for hyperspectral imaging-based authentication of wholemeal bread. *Food Control*, **2024**, *166*, 110715. DOI: 10.1016/j.foodcont.2024.110715.

2. Participación en proyectos

1. CA19145, European Network for assuring food integrity using non-destructive spectral sensors (SensorFINT). IP: Dolores Pérez Marín (Universidad de Córdoba). COST Action, Programa Marco de Horizonte 2020, 30/09/2020 - 29/09/2024. 600.000 €.

   Función: participante ("action participant")
2. PPJIA2021.09, Evaluación de la autenticidad de productos alimentarios mediante el empleo de técnicas espectroscópicas. Hacia el desarrollo de métodos analíticos 'verdes'. IP: Ana M. Jiménez Carvelo (Universidad de Granada), 01/01/2022 - 31/12/2022. 1.500 €.

   Función: investigadora
3. PIDB22-37, Orientación académica y divulgación en asignaturas de posgrado – Hacia la *Open Science* desde el Trabajo de Fin de Máster. IP: Ana M. Jiménez Carvelo (Universidad de Granada), 3/10/2022 - 30/05/2023. 766 €.

   Función: miembro del equipo de trabajo

4. CPP2021-008672, Implantación de la resonancia magnética nuclear de baja frecuencia de campo (LF-NMR) en laboratorios de control para estudios cuantitativos y de clasificación de productos alimenticios y de otros sectores industriales (NMR-CONTROL). IP: Antonia Garrido Frenich (Universidad de Almería) / Luis Cuadros Rodríguez (Universidad de Granada), 1/12/2022 - 30/09/2025. 1.319.989 €.

   Función: colaboradora

5. PPJIB2023-042, Desarrollo de una metodología analítica global basada en un enfoque metabolómico no dirigido para la caracterización de las diferencias en el metabolismo de los alimentos por parte de la microbiota intestinal de niños con diferentes patologías. IP: Alejandra Arroyo Cerezo (Universidad de Granada), 01/01/2024 - 31/12/2024. 1.250 €.

   Función: investigadora principal

3. Tareas docentes complementarias a la docencia reglada

Además de realizar las tareas docentes en enseñanzas de grado como parte del contrato FPU (120 horas), durante la tesis doctoral se ha colaborado de forma complementaria en las siguientes actividades docentes, todas ellas realizadas en la Universidad de Granada:

1. Co-tutorización de Trabajo de Fin de Grado (curso 2022-2023)

   Título del trabajo: Uso de la resonancia magnética nuclear de baja frecuencia de campo para el control de calidad y autenticidad del aceite de oliva virgen

   Grado: Grado en Ciencia y Tecnología de los Alimentos

   Tutores: Ana María Jiménez Carvelo y Alejandra Arroyo Cerezo

   Nota. Este trabajo fue presentado, pero finalmente no fue defendido por parte del estudiante.

2. Co-tutorización de Trabajo de Fin de Grado (curso 2023-2024)

   Título del trabajo: Determinación de parámetros de calidad de aceites de oliva: control de la conformidad del etiquetado

   Grado: Grado en Química

   Tutores: Ana María Jiménez Carvelo y Alejandra Arroyo Cerezo

3. Co-tutorización de Trabajo de Fin de Grado (curso 2023-2024)

   Título del trabajo: ¿Las mieles comercializadas en España cumplen la normativa de calidad nacional?

   Grado: Grado en Química

   Tutores: Antonio González Casado y Alejandra Arroyo Cerezo

4. Mentorización de Trabajos de Fin de Grado y Máster

   Como parte de la formación predoctoral, se han llevado a cabo actividades de mentorización de (3) Trabajos de Fin de Grado del grado en Química y de (2) Trabajos de Fin de Máster del máster universitario en Ciencias y Tecnologías Químicas (KHEMIA) durante los cursos académicos 2022-2023 y 2023-2024.

5. Comunicación en congreso docente

   A. Arroyo Cerezo, M.G Bagur-González, A. González Casado, A.M. Jiménez Carvelo. **Orientación académica y divulgación en asignaturas de posgrado – Hacia la Open Science desde el Trabajo de Fin de Máster**. [Oral 7']. *Foro de Innovación Docente*. *Granada (España), diciembre 2022.*

Conclusiones y perspectivas futuras

Conclusiones y perspectivas futuras

Finalmente, en esta sección se resumen las conclusiones derivadas de los estudios de investigación llevados a cabo durante la tesis doctoral.

I. Se ha participado en la redacción de dos capítulos que formarán parte de dos libros publicados próximamente, en los que (i) se describe el estado del arte de la química analítica 'verde' aplicada al análisis de los alimentos, y (ii) se sientan las bases del análisis de la similitud entre señales analíticas como parte fundamental de la quimiometría. Además, con el segundo se introdujo la aplicación de la teoría de la información a la química analítica, que será objeto de estudio durante la investigación posdoctoral.

II. Se han desarrollado modelos supervisados para formar parte de métodos analíticos de cribado en la autentificación de productos lácteos. La aplicación de la modalidad SORS permitió realizar análisis rápidos y no invasivos a través del envase original de dos productos alimenticios, lo que derivó en el uso de las señales analíticas adquiridas como huellas instrumentales para los siguientes fines:

   a. Clasificación de muestras de margarinas atendiendo a su origen geográfico de fabricación y a la presencia/ausencia de ingredientes de interés como son el aceite de oliva o fitoesteroles.

   b. Clasificación de muestras de queso en lonchas atendiendo al origen animal de la leche empleada para su fabricación.

   c. Cuantificación del contenido graso de muestras de margarina, y del contenido graso y proteico de las muestras de queso.

   Todo ello dio lugar a la publicación de dos artículos científicos.

III. Se ha formulado una propuesta para la evaluación *ex-ante* de la 'blancura' de nuevos métodos analíticos no destructivos basados en un enfoque no dirigido y orientados al análisis de la calidad y autenticidad de alimentos. Esta propuesta quedó recogida en forma de artículo científico, que condujo a su publicación en una revista de alto impacto.

IV. Se examinó el potencial de la instrumentación miniaturizada en el desarrollo de métodos analíticos de cribado para el análisis de la calidad y/o autenticidad de aceites de oliva virgen, resultando en dos estudios:

   a. Se evaluó el uso de dos instrumentos portátiles aplicando la espectrometría NIR para el desarrollo de modelos predictivos de diversos criterios de calidad y pureza de aceites de oliva virgen. Ello permitió detectar tres aceites declarados como aceites de oliva virgen extra, que sin embargo no cumplían con los límites establecidos en la legislación, demostrando así el potencial de la huella instrumental NIR para el fin propuesto. Este estudio fue fruto de una colaboración internacional establecida en la primera estancia de investigación realizada durante la tesis, y asimismo derivó en una publicación científica.

b. Se llevó a cabo un estudio preliminar empleando las huellas instrumentales de aceites vegetales adquiridas mediante espectrometría NMR de bajo campo. Esto permitió el hallazgo de resultados que incitan al optimismo, y que continuarán su desarrollo como parte de la investigación posdoctoral.

V. Se desarrolló una metodología para la resolución de la mezcla de señales analíticas resultado del uso de la modalidad SORS para la adquisición de espectros Raman. Ello demostró la capacidad de la quimiometría para mejorar el tratamiento de datos posterior a la adquisición de señales. Asimismo, este estudio se complementó con un profundo análisis de diferentes índices de similitud para la comparación entre señales analíticas. Todo ello se materializó en forma de articulo científico.

VI. Se optimizaron las condiciones experimentales para la adquisición de señales LF-NMR de alta calidad y en el menor tiempo de muestras de aceites vegetales, poniendo el broche final al procedimiento experimental para el desarrollo de métodos analíticos 'blancos' basados en esta técnica. Se empleó una metodología de diseño experimental poco explorada en química analítica que permitió generar un sistema robusto. Con este estudio también se propuso una forma de evaluar la calidad informativa de una señal analítica *ex-ante,* dando respuesta a un reto no resuelto en la química analítica actual mediante la aplicación de la teoría de la información, y cuyo desarrollo proseguirá durante la investigación posdoctoral.

El trabajo llevado a cabo durante el transcurso de la tesis doctoral ha permitido cumplir con los objetivos establecidos inicialmente, y al mismo tiempo, dar paso a nuevos desafíos que serán objeto de estudio de futuras investigaciones.

Los resultados preliminares obtenidos con la técnica analítica LF-NMR y presentados en el **Capítulo 3** serán ampliados para finalmente desarrollar un método analítico de cribado rápido, poco invasivo y que arroje resultados precisos para el control de la calidad y autenticidad del aceite de oliva. Asimismo, se trabajará en la transferencia e implementación del método.

La propuesta para evaluar la calidad informativa de una señal analítica *ex-ante* será ampliada para dar respuesta a otros desafíos presentes y que puedan ser resueltos de esta forma. Consecuentemente, se tiene en proyecto escribir un tutorial presentando las diferentes métricas que podrían aplicarse, a fin de trasladar a la comunidad analítica esta herramienta hasta la fecha no utilizada, a pesar de su enorme potencialidad.

Se continuará priorizando el enfoque verde de la química analítica para el análisis de alimentos, explorando técnicas analíticas no destructivas / no invasivas y siguiendo un enfoque no dirigido mediante la aplicación de la metodología de huellas instrumentales.