UNIVERSIDAD DE GRANADA

Dpto. de Química Analítica
Prof. Fermín Capitán García

Métodos multivariables cualimétricos para el aseguramiento y control de la calidad en el ámbito de la soberanía y calidad alimentaria

Tesis Doctoral

Programa de Doctorado en Química

Christian Hazael Pérez Beltrán

Directores:

Dr. Luis Cuadros Rodríguez

Dra. Ana María Jiménez Carvelo

Para optar por el título de:

Doctor por la Universidad de Granada

Granada, España, 2023

Editor: Universidad de Granada. Tesis Doctorales
Autor: Chistian Hazael Pérez Beltrán
ISBN: 978-84-1117-792-4
URI: https://hdl.handle.net/10481/81249

ÍNDICE GENERAL

Pág.

Resumen 1

Abstract 3

Presentación 5

Capítulo 1 INTRODUCCIÓN 9

1.1 Huellas instrumentales 13

1.2 La era del *Big-Data* y el análisis de datos multivariable en la calidad alimentaria 16

1.2.1 Reconocimiento de pautas no supervisado 23

1.2.2 Reconocimiento de pautas supervisado 25

1.3 El trayecto hacia métodos analíticos multivariable universales 32

1.4 Desarrollo de métodos analíticos multivariable aplicados *in-situ* 36

1.5 Justificación e hipótesis 39

1.6 Objetivos 41

Capítulo 2 CONTROL DE PROCESOS EN LA INDUSTRIA ALIMENTARIA 43

2.1 Resumen 45

2.2 Artículo científico I 48

2.3 Comunicación a congresos 89

Capítulo 3 TRANSFERENCIA DE MÉTODOS ANALÍTICOS MULTIVARIABLE BASADOS EN CROMATOGRAFÍA DE LÍQUIDOS 91

3.1 Resumen 93

3.2 Artículo científico II 101

3.3 Artículo científico III 139

3.4 Comunicación a congresos 173

Capítulo 4 **MÉTODOS ANALÍTICOS MULTIVARIABLE BASADOS EN TÉCNICAS NO INVASIVAS** 175

4.1 Resumen 177

4.2 Artículo científico IV 182

4.3 Estudio 2 205

4.4 Artículo científico V 221

4.5 Comunicación a congresos 248

Capítulo 5 **DISCUSIÓN INTEGRADA** 249

5.1 Visión general 251

5.2 Nuevos métodos analíticos multivariable para el aseguramiento y control de calidad alimentaria 253

5.2.1 Desarrollo y aplicación de nuevos métodos analíticos multivariable basados en cromatografía de líquidos y su transferencia mediante agnostización instrumental 253

5.2.2 Desarrollo de nuevos métodos analíticos multivariable complementarios para el control de calidad alimentario basados en técnicas no destructivas 264

5.3 Implicación en la industria alimentaria 271

Capítulo 6 **CONCLUSIONES GENERALES** 277

RESUMEN

La actual Tesis Doctoral desarrolla y aplica diversos métodos analíticos multivariable basados en la técnica de cromatografía de líquidos y en técnicas espectroscópicas vibracionales, conjuntamente con herramientas quimiométricas, los cuales se enfocan en el aseguramiento y control de calidad alimentaria, específicamente, en la autentificación de la calidad del aceite de oliva virgen y virgen extra, y Tequila Blanco.

El objetivo primordial que se persigue en esta Tesis es el desarrollo de distintos métodos analíticos multivariable, que sean rápidos, confiables, eficientes y de utilidad para la industria alimentaria para mejorar y aumentar la detección de fraude alimentario. Dichos métodos implican el uso de la metodología de huellas instrumentales, particularmente, huellas espectroscópicas y cromatográficas, y el uso de la metodología de 'agnostización instrumental', empleando herramientas quimiométricas para extraer de ellas la información analítica útil.

Inicialmente, se explora la utilidad de la agnostización instrumental sobre huellas cromatográficas de aceite de oliva, obtenidas mediante HPLC-UV/Vis, para construir modelos matemáticos con PLS-DA y SIMCA capaces de detectar adulteraciones en dichos aceites. Seguidamente, se demuestra el potencial que posee la agnostización instrumental para transferir señales analíticas entre laboratorios utilizando para ello huellas cromatográficas agnostizadas de Tequila Blanco, obtenidas mediante HPLC-DAD en laboratorios diferentes. Con dichas señales se desarrolló una base de datos global y un modelo matemático único para diferenciar entre las categorías '100 % de agave' y 'mixto' de muestras de Tequila Blanco, analizadas en España y México.

Posteriormente, las técnicas espectroscópicas FTIR, NIR y SORS, y determinadas herramientas quimiométricas fueron aplicadas para desarrollar distintos métodos analíticos multivariable capaces de autentificar la calidad de las categorías de muestras de Tequila Blanco mediante técnicas de reconocimiento de pautas supervisado, como PLS-DA, SVM, SIMCA, kNN, así como para cuantificar su contenido alcohólico con las técnicas de calibración multivariable PLSR y SVMR. Finalmente, se propone la aplicación de estos métodos analíticos multivariable en las cadenas alimentarias de cada una de las industrias olivarera y tequilera.

ABSTRACT

The current Doctoral Thesis develops and applies diverse multivariate analytical methods based on the liquid chromatography and spectroscopic vibrational analytical techniques, together with chemometric tools, which are focused on the assurance and quality food control, specifically, in the quality authentication of virgin and extra virgin olive oil, and White Tequila.

The paramount objective of this Doctoral Thesis is the development of different multivariate analytical methods, that are fast, reliable, efficient and useful for the food industry to improve and increase the food fraud detection. These methods imply the use of the fingerprinting methodology, particularly, spectroscopic and chromatographic fingerprints, and the use of the 'instrument-agnostizing' methodology, using chemometric tools to mine the useful analytical information from them.

Initially, the utility of the instrument-agnostizing methodology is explored on chromatographic fingerprints of olive oil, obtained through HPLC-UV/Vis, to build mathematic models with PLS-DA and SIMCA capable to detect adulterations in such oils. Afterwards, the intrinsic potential of the instrument-agnostizing methodology to transfer multivariate analytical signals among laboratories is proved, using 'agnostic' chromatographic fingerprints of White Tequila, obtained through HPLC-DAD in different laboratories. From such signals, a global database and a single mathematic model were developed in order to differentiate among the '100% agave' and 'mixed' categories of White Tequila samples, analyzed in Spain and México.

Thereupon, the FTIR, NIR and SORS spectroscopic techniques, and certain chemometric tools were applied together to develop different multivariate analytical methods capable to authenticate the quality of the categories of White Tequila samples using pattern recognition techniques, such as PLS-DA, SVM, SIMCA, kNN, as well as to quantify their alcoholic content using the PLSR and SVMR multivariate calibration techniques. Finally, the application of these multivariate analytical methods is suggested along the food chain of each of the olive oil and tequila industries.

PRESENTACIÓN

La actual Tesis Doctoral se rige por el Real Decreto 99/2011, del 28 de enero, del Ministerio de Educación (Gobierno de España) mediante el cual se regulan las enseñanzas oficiales de doctorado, y que está modificado por el Real Decreto 534/2013 del Ministerio de Educación, Cultura y Deporte (Gobierno de España) en lo relativo a la evaluación y defensa de la tesis doctoral.

El artículo 5 del Real Decreto 99/2011 establece las competencias que el doctorando debe adquirir durante sus estudios de doctorado, las cuales son:

a) *Comprensión sistemática de un campo de estudio y dominio de las habilidades y métodos de investigación relacionados con dicho campo.*
b) *Capacidad de concebir, diseñar o crear, poner en práctica y adoptar un proceso sustancial de investigación o creación.*
c) *Capacidad para contribuir a la ampliación de las fronteras del conocimiento a través de una investigación original.*
d) *Capacidad de realizar un análisis crítico y de evaluación y síntesis de ideas nuevas y complejas.*
e) *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.*
f) *Capacidad de fomentar, en contextos académicos y profesionales, el avance científico, tecnológico, social, artístico o cultural dentro de una sociedad basada en el conocimiento.*

Del mismo modo, este Real Decreto también establece que la obtención del título de doctor debe proporcionar una alta capacitación profesional en diversos ámbitos, como los relacionados con la creatividad e innovación, de manera que los doctores puedan:

a) *Desenvolverse en contextos en los que hay poca información específica.*
b) *Encontrar las preguntas claves que hay que responder para resolver un problema complejo.*

c) *Diseñar, crear, desarrollar y emprender proyectos novedosos e innovadores en su ámbito de conocimiento.*
d) *Trabajar tanto en equipo como de manera autónoma en un contexto internacional multidisciplinar.*
e) *Integrar conocimientos, enfrentarse a la complejidad y formular juicios con información limitada.*
f) *La crítica y defensa intelectual de soluciones.*

Además, en el artículo 13 se precisa que

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

REALIZACIÓN DE LA TESIS DOCTORAL

Esta Tesis Doctoral está enmarcada en el Programa de Doctorado en Química, bajo la línea de investigación "Metodologías de obtención de información analítica en sistemas reales" en la Escuela de Doctorado en Ciencias, Tecnologías e Ingenierías (EDCTI) de la Universidad de Granada, España.

La mayoría de las actividades realizadas durante esta Tesis Doctoral se han 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 dirección y supervisión del Dr. Luis Cuadros Rodríguez y la Dra. Ana María Jiménez Carvelo.

Las muestras empleadas para el trabajo de investigación realizado en la Universidad de Granada proceden de diversos proyectos científicos concedidos al grupo de investigación. Asimismo, se obtuvieron muestras a partir de la colaboración con la Unidad de Investigación Multidisciplinaria (UIM) de la Facultad de Estudios Superiores (FESC) de la Universidad Nacional Autónoma de México (UNAM), en Cuautitlán, Estado de México, y el Consejo Regulador de Tequila (CRT), en Zapopan, Estado de Jalisco, México.

Finalmente, hay que señalar que una parte importante del estudio experimental ha sido desarrollado durante un periodo de estancia de 3 meses en la UIM-FESC de la UNAM, el cual forma parte de esta Tesis Doctoral. Dicha estancia estuvo bajo la supervisión de la Dra. Guadalupe Pérez Caballero, el Dr. José de Jesús Olmos Espejel y la Dra. Alma Luisa Revilla Vázquez, y ha sido posible gracias a la financiación recibida por parte de la Asociación Universitaria Iberoamericana de Posgrado (AUIP) durante el curso académico 2021/2022.

ESTRUCTURA DE LA TESIS DOCTORAL

La presente Tesis Doctoral tiene como objetivo principal el desarrollo de métodos analíticos multivariable de utilidad para el aseguramiento cualimétrico y control de la calidad en el ámbito de la soberanía y calidad alimentaria. Para ello, esta Tesis Doctoral se ha dispuesto en seis capítulos en los cuales se plasman los resultados obtenidos para lograr dicho objetivo.

El **capítulo 1** presenta una introducción general de esta Tesis Doctoral, poniendo de manifiesto el contexto y campo de aplicación en el cual se enmarca. Siguiendo con la justificación, hipótesis y objetivos, tanto generales como específicos, los cuales son descritos al final de este capítulo.

El **capítulo 2** aborda el uso de herramientas quimiométricas para el control de procesos en la industria alimentaria dentro del marco de la Calidad mediante Diseño (QbD, *Quality by Design*) y la Tecnología Analítica de Procesos (PAT, *Process Analytical Technology*). Tras una revisión bibliográfica exhaustiva, se detalla el estado del arte de QbD/PAT en la industria alimentaria y se describen estudios a nivel industrial y/o de planta piloto enmarcados en QbD/PAT. Especial atención se presta al uso de herramientas quimiométricas como el diseño y optimización multivariable de experimentos, el análisis de datos multivariable y el control multivariable de procesos.

El **capítulo 3** sienta las bases para la transferencia de señales analíticas (huellas instrumentales) obtenidas con instrumentos de cromatografía de líquidos, sobre las cuales se ha aplicado la metodología de agnostización instrumental, desarrollada previamente por este grupo de investigación. En un primer estudio se explora la aplicación de la agnostización

instrumental de huellas cromatográficas de aceite de oliva virgen y virgen extra para el desarrollo de un método analítico multivariable capaz de detectar sus adulteraciones con aceites de orujo de oliva y aceites refinados de oliva, comprobando que la metodología de agnostización arroja resultados similares o mejores a la metodología tradicional en la cual se aplican herramientas convencionales de alineamiento de cromatogramas. Posteriormente, se presenta un segundo estudio en colaboración internacional utilizando la misma metodología, pero esta vez sobre huellas cromatográficas de Tequila Blanco, obtenidas en dos equipos analíticos diferentes en laboratorios distintos, en España y en México, y periodos de tiempos diferentes, con objeto de transferir y verificar dichas huellas cromatográficas en condiciones de reproducibilidad interlaboratorio.

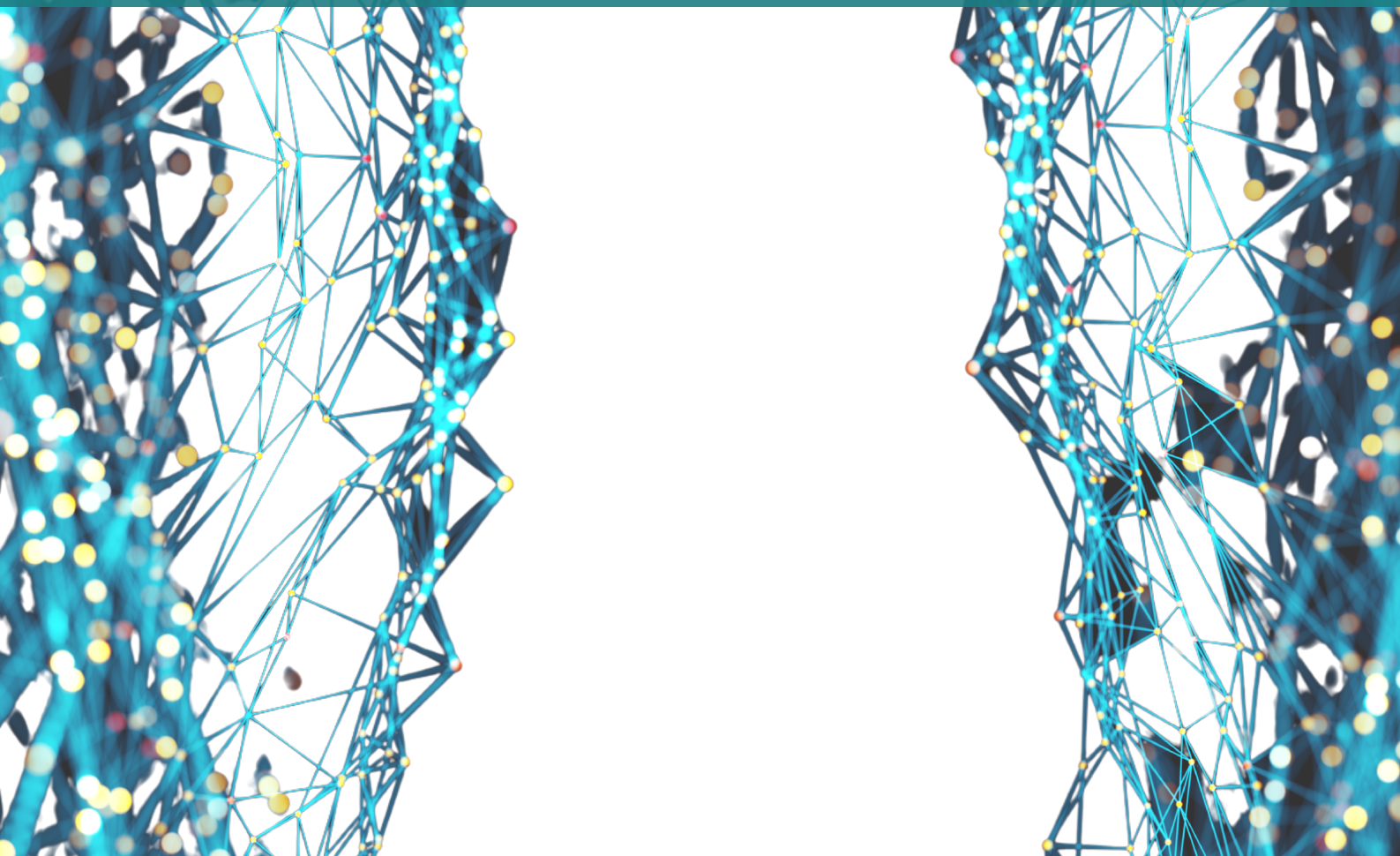
En el **capítulo 4** se exploran técnicas analíticas no destructivas y no invasivas con la finalidad de desarrollar métodos analíticos multivariable rápidos y confiables basados en técnicas espectroscópicas y herramientas quimiométricas para el control y aseguramiento de calidad del Tequila Blanco. Se evalúa la utilidad y ventajas de técnicas espectroscópicas tradicionales, como infrarrojo con transformada de Fourier (FTIR) e infrarrojo cercano (NIR), así como de una técnica novedosa y emergente en el ámbito alimentario, la cual es la espectroscopia Raman con sistema de compensación espacial (SORS).

En el **capítulo 5** se realiza una discusión integrada de los resultados de los **capítulos 2**, **3** y **4**. En este capítulo se lleva a cabo una comparación de las técnicas analíticas empleadas y su desempeño en las aplicaciones para las cuales fueron dispuestas. Además, se evalúa y discute la potencial aplicabilidad de los métodos analíticos multivariable desarrollados en esta Tesis Doctoral, los cuales están basados en cromatografía liquida y técnicas espectroscópicas rápidas y no invasivas en conjunto con la metodología de agnostización instrumental, para el aseguramiento y control de calidad de productos alimenticios a una escala global. Finalmente, se plantean recomendaciones y perspectivas futuras para investigaciones venideras.

Por último, el **capítulo 6** pone de manifiesto las conclusiones generales a las cuales se ha arribado a través de las investigaciones realizadas durante esta Tesis Doctoral.

Capítulo 1

INTRODUCCIÓN



1. INTRODUCCIÓN

En 1948 fue reconocido el concepto Derecho Alimentario (del inglés, *Right to Food*) en la Declaración de Derechos Humanos de las Naciones Unidas, pero no fue hasta 1996 cuando se formalizó la adopción del término "Derecho a una Alimentación Adecuada" (*Right to Adequate Food*) durante la realización de la Cumbre Mundial sobre Alimentación. Allí, se adoptó la definición de **seguridad alimentaria** y expresa que *existe seguridad alimentaria cuando todas las personas tienen en todo momento acceso físico y económico a suficientes alimentos inocuos y nutritivos para satisfacer sus necesidades y preferencias alimenticias a fin de llevar una vida activa y sana* [**1**].

Unos años más tarde, durante el año 2003, se erigió el término de **soberanía alimentaria**, el cual se define como *el derecho de las personas a definir su propia alimentación y agricultura; de proteger y regular su producción agrícola doméstica y de comercializar para lograr objetivos de desarrollo sostenible; de determinar el grado en el cual quieren ser autosuficientes; y de restringir el ingreso de productos no deseados en sus mercados* [**2**]. La soberanía alimentaria no niega la comercialización, al contrario, promueve la formulación de políticas y prácticas de comercio que sirvan para los derechos de las personas para una producción alimentaria segura, sana y ecológicamente sostenible.

Para lograr los objetivos planteados por la seguridad y soberanía alimentaria, mostrados en la **Figura 1**, es necesario implementar puntos de control de calidad durante los procesos de elaboración de los productos alimenticios, así como al producto final. Estos controles de calidad consideran los aspectos de seguridad (p.ej., física, química y microbiológico) y, sensorial y nutricional, siendo el color, sabor, apariencia y textura factores críticos para la calidad sensorial de los alimentos [**3**].

[1] Rome Declaration on World Food Security and World Food Summit Plan of Action, WFS 96/3, 1996.

[2] P. Rosset, Food Sovereignty: Global rallying cry of farmer movements, Inst. Food Dev. Policy Backgrounder. 9 (2003) 1-4.

[3] M. Samad Khan, M. Shafiur Rahman, Techniques to measure food safety and quality. Microbial, chemical, and sensory, in M. Samad Khan, M. Shafiur Rahman (Eds.), Introduction on techniques to measure food safety and quality, Springer Nature, 2021, pp. 1-9.

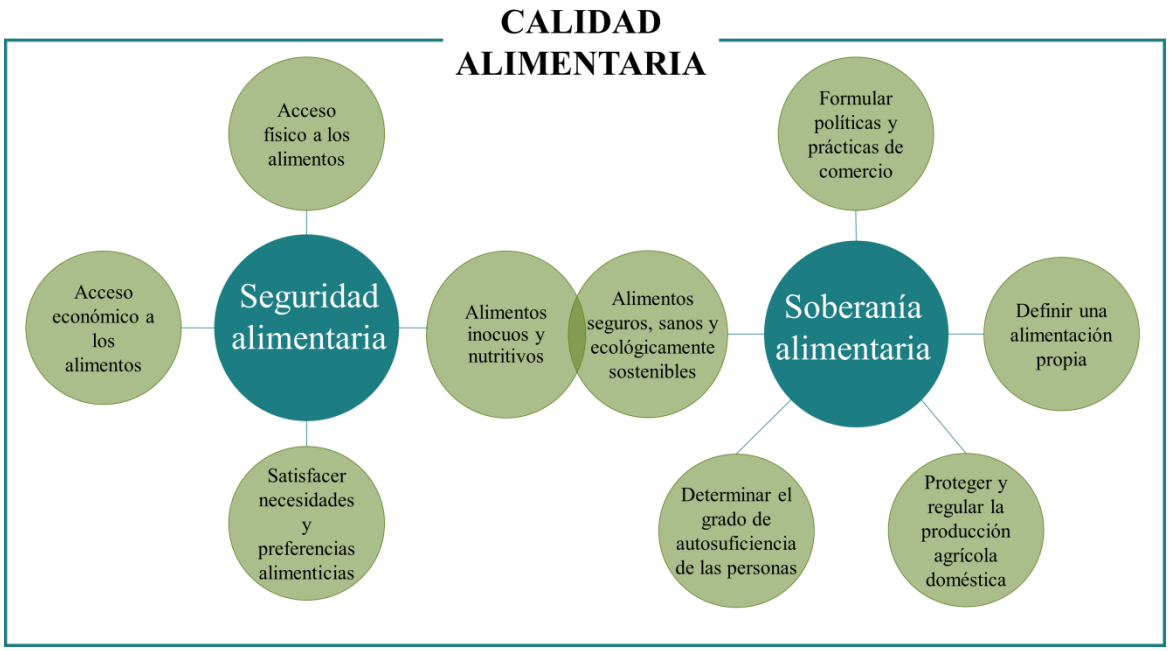


Figura 1. Objetivos de la seguridad y soberanía alimentaria necesarios para lograr una adecuada calidad alimentaria.

En sus inicios, el punto de control de calidad era implementado únicamente cuando el producto estaba finalizado, denominada como Calidad mediante el Ensayo (QbT, *Quality by Testing)*, obteniendo solamente un resultado por análisis (enfoque univariable), como se observa en la **Figura 2(a)**. Esta actividad se realiza en análisis químicos dirigidos para detectar posibles marcadores o contaminantes en los alimentos, los cuales pueden comprometer y/o poner en peligro la salud del consumidor, por lo que su uso es recomendado en estas situaciones.

No obstante, el enfoque univariable sigue siendo aplicado en la industria alimentaria para controlar la calidad de procesos y sistemas multifactoriales complejos, como lo son los alimentos, ocasionando pérdida de información al no considerar otras variables que podrían estar interaccionando entre sí, lo cual conlleva a tomar decisiones ineficientes o erróneas y a elaborar productos y subproductos alimenticios de baja calidad, así como llevar a cabo un gasto elevado de reactivos y disolventes. Por tanto, es necesario y aconsejable cambiar a un enfoque multivariable cuando se trata de recabar toda la información e interacciones

relacionadas con el proceso de producción y el alimento para asegurar y controlar la calidad de una manera más sencilla y eficiente, tal como se observa en la **Figura 2(b).**

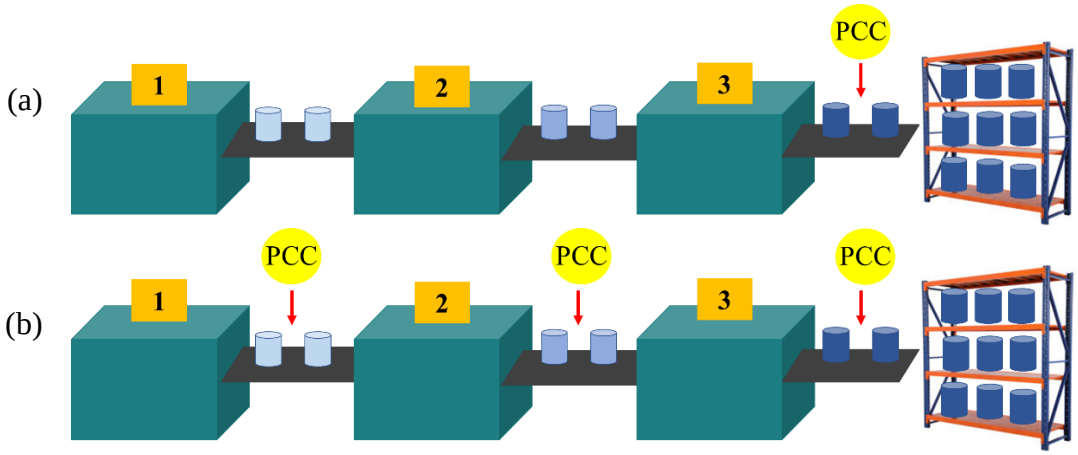


Figura 2. Proceso de producción de flujo continuo (a) con un punto crítico de control (PCC) de calidad de enfoque univariable y (b) con varios PCC de calidad de enfoque multivariable.

Una metodología centrada en un enfoque multivariable con gran potencial, pero que no ha sido explotada al máximo por la industria alimentaria, es la metodología denominada huellas instrumentales en conjunto con la aplicación de herramientas de minería de datos, las cuales han sido empleadas a lo largo de esta tesis doctoral para crear métodos analíticos multivariable rápidos y confiables para el aseguramiento y control de calidad alimentario.

1.1. Huellas instrumentales

Normalmente, la aplicación del enfoque univariable utiliza como dato de partida un número o escalar (tensor de orden 0), relacionado con compuestos específicos en la muestra, tal como la concentración de un contaminante obtenida a través de una calibración univariable o la temperatura de una sustancia obtenida con un termómetro; mientras que el enfoque multivariable hace uso de vectores (tensor de orden 1), matrices (tensor de orden 2) o cubos (tensor de orden 3) [**4**].

[4] M.G. Bagur González, A.M. Jiménez-Carvelo, F. Ortega-Gavilán, A. González-Casado, Chromatographic methods, in M. Galanakis (Ed) Food Authentication and Traceability, Elsevier, 2021, pp 65–99.

Por un lado, un vector de datos es el conjunto de datos o variables adquiridas de una sola medición, las cuales se disponen en una misma dimensión para conforman la señal de la muestra analizada, tal como los obtenidos mediante las diferentes técnicas de espectroscopía de infrarrojo, Raman, resonancia magnética nuclear (NMR) o cromatografía de líquidos de altas prestaciones con detector de absorción molecular en el ultravioleta/visible (HPLC-UV/Vis), entre otras. Por otro lado, una matriz de datos es obtenida mediante equipos instrumentales más complejos en los cuales se tiene un espectro de masas determinado (HPLC-espectrometría de masas (MS) o cromatografía de gases (GC)-MS)) o un espectro de absorción determinado (LC-detector de fila de diodos (DAD)) a un tiempo específico para cada muestra, obteniendo resultados gráficos en tercera dimensión [**5**].

En cromatografía de líquidos, los cromatogramas son considerados señales inespecíficas cuando se obtienen en un periodo corto de tiempo y la resolución de los picos y su separación no es el objetivo principal y, en técnicas espectroscópicas cuando se realizan las mediciones directamente sobre las matrices alimenticias. Estas señales inespecíficas contienen información de los compuestos que conforman la muestra, sobre los cuales se desconoce su identificación *a priori* [**6**]. La utilización de estas señales inespecíficas ha generado la aplicación de la metodología de trabajo conocida como 'huellas instrumentales' en el ámbito de la investigación y de la soberanía y calidad alimentaria, la cual es de rápida y sencilla aplicación, presentando un gran potencial para el control de calidad alimentaria.

La metodología de huellas instrumentales consiste en una obtención rápida y simple de una señal única y característica para cada muestra analizada, relacionada con sus propiedades y composición química, mediante un instrumento analítico capaz de generar, captar y almacenar enormes cantidades de datos en forma de espectros, cromatogramas, termogramas, voltamperogramas, electroferogramas o imágenes, entre otros [**6**].

[5] G. Escandar, H.C. Goicoechea, A. Muñoz de la Peña, A.C. Olivieri, Second- and higher- order data generation and calibration: A tutorial, 2014, Analytical Chimica Acta, 806, 8-26.

[6] A.M. Jiménez-Carvelo, S. Martin-Torres, L. Cuadros-Rodríguez, A. González-Casado, Non-targeted fingerprinting approaches, in: C.M. Galanakis (Ed.), Food Authentication and Traceability, Academic Press /Elsevier, 2021, pp. 163–193.

Un ejemplo de una huella instrumental cromatográfica se puede observar en la **Figura 3**, la cual fue obtenida mediante la técnica de cromatografía de líquidos de ultra altas prestaciones (UHPLC) en su modalidad de trabajo de 'fase normal' acoplada a un detector de absorción molecular UV/Vis.

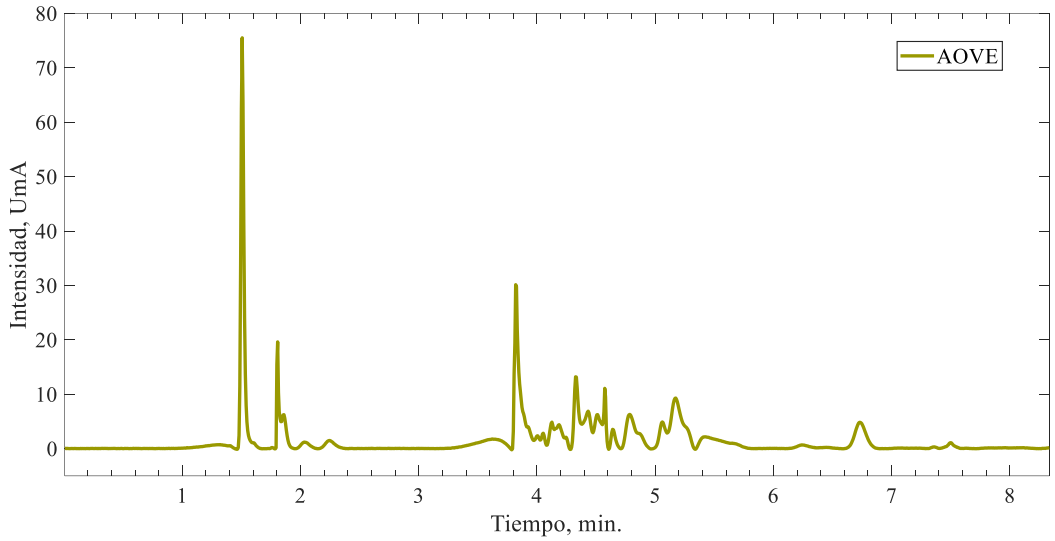


Figura 3. Huella instrumental cromatográfica de una muestra de aceite de oliva virgen extra (AOVE), obtenida mediante cromatografía de líquidos de ultra altas prestaciones acoplada a un detector de absorción molecular en el ultravioleta/visible (UHPLC-UV/Vis).

Al usar técnicas analíticas de separación como la cromatografía de líquidos (LC) de manera convencional, se esperan picos con buena resolución para identificar cada compuesto o familia de compuestos. Sin embargo, el uso de esta técnica analítica en conjunto con la metodología de huellas instrumentales se aleja de estos esquemas convencionales. La metodología de huellas instrumentales, no tiene como objetivo obtener una buena resolución de la señal, sino obtener la mayor cantidad de información en el menor tiempo posible.

A causa de la complejidad de las señales obtenidas, es necesario recurrir a técnicas avanzadas de tratamiento de datos para lograr la extracción de información útil y relevante que ayuden a autentificar de forma inequívoca la muestra analizada.

1.2. La era del *Big-Data* y el análisis de datos multivariable en la calidad alimentaria

La actual era digital, ha traído consigo un gran desarrollo de tecnologías capaces de generar y almacenar abundante información. Tanto es así que este desarrollo se ha dado también en los laboratorios analíticos en los cuales, debido a la existencia de instrumentos automatizados capaces de analizar grandes cantidades de muestras de alimentos y, por consiguiente, capaces de generar y adquirir grandes volúmenes de datos, ha dado lugar a una nueva disciplina científica conocida como "ciencia de los datos", caracterizada por la capacidad de extraer información útil a partir de grandes conjuntos de datos (*Big-Data*).

Estos datos deben ser tratados e interpretados de manera adecuada con distintas herramientas matemáticas/estadísticas para que puedan ser de utilidad para el fin previsto. Dependiendo el área de estudio en la cual se empleen estas herramientas recibirá un nombre en particular, aunque de manera general se refiere a ellas como minería de datos (*data mining)* o aprendizaje automático (*machine learning*). En el área de la salud, como medicina, biología, farmacia y/o biotecnología, se le conoce como bioinformática (*bioinformatics*) [**7**]; en el campo de la ingeniería como inteligencia computacional o artificial (*artificial intelligence*) [**8**]; mientras que en el área de la química analítica se le conoce como métodos de reconocimiento de pautas (*pattern recognition*), métodos de análisis de datos multivariable (*multivariate data analysis*), pero, principalmente, como quimiometría (*chemometrics*) [**9**].

[7] J.H. Duffus, M. Nordberg, D.M. Templenton, Glossary of terms used in toxicology (IUPAC recommendations 2007), 2007, Pure and Applied Chemistry, 79, 1153-1344.

[8] H.M. Kingston, M.L. Kingston, Nomenclature in laboratory robotics and automation (IUPAC recommendations 1994), 1994, Pure and Applied Chemistry, 66, 609-630.

[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, 2019, Food Research International, 122, 25-39.

De acuerdo a la Unión Internacional de Química Pura y Aplicada (IUPAC, *International Union of Pure and Applied Chemistry*) la **quimiometría** se define como *la ciencia que utiliza métodos matemáticos y estadísticos para extraer la mayor información de un grupo de señales analíticas complejas* [**10**]. Paralelamente, el organismo de normalización estadounidense, actualmente denominado ASTM International, (que proviene de la denominación originaria, *American Society for Testing and Materials*) define a los métodos de análisis multivariable como *una herramienta apropiada para explorar y manejar grandes conjuntos de datos heterogéneos, mapear datos de alta dimensionalidad en representaciones de menos dimensión, exponer correlaciones significativas entre variables multivariadas dentro de un solo conjunto de datos o correlaciones significativas entre variables multivariadas a través de conjuntos de datos* [**11**].

Entre estas herramientas quimiométricas, destacan aquellas conocidas como de reconocimiento de pautas, que se dividen en dos grupos principales: (i) **técnicas de reconocimiento de pautas no supervisado** y (ii) **técnicas de reconocimiento de pautas supervisado**, dentro de las cuales se pueden encontrar distintos y variados modelos estadísticos/matemáticos, tal como se puede observar en la **Figura 4**.

[10] D.B. Hibbert, Vocabulary of concepts and terms in chemometrics (IUPAC Recommendations 2016), 2016, Pure and Applied Chemistry, 88, 407-443.

[11] ASTM International, ASTM ES891-20: Standard guide for multivariate data analysis in pharmaceutical development and manufacturing applications, 2020.

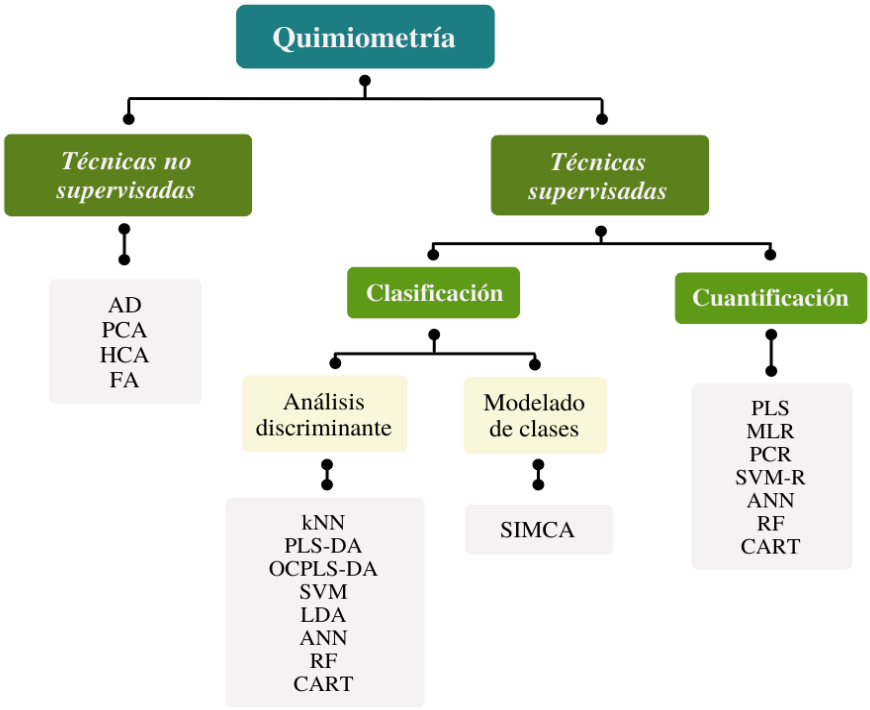


Figura 4. Esquema general de los tipos de herramientas quimiométricas de reconocimiento de pautas más utilizadas en la industria alimentaria.

Por orden alfabético y significado en español: **AD**: Agrupamiento difuso, **HCA**: Análisis de agrupamiento jerárquico, **AF**: Análisis factorial, **PCA**: Análisis de componentes principales, **kNN**: k-vecinos más cercanos, **OCPLS-DA**: análisis discriminante de una clase mediante regresión parcial por mínimos cuadrados, **PLS-DA**: análisis discriminante mediante regresión parcial por mínimos cuadrados, **LDA**: análisis discriminante lineal, **CART**: árbol de clasificación y regresión, **RF**: bosque aleatorio (árboles de decisión), **SIMCA:** modelado flexible e independiente por analogía de clases, **ANN**: redes neuronales artificiales, **SVM-R**: regresión por sistema de aprendizaje automático mediante vectores soporte, **MLR**: regresión lineal múltiple, **PLS**: regresión parcial por mínimos cuadrados, **PCR**: regresión por componentes principales.

Normalmente, el uso de herramientas quimiométricas, que dan lugar al desarrollo de un modelo matemático multivariable, está organizado en tres etapas generales (véase **Figura 5**): (i) obtención de información sobre el agrupamiento natural de los datos/objetos mediante el uso de herramientas quimiométricas de reconocimiento de pautas no supervisado, (ii) empleo de herramientas quimiométricas de reconocimiento de pautas supervisado para desarrollar modelos matemáticos/estadísticos que permitan clasificar las

muestras analizadas y/o cuantificar una propiedad característica de dichas muestras, basándose en valores característicos de un atributo o de una propiedad de interés previamente establecidos y, por último, (iii) predicción de la clases, a las cuales pertenecen las nuevas muestras desconocidas, o del valor de la propiedad característica en estudio.

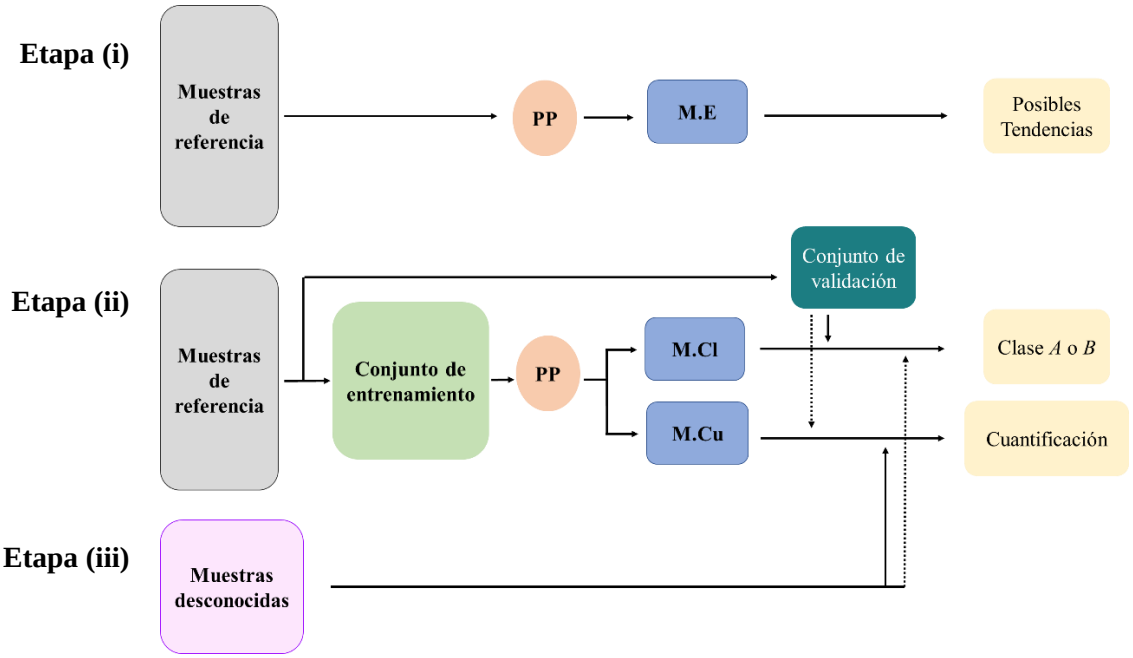


Figura 5. Esquema general para el desarrollo de modelos matemáticos/estadísticos aplicando herramientas quimiométricas de reconocimiento de pautas no supervisado y supervisado.
(i) Etapa de análisis exploratorio, (ii) desarrollo de modelos quimiométricos de clasificación y/o cuantificación, y (iii) predicción de clases y/o características de nuevas muestras desconocidas en base al modelo desarrollado.
PP: pre-procesamiento, **M.E**: Modelo exploratorio, **M.Cl**: Modelo de clasificación, **M.Cu**: Modelo de cuantificación.

Es importante mencionar que, la aplicación de los métodos de pre-procesamiento de datos previo a la generación de los modelos matemáticos permite que estos vean mejorada su robustez e interpretabilidad al remover o reducir fuentes de variación no deseadas que están presente en los datos, como puede ser la corrección de la línea base. Con el uso de los **métodos de pre-procesado** adecuados se logra una mayor extracción de información de interés y relevante, que se encuentra oculta en los datos de señales complejas, p.ej. huellas instrumentales.

Estos métodos de pre-procesado de datos pueden ser de diferentes tipos, tales como de: transformación, filtrado, normalización, escalado o centrado en los datos [**12**].

Entre los métodos de pre-procesamiento más empleados para el tratamiento de huellas instrumentales cromatográficas y espectroscópicas se encuentran:

a) Alineamiento de los picos

Principalmente aplicado sobre señales cromatográficas, en donde el tiempo de retención sufre variaciones de posición a lo largo del tiempo.

Los métodos más comunes para mitigar este fenómeno son (i) deformación optimizada de correlación (COW, *correlation optimized warping*) y (ii) el algoritmo de cambio optimizado de correlación de intervalo (icoshift, *Interval Correlation Optimised Shifting algorithm*). El método COW está basado en una corrección lineal continua por partes o segmentos de muestra a la vez, destinado a alinear un vector de datos de muestra en base a un vector de referencia al permitir cambios limitados en las longitudes de los segmentos en el vector de muestra [**13**,**14**]. El método icoshift, a diferencia del COW, utiliza una función de corrección lineal por segmentos basado en un modelo de inserción/eliminación (I/D, *insertion/deletion*), en el cual los intervalos definidos son desplazados para maximizar su correlación cruzada con el segmento correspondiente en donde se encuentra el pico de referencia [**15**].

b) Corrección de la línea base

Los métodos de corrección de la línea base tienen como objetivo substraer o eliminar derivas u otras desviaciones sistemáticas de la línea base de la señal de las muestras analizadas, las cuales se pueden dar debido a variaciones propias del equipo instrumental, temperatura, presión, entre otras.

[12] B.M. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Winding, R.S. Koch, Chemometrics Tutorial for PLS_Toolbox and Solo, Eigenvector Research, Inc., WA, USA, 2006.

[13] N.P.V. Nielsen, J.M. Carstensen, I. Smedsgaard, Aligning of single and multiple wavelength chromatographic profiles for chemometrics data analysis using correlation optimized warping, 1998, Journal of Chromatography A, 805, 17-35.

[14] G. Tomasi, F. Van de Berg, C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data, 2004, Journal of Chemometrics, 18, 231-241.

[15] G. Tomasi, F. Savorani, S.B. Engelsen, Icoshift: An effective tool for the alignment of chromatographic data, 2011, Journal of Chromatography A, 1248, 7832-7840.

Un ejemplo de este tipo de pre-procesamiento es el basado en el filtro Whitaker, el cual substrae la línea base usando el método de mínimos cuadrados penalizados, mismo que busca una combinación balanceada entre la aspereza de los datos originales y la falta de ajuste de los datos originales a los nuevos datos [**16**].

c) Filtro ponderación de mínimos cuadrados generalizados

En ocasiones, el ruido de fondo puede afectar a la elaboración de un modelo matemático, en estos casos, el filtro de ponderación de mínimos cuadrados generalizados (GLSW, *generalized least squares weighting)* puede ayudar a simplificarlo antes de desarrollar el modelo matemático, ya que identifica algunas estructuras de covarianza no deseada y las remueve de los datos [**12**].

d) Centrado en la media

El centrado en la media (*mean centering*) es una transformación aditiva de variables continuas, la cual consiste en calcular la media a partir de las variables que conforman la señal, en este caso dispuestas en columnas, y a continuación substraerla a cada una de esas mismas variables [**12**,**17**], llevando a cabo el centrado de los datos por variables, tal como se expresa en la siguiente ecuación:

$$m_{ij} = x_{ij} - \bar{x}_j \quad \text{(Ec. 1)}$$

donde m representa los datos centrados en la media, x las variables de la señal y $\bar{x}$ la media de las variables (siendo *i* las filas y *j* las columnas).

e) Autoescalado

El método de autoescalado (*autoscaling*) es empleado para corregir diferentes escalas y/o unidades de las variables con el objetivo de que la información contenida en cada variable sea de igual importancia. Esto se lleva a cabo al centrar en la media los datos y, posteriormente, dividir cada variable (columna) por la desviación estándar de la columna, tal como se muestra en esta ecuación:

$$a_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad \text{(Ec. 2)}$$

[16] P.H.C Eilers, A perfect smoother, 2003, Analytical Chemistry, 75, 3631-3636.

[17] M. Hofer, Mean centering, in J. Matthes, R. Potter, C.S. Davis (Eds.), The international encyclopedia of communication research methods, Wiley Blackwell, 2017, pp. 248-269.

donde a representa los datos autoescalados, x las variables de la señal, $\bar{x}$ la media de las variables y s la desviación estándar de las variables de cada columna [**12**,**18**].

f) Decimación

La decimación (*decimation*/*downsampling*) es utilizado cuando se trabaja con grandes volúmenes de datos/variables, ayudando a reducir los requerimientos digitales computacionales, ya que este método de pre-procesamiento consiste en disminuir la frecuencia de muestreo de una señal por un número entero, conocido como factor de submuestreo [**19**]. Por tal motivo, es importante asegurar que el perfil de la señal se mantendrá después de aplicar este método.

g) Derivadas

El uso de derivadas (*derivatives*) es comúnmente empleado para eliminar señales sin importancia de la línea base tomando la derivada de las respuestas medidas con respecto al número de variable u otra escala de eje relevante. Las derivadas son una forma de filtro de alto paso y escalado dependiente de la frecuencia y son, normalmente, usadas cuando características de más baja frecuencia contienen la señal de interés, por lo que este método debe de ser usado sólo cuando las variables estén fuertemente relacionadas entre sí y las variables adyacentes contengan señales similares correlacionadas [**12**].

h) Suavizado

Los métodos de suavizado (*smoothing*) son empleados con la finalidad de eliminar el ruido que acompaña a la señal analítica de interés, asumiendo que las variables que están próximas unas de otras en la matriz de datos están relacionadas entre sí y contienen información similar que puede ser promediada conjuntamente para reducir el ruido sin una pérdida significativa de la señal de interés.
El algoritmo más conocido y usado es el propuesto por Savitzky-Golay, el cual,

[18] R. Bro, A.K. Smilde, Centering and scaling in component analysis, 2003, Journal of Chemometrics, 17, 16-33.

[19] M. Parker, Decimation and interpolation, in M. Parker (Ed.), Digital signal processing 101, Everything you need to know to get started, Elsevier, 2022, 65-74.

esencialmente, ajusta polinomios individuales a ventanas alrededor de cada variable en la señal, los cuales son usados para suavizar la señal [**12**,**20]**.

i) Remuestreo

Este método de pre-procesamiento (*resampling*) es utilizado para ajustar a un mismo número de variables las señales cuando estas contienen diferentes números de variables, ya sea por provenir de distintos instrumentos analíticos o haber sufrido un procesado de datos previo. Para ello, se realiza una interpolación lineal la cual crea un nuevo vector de para cada señal con el mismo perfil, pero con un número de variables común para todas ellas [**21**].

1.2.1. Reconocimiento de pautas no supervisado

Este tipo de herramientas quimiométricas son utilizadas para evaluar la estructura natural de los datos con el objetivo de realizar análisis exploratorios para identificar posibles tendencias, relaciones entre muestras o la existencia de muestras anómalas en el conjunto de datos. Las herramientas quimiométricas más comunes para este propósito en el control de calidad alimentario se pueden observar en la **Figura 6** y se describen a continuación:

a) Análisis de componentes principales (PCA – *Principal component analysis*)

Esta herramienta quimiométrica es, quizás, la más empleada en las distintas áreas de estudio para extraer información de relevancia dada su facilidad de uso e interpretabilidad. La herramienta de PCA encuentra las variables o factores que mejor describan las tendencias principales en los datos. Estas variables o factores son proyectados en un nuevo sistema de ejes ortogonales llamado "componentes principales" (PCs, *principal components*), definido por la siguiente ecuación:

$$\mathbf{X} = \mathrm{a}_1\mathrm{b}_1^{\mathrm{T}} + \mathrm{a}_2\mathrm{b}_2^{\mathrm{T}} + \cdots + \mathrm{a}_\mathrm{c}\mathrm{b}_\mathrm{c}^{\mathrm{T}} + \mathrm{E} \quad \text{(Ec. 3)}$$

[20] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, 1964, Analytical Chemistry, 36, 1627-1639.

[21] The MathWorks, Inc. https://es.mathworks.com/help/signal/ug/resampling.html. Acceso: 16-01-2023.

donde **X** es la matriz de datos; a son las puntuaciones (*scores*) que constituyen las nuevas coordenadas de las muestras en el espacio de las componentes principales, y que contienen información de cómo las muestras se relacionan entre sí; b son los coeficientes de peso o ponderales (*loadings*) que contienen información de cómo las variables se relacionan una con otra; c hace referencia al número final de componentes principales seleccionadas para conformar el PCA, quedando factores de varianza pequeños sin explicar que son consolidados en la matriz de residuales E. Usualmente, los datos con los que se trabaja pueden llegar a ser explicados con una cantidad muy inferior de factores a comparación del número original de variables, reduciendo su dimensionalidad sin una pérdida significativa de información [**22**,**23**].

b) Análisis de agrupamiento jerárquico (HCA – *Hierarchical cluster analysis*)

El HCA es usado para agrupar diferentes muestras u objetos utilizando alguna/s características comunes entre ellos. Las dos categorías principales del HCA son la aglomerativa y la divisiva. Los métodos aglomerativos comienzan con un solo objeto y continúa agregando (aglomerando) objetos y/o grupos de objetos existentes para formar otros agrupamientos más grandes, mientras que los métodos divisivos comienzan en el sentido opuesto, con un gran aglomerado que contiene todos los objetos y continua con el fraccionamiento de este grupo en otros más pequeños. Todos los agrupamientos mencionados anteriormente están condicionados por la medida de distancia que se elija para desarrollarlos, entre las cuales están: el vecino más cercano, vecino más lejano, mediana o, el más comúnmente utilizado, el método de Ward. Usualmente, el resultado de este tipo de análisis es una gráfica denominada dendrograma (similar a un árbol genealógico descendente) en el cual los objetos se organizan en una fila de acuerdo a sus similitudes y se van uniendo en cierto punto en

[22] H. Abdi, L.J. Williams, Principal component analysis, 2010, Wiley Interdisciplinary Reviews: Computational Statistics, 2, 433–459

[23] J.E. Jackson, Principal Components and Factor Analysis: Part 1-Principal Components, 1981, Journal of Quality Technology, 13, 201-213.

el eje vertical, mismo que representa la medida de semejanza a la cual cada objeto se une a un grupo [**24**,**25**,**26**].

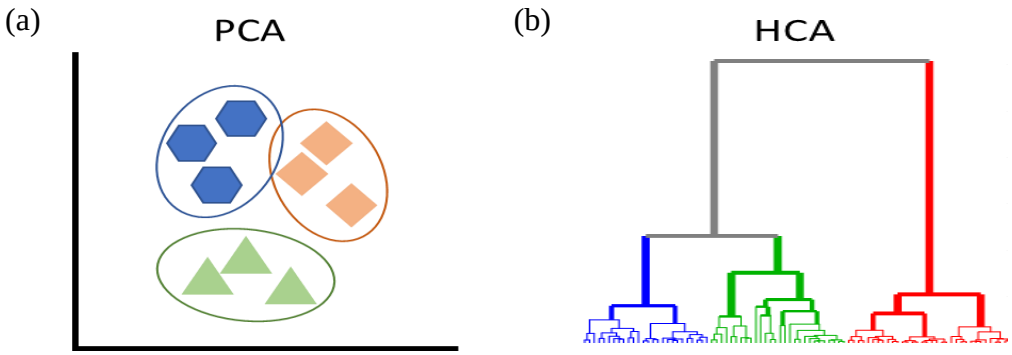


Figura 6. Descripción gráfica de las herramientas quimiométricas de reconocimiento de pautas no supervisado más utilizadas para el análisis exploratorio de los datos en el control de calidad alimentario. (a) Análisis de componentes principales (PCA) y (b) análisis de agrupamiento jerárquico de tipo aglomerativo (HCA).

1.2.2. Reconocimiento de pautas supervisado

Las herramientas de reconocimiento de pautas supervisado son aplicadas para desarrollar modelos de clasificación y/o de cuantificación, por lo que pueden ser de tipo cualitativo o cuantitativo. A diferencia de las herramientas de pautas no supervisadas, en éstas el modelo matemático/estadístico se construye conociendo la clase a la cual pertenece el objeto, la cual puede estar conformada por una o varias características de interés a ser evaluadas, como podrían ser: categoría del alimento, composición química, origen geográfico, variedad del cultivar, año de recolección del cultivo y/o año de producción, entre muchas otras más según el objetivo del estudio.

[24] M. Otto, Chemometrics. Statistics and Computer Application in Analytical Chemistry, third ed., Wiley-VCH, 2017.
[25] R.G. Brereton, Applied Chemometrics for Scientists, Chichester, John Wiley and Sons Ltd. 2007.
[26] Eigenvector Research Inc. https://www.wiki.eigenvector.com/index.php?title=Cluster. Acceso: 09.08.2022

❒ ***Desarrollo del modelo matemático/estadístico***

Para comenzar con el desarrollo de los distintos modelos matemáticos, es necesaria la etapa previa de asignación de clases específicas o del valor de una propiedad o atributo a cada una de muestras bajo estudio, las cuales serán usadas por los modelos matemáticos con fines de diferenciación, clasificación y/o cuantificación. Esta asignación se hace en base a los criterios del estudio y características propias de la muestra que son previamente conocidas de forma fiable. Una vez establecidas y definidas las clases o categorías, o bien asignados los valores numéricos a la propiedad/atributo diana característica de cada muestra, se desarrolla el correspondiente modelo matemático/estadístico, el cual consta de las siguientes etapas: (i) **entrenamiento o calibración del modelo**, y (ii) **validación o evaluación del modelo**.

En la etapa de entrenamiento es necesario, cuando existan muestras suficientes, seleccionar muestras de referencia pertenecientes a una o varias clases conocidas o que contengan atribuida la cantidad correspondiente a la propiedad bajo estudio. Esto se lleva a cabo dividiendo el conjunto de datos original en dos subconjuntos: (i) **conjunto de entrenamiento** y (ii) **conjunto de validación**, los cuales suelen estar conformados por una proporción entre 60/40 - 80/20 %, respectivamente. Esta selección debe incluir las muestras más representativas, ya sea para determinar clases o predecir el valor de la propiedad diana, y puede realizarse de manera aleatoria por el analista o mediante algoritmos de selección, tales como Kennard-Stone (también conocido como CADEX) [**27**], CENTER, SELECT [**28**] o más recientemente *Onion* [**29**], entre otros. Por un lado, el algoritmo de selección CADEX encuentra primero las dos muestras (datos) más lejanas entre sí, considerando la distancia geométrica euclidiana; posteriormente, agrega otra muestra al conjunto previamente seleccionado, la cual debe de tener la distancia de separación más extensa de entre las muestras restantes.

[27] R.W. Kennard, L.A. Stone, Computer aided design of experiments, 1969, Technometrics, 11, 137-148.

[28] J.S. Shenk, M.O. Westerhaus, Population definition, sample selection, and calibration procedures for near infrared reflectance spectroscopy, 1991, Crop Science, 31, 469-474.

[29] N.B. Gallagher, D.O. Sullivan, Selection of representative learning and test sets using the onion method, 2020, Eigenvector Research Incorporated.

Este proceso es repetido hasta que el número requerido de muestras (k) ha sido alcanzado para cada uno de los subconjuntos de entrenamiento y validación. Por otro lado, el algoritmo *Onion* selecciona las muestras a partir de un nuevo sistema de ejes dimensional (como el generado en PCA), basándose en la estructuración de capas sucesivas de forma similar a como se dispone el bulbo de la cebolla (de ahí su denominación). La capa externa es seleccionada para el conjunto de entrenamiento y el número de muestras elegidas de esta capa está definido por la denominada "fracción bucle" utilizando la distancia euclidiana o de Mahalanobis. El proceso de selección es repetido en cada capa asignando muestras a los subconjuntos de entrenamiento y validación, hasta llegar a las muestras restantes ubicadas en el centro del sistema de ejes dimensional, las cuales son asignadas al azar al conjunto de entrenamiento o validación [**29**].

Así mismo, en esta etapa es usualmente necesario realizar algún tipo de pre-procesamiento (PP), descritos en la subsección 1.2, a los datos originales con el fin de eliminar fuentes extrañas de variabilidad para no entorpecer la estimación del modelo matemático y, de esta manera, evidenciar la información más relevante de los datos que permitan realizar clasificaciones y cuantificaciones lo más adecuadas posible.

Durante **la etapa de validación** se evalúa la capacidad del modelo para clasificar o cuantificar, siguiendo las estrategias de validación cruzada interna o validación externa. Por un lado, la **validación cruzada interna** se realiza: (i) cuando hay un número pequeño de muestras y no pueden dividirse en conjuntos de entrenamiento y validación, (ii) para valorar la capacidad de clasificación y/o cuantificación del modelo sobre los datos usados en el conjunto de entrenamiento, y (iii) para seleccionar el número adecuado de componentes o variables necesarias para fijar el modelo. Por otro lado, la **validación externa** se lleva a cabo utilizando un conjunto de datos independiente al utilizado en la etapa de entrenamiento.

La clasificación realizada por los modelos se evalúa mediante **métricas de calidad en el desempeño**, siendo las cuatro principales: sensibilidad, especificidad, valor predictivo positivo y valor predictivo negativo. La **sensibilidad** (SENS, *sensitivity*) o **veracidad** (TRUE, *trueness*) de un modelo matemático de clasificación hace referencia a la relación de las muestras de la clase diana (comúnmente seleccionada como la clase objetivo)

correctamente clasificadas entre el total de muestras de la misma clase, mientras que la **especificidad** (SPEC, *specificity*) indica la relación de las muestras de la clase alternativa correctamente clasificadas. El **valor predictivo positivo** (PPV, *positive predictive value*) o **precisión** (PREC, *precision*) indica la proporción de muestras correctamente clasificadas de la clase diana entre todos los valores asignados a esa misma clase, incluyendo muestras de las clases diana y alternativa; en cambio, el **valor predictivo negativo** (NPV, *negative predictive value*) indica la proporción de muestras correctamente clasificadas de la clase alternativa entre todos los valores asignados a dicha clase [**30**]. A partir de la SENS, SPEC, PPV y NPV se pueden calcular otras métricas de calidad en el desempeño, por ejemplo: el ratio de falsos positivos y negativos, índice de Youden, relación de verosimilitud para resultados positivos y negativos, relación de diagnóstico, valor F, poder discriminante, eficiencia o exactitud, relación de clasificaciones incorrectas, área bajo la curva, coeficiente Gini, G-media, coeficiente de correlación de Matthew, relaciones de aciertos y errores debidos a la suerte, coeficiente Kappa y probabilidades condicionales de Bayes [**30**].

En cambio, las cuantificaciones predichas por los modelos son evaluadas mediante las métricas de calidad en el desempeño propuestas por la ASTM en su norma E2617 [**31**], las cuales son: error estándar de validación (SEV, *standard error of validation*) y desviación estándar de los residuos de validación (SDV, *standard deviation of validation residuals*), mismas que son medidas de la coincidencia esperada del modelo matemático empírico con el método de referencia. De igual manera, se utiliza el coeficiente de determinación (R^2) para evaluar el ajuste del modelo matemático a los datos en la etapa de entrenamiento, mientras que el error cuadrático medio (RMSE, *root mean square error*), error absoluto medio (MAE, *mean absolute error*), error absoluto de la mediana (MdAE, *median absolute error*) son utilizadas para la validación del modelo matemático [**32**,**33**].

[30] L. Cuadros-Rodríguez, E. Pérez-Castaño, C. Ruiz-Samblás, Quality metrics in multivariate classification methods for qualitative analysis, 2016, Trends in Analytical Chemistry, 80, 612-624.

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

[32] Y.C. Martin, R. Abagyan, G.G. Ferenczy, V.J. Gillet, T.I. Oprea, J. Ulander, D. Winkler, N.S. Zefirov, Glossary of terms used in computational drug design, part II (IUPAC Recommendations 2015), 2016, Pure Applied Chemistry, 88, 239-264.

[33] R.J. Hyndman, A.B. Koehler, Another look at measures of forecast accuracy, 2006, International Journal of Forecasting, 22, 679-688.

Una vez validado y evaluado el modelo, es aplicado sobre muestras desconocidas con el objetivo de predecir la clase a la que podrían pertenecen y/o el valor de la propiedad o atributo diana que haya sido modelado.

- ***Herramientas quimiométricas con fines de clasificación***

Las diferentes herramientas quimiométricas de reconocimiento de pautas supervisado más usadas en la calidad alimentaria con fines de identificación, clasificación y cribado se pueden observar en la **Figura 7** y se describen a continuación:

a) k-vecinos cercanos (kNN, *k-nearest neighbors*)

Esta herramienta de clasificación es considerada como una técnica "perezosa", ya que el modelo no "aprende" de los datos introducidos, sino que la clase que se le asignará a una nueva muestra será en base a la muestra o grupo de muestras existentes más cercanas a ella. Esta cercanía está definida en términos de la distancia euclidiana y la selección de una clase u otra se realizará de acuerdo al valor otorgado al parámetro 'k' (k = 1, 2, 3, ...,n.) [**12**,**34**].

b) Modelado flexible e independiente por analogía de clases (SIMCA, *soft independent modeling by class analogy*)

Este modelado matemático consiste en desarrollar un modelo PCA para cada una de las clases definidas inicialmente, las cuales pueden ser explicadas por un número igual o diferente de componentes principales (PCs) [**12**,**22**]. La aplicación tradicional de 1os modelos de clasificación ha sido llevar una clasificación binaria en la cual el modelo se entrena con dos clases de entrada (2iC, *two input-class*), siendo una la clase diana o clase objetivo, y la otra la clase alternativa. En los modelos matemáticos discriminantes las muestras a evaluar serían clasificadas en una de las dos clases. No obstante, en los modelos matemáticos basados en el modelado de clases, como lo es SIMCA, esta situación es distinta, ya que además de que las muestras pueden ser clasificadas en una de las dos clases de entrada, estas también se pueden clasificar como pertenecientes a ambas clases o a ninguna de ellas, obteniendo como resultado cuatro posibles escenarios.

[34] M.A. Sharaf, D.L. Illman, B.R. Kowalski, 1986, Chemometrics, John Wiley & Sons.

De igual manera, también es posible desarrollar modelos matemáticos con una sola clase de entrada (1iC, *one input-class*), la cual es determinada como la clase objetivo [**35**]. La aplicación de la metodología 1iC es exclusiva de modelos matemáticos basados en modelado de clases, ya que son generados de manera independiente para cada clase. Esta metodología ha sido de gran utilidad para autentificar y asegurar la calidad de distintos alimentos [**36**,**37**], ya que el modelo matemático puede ser entrenado sólo con muestras representativas del alimento de interés, obteniendo como posibles resultados que la muestra pertenece o no pertenece a la clase objetivo.

c) Análisis discriminante mediante regresión parcial de mínimos cuadrados (PLS-DA, *partial least squares-discriminant analysis*)

El objetivo principal de PLS-DA es determinar si las muestras pueden ser separadas en una clase u otra. Esta diferenciación se logra llevando a cabo, en primer lugar, un modelo de regresión PLS en el cual se selecciona un número de variables latentes (LVs, *latent variables*) con las cuales se establecen los límites de las clases. Enseguida, se desarrolla un análisis discriminante (DA, *discriminant analysis*) para clasificar las muestras en una clase en particular, basado en el número obtenido para cada muestra, siendo 1 la clase de interés (clase diana) y 0 si la muestra no pertenece a esta, en caso de tratarse de una notación binaria [**12**,**38**].

d) Sistema de aprendizaje automático mediante vectores soporte (SVMs, *support vector machines*)

Normalmente, las muestras que se encuentran superpuestas en distintos sistemas de ejes dimensionales del PCA o PLS (gráficas de puntuaciones o *score plots*) son

[35] O.Y. Rodionova, P. Oliveri, A.L. Pomeranstev, Rigorous and compliant approaches to one-class classification, 2016, Chemometrics and Intelligent Laboratory Systems, 159, 89-96.

[36] A.M. Jiménez-Carvelo, E. Pérez-Castaño, A. González-Casado, L. Cuadros Rodríguez, One-input and two-input class classifications for differentiating olive oil from other edible oils by use of the normal phase liquid chromatography fingerprint of the methyl-transesterified fraction, 2017, Food Chemistry, 221, 1784-1791.

[37] G. Campmajó, J. Saurina, O. Núñez, FIA-HRMS fingerprinting subjected to chemometrics as valuable tool to address food classification and authentication: Application to red wine, paprika, and vegetable oil samples, 2022, Food Chemistry, 373, 131491.

[38] R.G. Brereton, Chemometrics for Pattern Recognition, 2009, John Wiley & Sons.

clasificadas erróneamente haciendo uso de herramientas quimiométricas de reconocimiento de pautas lineales, tales como kNN, SIMCA o PLS-DA.

Para solventar este problema, SVMs genera uno o varios espacios nuevos de alta dimensión (hiperplanos) a partir de la distribución original de las muestras en los cuales se generan nuevos límites de separación (lineales o curvos), que ahora son capaces de clasificar correctamente las muestras [**24**,**25**].

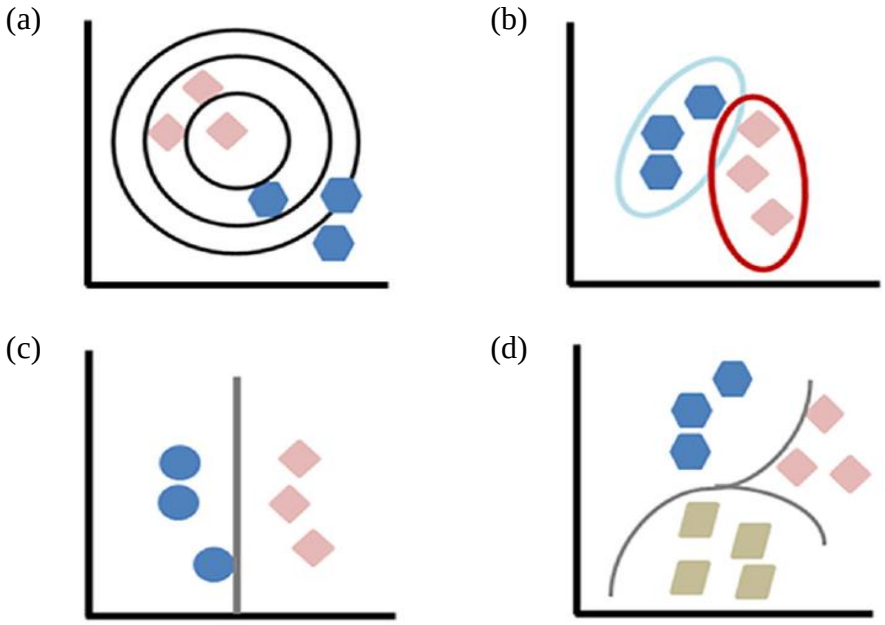


Figura 7. Descripción gráfica de algunas herramientas quimiométricas de reconocimiento de pautas supervisado más utilizadas para clasificación en el control de calidad alimentario. (a) kNN: k-vecinos cercanos, (b) SIMCA: modelado flexible e independiente por analogía de clases, (c) PLS-DA: análisis discriminante mediante regresión parcial de mínimos cuadrados, (d) SVM: sistema de aprendizaje automático mediante vectores soporte.

*Figura adaptada de: 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, 2019, Food Research International, 122, 25-39 [**9**].*

❒ ***Herramientas quimiométricas con fines de cuantificación***

Otro de los objetivos en el análisis de datos multivariable es la determinación del valor de una o varias propiedades físico-químicas características de las muestras bajo estudio, en términos cuantitativos, a partir de la señal adquirida, p.ej., predicción del contenido total graso de margarinas y otros productos para untar a partir de sus huellas instrumentales obtenidas mediante espectroscopía Raman convencional y Raman con sistema de compensación espacial (SORS) [**39**].

Uno de los modelos más usados para este fin es el descrito a continuación:

e) Regresión parcial lineal mediante mínimos cuadrados (PLSR, *partial least-squares regression*)

Esta herramienta quimiométrica de cuantificación o predicción busca factores que capturen la mayor cantidad de varianza en las variables predictoras (p.ej. espectro o huella instrumental) y al mismo tiempo que correlacione de mejor manera las variables predictoras con las variables a predecir (p.ej. concentraciones o porcentajes de contenido alcohólico), maximizando la covarianza entre ellas [**12**].

Dada la característica de PLSR para capturar la varianza en las variables predictoras, se utilizó en esta tesis doctoral de una manera no convencional para el análisis exploratorio de los datos, tal como se hace en PCA.

1.3. El trayecto hacia métodos analíticos multivariable universales

El potencial de las herramientas quimiométricas para el desarrollo de métodos analíticos multivariable en el ámbito de la calidad alimentaria es ampliamente conocido y prueba de ello es la multitud de publicaciones científicas que pueden ser encontradas en bibliografía. Sin embargo, aún existe una laguna sobre el empleo y transferencia de los métodos analíticos multivariable hacia laboratorios que realizan análisis de rutina ya que, al ser implementados en laboratorios analíticos de control específicos, presentan una alta

[39] A.M. Jiménez-Carvelo, A. Arroyo-Cerezo, S. Brikani, W. Jia, A. Koidis, L. Cuadros-Rodríguez, Rapid and non-destructive spatially offset Raman spectroscopic analysis of packaged margarines and fat-spread products, 2022, Microchemical Journal, 178, 107378.

dependencia con la plataforma analítica empleada para obtener la señal, lo cual origina una escasa o nula transferibilidad de los mismos.

Uno de los principales inconvenientes de su transferencia se encuentra cuando se aplican técnicas analíticas cromatográficas, en especial, la cromatografía de líquidos. Esta técnica puede llegar a presentar variaciones en los tiempos de retención e intensidades de las señales obtenidas debido, principalmente, a las interacciones entre la fase móvil, la fase estacionaria y la matriz que estemos analizando, ya sea entre una misma tanda de análisis realizada durante el mismo día, entre tandas analizadas en diferentes días y, especialmente, entre laboratorios que empleen equipos instrumentales similares. Esta situación hace que las huellas instrumentales cromatográficas sean dependientes de las condiciones particulares del instrumento analítico y del laboratorio analítico en el cual se obtuvieron, lo cual imposibilita la transferencia de los métodos analíticos multivariable desarrollados a otros laboratorios analíticos.

Con la finalidad de generar métodos analíticos multivariable transferibles a otros laboratorios, este grupo de investigación ha desarrollado una metodología novedosa denominada "**agnostización instrumental**" [**40**,**41**], la cual tiene potencial aplicación para el control de calidad alimentario en cualquier laboratorio del mundo, facilitando la comparación de resultados y comprobación de la calidad y autenticidad de los alimentos.

A pesar de su significado filosófico, el término de "**agnóstico**" es adecuado en este contexto científico-técnico. De hecho, el diccionario Cambridge "en línea" incluye entre los significados aplicables al adjetivo '*agnostic*' el siguiente: '*relating to hardware or software that can be used with many different types of platform or system*' [**42**] que traducido viene a indicar que es aplicable al hardware o software que puede utilizarse con muchos tipos diferentes de plataformas o sistemas.

[40] L. Cuadros-Rodríguez, F. Ortega-Gavilán, S. Martín-Torres, S. Medina-Rodríguez, A.M. Jiménez-Carvelo, A. González-Casado, M.G. Bagur-González, Standardization of chromatographic signals – Part I: Towards obtaining instrument-agnostic fingerprints in gas chromatography, 2021, Journal of Chromatography A, 1641, 461983.

[41] L. Cuadros-Rodríguez, S. Martín-Torres, F. Ortega-Gavilán, A.M. Jiménez-Carvelo, R. López-Ruíz, A. Garrido-Frenich, M.G. Bagur-González, A. González-Casado, Standardization of chromatographic signals – Part II: Expanding instrument-agnostic fingerprints to reverse phase liquid chromatography, 2021, Journal of Chromatography A, 1641, 461973.

[42] https://dictionary.cambridge.org/dictionary/english/agnostic. Acceso: 09-10-2022

Por tanto, la **agnostización instrumental** hace referencia a la obtención de huellas instrumentales estandarizadas que puedan ser comparables aun cuando: (i) hayan sido obtenidas en dos o más instrumentos analíticos similares, (ii) las condiciones de los instrumentos analíticos varíen debido a su inherente particularidad, o (iii) los tiempos e intensidades de las señales sean distintos [**40**]. Esta metodología se basa en dos fases: (1) establecimiento previo de un conjunto de valores estandarizados de tiempos, constantes independientes del sistema (SRS, *standard retention scores*) definidos a partir del análisis de un estándar externo constituido por una mezcla de compuestos químicos específicamente seleccionados para este fin, y (2**)** estandarización tanto de las intensidades como de los tiempos de retención, utilizando los SRS previamente establecidos, que constituye la etapa de agnostización propiamente dicha, tal como se observa en la **Figura 8** [**41**].

La primera fase del proceso, **establecimiento de los SRS**, se realiza una única vez al inicio del estudio y son utilizados en la segunda fase de la agnostización instrumental. La obtención de los SRS se hace a partir de un sistema químico de referencia invariante (MPE, mezcla patrón externa), el cual debe, idealmente, cumplir los siguientes requisitos: tener un comportamiento químico similar a los componentes endógenos de interés presentes en la muestra, ser analizado mínimamente 10 veces, cubrir todo el rango de tiempo del análisis cromatográfico y mostrar un perfil de elución regular bajo las condiciones experimentales seleccionadas.

En la segunda fase se realiza **la agnostización de la señal instrumental**, para lo cual es necesario, primeramente, (*a*) la **normalización de la intensidad**, basándose en la intensidad de un patrón interno (PI). El PI es agregado a cada una de las muestras a ser analizadas en una cantidad propiamente establecida y constante, cuya concentración debe ser seleccionada de modo que genere un pico cuya altura sea equivalente o ligeramente superior a la intensidad más alta observada en las señales instrumentales cromatográficas; de esta manera, la máxima intensidad será debido al PI y obtendrá el valor de 1 tras realizar la normalización.

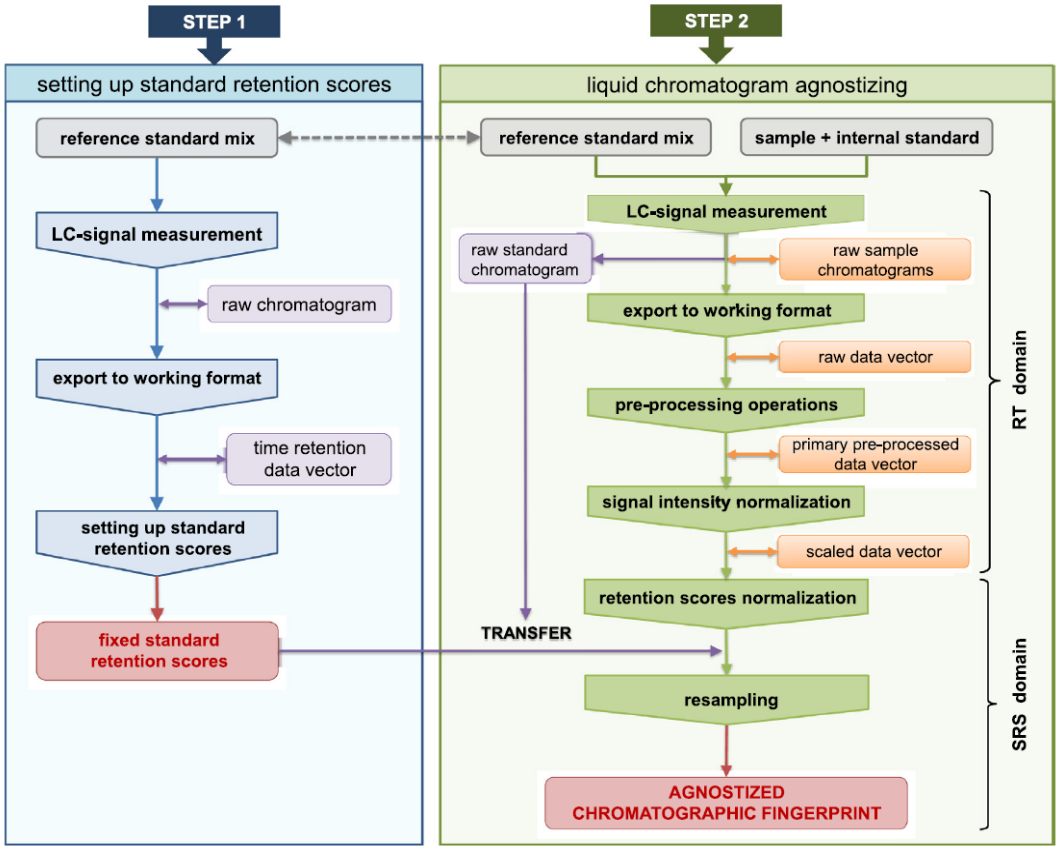


Figura 8. Diagrama de flujo del proceso general para realizar la agnostización instrumental en cromatografía de líquidos. Paso 1: Establecimiento del conjunto de valores estandarizados de tiempos, constantes independientes del sistema (SRSs, *standard retention scores*). Paso 2: Agnostización de la señal cromatográfica.

*Figura adaptada de: L. Cuadros-Rodríguez, S. Martín-Torres, F. Ortega-Gavilán, A.M. Jiménez-Carvelo, R. López-Ruíz, A. Garrido-Frenich, M.G. Bagur-González, A. González-Casado, Standardization of chromatographic signals – Part II: Expanding instrument-agnostic fingerprints to reverse phase liquid chromatography, 2021, Journal of Chromatography A, 1641, 461973 [**41**].*

Para completar el proceso de agnostización de la señal instrumental cromatográfica, se continúa con (*b*) la **normalización del tiempo de retención** mediante la transferencia de los SRS de la MPE, obtenidos en la fase 1, a las señales instrumentales cromatográficas después de haber normalizado sus intensidades en base al PI. Para lograr este objetivo, la MPE debe ser analizada al inicio y final de cada tanda de análisis de las muestras, llevando a cabo el registro de los tiempos de retención para cada compuesto químico que conforma la MPE. Tras ello, se crea un vector de datos reducido que contiene los elementos del tiempo de elución de cada señal cromatográfica, al cual se le asignan, dato a dato, sus

valores correspondientes de SRS mediante una función a intervalos que es transferida mediante interpolación lineal por segmentos (*spline*). De este proceso, se obtiene como resultado un nuevo vector en el cual se establece un nuevo dominio para todas las señales instrumentales cromatográficas, el cual se basa en SRSs independientes del instrumento y de los tiempos de retención. Por último, se implementa el método de pre-procesamiento 'remuestreo' (*resampling*) para ajustar el número de variables de los SRSs de cada muestra a un único número de variables común para todas ellas [**40**,**41**]. Como resultado de este proceso, se obtienen las huellas instrumentales agnostizadas de cada una de las matrices alimentarias analizadas.

Dentro de este marco contextual, a lo largo de esta tesis doctoral se busca desarrollar métodos analíticos multivariable basados en la metodología de huellas instrumentales obtenida por cromatografía de líquidos y herramientas quimiométricas, con la finalidad de establecer una metodología eficiente para transferir métodos analíticos multivariable entre distintos laboratorios dedicados al aseguramiento y control de calidad alimentaria y, así, lograr comparaciones interlaboratorio de resultados con mayor reproducibilidad y confiabilidad.

1.4. Desarrollo de métodos analíticos multivariable aplicados *in-situ*

Las técnicas analíticas tradicionales, tales como la cromatografía de gases y líquidos, han sido utilizados amplia y satisfactoriamente para el control de calidad en la industria alimentaria, debido a la exactitud y confianza de sus resultados [**43**]. Sin embargo, este tipo de métodos analíticos suelen ser llevados a cabo en instalaciones específicas (*off-line*), son generalmente caros y con tiempos de análisis de larga duración, requieren personal técnico altamente cualificado y, en la mayoría de los casos, reactivos no amigables con el medio ambiente. Para mejorar la eficiencia de estos métodos analíticos tradicionales en el control de la calidad alimentaria, se han desarrollado métodos analíticos alternativos (que fueron

[43] A.F. El Sheikha, Food authentication: Introduction, techniques, and prospects, in: C.M. Galanakis (Ed.), Food Authentication and Traceability, Academic Press / Elsevier, 2021, pp. 1–34.

denominados por el Prof. Miguel Valcárcel como métodos de vanguardia [**44**]), que se caracterizan por ser simples, rápidos, no destructivos ni invasivos, capaces de recolectar toda la información analítica de la muestra de manera confiable y representativa mediante instrumentos que suelen ser de pequeño tamaño y fácilmente transportables.

Dichos métodos de vanguardia, suelen estar basados en técnicas espectroscópicas que se caracterizan por la interacción directa y no destructiva de la radiación electromagnética (REM) con la muestra o materia bajo estudio, en las cuales es común obtener un vector conocido como "*espectro*", mismo que contiene toda la información de los compuestos de la muestra de manera inespecífica. Algunas de estas técnicas son las espectroscopías de infrarrojo medio con transformada de Fourier (FTIR, *Fourier-transform infrared*), de infrarrojo cercano (NIR, *near infrared*), Raman, de absorción molecular ultravioleta/visible (Uv/Vis), o de resonancia magnética nuclear (NMR, *nuclear magnetic resonance*), entre otras [**45**].

Una técnica emergente de vanguardia, según lo descrito anteriormente, con gran potencial para el aseguramiento y control de la calidad alimentaria es la denominada **espectroscopía Raman con sistema de compensación espacial** (**SORS**, ***spatially offset Raman spectroscopy***) [**46**], cuya utilidad ha sido comprobada ampliamente en la industria farmacéutica desde el 2007 [**47**], ya que permite realizar análisis *in-situ* rápidos y confiables de materiales envasados a través de contenedores opacos o poco transparentes. No obstante, su uso en la industria alimentaria ha sido escaso, siendo en 2011 cuando se utilizó por primera vez para estudiar el contenido de licopeno del tomate en sus distintas etapas de maduración [**48**] y, más recientemente, para el análisis y control de calidad de muestras de margarinas a través de sus contenedores originales [**39**].

---

[44] M. Valcárcel, S. Cárdenas, Vanguard-rearguard analytical strategies, 2005, Trends in Analytical Chemistry, 24, 67–74.

[45] A.S. Franca, L.M.L. Nollet, Spectroscopic Methods in Food Analysis, 2017, CRC Press.

[46] 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, 2005, Applied Spectroscopy, 59, 393–400.

[47] P. Matousek, A.W. Parker, Non-invasive probing of pharmaceutical capsules using transmission Raman spectroscopy, 2007, Journal of Raman Spectroscopy, 38, 563–567.

[48] J. Qin, K. Chao, M. S. Kim, Investigation of Raman chemical imaging for detection of lycopene changes in tomatoes during postharvest ripening, 2011, Journal of Food Engineering, 107, 277–288.

Las mediciones analíticas realizados con la técnica SORS se basan en los fundamentos de la espectroscopía Raman convencional [**49**,**50**], con la diferencia primordial que en la técnica SORS se obtiene un conjunto de espectros Raman de la muestra a una distancia determinada (medida en milímetros, mm) del punto de iluminación del láser, tal como se puede apreciar en la **Figura 9**. Dicho conjunto de espectros pertenece a las capas superficiales y subsuperficiales tanto del contenedor como de la muestra, mismos que, después de ser sometidos a un proceso de resolución espectral, son unificados en un solo espectro Raman, el cual pertenece a la muestra que se encuentra al interior del contenedor opaco [**46**,**51**].

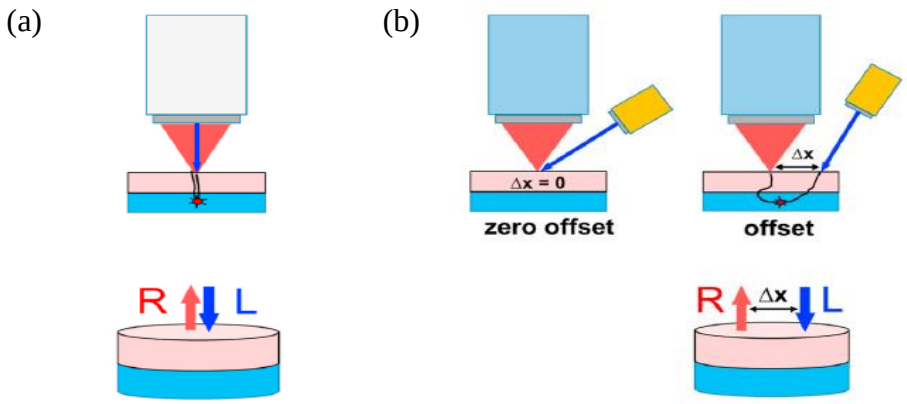


Figura 9. Esquema del funcionamiento de las técnicas de espectroscopia Raman (a) convencional y (b) con sistema de compensación espacial (SORS, *spatially offset Raman spectroscopy*)

*Figura adaptada de: 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, 2021, LWT-Food Science and Technology, 2021, 149, 111822 [**51**].*

[49] C.V. Raman, K.S. Krishnan, A new type of secondary radiation, 1928, Nature, 501-502.

[50] E. Smith, G. Dent, Modern Raman spectroscopy: a practical approach, second ed., Jhon Wiley & Sons, Inc., 2019.

[51] 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, 2021, LWT-Food Science and Technology, 2021, 149, 111822.

Con el uso de este tipo de nuevas tecnologías y metodologías descritas en los apartados anteriores, será posible mejorar y complementar los actuales procesos de control y aseguramiento de calidad de los productos alimenticios, los cuales están basados, mayoritariamente, en técnicas analíticas tradicionales. Del mismo modo, será posible el desarrollo de nuevos métodos analíticos multivariable que puedan ser implementados *in-situ*, aumentando la eficiencia en la detección de productos alimenticios auténticos y/o adulterados en puntos de consumo y/o venta. Así mismo, podrán desarrollarse métodos analíticos multivariable para implementarse *on-* o *in-line*, optimizando, modernizando y aumentando el control de calidad en la industria alimentaria.

Es por lo anterior que, el potencial de técnicas analíticas no invasivas, como lo son las técnicas SORS, FTIR y NIR, han sido exploradas en esta tesis doctoral para el posible desarrollo de métodos analíticos multivariable a ser aplicados de manera *in-situ*, *at-*, *on-* o *in-line* para el aseguramiento y control de calidad alimentario, específicamente, de bebidas destiladas.

1.5. Justificación e hipótesis

El uso de herramientas estadísticas multivariable o quimiométricas para el control y aseguramiento de la calidad de sus procesos y productos es una rutina hoy en día implantada en la industria farmacéutica. Estos procedimientos estadísticos pueden involucrar el tratamiento de datos basados en el uso de números individuales (enfoque univariable) o de vectores/matrices de datos (enfoque multivariable) relacionados directamente con la calidad del producto final. El enfoque multivariable es el de mayor uso y utilidad debido a la gran cantidad de información relevante que se obtiene del control de procesos. Sin embargo, a pesar de su recomendable uso, la industria alimentaria en general, y los laboratorios analíticos en particular lo utilizan en muy poca o nula medida para el aseguramiento de calidad de productos alimenticios. Esta falta de aplicación puede deberse al desconocimiento de su existencia, desconfianza o por la creencia infundada que son herramientas bastante complejas.

Este déficit de uso podría estar ocasionando una pérdida de valiosa información analítica que podría contribuir al mejoramiento continuo, aseguramiento de la calidad, aumento en la confianza de sus métodos analíticos y de sus resultados.

Contrariamente a las farmacéuticas, los laboratorios analíticos que trabajan en el ámbito alimentario aún siguen haciendo uso del enfoque univariable para el control y aseguramiento de la calidad de sus productos alimenticios. Dichas actividades suelen ser realizadas y analizadas de manera individual, cuando es posible tomar en cuenta múltiples datos para su mejor tratamiento y entendimiento mediante acciones sugeridas por la *Tecnología Analítica para Procesos* (PAT, *'Process Analytical Technology'*) [**52**] y la *Calidad Mediante el Diseño* (QbD, *'Quality by Design'*) [**53**], logrando obtener datos más significativos, métodos analíticos optimizados y productos alimenticios de mayor calidad.

Dado este contexto, es sumamente importante el desarrollo, actualización y adopción del enfoque multivariable en los laboratorios analíticos dentro de sus actividades de aseguramiento y control de calidad de productos alimenticios, por lo que esta tesis doctoral toma esa dirección y parte de la siguiente **hipótesis:**

> Es posible implementar el uso de herramientas matemáticas/estadísticas, recomendadas por organismos internacionales y utilizadas por la industria farmacéutica, en laboratorios del ámbito alimentario para desarrollar métodos analíticos multivariable capaces de facilitar el proceso de control y aseguramiento de calidad de productos alimenticios. La razón principal de ello reside en que, mediante el uso de un enfoque multivariable los procedimientos de producción y/o evaluación de calidad se podrían agilizar y optimizar.

[52] US FDA, Guidance for industry: PAT – A framework for innovative pharmaceutical development, manufacturing, and quality assurance, U.S. Food and Drug Administration, 2004.

[53] ICH Q8(R2), Pharmaceutical development, International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, 2009.

1.6. Objetivos

La presente tesis doctoral tiene como objetivo general:

> Desarrollar nuevos métodos analíticos multivariable cualimétricos que sean rápidos, confiables y eficientes para el aseguramiento y control de calidad de productos alimenticios, demostrando su aplicabilidad con el tequila, aceite de oliva virgen y virgen extra.

Este objetivo general se extiende en diferentes objetivos específicos que se plantean seguidamente:

(1) Desarrollar un método analítico multivariable basado en mediciones mediante espectroscopía de infrarrojo medio con transformada de Fourier (FTIR), fusión de datos y quimiometría para la diferenciación de categorías 100 % agave y mixto del Tequila Blanco.

(2) Desarrollar un método analítico multivariable basado en espectroscopía de infrarrojo cercano (NIR) y herramientas quimiométricas para diferenciar entre categorías 100 % agave y mixto del Tequila Blanco y predecir su contenido alcohólico.

(3) Desarrollar un método analítico multivariable para el control de calidad del aceite de oliva virgen y virgen extra, basado en huellas instrumentales cromatográficas agnostizadas y herramientas de reconocimiento de pautas supervisado y no supervisado.

(4) Desarrollar un método analítico multivariable (global) para el control de calidad del Tequila Blanco, basado en huellas instrumentales cromatográficas agnostizadas, obtenidas en diferentes laboratorios (España y México) en condiciones de reproducibilidad.

(5) Desarrollar un método analítico *in situ* rápido, confiable y no invasivo, basado en una técnica novedosa de espectroscopia Raman con sistema de compensación espacial (SORS) y herramientas estadísticas multivariable para la diferenciación entre Tequilas Blancos 100 % agave y mixto, y la predicción de su contenido alcohólico.

Capítulo 2

CONTROL DE PROCESOS EN LA INDUSTRIA ALIMENTARIA

2. CONTROL DE PROCESOS EN LA INDUSTRIA ALIMENTARIA

2.1. Resumen

Entre los años 80 y finales de los años 90 se suscitaron distintas crisis alimentarias dentro de la Unión Europea (UE) que tuvieron grandes repercusiones negativas, tanto económicas como humanitarias, de las cuales algunas de las más representativos fueron el síndrome del aceite tóxico o aceite de colza desnaturalizado (España, 1981) [**1**], encefalopatía espongiforme bovina, mejor conocida como enfermedad de las vacas locas (Reino Unido, 1986-1996) [**2**] o la presencia de dioxinas en pollos y huevos destinados al consumo humano (Bélgica, 1999) [**3**]. Debido a este tipo de incidentes alimentarios, la UE llevó a cabo el establecimiento de requisitos y principios generales relacionados con los alimentos y la alimentación, la cual fue denominada Legislación Alimentaria General por el Parlamento Europeo y del Consejo a través del Reglamento (CE) Nº 178/2002 [**4**]. Dicha legislación promueve *un nivel elevado de protección de la vida y la salud de las personas, así como de proteger los intereses de los consumidores, incluidas unas prácticas justas en el comercio de alimento;* la cual es aplicada mediante el control de procesos en todas las etapas de producción, transformación y distribución de los alimentos.

El control de procesos en la industria alimentaria engloba un conjunto de actividades enfocadas a mantener sus procedimientos en correcto orden y funcionamiento, conservando la calidad deseada de sus productos. Este control debe extenderse sobre la instrumentación utilizada, los métodos de análisis, la ingeniería, aspectos bioquímicos del alimento e incluso sobre los propios operarios.

[**1**] A. Segura Benedicto, J. Oñorbe de Torre, El síndrome del aceite tóxico, 2006, Revista de Administración Sanitaria, 4, 599-606.

[**2**] S. Torrades, La enfermedad de las vacas locas, 2001, OFFARM, 20, 110-116.

[**3**] J.J.F. Polledo, Contaminación por dioxinas en 1999: un fantasma atraviesa Europa, 2006, Revista de Administración Sanitaria, 4, 643-653.

[**4**] Reglamento (CE) Nº 178/2002 del Parlamento Europeo y del Consejo de 28 de enero de 2002, por el que se establecen los principios y los requisitos generales de la legislación alimentaria, se crea la Autoridad Europea de Seguridad Alimentaria y se fijan procedimientos relativos a la seguridad alimentaria.

Dichas actividades de control permiten disminuir la perdida de productos, incrementar la productividad de trabajadores y maquinas, incrementar la higiene del proceso y producción, disminuir los efectos de la variabilidad natural y caducidad de los alimentos, así como incrementar su calidad global [**5**].

Una manera de controlar procesos complejos de manera más eficiente es a través de un enfoque multivariable aplicando herramientas quimiométricas, ya que permiten controlar y monitorear un mayor número de variables (parámetros) del proceso y/o producto simultáneamente, tal como lo hace la industria farmacéutica desde inicios del siglo XXI. Dichas herramientas fueron sugeridas inicialmente por la Administración de Alimentos y Medicamentos de los Estados Unidos de América (USFDA, *Food and Drug Administration*) a través de una guía recomendada para la industria farmacéutica en la cual se sugería el uso de un marco novedoso denominado Tecnología Analítica de Procesos (**PAT**, ***Process Analytical Technology***) [**6**]. Posteriormente, las mismas herramientas quimiométricas fueron recomendadas por la Conferencia Internacional sobre Armonización de Requisitos Técnicos para el Registro de Productos Farmacéuticos para Uso Humano (ICH, *International Conference on Harmonization of Technical Requirements for Pharmaceutical for Human Use*) a través de su guía ICH Q8(R2) [**7**] en la cual se incitaba al sector farmacéutico a la obtención de calidad de sus productos mediante el diseño (**QbD**, ***Quality by Design***) y a adoptar el PAT como concepto propiciador de la QbD.

Una de las herramientas quimiométricas recomendadas en las guías citadas anteriormente y de las más utilizadas en la industria farmacéutica es el análisis de componentes principales (PCA, *principal component analysis*), la cual es empleada para monitorear en tiempo real diferentes variables o parámetros de procesos. No obstante, su aplicación real en la industria alimentaria ha sido poco explorada y las aplicaciones que existen se realizan de manera confidencial en la mayoría de las empresas, por lo que hay verdaderamente escasos casos reportados en literatura.

[**5**] G. Trystram, F. Courtois, Food processing control – Reality and problems, 1994, Food Research International, 27, 173-185.

[**6**] US FDA, Guidance for industry: PAT – A framework for innovative pharmaceutical development, manufacturing, and quality assurance. U.S. Food and Drug Administration, 2004.

[**7**] ICH Q8(R2), Pharmaceutical development, International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, 2009.

En este sentido, en el marco de la tesis doctoral se desarrolló un artículo de revisión en el cual se pone de manifiesto el estado actual de distintas herramientas quimiométricas aplicadas en la industria alimentaria enmarcadas en un contexto de QbD y PAT, mismo que se detalla a continuación.

2.2. Artículo científico I

Food Engineering Reviews (2023) 15:24–40
https://doi.org/10.1007/s12393-022-09324-0

QbD/PAT—State of the Art of Multivariate Methodologies in Food and Food-Related Biotech Industries

Christian H. Pérez-Beltrán[1] · Ana M. Jiménez-Carvelo[1] · Anabel Torrente-López[1,2] · Natalia A. Navas[1,2] · Luis Cuadros-Rodríguez[1,2]

Received: 26 May 2022 / Accepted: 7 October 2022 / Published online: 24 October 2022

Abstract
Several investigations have been made at lab scale considering the quality by design (QbD) and process analytical technology (PAT) approaches. Nonetheless, such applications have been focused on the analyzers or multivariate tools used at small scales. This comprehensive review presents the state of the art of both QbD and PAT. In addition, the key historical events since 1940 which have influenced the development of the QbD/PAT system are also highlighted. Moreover, the application of the recommended PAT tool of *multivariate tools for design*, *data acquisition*, *and analysis* (design of experiments, multivariate data analysis, and multivariate process control) is revised for the food and food-related biotechnology industries and describes the applications reported over the last 20 years. On this subject, only 34 studies were found in literature whose relation was close with both industries at industrial or pilot plant scales; a description of each of them focusing on multivariate tools is presented. Finally, some conclusions and future perspectives on this topic are given, with the aim of initiating a change in the field.

Keywords Quality by design · Process analytical technology · Design of experiments · Multivariate data analysis and process control · Food and food-related biotech industries

Introduction

During the initial steps in the design of the infrastructure for industrial quality, it was assessed once the product was finished, and, paradoxically, this goal is still common in many cases today. According to Koch [1], it was around the 1940s in Germany, after the World War II, when the concept of *process analytics* (PA) or *process analytical chemistry* (PAC) were proposed in the framework of the chemical and petrochemical industries. The distinctive features of this new approach were explained in detail in an outstanding and pioneering tutorial published by Callis et al. [2]. In such industries, PAC was implemented as the chemical or physical analyses of materials carried out during the elaboration process to understand the composition of molecules of interest, which was popularized for the following twenty years and adopted by refineries and nuclear plants.

Whilst PAC was being implemented, the quality trilogy was stated for the first time by Juran [3], work in which a new direction for managing quality was proposed through quality planning, control, and improvement. Later on, such processes gave rise to the *quality by design* (QbD) concept in 1992 by Juran [4], where the importance of planning the quality sought after by customers and consumers was outlined and explained through the use of the quality trilogy. After the development of the QbD Juran's approach, the Food and Drug Administration of the United States of America (US FDA) introduced in 2004 the concept of *process analytical technology* (PAT) [5] for its use in the pharmaceutical industry due to the loss of credibility of the industry before this time. The introduction of PAT was based on PAC, modifying the term *chemical* to *technology* and including more features, like microbiological, mathematical, and risk analysis, to enhance understanding and to control the production process. According to the above FDA guidance, PAT is considered *a system for designing*,

✉ Christian H. Pérez-Beltrán
christianpb@correo.ugr.es

[1] Department of Analytical Chemistry, Faculty of Science, University of Granada, c/ Fuentenueva s/n, E-18071 Granada, Spain

[2] Biohealth Research Institute (ibs.GRANADA), University of Granada, Granada, Spain

QbD/PAT—State of the Art of Multivariate Methodologies in Food and Food-Related Biotech Industries

Christian H. Pérez-Beltrán[1*], Ana M. Jiménez-Carvelo[1], Anabel Torrente-López[1,2], Natalia A. Navas[1,2], Luis Cuadros-Rodríguez[1,2]

[1] Department of Analytical Chemistry, Faculty of Science, University of Granada, c/ Fuentenueva s/n, E-18071, Granada, Spain.

[2] Biohealth Research Institute (ibs.GRANADA), University of Granada, Granada, Spain.

*Corresponding author email: christianpb@correo.ugr.es

Keywords:

Quality by design; Process analytical technology; Design of experiments; Multivariate data analysis and process control; Food and food-related biotech industries

1. Introduction

During the initial steps in the design of the infrastructure for industrial quality, it was assessed once the product was finished, and, paradoxically, this goal is still common in many cases today. According to Koch (1999), it was around the 1940s in Germany, after the World War II, when the concept of *process analytics* (PA) or *process analytical chemistry* (PAC) were proposed in the framework of the chemical and petrochemical industries. The distinctive features of this new approach were explained in detail in an outstanding and pioneering tutorial published by Callis et al. (1987). In such industries, PAC was implemented as the chemical or physical analyses of materials carried out during the elaboration process to understand the composition of molecules of interest, which was popularized for the following twenty years and adopted by refineries and nuclear plants.

Whilst PAC was being implemented, the quality trilogy was stated for the first time by Juran (1986), work in which a new direction for managing quality was proposed through quality planning, control, and improvement. Later on, such processes gave rise to the *quality by design* (QbD) concept in 1992 by Juran (1992), where the importance of planning the quality sought after by customers and consumers was outlined and explained through the use of the quality trilogy. After the development of the QbD Juran's approach, the Food and Drug Administration of the United States of America (US FDA) introduced in 2004 the concept of *process analytical technology* (PAT) (US FDA, 2004) for its use in the pharmaceutical industry due to the loss of credibility of the industry before this time. The introduction of PAT was based on PAC, modifying the term *chemical* to *technology* and including more features, like microbiological, mathematical, and risk analysis, to enhance understanding and to control the production process. According to the above FDA guidance, PAT is considered *a system for designing, analyzing, and controlling manufacturing through timely measurements (i.e., during processing) of critical quality and performance attributes of raw and in-process materials and processes, with the goal of ensuring final product quality*. The same definition is also included in the ASTM E2363-14 (2014) standard.

The early PAT implementation was based mainly on analytical chemistry and was focused on the development and application of at-line/on-line analytical instruments, particularly process analyzers, as evidenced in Koch et al. (2007) and Bakeev (2010). Thus, the PAT concept was more linked to the use of analyzers than with the implementation of their paradigms. This fact caused confusion for some time, to the extreme of erroneously naming many analytical instruments as PAT analyzers or PAT tools. This is why a description of such instruments or their applications is not included in the paper as it is out of the scope of this review.

Few years after the acceptance of the PAT approach by the FDA, the International Conference on Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH) produced the ICH Q8(R2) guideline (ICH 2009), which retrieves the term QbD and adopted PAT definition and concept for the pharmaceutical industry. The ICH Q8(R2) defines QbD as *a systematic approach to development that begins with predefined objectives and emphasizes on product and process understanding and process control, based on sound science and quality risk management*. One important aspect to consider when using QbD is the design space, defined as *multidimensional combination and interaction of input variables and process parameters that have been demonstra*ted to provide assurance of quality (ICH 2009; Orlandini et al. 2013). The process should constantly work under the designed space, which allows for continuous quality improvement; but when deviations occur out of it, there is considered to be a change, and possible sources of no conformities should be valued (ICH 2009). This usually launches a post-approval requirement change process, which is aimed at establishing a new design space.

The aforementioned key historical events are summarized in Fig. 1 in which each action is represented by a black horizontal line in the rising arrow, representing a timeline and at the same time the way knowledge in quality has been increasing through the coming years represented by the shape of the arrow. A similar complementary historical review of PAT during its beginnings was performed by Chew and Sharratt (2010).

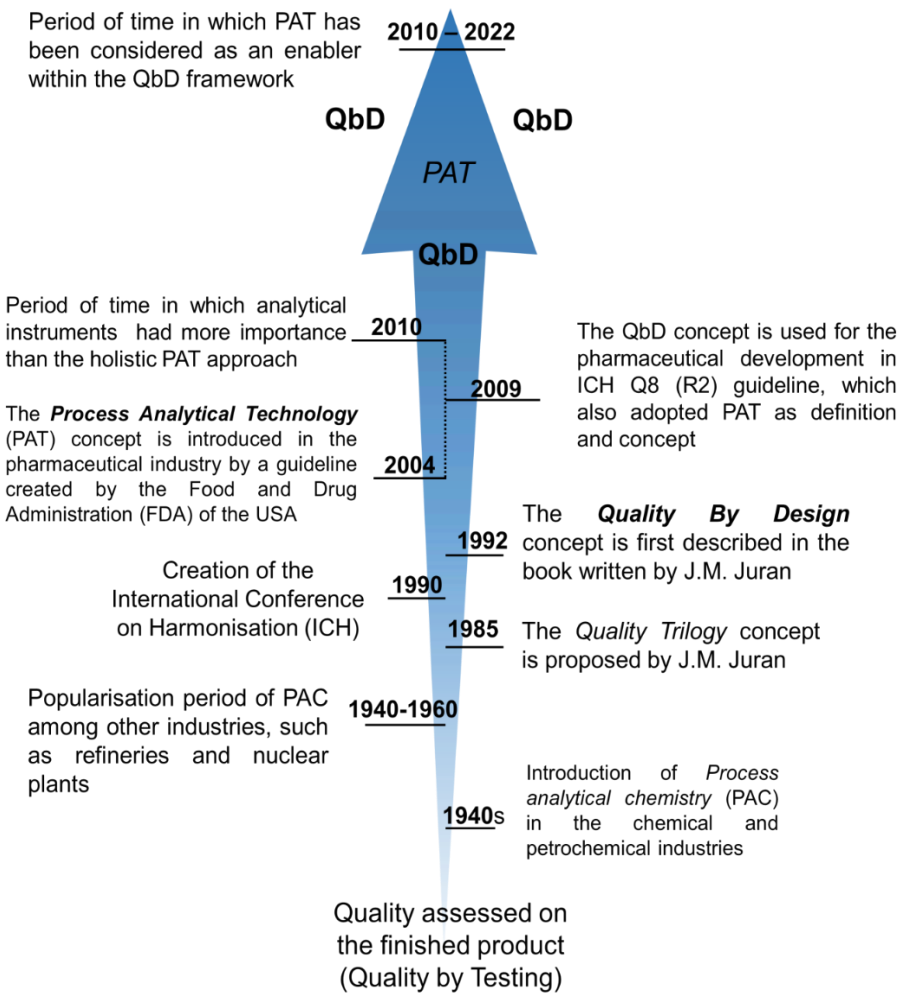


Fig. 1 Creation and evolution of the quality by design (QbD) and Process Analytical Technology (PAT) concepts

In this regard, QbD and PAT share similar targets, being product and process understanding and process control. Nonetheless, QbD is considered the framework in which PAT is applied to achieve the goals previously mentioned. The QbD approach starts with the definition of the quality target product profile (QTPP). Thus, PAT is currently considered as an enabler whose analytical instruments and strategies make possible the creation of easier and smarter control plans, identification of critical process parameters (CPP) and detection of important interactions between process parameters, and relevant critical quality attributes (CQA) of the material to obtain the desired quality by designing it through the process with real-time or near real-time measurements.

These measuring strategies are characterized by being performed during the process using analytical devices capable of non-destructive analysis, which is one of the PAT elements.

Besides the implementation of analytical devices, the PAT system encourages the inclusion of other elements (FDA 2004), such as (i) multivariate tools for design, data acquisition, and analysis; (ii) process control tools; and (iii) continuous improvement and knowledge management tools. A conceptual diagram displaying the QbD framework and the different PAT tools used to achieve the ultimate goal can be observed in Fig. 2.

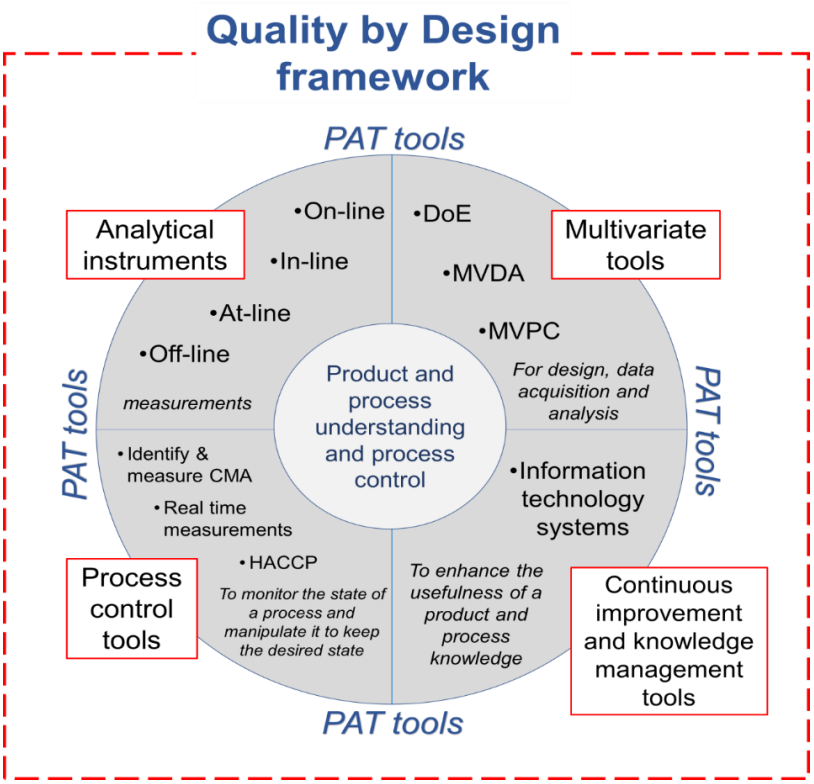


Fig. 2 Conceptual diagram of the quality by design (QbD) framework in which different process analytical technology (PAT) elements and tools are considered to achieve the main central objective of understanding the process and product, and the control of the process. DOE, design of multivariate experiments; MVDA, multivariate data analysis; MVPC, multivariate process control; CMA, critical material attributes; HCCP, hazard analysis critical control points

In this sense, the analytical instruments and process analyzers for analytical control used in the PAT system can provide different measurements, which are divided in on-line, in-line, at-line and off-line analyses according to the way in which they are performed (FDA 2004; ASTM E2363 2014).

On-line measurements consist on those determinations where the sample is deviated from the manufacturing process line and then returned to the process stream. For the in-line measurements, the sample is not removed from the manufacturing line, and it can be invasive or non-invasive. These types of measurements are performed by different analyzers or sensors, which can be implemented individually or in conjunction during the same process, producing different types of data that need to be treated and fused. This data fusion (DF) procedure is the current challenge and the next step in the evolution of PAT that could provide a more comprehensive understanding of the system and the opportunity to predict complex quality attributes, as described by Casian et al. (2022). In fact, the authors sate that most of DF applications can be found in the food sector, where it was used for food quality authentication, food and beverage characterization, and food quality assessment.

As it may be inferred, the aforementioned measurements produce enormous volumes of data, which need to be treated with different multivariate tools to mine the relevant and non-evident information. The main multivariate tools to be applied are design of multivariate experiments (DOE), multivariate data analysis (MVDA), and multivariate process control (MVPC), which are briefly described below.

The first described tool is DOE, which according to the ASTM (ASTM E2363 2014), it makes reference to *the arrangement in which an experimental program is to be conducted, and the selection of the levels (versions) of one or more factors or factor combinations to be included in the experiment.* The quality of a product is linked to its production process, which must be all the time under control and the defined design space. When the pre-defined design space is modified, new requirement approvals using characterization studies must be performed over the new process to confirm that it can operate and perform properly yielding the desired product quality. Thus, DOE promotes innovation, problem solving, and discovery, and it is used to screen a process or response, create a space design, optimize a response(s), and study the interrelationships among multiple factors of interest (ASTM E456-13A 2017; ISO 3534-3 2013). A similar approach for the process control, optimization and creation of future process trajectory is known as model predictive control (MPC). However, this model is not based on multivariate experimental designs, but in the

reduction of the squared error of the process trajectory over the time horizon, minimizing the short-term effects of unknown and erratic signals, manipulating multiple input variables (Boudreau and McMillan 2007). Nonetheless, only DOE is reviewed in this work, since it is the considered tool within the PAT system. The reader is kindly referred to literature (Boudreau and McMillan 2007) to find out more about MPC.

The second methodology, MVDA, is described by ASTM as *an appropriate tool for exploring and handling large sets of heterogenous data, mapping data of high dimensionality onto lower dimensional representations, exposing significant correlations among multivariate variables within a single data set or significant correlations among multivariate variables across data sets* (ASTM ES891-20 2020). MVDA is a key enabler for process understanding and decision making, and for the release of intermediate and final products after being appropriately validated using a science and risk-based approach. It involves the use of multivariate data regression and dimension reduction that promote the development of different unsupervised and supervised multivariate models to explore/screen the data, to analyze the similarity of data set, or to predict (classify/quantify) qualitative features or physicochemical characteristics of products from the data.

The third multivariate tool is MVPC, which is described by the ISO 7870-7 (2020) as *a process monitoring of problems in which many related variables are of interest using multivariate control charts as the main mechanism to statistically evaluate and control the process*. Based on this statistical evaluation and control, deviations can be identified through statistical signals, which make feasible the elimination of the possible cause of variation, bringing back the process to the state of statistical control to predict the performance and assess the capability to meet the future specifications.

From the different tools employed in the PAT system, special attention is paid to multivariate tools for design, data acquisition, and analysis. On this subject, this paper focuses on the multivariate tools for QbD/PAT applied both at industrial plant and pilot plant scales. Specifically, the application of DOE, MVDA, and MVPC in the food industry is described and reviewed in the “QbD/PAT Tools in the Food Industry” section, as well as in the food-related biotech industry in the “QbD/PAT Tools in the Food-Related Biotech Industry” section, considering the studies developed within a time span of 20 years.

The pharmaceutical industry is not considered here since the QbD and PAT principles were concisely developed for it, and thus, a huge number of reports, applications and discussions on the subject can be found in literature. Finally, the “Conclusions and Final Remarks” section intends to make evident the lack of QbD/PAT principles in the food and food-related biotech industries at industrial scale, encouraging a closer interaction among academic researchers and industry. This section also contains some ending remarks about the current review and future perspectives for both industrial sectors.

Note that the studies described in this review were selected considering, mainly (1) the scale at which they were performed—industrial or pilot plant scale—but some lab-scale studies were also considered because they involved the directly cooperation with the industry or company, but at once, (2) they had to explicitly mention that take into account QbD as a framework whose principles and goals must be achieved with PAT tools. Such studies were sought in the Web of Science and Google Scholar research tools, using “QbD”, “PAT”, “food”, “biotechnology”, “industry”, “industrial-scale” and/or “pilot plant” as principal search words.

2. QbD/PAT Tools in the Food Industry

The QbD/PAT bases and tools are still rarely implemented in the food industry. This may be because each company has its preferred data acquisition methods and processes, from which may assess the quality of products off-line using laboratory-based analytical methods, many of them destructive in nature, time-consuming, costly in terms of complex sample preparation, and require highly skilled operators, making the implementation of a PAT system difficult, as pointed out by Teixeira et al. (2014) and Hitzmann et al. (2015). Although QbD/PAT tools have been demonstrated to have a huge impact for improving the process understanding and control, and the end-quality of a product as evidenced in different research reports and review papers (Rathore and Kapoor 2017; Rifna et al. 2022; Khan et al. 2012; Misra et al. 2015), they have been applied to a limited extent at industrial scale in the food sector. The next subsections expose those cases in which PAT multivariate tools are applied in the food and food-related biotechnology industries, considering both industrial plant and pilot plant scales.

2.1. Design of Multivariate Experiments (DOE)

DOE is characterized by determining the relationships between input factors that promotes variation over the output responses. Only two studies with similar industrial-scale operating conditions were found in literature, which are presented in this subsection.

Fissore et al. (2014) investigated the design space of a freeze-drying process of coffee extract. The study was focused on the primary drying of the process, since it accounts for most of the energy consumption. The design space was constituted by a set of operating conditions named heating fluid temperature (T_{fluid}) and chamber pressure (P_c). This research group found that, if only drying time is considered for a cycle with constant values of T_{fluid} and P_c, the values that should be used, according to the design space, are high values of shelf temperature and low values of chamber pressure (−5°C and 5 Pa).

The second most similar study, in which DOE was applied at industrial scale, was recently performed by Tessarini et al. (2021). The authors developed and optimized the beer containing malted and non-malted substitutes under a QbD framework. For this study, a simplex-centroid mixture design was employed in which different proportions of maize, oat, malt, rice, rye malt, and sorghum cereals were tested. The starting QTPP of the Ale beer was 45% of barley malt substitutes as maximum, with 5.5 ± 0.55% v/v of alcohol content. CQA were color and pH, since they are directly related to the sensory part of the QTPP. Afterwards, a screening design took place where maize, oat, and rice were selected as barley malt substitutes for the optimization design. The optimization design consisted on performing a new mixture design with the three barely malt substitutes used to manufacture new formulations. The results reported that the following parameters should be considered to create an optimized formulation similar to the Ale beer: (i) employ high proportions of barley and oat malt (55 and 42%, respectively) and 3% of maize in the manufacturing process, (ii) use malt substitute for a greater level of color on the “European Brewery Convention” (EBC) scale or non-malted to lower it, and (iii) add higher concentrations of non-malted substitutes to reach optimal sensory scores.

The usefulness of DOE at the industrial scale in the food industry is demonstrated with the two previously discussed studies. In the same regard, it is also evident the lack of studies at industrial scale applying DOE, which shows that this PAT tool is still novel in the

development of food products or that industries do not require it, since they already have well-established quality characteristics of their products. Nonetheless, DOE could be used to enhance the quality features of existing products and to reduce costs.

2.2. Multivariate Data Analysis (MVDA)

The potential of MVDA lies in its ability to extract meaningful information from complex data in a fast and easy manner. In fact, more studies applying MVDA than DOE were found in literature. Different techniques and foodstuffs are discussed in the following paragraphs.

The first study described in this subsection was performed by Xing et al. (2007), who investigated the potential of visible spectroscopy for the classification and the color attribute prediction using as CQA the coordinates based on the CIE *L*A*B** color space (*L**, lightness of the color of the sample; *A**, red and green characteristics; and *B**, yellow and blue characteristics). The experiments were conducted with a total of 189 samples at room temperature between 20 and 22°C.

A stepwise discriminant analysis (STEPDISC) is performed to select a subset of the quantitative variables for use in discriminating among the classes. Canonical discriminant analysis (CDA) method was used to classify the samples into different groups. The partial least squares (PLS) regression was used to predict color attributes based on the reflectance spectra. Preprocessing methods such as Norris 1st derivative, Savitzky–Golay (SG) 2nd derivative, mean normalization (MN), and range normalization were tried. The authors found significant difference among pale and red meat classes according to their mean reflectance spectra and to the Duncan test. Thus, they decided to classify the meat samples into pale and red classes. For the development of the first CDA model, four wavelengths (420, 440, 580, and 620 nm) were used to differentiate among the two classes of meat, obtaining a classification accuracy of 85%. Additionally, a discrimination model to differentiate between pale, firm, and nonexudative (PFN) meat and pale, soft, and exudative (PSE) meat classes was created with five wavelengths (420, 580, 440, 550 and 600 nm). The accuracy of the model for the PFN class was 82% whilst for the PSE class was 84%. Finally, a PLS model was developed using the full wavelength of the reflectance spectra to predict the color of the meat classes. Indeed, the model was capable to predict better the *L** coordinate than *A** and *B** ones, using as preprocessing technique the SG 2nd derivative.

Nonetheless, the authors commented that further testing was needed to assess the effectiveness of the prediction model.

In the same subject of meat, Sørensen et al. (2012) reported a method using spatially resolved near-infrared (NIR) spectroscopy to determine the fat quality of porcine carcasses by estimating the iodine value (IV). The study included 35 carcasses from a slaughterhouse, which were from a daily production stock and belonged to three different categories based on feeding regimes to ensure variability in the composition of the fatty acids. The NIR spectra were obtained 5 cm from the belly split line and 5 cm from the neck separation cut. Reference analyses were performed using gas chromatography of 287 disks of fat and skin collected close to the spot of spectroscopic measurements. Principal component analysis (PCA) was used to assess the GC results and screen whether the feeding groups could be distinguished based on their fatty acid composition. PLS was applied and used to predict the measured IV from the recorded spectra after mean centering and extended multiplicative signal correction (EMSC) and interval partial least squares (iPLS). The authors found that, combining EMSC and iPLS, the model yielded good results for IV predictions. Finally, the model was applied to predict the IV as a function of fat layer depth. The authors were able to measure the fat quality and fat layer quality differences of porcine meat at full abattoir processing speed with an on-line NIR transmission spectroscopy method refined with PCA, preprocessing techniques, and PLS.

Moreover, Achata et al. (2019) also studied pork meat, but hyperspectral imaging (HSI) was used in this occasion. The authors used HSI with chemometrics to develop classification models of brined and non-brined pork loins and to predict the salt concentration used to brine the samples. For this study, 144 fresh pork loins (FPL), which were obtained from 16 animals, were brined using salt solutions concentrated at 5, 10, and 15% (w/v). The hyperspectral images were obtained in the Vis-NIR and NIR ranges at 20 ºC from both sides of the samples. The authors used PCA, PLS-DA and PLS in combination with several pre-treatments on both reflectance and logarithm transformed data. In this sense, researchers could observe that samples grouped according to the experimental treatment they suffered; moreover, the authors obtained a classification model capable to correctly discriminate all brined and non-brined samples using the 957–1664 nm

spectral range and quite good prediction models to estimate the salt concentration of raw and cooked FPL using the 957–1664 nm spectral range.

Furthermore, Pullanagari et al. (2015) performed on-line quantifications of fatty acids (FA) of lamb meat under commercial abattoir conditions by means of visible-near-infrared spectroscopy (Vis-NIR), since they influence in the determination of the meat quality. The experiment consisted of the Vis-NIR analyses of 500 lambs from two commercial lamb processing plants. A PLS-based genetic algorithm (PLS-GA) was applied and is capable to compute and find optimization solutions useful for informative wavelength selection. Half of the 500 samples were randomly included in the calibration of the GA-PLS model; the rest were employed to validate it. The authors found that the majority of the studied FA and monounsaturated FA (MUFA) could be predicted with moderate accuracy. They concluded that, although the exact quantification of FA was not absolutely reliable with such procedure, it is evident that it could be adequate and implemented for screening purposes as part of a quality control process in the food industry.

Moreover, Lintvedt et al. (2022) performed in-line measurements using Raman spectroscopy and PLS regression to estimate the concentration of eicosapentaenoic + docosahexaenoic acids in 63 salmon samples from three farming locations and the residual bone concentration (% ash) in 66 samples of mechanically recovered ground chicken, provided by a processing plant. The samples were measured with similar operating conditions to those used in the industry. Each sample was placed on plate covered with aluminum foil and passed through a dark cabinet using a conveyor belt at speeds 0.30, 0.15, 0.075, and 0.03 m/s, corresponding to exposure times of 1, 2, 4, and 10 s, respectively. The authors were able to obtain a model with good and acceptable results of prediction, they found that appropriate data preprocessing is fundamental to reduce the noise of the spectra, allowing them to acquire good results with models built using exposure times of 1 or 10 s.

Ørnholt-Johansson et al. (2017) interpreted the data from companies to evaluate if MVDA could increase their production yield. For this research, 60 Atlantic salmons (Salmo solar) from three slaughterhouses in Norway were used. The authors found a meaningful weight loss through the processing of the fishes. They detected that mechanical filleting provided the initial weight difference. However, through the study of the data, the authors concluded

that the main cause for the weight loss was the cut that divided the fillets from the skeletal frame. From these observations, three different PLS prediction models were built to estimate the yield after mechanical filleting. After comparing the three models, the authors decided to keep model two, which was built with variables shape ratio, W/LT, weight divided by length and thickness; K factor (W/L^3), weight divided by the cubed length, length, thickness, and weight. This model allowed the authors to differentiate salmons according to the slaughterhouse, using score and correlation loading plots, and allowed them to find that salmons from companies 2 and 3 were more similar between them than salmon from company 1. Finally, PCA was applied to study 13 deviating samples from the PLS model, discovering that two groups of salmon could be found according to the cut on their belly (straight or angled belly cut). The findings of this research ended up with valuable information that can help production companies in the decision-making process.

MVDA has also been particularly applied in the dairy sector. Lyngard et al. (2012) performed a real time modeling of milk coagulation during the coagulation of twelve cheese batches to attain meaningful insights on this process through the application of near infrared (NIR) reflection measurements, formulation and testing of models. Twelve milk coagulation experiments were performed using 5 L of reconstituted milk, which were transferred to a 6 L of cheese vat imitating an industrial setting with normal operation conditions (NOC). Two PCA models were created and assessed according to the NIR measurements. The first model monitored the entire coagulation process with an s-shaped profile, involving enzymatic proteolysis of k-casein, paracasein aggregation, and gelation. The second model focused on each coagulation phase to obtain a more robust model for real-time use. The first model presented an excellent performance, but it was not appropriate for industrial screen of on-line parameters due to large variations in some estimation, caused by small alterations in the NIR measurements. The second one also performed well, and it led to acceptable parameters fittings, which make it a good option for a real-time application.

Following up with the dairy industry, Rimpiläinen et al. (2015) tested a data-driven approach in an industrial-scale powder plant to predict and evaluate the end-point properties of milk powder through one function property known as sediment.

A standard offline laboratory sediment test was performed to four consecutive production processes with 339, 300, 273, and 284 samples. Thirteen plant variables were examined to make possible the evaluation and implementation of a real-time quality control process. A prediction model was built based on conditional probability distributions (CPDs) using the first 75 sediment samples as training data set and then used to predict only the next sediment value, updating the model after every new sediment result. The prediction results showed that production processes 1, 3, and 4 were consistent with the established Gaussian assumption, but for process 2, the sediment results did not fit very well. It was observed that process pairs 1–2 and 3–4 were similar, indicating a possible change in the operating conditions of the procedure. According to the obtained results, the authors claim that sediment values could be controlled and decreased in each process. For production process 1, dryer temperatures (T_{D1} and T_{D2}) should be increased; in process 2, direct contact heater temperature (T_{DC}) should be also increased; in process 3, steam injector temperature (T_{SI}) should be decreased and T_{DC} increased; and for process 4, T_{SI} should be decreased. However, when all processes were considered together, T_{SI} and milk concentrate temperature (T_C) had the highest influence on sediments. The authors found that, the average prediction error levels of the CPD model were comparable with the PLS model. Nonetheless, they decided to keep the CPD model since it offered them a good manner to evaluate the influence of each predictor variable over the sediments. In this sense, it was verified that MVDA can be used to develop prediction models to find suggestions on how to adjust plant variables to improve the sediment values.

MVDA has also been applied in the olive oil industry, where Tamborrino et al. (2017) studied the effect of calcium carbonate during the extraction process over the olive oil quality, energy consumption, and rheological properties to improve the extraction process adjusting malaxation parameters. The olive pastes and olive oil were obtained with a continuous olive oil mill plant. The energy consumption was assessed with measurements of the active, reactive, and apparent power. The electricity consumption was calculated regarding the rates of operation. Additionally, the viscosity of the pastes was measured to obtain their rheological properties. Once the chemical results were obtained, they applied PCA to make evident the main variables that affected oil samples.

Researchers decided to use the effect plot on the volatile compounds and trans-2-hexanal to show the experimental trials trend.

Furthermore, Picouet et al. (2019) used the QbD approach to predict the final acrylamide content of deep-fried potatoes “chips”, which is a parameter related with safety. The final chips were defined according to twelve quality target parameters (QTPs), including 3 color coordinates (CIE *L*A*B**), 5 sensory attributes (odor roast, flavor rancid, flavor roast, crunchy, and oil mouth feel), 3 concentrations of volatile compounds (hexanal, pentylfuran, and 2.4diecadenal), and total acrylamide content. In this case, the utilized MVDA tool was a classical least-squares multilinear regression (MLR) coupled to a step-wise model, which was created with 65 frying experiments for the calibration set and 33 for the validation set, using a mid-level fusion approach of four different QTPs parameters (color coordinate *A**, “flavor roast” sensory descriptor, acrylamide and pentylfuran contents). The results suggested that the predictive models for the acrylamide content were unsatisfactory, since it is still not clear some complex mechanisms and factors that influence the quality parameters of the potato chips.

Two recent applications of MVDA in the food industry were performed using hyperspectral imaging (HSI). Liu et al. (2017) evaluated the potential use of HSI to detect sucrose adulteration in tomato paste, to compare the detection and prediction performance of different chemometrics methods, and to identify the lowest proportion of sucrose in tomato paste that can be safely detected. The study consisted of the multispectral analysis at 19 wavelengths of two batches of pure concentrated tomato paste provided by the industry. The tomato pastes and sucrose mixtures were made at 1–9% proportion levels (w/w). The authors were able to observe two clear groupings of batch 1 and batch 2 of tomato pastes using PCA. Furthermore, authors quantified the level of adulteration in tomato paste using calibration models generated by PLS, least squares-support vector machine (LS-SVM), and back propagation neural network (BPNN). Researchers found that LS-SVM provided the best predictive results for both batches of tomato pastes. In this study, 100% accuracy was obtained in the prediction set with a detection limit of sucrose of 1% using HSI with multivariate methods.

Additionally, HSI was applied over milk powders by Munir et al. (2018) to determine if it could be used as a process analyzer for the real-time quality control, coupled to a predictive regression model. The whole milk powder samples were obtained from three different factories with the same specifications, but with some equipment differences.

The authors studied the effects of pre-processing the signal either by smoothing and/or differentiating it, and they applied PCA and PLS as multivariate analysis methods to find possible trends among the powders and to construct a model capable to predict the origin of an unknown powder, respectively. In this sense, the authors demonstrated through PCA that a high degree of smoothing is a suitable preprocessing step capable to maximize the differentiation performance among the three factories, and between "poor" and "good" quality milk powders. Moreover, the constructed PLS model yielded 79–87% of accuracy regarding the dispersibility predictions related with "good" and "poor" powders. With the current study, the authors developed an analytical method using hyperspectral imaging and MVDA capable to distinguish among milk powders from different factories with diverse qualities and properties, which could be implemented on-line by the industry.

Continuing with the application of MVDA in the food industry, Moscetti et al. (2019) developed, optimized, and predicted the desalting process by electrodialysis of soy sauce through the application of on-line NIR spectroscopy, level and conductivity probes, and a control strategy of the electric current generator. The raw soy sauce was analyzed in a laboratory-scale electrodialyzer plant in which monitoring and controlling activities were performed. The experiments were performed at four different electric current profiles, where only the first one was employed as the calibration set and the other three as prediction sets. The concentration of salt, non-salt solids, and amino nitrogen were determined with prediction models based on NIR spectroscopy to completely constitute the desalting process. The PLS model was performed with the 1100–1925nm spectral range, over which standard normal variate (SNV), SG, and mean centering (MC) pre-processing steeps were applied, obtaining good predictions for the salt and non-salt concentration of the desalting process of soy sauce, as well as for the conductivity, osmotic pressure, and density.

Another study was performed by Lan et al. (2022), who compared the ability of NIR, mid-infrared (MIR), and Raman spectroscopies, and HSI to assess the composition and texture characteristics of apple purees produced in-house that mimic an industrial process. In this study, 62 samples from two different processes were analyzed and further used to build predictive and discriminative models, which were PLS, SVM, and random forest (RF).

Several classification models were developed to discriminate among five characteristics of apple puree; prediction models were built to foresee eight rheological and structural properties and nine biochemical properties of the apple purees. Researchers found that the MIR technique coupled with RF and SVM had a higher discrimination accuracy of purees than the PLS-discriminant analysis (PLS-DA) and that NIR coupled with PLS resulted in better predictions of the quality puree parameters than the SVM and RF quantitative models.

MVDA has also been applied for the production of beer. Tessarini et al. (2020) proposed a real-time monitoring process of beer parameters applying infrared spectroscopy with MVDA to predict the final quality of beer. The beer formulations were manufactured in a pilot plant where samples were collected from mashing, fermentation, and maturation processes for further physicochemical analysis using attenuated total reflection-FTIR (ATR-FTIR) spectroscopy. In this case, the PLS regression model was used to predict the final alcohol content, density, pH, and color, obtaining acceptable results. Furthermore, the predicted values for each manufacturing stage were contrasted with their corresponding experimental value using analysis of variance (ANOVA), which indicated that no statistical difference was observed among them. As a matter of fact, the authors could predict the desired quality of the finished beer through the manufacturing process, making possible the identification of deviations in the system, taking preventive or corrective actions if necessary.

Moreover, Schorn-García et al. (2021) developed a PAT-based methodology to monitor and control a wine alcoholic fermentation process using spectroscopy and MVDA. Five alcoholic fermentations of *Saccharomyces cerevisiae* were performed and analyzed on-line and at-line to obtain the reference values. The authors built a PCA model to observe the trend of the density evolution during the fermentation process, which was easily followed

by the plot elaborated with the first PC against time. Moreover, a PLS model was performed to predict the density (g/mL) along the alcoholic fermentation and to predict the biological time of the process, obtaining high and good correlation values between the predicted and the measured values. Furthermore, the authors developed a PLS-DA model to determine if the samples were under or out of control, which was built using five normal alcoholic fermentations and five contaminated fermentations. Results of this model allowed the authors to properly classify the samples in the corresponding classes with satisfactory values of sensitivity and specificity.

Another study performed by Wei et al. (2022), consisted on achieving on-line measurements in continuous acquisition mode with an optical fiber probe system of 2300 tobacco leaves, was collected in three different years. The spectral data obtained from the NIR analyses were used to develop PLS, SVM, and convolutional neural network (CNN) quantitative models to predict changes in moisture, starch, protein and soluble sugars of the samples during a flue-curing process, which lasted 7–8 days. In this regard, the authors were able to demonstrate that, for this particular case, CNN model performed better in the monitoring process than PLS and SVM. Moreover, the authors also created a strategy to include seasonal and temperature variability into the model to predict samples from a new harvest season in a curing barn, providing a potential and practical method to overcome performance degradation by seasonal differences and temperature oscillations.

Finally, Upadhyay et al. (2022) studied ready-to-cook (RTC) food products, specifically, instant noodles. The authors performed in-line and at-line measurements of NIR and visible spectra at a pilot plant noodle manufacturing line of Nestle R&D Centre India Private Limited (Haryana, India). The spectra obtained from the in-line measurements were used to monitor some quality parameters during the process, such as moisture, crude protein, total fat, and total ash, whilst the at-line measurements perform the prediction of their content in the final product. The authors took advantage of some MVDA tools, such as PCA to study sample distribution patterns and detect probable outliers and PLS and SVM to perform the prediction/calibration models, together with different preprocessing techniques and competitive adaptive reweighted sampling (CARS) selection algorithm to improve the results of the models. According to the results presented in this work, the authors were able

to obtain excellent prediction models with SVM under full wavelength for all the quality parameters, except for total ash, demonstrating that the quality monitoring of instant noodles produced under pilot plant facility is effectively achievable using NIR on the manufacturing lines.

As it could be observed in this subsection, MVDA analysis has been more applied than DOE in the food industry, which indicates its importance to monitor processes and gain knowledge about them. Nonetheless, the number of studies is still low, showing a good opportunity of improvement for industries.

2.3. Multivariate Process Control (MVPC)

Multivariate process control (MVPC) increased its popularity within the statistical process control, since the applied techniques reduce the amount of information contained in the variables of the process down to two or three metrics through the application of statistical modeling, according to Bersimis et al. (2005). Despite of its well-known benefits, only three applications were found in literature in which MVPC has been applied under very similar operating conditions to the food industry.

The first study was reported by Tokatli et al. (2005) in which the critical control points (CCP) of a continuous food pasteurization process were monitored with MVPC, and fault detection and diagnosis methods were developed. The study was performed in a high-temperature short-time pasteurization pilot plant. According to the authors, they found that the studied monitoring and diagnosis charts were able to show deviations in the holding tube-outlet temperature measurements caused by variations in the holding tube-inlet temperature sensor, in the preheater temperature sensor, and in the steam valve of the plant. From this information, corrective actions can be performed in advance and avoid undesired effects on the pasteurized product temperature.

The second study was recently reported by França et al. (2021), who monitored the whole production process of craft beer, using NIR spectroscopy and MVPC. In this study, seven batches of Belgian Pale Ale (BPA) craft beer were produced using the same standard machinery (32-L capacity) that most of the home brewers employ.

Four of the seven batches with NOC were used to establish the control chart and to study the variability within and among batches; the validation of the model was done using the three remaining batches, two were out of order and only one was under NOC. The control chart was created using PCA of the NOC batches of beer with Hotelling's T^2 and sum of square residuals Q statistics, established at 95% of confidence interval. The authors created a PCA model that successfully associated NIR information with the different steps of the beer production, since the PCs provided essential information concerning biochemical changes in the saccharification process, appearance of fermentable sugar, fermentation and ethanol transformation by the yeast. Moreover, the authors used the PCA information to build the calibration control chart in which most of the observations were within the established T^2 and Q established limit. When external and validation batches were analyzed with the developed calibration control chart, it was observed that two batches were extremely out of the NOC, specifically in the fermentation step. In this regard, researchers could monitor and control the overall process of beer production in each step of its production through the combination of NIR and MVPC methodologies.

The third reported study in which MVPC was used is the work presented by Schorn-García et al. (2021). In this case, the authors monitored and controlled the possible contaminations with lactic acid bacteria in a wine alcoholic fermentation process using ATR-MIR spectroscopy. Mainly, the authors utilized 10 normal alcoholic fermentations and 4 contaminated fermentations, whose evolutions were properly monitored and detected through a Q-residuals plot, obtained from a previously elaborated PCA model. Thus, the authors were able to follow this process and point out the fermentations out of control, and to detect process deviations using Q-residuals plots and contribution plots, which allowed to assign the cause of such deviations to specific regions of the spectra that are used to differentiate normal and abnormal process samples.

The current section demonstrates that PAT tools, such as DOE, MVDA, and MVPC, are being used in the food industry and have demonstrated their usefulness in the sector, as summarized in Table 1. However, the lack of these studies at industrial and pilot plant scales is notorious.

Table 1. QbD/PAT implementation in the food industry on a pilot plant and industrial scale.

QbD/PAT tool	Food or related item	Goal	Multivariate methodology	Ref.
DOE	Coffee	Definition of the design space to optimize the process in terms of energy losses and efficiency	NM	Fissore et al. (2014)
	Beer	Manufacturing and optimization of the formulation process	S-CMD	Tessarini et al. (2021)
MVDA	Pork meat	Potential of visible spectroscopy to classify and predict meat quality	CDA, PLS	Xing et al. (2007)
	Porcine carcasses	Determination of fat quality using spatially resolved NIR spectroscopy	PCA, PLS	Sørensen et al. (2012)
	Brined pork	Classification of brined and non-brined pork loins and prediction of salt concentration with Vis-NIR Hyperspectral imaging	PCA, PLS-DA, PLS	Achata et al. (2019)
	Lamb meat	On-line quantification of fatty acids by Vis-NIR spectroscopy	PLS	Pullanagari et al. (2015)
	Deboned chicken and salmon	In-line determination of fatty acids in Salmon and residual bone concentration in chicken using Raman spectroscopy	PLS	Lintvedt et al. (2022)
	Salmon	Optimization of the production process to increase the yield	PLS, PCA	Ørnholt-Johansson et al. (2017)
	Cheese	Real time modeling of milk coagulation	PCA	Lyngaard et al. (2012)
	Milk powder	Prediction of functional properties based on manufacturing data	CPDs, PLS	Rimpiläinen et al. (2015)
	Olive oil	Improvement of the olive oil extraction process	PCA	Tamborrino et al. (2017)
	Potato 'chips'	Identification of main quality and process parameters	MLR-SW	Picout et al. (2019)
	Tomato paste	Qualitative and quantitative detection of sucrose adulteration	PCA, PLS, LS-SVM, BPNN	Liu et al. (2017)
	Milk powder	Development of a real time quality control process using hyperspectral imaging spectroscopy (HIS)	PCA, PLS	Munir et al. (2018)
	Soy sauce	Development, optimization and prediction of the desalting process by electrodialysis	PLS	Moscetti et al. (2019)

		using on-line NIR spectroscopy		
	Apple puree	Prediction of the composition and texture characteristics using NIR, MIR, Raman spectroscopies and HIS.	PLS, RF, SVM	Lan W. et al. (2022)
	Beer	Prediction of the final quality of beer through a real-time monitoring process using ATR-FTIR spectroscopy	PLS	Tessarini et al. (2020)
	Wine	Monitoring and control process of the alcoholic fermentation using ATR-MIR spectroscopy	PCA, PLS, PLS-DA	Schorn-García et al. (2021)
	Tobacco leaves	Monitoring and prediction of different parameters involved in the flue-curing process using NIR spectroscopy	PLS, SVM, CNN	Wei et al. (2022)
	Instant noodles	Monitoring and prediction of several quality parameters during the production process using NIR-vis spectroscopy.	PCA, PLS, SVMR	Upadhyay et al. (2022)
MVPC	Milk	Monitoring and control of critical points to detect faults in sensors during early stages of the process	NM	Tokatli et al. (2005)
	Craft beer	Monitoring of the whole production process with NIR spectroscopy and control possible deviations of the process	PCA	França et al. (2021)
	Wine	Monitoring and identification of contaminated alcoholic fermentations using ATR-MIR spectroscopy	PCA	Schorn-García et al. (2021)

Abbreviations (a.o): BPNN back propagation neural network, CDA canonical discriminant analysis, CPDs conditional probability distributions, CNN convolutional neural networks, DOE design of experiments, LS-SVM least squares-support vector machines, MLR-SW multilinear regression step-wise model, MVDA multivariate data analysis, MVPC multivariate process control, NM not mentioned, PLS partial least squares, PLS-DA partial least squares-discriminant analysis, PCA principal component analysis, PAT process analytical technology, QbD quality by design, S-CMD simplex-centroid mixture design

3. QbD/PAT Tools in the Food-Related Biotech Industry

The previous section dealt with the multivariate tools recommended by the QbD and PAT system in the food industry. The same three tools are identified in diverse studies at industrial or pilot plant scales in the food-related biotechnology industry and further described in this section. One of the most important features applied within this industry are the bioreactors, which are the key unit operation to perform different processes.

As well noted by Boudreau and McMillan (2007), the process control of bioreactors tries to influence the reactions inside the cell by regulating the environment that surrounds it, in order to obtain a specific product. Please note that bioreactors are mainly related to biopharmaceutical and biochemical industries; however, these industries are out of the scope of this work, since the use of the PAT multivariate tools are well established and used within them. Instead, emphasis is made on food-related biotech processes and products, such as in the fermentation process, which was the most common topic of research, as it is shown in the following subsections.

3.1. Design of Multivariate Experiments (DOE)

Despite of the well-known benefits of DOE to promote innovation and solve problems, only these three studies were found, as shown in this section.

Harms et al. (2008) developed a stepwise approach for defining the design space for the production of a protein, which involved the fermentation of a methylotrophic yeast *Pichia pastoris* in a pilot plant facility executing two 300 L runs. The authors designed three different studies taking into account the (i) absorbance and feed rate screening, (ii) culture parameters, and (iii) protein stability. During the first study, the authors used a DOE named fractional factorial screening design with a resolution of IV. Eight factors were tested at two levels in four blocks with one center point per experimental block. Despite of finding statistically significant effects over the final absorbance, the authors considered them as non-key operating parameters, since they were of small magnitude. Regarding the second study, pH and temperature were characterized using a two-level full factorial design (FFD). Results showed that only temperature had a statistically significant effect on titer. The third study was performed with a two-level FFD to estimate all main and interaction effects for

the growth and productivity in the induction phase. According to the reported results, the interaction among temperature and dissolved oxygen had a statistically significant effect on the percentage of solids and titer. Additionally, temperature also had a statistically significant effect on titer, considering temperature, pH, and absorbance as key parameters for the growing process. The second part of this study involved the characterization of temperature and pH for the post-induction process using again a two-level FFD. With this study, the authors found that neither pH, temperature, nor their interaction were considered to have a significant effect on post-induction product protein concentration, demonstrating that the product is stable and there was no proteolytic degradation. In this sense, the authors established the design space for the fermentation process and identified temperature, pH, and absorbance as key operating parameters for process characterization through the use of risk analysis and DOE.

Moreover, Bayer et al. (2020) proposed the use of DOE with hybrid modeling for process characterization, using 20 L cultivations of *Escherichia coli* fed-batch. The study was split in two phases: (i) finding the model with the best performance in describing the biomass concentration and soluble product titer and (ii) determining which model was most accurate to predict the entire process. The studied models in the first phase were response surface model (RSM) with a FFD, artificial neural network (ANN), and hybrid model (RSM+ANN) and only AAN and hybrid models in the second phase. The authors demonstrated that the hybrid model was superior to the ANN model in predicting the biomass concentration and the soluble product titer. In most of the cases, the hybrid model correctly matched the predicted values with the analytical measurements with small prediction intervals. Regarding these results, an approach was developed to characterize and optimize the entire process using a dynamic hybrid model, making possible to obtain the desired product at the end of the process controlling the CPP. Such results were achievable due to the structure of the model, which differentiate if the variations of the process are caused by the metabolism of the bacteria or due to the process operations.

Lastly, the control and optimization of lactose production through its crystallization process at industrial scale was studied by Galvis et al. (2022), who developed a novel strategy based on retrospective QbD approach and new experiments coming from DOE.

After the use of MVDA, the authors identified 4 out of 32 variables as critical process parameters (CPPs). These four CPPs were included in a face-centered DOE considering low, medium, and high levels each and with three replications of the design center. The results of these experiments allowed researchers to analyze the effects of the 4 CPPs and their different interactions over the mass percentage of total fines. However, the authors mentioned that the experimental design was not fully performed as it was planned and the experimental factors were not completely independent; thus, the statistical significance of the factors was not reliably quantified. Nonetheless, the authors were able to compare the historical data with the new obtained data using contour plots, finding important insights that allowed them to improve the quality of the final product by up to 7%.

As already outlined in the earlier section focused to the food industry, there are few reported instances using DOE. Thus, more awareness of this tool and effort are needed within the food-related biotech industry to apply DOE.

3.2. Multivariate Data Analysis (MVDA)

Due to the importance of MVDA, it has also been applied in food-related biotechnology sector. In fact, this section deals with five studies at industrial plant or pilot plant scale in which MVDA is applied. In this sector, it is common to produce or utilize cell cultures, which need to be monitored during a complex process.

According to this, Abu-Absi et al. (2010) decided to assess and monitor different parameters in a cell cultivation process in 500 L bioreactors, using off-line and in-line Raman spectroscopy coupled to MVDA. Samples and measurements were taken and performed from four bioreactor runs; the data from the first three were used for the calibration data set and the last one for the validation data set. The authors intended to predict parameters with PLS and some preprocessing techniques, such as 1st and 2nd derivatives, variance scaling, and SNV path length correction. As reported by the authors, the calibration of the models was good for glutamine, glutamate, glucose, lactate, ammonium, viable cell density (VCD), and total cell density (TCD). Afterwards, these models were validated, and predictions were contrasted with the measured values. Researchers found a model using the three lots for glutamine which differed 30% between the measured and predicted values. For glutamate, the average difference was 12%, for

glucose was 15%, for lactate was 13%, and for VCD and TCD the difference was 15%. The only model that did not match the predicted values with the measured values was the ammonium model. Nonetheless, the authors considered that these performances were good enough for the purpose of their study. In this sense, researchers were able to provide immediate feedback and control the process performance using real-time measurements, ensuring consistent manufacture of the mammalian cell cultures using MVDA.

Mercier et al. (2013) also used MVDA to monitor the early development of a cell cultivation process. This study consisted on 17 and 10 cell cultivation runs of 2 L and 10 L, respectively. During the initial steps of the process, the evolution of the behavior of the batches was checked, and PLS was employed to relate the data of the process to a response variable which represented the run maturity. Then, batch level modeling was created, considering each batch as single unit. At this point, PCA was used to explore the data and then PLS was employed to understand how the initial conditions of the process influenced over it. When the authors analyzed the score plots of the off-line and on-line variables model, they realized that clusters were clearly observed according to the scale of the cultivations, causing operational differences in both on and off-line process variables. The authors attributed this behavior to the consumption of O_2 and CO_2, which was higher for the 2 L cultures than for 10 L cultures that was associated to the aeration strategy, which was not linearly scaled between the two bioreactor volumes. Additionally, analysts found that the cell diameter for the 10 L cultivations was on average 1.1 µm smaller than the 2 L cultivations that was attributed to the cross flows inside the fibers of the equipment, which were distinct in the two bioreactor scales. PCA was also employed for batch diagnosis in which 7 batches were further analyzed, which showed deviations due to the concentrations of additives in feed medium during the perfusion, inoculation at twice the target cell density for the off-line variables, a change in the procedure for medium preparation, and due to a deviation in absorbance probe calibration for the on-line variables. Moreover, PLS was used to establish correlations between process parameters and process responses; however, the authors reported that the generated PLS models showed a poor fit. In this sense, the authors proved that PCA could be used as valuable tool to identify deviations in early development of cell cultivation processes, being the scale effect a relevant factor to take in to consideration when developing a process as the presented here.

The third study was performed by Ferreira et al. (2007), who studied if multiway PCA (MPCA) and multiway PLS (MPLS) could be used to (i) model 16 industrial fermentation processes of *Streptomyces clavuligerus* strain for the production of clavulanic acid using a pilot plant and (ii) to predict the fermentation yield. The acquired data were preprocessed using SG filter and then explored using MPCA. The authors were able to differentiate among batches based on the trajectories of variables measured on-line, being the most different batches 3, 5, 6, and 7 from the rest. This difference was caused in batch 3 for its high conductance profile and for keeping low values of temperature for a long period of time. Batches 5–7 presented different conditions for the substrate addition, producing changes in the quality variables (biomass, absorbance, and conductance). The MPLS was performed to predict the final concentration of the clavulanic acid for each batch and also to evaluate what variables influenced the most over the productivity. With this model, batches 3, 5, 6, and 7 were no different from the others, attributing this behavior to the basis of each method, since MPCA focused on the covariance of the variance, whilst MPLS focused on the covariance of the X-block (process variables) that is more correlated with the Y-block (response variables). Moreover, the authors found that capacitance was the principal variable for the prediction of the final product concentration using the weight contribution plot. In this regard, researchers could improve the knowledge through MVDA on a fermentation process carried out in a pilot plant in which dissimilarities were detected according to abnormal changes in quality variables, predicting the final product with moderate accuracy and detecting the most important variables that influenced the most over the productivity prediction (capacitance and absorbance).

Furthermore, Alves-Rausch et al. (2014) performed a real-time multiparameter monitoring during a fermentation process in a 50 L bioreactor, intending to produce *Bacillus* spores and introducing into the study floor vibrations and high humidity as in the industrial environment. The fermentation process was divided in 5 batches, where temperature was controlled at 39°C and pH at a specific set point by addition of NaOH (50% v/v) during the growth phase or H_2SO_4 (38%) during the sporulation phase. In this case, PCA was performed with spectra collected directly from the reactor at 0.25 and 0.5 h before inoculation. The media formulation was considered “good” or “bad” according to the final production yield of the fermentations.

Two models were tried; without any preprocessing and applying SNV, finding that SNV removed most physical effects from the spectra. The SNV-PCA model was used together with the distance to the model in the X-space (DModX), allowing to identify that one batch was different from the other four, which was mainly attributed to a reduction in the content of yeast extract. Regarding the use of XLS (extension of PLS implemented in a specific software), five calibration models with no preprocessing were performed for acetoin, absorbance at $\lambda = 600$ nm (A_{600}), dry mass, and two sum parameters for sugar and analytes, which were considered as indicators of sugar consumption and overall metabolism. With these models, the authors were able to monitor a large-scale industrial fermentation process getting important insights of the process, such as the in-line prediction of acetoin concentrations that gives information of the metabolic state of bacillus, and that A_{600} and biomass in-line values provide important information about the growth and sporulation of the culture growing for further process and medium optimization.

Lastly, Galvis et al. (2022) developed a novel strategy based on retrospective QbD approach to control and optimize the crystallization process for lactose production. Such strategy was developed using long-running historical data of 2 years obtained from an industrial production facility, using expert knowledge and including new experiments. The authors intended to improve production quality by reducing the mass percentage of small crystal fines produced as critical quality attributes, and they used different MVDA along the process to achieve this goal. In fact, PCA was used in the first place to detect and remove outlying samples, and PLS was applied to identify the variables of the process that were more critical for the production quality. In this sense, the authors were able to properly identify 4 critical process parameters out of 32 studied variables, using PLS and variable importance in projection (VIP) that once they were optimized, the product quality improved up to 7%.

From this subsection, is evident how MVDA is of great importance to monitor, optimize, and get important information of the fermentation and crystallization processes. Additionally, it is evident that such studies need to be performed and adopted by the industry to take all the advantages of these methodologies.

3.3. Multivariate Process Control (MVPC)

From the previous subsection, Alves-Rausch et al. (2014) performed a real-time multiparameter monitoring during a fermentation process, which intended to produce *Bacillus* spores. The authors monitored the process by performing a batch evolution model (BEM) based on the NIR data, which captured the variations in the spectra over time and a reference batch trajectory was built including process control limits based on ±3 SD.

The BEM was based on the PLS models in which the PLS scores were averaged for each time point. Four well-behaved batches were used to build the model, leaving one batch out for validation. The BEM on the SNV preprocessing data gave them better insights of the three different metabolic stages identified in the BEM plot than the BEM with no preprocessing. In the first stage, microorganisms started to grow and consume the sugar sources, then, microorganisms started to consume the metabolite produced in the first stage, and finally, the metabolites in the media were completely consumed, and the spectral changes were smaller, which may be an indicator that cell growth stopped and microorganisms started to sporulate. In this sense, researchers could monitor a fermentation process using BEM, making possible to create a reference batch trajectory and to detect future deviations for the coming batches.

Another study, performed by Krause et al. (2015), monitored seven aerobic fermentation batches of *Saccharomyces pastorianus*, variety *calsbergensis*. This process was carried out at pilot scale in an industrial fermentation tank with 70-L capacity in which information of seven sensors was studied through MVPC and "particle swarm optimization" (PSO). MVPC was based on "unfold-PLS" and was used to create statistically supported process trajectories for process control. Two levels of MVPC were used: level 1 (maturity prediction) and level 3 (residual standard deviation (RSD)). The seven initial input data coming from the sensors were extended by fully polynomial extension of second order including mixed terms, obtaining a total of 35 input variables, but variable importance in the projection (VIP) was applied since not all variables were of the same importance to model the target of interest. Once the VIP was applied, twelve variables were considered for the elaboration of the multivariate process trajectory control charts. In this sense, the authors showed the result for three-time sector through the use of MVPC charts, where all

sensors demonstrated to work properly within the established boundaries, showing good similarity among each individual input trend and the historical data. Researchers reported that 12 inputs were used in 90.8% of the modeled cases and 11 in only 8.7% of the other ones. Scientists reported that all trajectories always kept the direction between the established 3 σ limits, developing a successful approach to monitor the trajectory progress of the fermentation process and capable to predict false input information.

The last study found in literature in which MVPC was applied at industrial scale was performed by Gunther et al. (2007), who applied PCA and MVPC to industrial fermentation data obtained from the industry. These methodologies were used to detect and diagnose possible abnormal conditions from both on-line and off-line analyses. A total of ten batches with 1084 samples each monitored through 11 process variables were analyzed from 300 L reactors. Batches 1–8 with NOC were used to develop the PCA model, batch 9 to validate it, and batch 10 to detect problems within the process. In fact, the score plot of the PCA model showed a similar trend of the first 9 batches, but batch 10 was clearly different. Furthermore, the authors performed the monitoring process of these batches using T^2 and Q statistics, as part of the MVPC, from the off-line analyses, resulting in the same results as in PCA. These results were further confirmed studying the on-line data of the fermentation batches 9 and 10 on which T^2 and squared prediction error (SPE) were applied as part of the MVPC. Results led to the same conclusions as PCA and T^2 and Q plots; however, in this comparison, SPE evidenced more clearly the fault detections than the T^2 plot. Hence, it was demonstrated that MVPC could help in the monitoring process of a fermentation process identifying NOC and abnormal batches. All the discussed studies in which PAT tools are applied in the food-related biotech industry are summarized in Table 2.

Table 2 QbD/PAT implementation in the biotechnology industry on a pilot plant and industrial scale

QbD/PAT tool	Process	Goal	Multivariate methodology	Ref.
DOE	Fermentation	Definition of the design space for a product	FFD, 2L-FFD,	Harms et al. (2008)
	Cultivation of *Escherichia coli*	Fast characterization and optimization of the process	RSM, FFD	Bayer et al. (2020)
	Crystallization	Control and optimization of the process using historical and new data	FC	Galvis et al. (2022)
MVDA	Cultivation	Assessment and monitoring of the process using different parameters	PLS	Abu-Absi et al. (2010)
	Cultivation	Monitoring of the early stage of the process	PCA, PLS	Mercier et al. (2013)
	Fermentation	Model the industrial fermentation process and predict its yield	MPCA, MPLS	Ferreira et al. (2007)
	Fermentation	Real-time multiparameter monitoring of a fermentation process	PCA, PLS	Alves-Rausch et al. (2014)
	Crystallization	Selection of critical process parameters	PCA, PLS	Galvis et al. (2022)
MVPC	Fermentation	Real-time multiparameter monitoring of a fermentation process	PLS	Alves-Rausch et al. (2014)
	Fermentation	Monitoring of fermentation process	U-PLS	Krause et al. (2015)
	Fermentation	Detection of possible abnormal batches	PCA	Gunther et al. (2007)

Abbreviations (a.o): 2L-FFD 2-level full factorial design, BBD Box-Behnken design, DOE design of experiments, FC face centered, FFD fractional factorial design, JY-PLS Join-Y partial least squares, MVDA multivariate data analysis, MVPC multivariate process control, MPLS multiway partial least squares, MPCA multiway principal component analysis, NM not Mentioned, PLS partial least squares, PBD Plackett-Burman design, PCA principal component analysis, PAT process analytical technology, QbD quality by design, RSM response surface model, U-PLS unfold partial least squares

4. Conclusions and Final Remarks

As observed from the "QbD/PAT Tools in the Food Industry" section to the "QbD/PAT Tools in the Food-Related Biotech Industry" section, the multivariate tools for design, data acquisition, and analysis (DOE, MVDA, and MVPC) recommended by the PAT system under a QbD framework were described and discussed within the food and food-related biotech industries.

In the case of the food industry, 23 studies were addressed and 11 in the food-related biotech industry. This makes evident the narrow circumstances where DOE, MVDA, and MVP are used at industrial and pilot plant scales. Moreover, several different studies applying similar analytical techniques that the ones exposed here at the industrial, pilot plant or lab scale can be found in literature, but they are not reported in this review because they were not performed under the QbD/PAT framework nor follow the QbD/PAT principles, such as the investigations gathered by Grassi and Alamprese (2018). The same observation was noted by Djekic et al. (2019) when performing a survey to 203 companies from the European Union and abroad. The authors found that the application of models in the food industry consists of simplified models that do not evaluate the processes, quality, or safety conditions and environmental impact, thus, revealing that the application of mathematical models in food companies has not been a matter of interest yet, identifying it as a research gap.

This absence of multivariate tools should be given by different factors, such as (i) poor knowledge on modeling, (ii) not user-friendly models/software, (iii) instability of processes when introducing experimental tests, (iv) additional cost of new experiments that the company is not willing to assume (Pietraszek et al. 2020), or (v) the high confidentiality of the studies which hinders the free publication of the results in scientific journals. In this regard, it has been noticed that QbD and PAT tools have been applied in the academy mainly for research purposes at lab scale and in some companies to improve quality control and rapidly evaluate the final product to increase productivity, losing the holistic view and devaluating the real purpose of the QbD/PAT system, which is to ensure quality through continuous and real-time feedback (on-, in-line analysis). Researchers have made a great labor in proving the application of QbD/PAT at lab scale, in pilot plants, and, in some cases, at industrial scale, as exposed within this review paper.

However, it is time to stop basing, associating, and focusing QbD and PAT only to the use of analytical instruments and to start sharing, as much as the confidentiality of each company allows results on the use of DOE, MVDA, and MVPC. QbD and PAT have a more profound meaning than using novel analytical analyzers and multivariate data analysis to increase productivity; both approaches intend to help companies at getting better

with their manufacturing and quality assurance processes, products, and final customer to whom is directed. For this to start changing in the food and food-related biotech industries, both of them with their corresponding academia and regulatory organizations should cooperate more to bring together all the diffused work produced by each of them to create a more solid QbD framework with PAT as an enabler. As noted by O'Donell et al. (2014), adopting this strategy might create a society for the both industries in which QbD/PAT will be the core center of their activities, assembling chief executive officers and associated companies, government representatives, process engineers, scientists, and technicians, aiming to provide to these industries with a stronger, smarter, and more efficient working framework for the upcoming years.

Summarizing, the key historical aspects and fundamentals of the QbD framework and PAT system were reviewed, as well as their application within the food and food-related biotechnology industries. Special attention was given to the use of multivariate tools for design, data acquisition, and analysis in these industries. A total of 34 case studies were found in literature in which DOE, MVDA, and MVPC at lab scale, pilot plant, and industrial scale were applied for both industries. From this revision, it was observed that the implementation of these tools is still under research and that food and food-related biotech companies are not applying them within their processes, with the exception of some studies reported in this work. It was also noticed that QbD and PAT are being used indistinctly by these industries with emphasis on analytical instruments and multivariate tools to make their analyses and processes faster. In this sense, the authors make an appeal of encouragement to both industries, to researchers, and academia to work closer and improve the current practices, aiming to start a new direction for QbD and PAT in order to adopt them as the leading rules within their processes and industries.

Acknowledgements

C.H. Pérez-Beltrán acknowledges the scholarship provided by the Autonomous University of Sinaloa (México). A. Torrente-López acknowledges the FPU predoctoral grant (ref.: FPU18/03131), which is currently receiving from the Ministry of Universities, Spain.

Author Contribution

All the authors contributed to the review conception and design. Literature search was performed by Christian H. Pérez-Beltrán and Luis Cuadros-Rodríguez. The first draft of the manuscript was written by Christian H. Pérez-Beltrán, and all the authors revised and commented on subsequent versions. All the authors read and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

References

ASTM E2363-14 (2014). Standard terminology relating to process analytical technology in the pharmaceutical industry. ASTM International, West Conshohocken, USA.

ASTM E456-13A (2017)e6. Standard terminology relating to quality and statistics, ASTM International, West Conshohocken, USA.

ASTM ES891-20 (2020). Standard guide for multivariate data analysis in pharmaceutical development and manufacturing applications. ASTM International, West Conshohocken, USA.

Abu-Absi NR, Kenty BM, Ehly Cuellar M, Borys MC, Sakhamuri S, Strachan DJ, Hausladen MC, Jian Li Z (2010). Real time monitoring of multiple parameters in mammalian cell culture bioreactors using an in-line Raman spectroscopy probe. Biotechnol. Bioeng.
https://doi.org/10.1002/bit.23023

Achata EM, Inguglia ES, Esquerre CA, Tiwari BK, O'Donnell CP (2019) Evaluation of Vis-NIR hyperspectral imaging as a process analytical tool to classify brined pork samples and predict brining salt concentration. J. Food Eng.
https://doi.org/10.1016/j.jfoodeng.2018.10.022

Alves-Rausch J, Bienert R, Grimm C, Bergmaier D (2014) Real time in-line monitoring of large-scale Bacillus fermentations with near-infrared spectroscopy. J. Biotechnol.
https://doi.org/10.1016/j.jbiotec.2014.09.004

Bakeev KA (2010) Process analytical technology. Chichester, England.

Bayer B, Von Stosch M, Striedner G, Duerkop M (2020) Comparison of Modeling Methods for DoE-Based Holistic Upstream Process Characterization. Biotechnol. J.
https://doi.org/10.1002/biot.201900551

Bersimis S, Panaretos J, Psarakis S (2005) Multivariate statistical process control charts and the problem of interpretation – A short overview and some applications in industry. In Proceedings of the 7th Hellenic European Conference on Computer Mathematics and Its Applications. Athens, Greece.

Boudreau MA, McMillan GK (2007) New directions in Bioprocess Modeling and control: Maximizing process analytical technology benefits. North Carolina, U.S.A.

Callis JB, Illman DL, Kowalski BR (1987) Process analytical chemistry. Anal Chem.
https://doi.org/10.1021/ac00136a001

Casian T, Nagy B, Kovács B, Galata DL, Hirsch E, Farkas A (2022) Challenges and opportunities of implementing data fusion in process analytical technology – a review. Molecules.
https://doi.org/10.3390/molecules27154846

Chew W, Sharratt P (2010) Trends in process analytical technology. Anal. Methods.
https://doi.org/10.1039/C0AY00257G

Djekic I, Mujčinović A, Nikolić A, Jambrak AR, Papademas P, Feyissa AH, Kansou K, Thomopoulos R, Briesen H, Kavallieratos NG, Athanassiou CG, Silva CLM, Sirbu A, Moisescu AM, Tomasevic I, Brodnjak UV, Charalambides M, Tonda A (2019) Cross-European initial survey on the use of mathematical models in food industry. J. Food Eng.
https://doi.org/10.1016/j.jfoodeng.2019.06.007

Ferreira AP, Almeida Lopes J, Cardoso de Menezes J (2007) Modelling industrial fermentation data with multiway multivariate techniques. IFAC Proceedings Volumes.
https://doi.org/10.3182/20070606-3-MX-2915.00044

Fissore D, Pisano R, Barresi AA (2014) Applying quality-by-design to develop a coffee freeze-drying process. J. Food Eng.
http://dx.doi.org/10.1016/j.jfoodeng.2013.09.018

França L, Grassi S, Pimentel MF, Amigo JM (2021) A single model to monitor multistep craft beer manufacturing using near infrared spectroscopy and chemometrics. Food Bioprod. Process.
https://doi.org/10.1016/j.fbp.2020.12.011

Galvis L, Offermans T, Bertinetto CG, Carnoli A, Karamujic E, Li W, Szymanska E, Buydens LMC, Jansen JJ (2022) Retrospective quality by design r(QbD) for lactose production using historical process data and design of experiments. Comput. Ind.
https://doi.org/10.1016/j.compind.2022.103696

Grassi S, Alamprese C (2018) Advances in NIR spectroscopy applied to process analytical technology in food industries. Curr. Opin. Food Sci.
https://doi.org/10.1016/j.cofs.2017.12.008

Gunther JC, Conner JS, Seborg DE (2007) Fault detection and diagnosis in industrial fed-batch cell culture, Biotechnol. Prog.
https://doi.org/10.1021/bp070063m

Harms J, Wang X, Kim T, Yang X, Rathore AS (2008) Defining Process Design Space for Biotech Products: Case Study of *Pichia pastoris* Fermentation. Biotechnol. Prog.
https://doi.org/10.1021/bp070338y

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

ICH Q8(R2) (2009). Pharmaceutical development, International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, Geneva, Switzerland.

ISO 3534-3:2013 (2013). Statistics – Vocabulary and symbols – Part 3: Design of experiments. International Organization for Standardization, Geneve, Switzerland.

ISO 7870-7:2020 (2020). Control charts – Part 7: Multivariate control charts. International Organization for Standardization, Geneve, Switzerland.

Juran JM (1986) The quality trilogy: A universal approach to managing for quality. Qual. Prog. 19(8), 19-24.

Juran JM (1992) Juran on quality by design: the new steps for planning quality into goods and services. New York, USA.

Khan IA, Smillie T (2012) Implementing a “quality by design” approach to assure the safety and integrity of botanical dietary supplements. J. Nat. Prod.
https://doi.org/10.1021/np300434j

Koch KH (1999) Process analytical chemistry – Control, optimization, quality, economy. Berlin, Germany.

Koch MV, VandenBussche KM, Chrisman RW (2007) Micro instrumentation for high throughput experimentation and process intensification – A tool for PAT. Weinheim, Germany.

Krause D, Hussein MA, Becker T (2015) Online monitoring of bioprocesses via multivariate sensor prediction within swarm intelligence decision making. Chemom. Intell. Lab. Syst.
https://doi.org/10.1016/j.chemolab.2015.04.012

Lan W, Baeten V, Jaillais B, Renard CMGC, Arnould Q, Chen S, Leca A, Bureau S (2022) Comparison of near-infrared, mid-infrared, Raman spectroscopy and near-infrared hyperspectral imaging to determine chemical, structural and rheological properties of apples purees. J. Food Eng.
https://doi.org/10.1016/j.jfoodeng.2022.111002

Lintvedt TA, Andersen PV, Afseth NK, Marquardt B, Gidskehaug L, Wold JP (2022) Feasibility of in-line Raman spectroscopy for quality assessment in food industry: How fast can we go? J. Food Eng.
https://doi.org/10.1177/00037028211056931

Liu C, Hao G, Su M, Chen Y, Zheng L (2017) Potential of multispectral imaging combined with chemometric methods for rapid detection of sucrose adulteration in tomato paste. J. Food Eng.
http://dx.doi.org/10.1016/j.jfoodeng.2017.07.026

Lyndgaard CB, Engelsen SB, Van den Berg FWJ (2012) Real-time modelling of milk coagulation using in-line near infrared spectroscopy. J. Food Eng.
http://doi:10.1016/j.jfoodeng.2011.07.029

Mercier SM, Diepenbroek B, Dalm MCF, Wijffels RH, Streefland M (2013) Multivariate data analysis as a PAT tool for early bioprocess development data. J. Biotechnol.
https://doi.org/10.1016/j.jbiotec.2013.07.006

Misra NN, Sullivan C, Cullen PJ (2015) Process analytical technology (PAT) and multivariate methods for downstream process. Curr. Biochem. Eng.
10.2174/2213385203666150219231836

Moscetti R, Massantini R, Fidaleo M (2019) Application on-line NIR spectroscopy and other process analytical technology tools to the characterization of soy sauce desalting by electrodialysis. J. Food Eng.
https://doi.org/10.1016/j.jfoodeng.2019.06.022

Munir MT, Wilson DI, Yu W, Young BR (2018.) An evaluation of hyperspectral imaging for characterising milk powders. J. Food Eng.
https://doi.org/10.1016/j.jfoodeng.2017.10.001

O'Donnell CP, Fagan C, Cullen PJ (2014) Process analytical technology for the food industry. New York, USA.

Orlandini S, Pinzauti S, Furlanetto S (2013) Application of quality by design to the development of analytical separation methods. Anal. Bioanal Chem.
https://doi.org/10.1007/s00216-012-6302-2

Ørnholt-Johansson G, Gudjónsdóttir M, Engelbrecht Nielsen M (2017) Analysis of the production of salmon fillet – Prediction of production yield. J. Food Eng.
http://dx.doi.org/10.1016/j.jfoodeng.2017.02.022

Picouet PA, Gou P, Pruneri V, Diaz I, Castellari M (2019) Implementation of a quality by design approach in the potato chips frying process. J. Food Eng.
https://doi.org/10.1016/j.jfoodeng.2019.04.013

Pietraszek J, Radek N, Goroshko AV (2020) Challenges for the DOE methodology related to the introduction of Industry 4.0. Prod. Eng. Arch.
https://doi.org/10.30657/pea.2020.26.33

Pullanagari RR, Yule IJ, Agnew M (2015) On-line prediction of lamb fatty acid composition by visible near infrared spectroscopy. Meat Sci.
http://dx.doi.org/10.1016/j.meatsci.2014.10.008

Rathore AS, Kapoor G (2017) Implementation of quality by design toward processing of food products. Prep. Biochem. Biotechnol.
https://doi.org/10.1080/10826068.2017.1315601

Rifna EJ, Pandiselvam R, Kothakota A, Subba Rao KV, Dwived M, Kumar M, Thirumdas R, Ramesh SV (2022) Advanced process analytical tools for identification of adulterants in edible oils – A review. Food Chem.
https://doi.org/10.1016/j.foodchem.2021.130898

Rimpiläinen V, Kaipio JP, Depree N, Young BR, Wilson DI (2015) Predicting functional properties of milk powder based on manufacturing data in an industrial-scale powder plant. J. Food Eng.
https://doi.org/10.1016/j.jfoodeng.2014.12.010

Schorn-García D, Cavaglia J, Giussani B, Busto O, Aceña L, Mestres M, Ricard Boqué R (2021) ATR-MIR spectroscopy as a process analytical technology in wine alcoholic fermentation – A tutorial. Microchem. J.
https://doi.org/10.1016/j.microc.2021.106215

Sørensen H, Petersen KM, Engelsen SB (2012) An on-line NIT method for determining depth profiles of fatty acid composition and iodine value in porcine adipose fat tissue. Appl. Spectrosc.
https://doi.org/10.1366/11-06396

Tamborrino A, Squeo G, Leone A, Paradiso VM, Romaniello R, Summo C, Pasqualone A, Catalano P, Bianchi B, Caponio F (2017) Industrial trials on coadjuvants in olive oil extraction process – Effect on rheological properties, energy consumption, oil yield and olive oil characteristics. J. Food Eng.
http://dx.doi.org/10.1016/j.jfoodeng.2017.02.019

Teixeira JA, Vicente AA, Macieira da Silva FF, Azevedo Lima da Silva JS, da Costa Martins RM (2014) In Teixeira JA, Vicente AA (eds) Engineering Aspects of Food Biotechnology, 1st edn. CRC Press, Boca Raton, USA.

Tessarini ES, De Almeida e Silva JB, Rebello Lourenço F (2021) Development and optimization of beer containing malted and non-malted substitutes using quality by design (QbD) approach. J. Food Eng.
https://doi.org/10.1016/j.jfoodeng.2020.110182

Tessarini ES, Rebello Lourenço, F. (2020). Real-time monitoring of beer parameters using infrared spectroscopy - A process analytical technology approach. J. AOAC Int.
https://doi.org/10.1093/jaoacint/qsaa057

Tokatli F, Cinar A, Schlesser JE (2005) HACCP with multivariate process monitoring and fault diagnosis techniques: application to a food pasteurization process. Food Control.
https://doi.org/10.1016/j.foodcont.2004.04.008

Upadhyay R, Gupta A, Niwas Mishra H, Bhat SN (2022) At-line quality assurance of deep-fried instant noodles using pilot scale visible-NIR spectroscopy combined with deep-learning algorithms. Food Control.
https://doi.org/10.1016/j.foodcont.2021.108580

US FDA (2004). Guidance for industry: PAT – A framework for innovative pharmaceutical development, manufacturing, and quality assurance. U.S. Food and Drug Administration.

Wei K, Bin J, Wang F, Kang C (2022) On-line monitoring of the tobacco leaf composition during flue-curing by near-infrared spectroscopy and deep transfer learning. Anal. Lett.

https://doi.org/10.1080/00032719.2022.2046021

Xing J, Ngadi M, Gunenc A, Prasher S, Gariepy C (2007) Use of visible spectroscopy for quality classification of intact pork meat. J. Food Eng.

https://doi.org/10.1016/j.jfoodeng.2007.01.020

2.3. Comunicación a congresos

C.H. Pérez-Beltrán, A.M. Jiménez-Carvelo, A. Torrente-López, N.A. Navas-Iglesias, L. Cuadros-Rodríguez. QbD/PAT: moving from lab-scale analytics to their application in food and food-focused biotechnology industries. EuroPACT 2021. En línea, Sede Frankfurt, 2021. *Comunicación en formato Póster*.

Capítulo 3

TRANSFERENCIA DE MÉTODOS ANALÍTICOS MULTIVARIABLE BASADOS EN CROMATOGRAFÍA DE LÍQUIDOS

3. TRANSFERENCIA DE MÉTODOS ANALÍTICOS MULTIVARIABLE BASADOS EN CROMATOGRAFÍA DE LÍQUIDOS

3.1. Resumen

La transferencia de un método analítico es un proceso extenso y complejo que tiene lugar entre distintos laboratorios –normalmente, aunque no de forma exclusiva, entre laboratorios de investigación, desarrollo e innovación (I+D+i) y laboratorios de análisis de rutina o control de calidad– el cual consiste en implementar el método analítico desarrollado por un laboratorio (*A*) en otro laboratorio (*B*). Una vez hecha la transferencia del método analítico entre laboratorios, en el caso habitual de que el método a transferir esté basado en una estrategia univariable, se deben evaluar y comparar los resultados obtenidos con el laboratorio de referencia (*A*) mediante métricas de calidad en el desempeño (p.ej., veracidad, precisión, límites inferiores, robustez, etc.) para comprobar la efectividad del método y la adecuada cualificación del laboratorio (*B*) para ejecutarlo durante futuras aplicaciones de rutina [**1**].

En el caso de métodos analíticos multivariable (MAM), la transferencia no se centra en el procedimiento para llevar a cabo dichos métodos como ocurre con los métodos univariable, sino que presenta una singularidad importante, ya que la transferencia se centra en la propia señal analítica, a partir de la cual se establecen los modelos matemáticos de clasificación o cuantificación. Habitualmente, los resultados o señales instrumentales de ambos laboratorios se ven influenciadas y dependen en gran medida de cada equipo instrumental utilizado para su obtención, por lo que la transferencia resulta bastante compleja y podría estar limitada. Este hecho cobra especial importancia en el caso de transferencia de señales analíticas que deberán ser utilizadas en su totalidad, y es necesario una etapa de estandarización para encontrar resultados con un alto grado de armonización. Téngase en cuenta que el modelo multivariable es el mismo, y lo que caracteriza cada laboratorio es precisamente la señal que aporta a dicho modelo.

[**1**] E. Rozet, W. Dewé, E. Ziemons, A. Bouklouze, B. Boulanger, P. Hubert, Methodologies for the transfer of analytical methods: A review, 2009, Journal of Chromatography B, 877, 2214 – 2233.

Para lograr una transferencia de MAM basados en cromatografía de líquidos, es deseable la aplicación de una metodología que contribuya a la eliminación de la dependencia que presentan las señales instrumentales con la plataforma analítica empleada para su obtención. En esta línea, el grupo de investigación, en dónde se ha desarrollado esta tesis, ha propuesto una metodología denominada "agnostización instrumental" (véase **Introducción**, subsección **1.3**), con la cual es posible obtener señales instrumentales no dependientes del instrumento de medida, permitiendo realizar la transferencia de las señales instrumentales entre distintos laboratorios analíticos ubicados en diferentes zonas geográficas del mundo, promoviendo a su vez la posibilidad de generar bases de datos globales para facilitar la comprobación de calidad del producto alimenticio analizado.

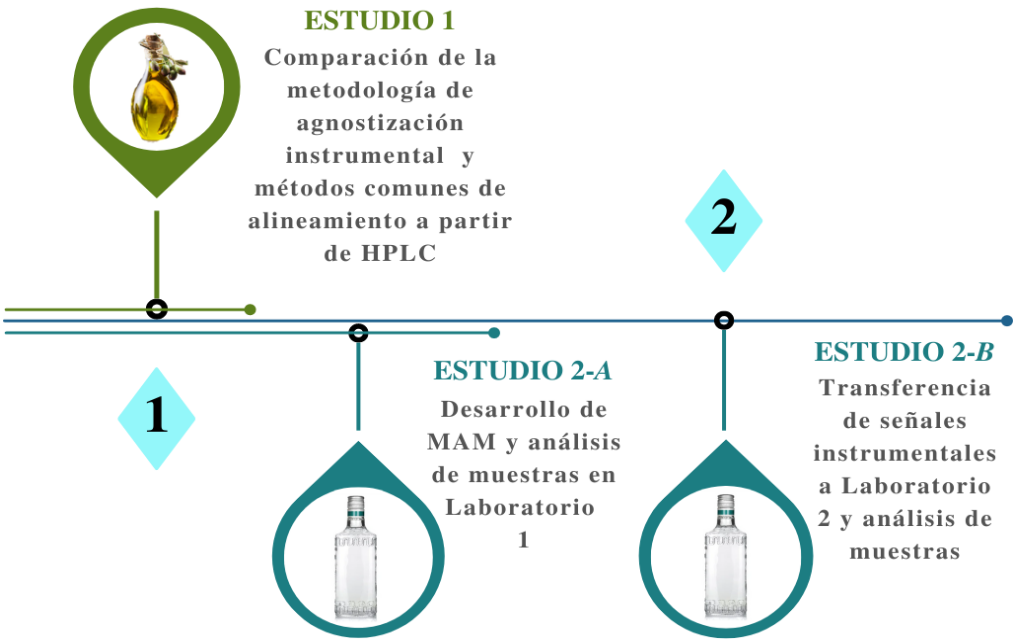


Figura 10. Esquema de los estudios llevados a cabo para lograr la transferencia de señales instrumentales entre laboratorios analíticos enfocados al aseguramiento y control de calidad. *Estudio 1*: Comparación de las metodologías de agnostización instrumental y métodos comunes de alineamiento de señales, utilizando cromatografía de líquidos de altas prestaciones (HPLC). *Estudio 2*: Transferencia interlaboratorio de señales instrumentales a nivel internacional. 2-*A*: Desarrollo del método analítico multivariable (MAM) y análisis de muestras de Tequila, 2-*B*: Implementación del MAM desarrollado en estudio 2-*A*, y análisis de las mismas muestras, así como de nuevas muestras problema.

En este sentido, se plantearon y desarrollaron dos estudios diferentes para lograr la transferencia de señales instrumentales obtenidas mediante cromatografía de líquidos de altas prestaciones, tal como se muestra en la **Figura 10.** En primer lugar, se realizó un estudio en los laboratorios del grupo de investigación en la Universidad de Granada con objeto de comprobar que la metodología de agnostización instrumental de huellas instrumentales cromatográficas provee los mismos resultados que al aplicar los métodos más comunes de alineamiento de señales, empleados en el desarrollo de métodos analíticos multivariable. Posteriormente, se llevó a cabo la ejecución de un segundo estudio interlaboratorio utilizando dos cromatógrafos de líquidos similares ubicados en España y México, respectivamente, en dos periodos de tiempo diferentes para comprobar la transferencia de huellas instrumentales cromatográficas agnostizadas.

Las matrices alimenticias objeto de estudio fueron el (**i**) **aceite de oliva**, y (**ii**) **Tequila**, en el primer y segundo estudio, respectivamente, las cuales representan un pilar fundamental en el sistema agroalimentario, así como en el sector socioeconómico en España y México.

(i) Aceite de oliva

España es el primer productor y exportador mundial de aceite de oliva, el cual conforma un sector de gran relevancia económica, social y comercial. El aceite de oliva es un alimento básico tradicional (obtenido del árbol *Olea europaea*), el cual conforma un componente fundamental de la dieta mediterránea. A diferencia de otros tipos de aceites comestibles, el aceite de oliva tiene la proporción más alta de ácidos grasos monoinsaturados a poliinsaturados, así como también otros nutrientes muy preciados que promueven, a través de sus efectos favorables, un buen estado de salud al consumidor, tales como los compuestos fenólicos antioxidantes hidroxitirosol y oleuropeína, vitaminas, lignanos, escualeno y terpenoides, entre otros [**2**,**3**].

[**2**] E. Tripoli, M. Giammanco, G. Tabacchi, D. Di Majo, S. Giammanco, M. La Guardia, The phenolic compounds of olive oil: structure, biological activity and beneficial effects on human health, 2005, Nutrition Research Reviews, 18, 98-112.

[**3**] C. Markellos, M.E. Ourailidou, M. Gavriatopoulou, P. Halvatsiotis, T.N. Sergentanis, T. Psaltopoulou, Olive oil intake and cancer risk: A systematic review and meta-analysis, 2022, PLoS One, 17, e0261649.

Existen distintas categorías de calidad de aceite de oliva, sin embargo, sólo la comercialización de cuatro de ellos está permitida al consumidor, de acuerdo al Reglamento (UE) nº1308/2013 [**4**], siendo estos (de mayor a menor calidad y precio): **aceite de oliva virgen extra** (**AOVE**), **aceite de oliva virgen** (**AOV**), **aceite de oliva** (**AO**) y **aceite de orujo de oliva** (**AOO**). En primer lugar, el AOVE y AOV son aceites obtenidos del fruto del olivo, exclusivamente por medios mecánicos u otros procedimientos físicos aplicados en condiciones que excluyan toda alteración del producto, y que no se ha sometido a ningún otro tratamiento que no sea su lavado, decantación, centrifugado o filtración, excluidos los aceites obtenidos con el uso de disolventes o de coadyuvantes de acción química o bioquímica, por un procedimiento de reesterificación o como resultado de cualquier mezcla con aceites de otros tipos. Se corresponden con lo que podría identificarse con un zumo de aceituna. Por tanto, la diferencia entre un AOVE y un AOV radica, por un lado, en la acidez libre máxima expresada en ácido oleico, siendo de 0.8 g por 100 g para el AOVE y de 2 g por 100 g para el AOV; y, por otro lado, los 'defectos' organolépticos que se le otorgan a los AOV por parte de paneles de cata reconocidos, disminuyendo la calidad de AOVE a AOV. Posteriormente, el AO es aquel aceite que contiene exclusivamente aceites de oliva refinados (AOR) mezclados con una baja proporción de aceites de oliva vírgenes (AOVE o AOV) con una acidez libre en ácido oleico de no más de 1 g por 100 g. Por último, el AOO es aquel aceite obtenido de la mezcla de aceite de orujo de oliva refinado (AOOR) y aceites de oliva vírgenes (AOVE o AOV) con una acidez libre en ácido oleico de no más de 1 g por 100 g [**4**], entendiéndose por AOOR aquel aceite que es obtenido del refino de aceite procedente de extraer la fracción grasa del orujo, es decir, de los residuos sólidos generados durante la molienda de la aceituna, mediante un tratamiento con disolventes seguido de un posterior tratamiento químico.

[**4**] Reglamento de Ejecución (UE) Nº 1308/2013 del Parlamento Europeo y del Consejo por el que se crea la organización común de mercados de los productos agrarios y por el que se derogan los Reglamentos (CEE) nº922/72, (CEE) nº234/79, (CE) nº 1037/2001 y (CE) nº1234/2007.

(ii) Tequila

México es el único país en el cual se permite llevar a cabo el proceso completo de producción de tequila (desde la cosecha de la planta de la cual se extrae hasta su destilación), ya que está protegido por la 'Denominación de Origen Tequila' desde 1974 [**5**] y posteriormente reconocida en la Unión Europea en 1997 [**6**]. El Tequila, según lo dispuesto en la NOM-006-SCFI-2012 [**7**], es una bebida espirituosa alcohólica regional obtenida por la doble destilación de mostos, preparados directa y originalmente del material extraído de las cabezas (*piñas*) de la planta de agave de la especie *Tequilana Weber variedad azul*, previa o posteriormente hidrolizadas o cocidas, y sometidas a fermentación alcohólica con levaduras. Los mostos, por disposición legal, son susceptibles a ser enriquecidos y mezclados conjuntamente en la formulación con otros azucares hasta en una proporción no mayor de 49 % de azucares reductores totales expresados en unidades de masa. Cuando se lleva a cabo este enriquecimiento, el tequila obtenido pertenece a la categoría '***Tequila mixto***' o, simplemente, '*Tequila*'; por el contrario, cuando existe sólo la presencia de azucares obtenidos del agave *Tequilana Weber variedad azul* en el proceso de fermentación, se obtiene un tequila perteneciente a la categoría '***Tequila 100 % agave***', '*100 % de agave*', '*100 % puro de agave*' o '*100 % puro agave*'.

Asimismo, de acuerdo a las características adquiridas en procesos posteriores a la destilación, el tequila se clasifica en las clases: **Blanco** o **Plata**, **Joven** u **Oro**, **Reposado**, **Añejo** y **Extra añejo**. El primero de ellos, el Tequila Blanco, es un producto transparente no necesariamente incoloro, sin ningún tipo de edulcorante (*abocante*) o suavizado de sabor (*abocamiento*), obtenido de la destilación añadiendo

[**5**] Secretaría de Patrimonio y Fomento Industrial—Dirección General de Invenciones y Marcas. Declaración General de Protección a la Denominación de origen 'Tequila', Diario Oficial de la Federación (Gobierno de México). Número del oficio: 16-I.-57348 (9 diciembre 1974).

[**6**] Acuerdo entre la Comunidad Europea y los Estados Unidos Mexicanos sobre el reconocimiento mutuo y la protección de las denominaciones en el sector de las bebidas espirituosas, 2020, DO L 152 de 11.6.1997, p.16

[**7**] Norma Oficial Mexicana NOM-006-SCFI-2012, Bebidas alcohólicas-Tequila-Especificaciones, Comité Consultivo Nacional de Normalización de Seguridad al Usuario, Información Comercial y Prácticas de Comercio (CCNNSUICPC), Gobierno de México.

únicamente agua de dilución para ajustar la graduación alcohólica comercial requerida, teniendo una maduración menor de dos meses en recipientes de roble o encino.

El Tequila Reposado es un producto a ser abocado (con ingredientes como color caramelo, extracto de roble o encino natural, glicerina o jarabe a base de azúcar), sujeto a un proceso de maduración de por lo menos dos meses en contacto directo con la madera de recipientes de roble o encino. Posteriormente, el Tequila Añejo es aquel que ha sufrido un proceso de maduración de por lo menos un año y puede ser abocado, mientras que el Tequila Extra Añejo está sujeto a un proceso de maduración de por lo menos tres años [**7**].

❒ ***Problemas de adulteración y falsificación del aceite de oliva y Tequila***

El 70% de la producción del aceite de oliva en la Unión Europea (UE) y el 45% de la producción mundial [**8**] es realizada en España. Es un alimento con gran demanda debido a las características nutricionales que presenta y por los beneficios que aporta a la salud del consumidor, siendo uno de los alimentos más exportados y, por ende, de los más regulados y controlados en España [**9**] y en la UE [**10**]. Lo mismo sucede con el tequila, el cual es otro producto con gran demanda a nivel internacional, del cual México exportó al extranjero el 64 % de la producción total en 2022, por lo que es altamente regulado por el Consejo Regulador del Tequila (CRT) [**11**].

La alta demanda del aceite de oliva y tequila despierta un gran interés por parte de personas y/u organizaciones que buscan obtener beneficios económicos de manera ilícita realizando fraude alimentario, lo cual convierte a estos productos en blancos claros y propensos a sufrir modificaciones mediante la adulteración y/o falsificación.

[**8**] Ministerio de Agricultura, Pesca y Alimentación. Gobierno de España. https://www.mapa.gob.es/es/agricultura/temas/producciones-agricolas/aceite-oliva-y-aceituna-mesa/aceite.aspx

[**9**] Real Decreto 760/2021, de 31 de agosto, por el que se aprueba la norma de calidad de los aceites de oliva y de orujo de oliva.

[**10**] Reglamento de Ejecución (UE) 2019/1604 de la Comisión por el que se modifica el Reglamento (CEE) nº 2568/91 de la Comisión relativo a las características de los aceites de oliva y de los aceites de orujo de oliva y sobre sus métodos de análisis.

[**11**] Consejo Regulador del Tequila (CRT), México. https://www.crt.org.mx/

Por un lado, la adulteración hace referencia a la sustitución y/o adición deliberada y no declarada de sustancias o compuestos con el fin de modificar la composición intrínseca original del alimento para obtener mayores rendimientos económicos y de producción [**12**]. Por otro lado, la falsificación se define como la operación de imitación, reacondicionamiento, reconstrucción, alteración de las características de calidad, con el objetivo de obtener productos similares a los originales [**13**,**14**].

Entre los fraudes más comunes con respecto al aceite de oliva son aquellos relacionados con los AOVE y AOV, ya que son los que alcanzan los precios más superiores en el mercado. Uno de los fraudes es su mezclado con otros aceites de menor calidad, los cuales son más baratos de producir, aumentando los márgenes de producción y ganancias económicas de manera ilegal. Otro fraude es su mal etiquetado, declarando que un aceite de oliva de menor calidad (p.ej., refinado o de orujo) es AOVE, o también ocultar el verdadero lugar de origen del aceite [**15**]. Mientras que los fraudes más comunes relacionados con el tequila también tienen que ver con su mal etiquetado, indicando que un tequila de una clase inferior y más económica (Tequila Reposado) pertenece a una clase superior de mayor precio (Tequila Extra-Añejo); adición de compuestos edulcorantes a tequilas de clases inferiores para darles la apariencia de tequilas añejos o extra añejos y, la más grave de todas por poner en riesgo la vida del consumidor, la producción de tequila o mezclado de tequilas auténticos con alcoholes no permitidos, como lo es el metanol en la mayoría de los casos, o propanol, etilenglicol, entre otros [**16**,**17**], lo cual sigue siendo un problema hoy en día a

[**12**] J. Rees, Food adulteration and food fraud, 2020, Reaktion Books Ltd.

[**13**] J.P. Battershall, Food adulteration and its detection, 2019, Good Press.

[**14**] C.M. Canja, A. Mazarel, M.I. Lupu, V. Padureanu, D.V Enache, Foodstuff falsification – a nowadays problem, 2016, Bulletin of the Transylvania University of Brasov, Series II, Vol. 9, 69-74.

[**15**] J. Lozano-Castellón, A. López-Yerena, I. Domínguez-López, A. Siscart-Serra, N. Fraga, S. Sámano, C. López-Sabater, R.M. Lamuela-Raventós, A. Vallverdú-Queralt, M. Pérez, Extra virgin olive oil: A comprehensive review of efforts to ensure its authenticity, traceability, and safety, 2022, Comprehensive reviews in food science and food safety, 21, 2639-2664.

[**16**] G. Pérez-Caballero, J.M. Andrade, P. Olmos, Y. Molina, I. Jiménez, J.J. Durán, C. Fernandez-Lozano, F. Miguel-Cruz, Authentication of tequilas using pattern recognition and supervised classification, 2017, Trends in Analytical Chemistry, 94, 117-129.

[**17**] D.G. Barceloux, R. Bond, E.P. Krenzelok, H. Cooper, J.A. Vale, American academy of clinical toxicology practice guidelines on the treatment of methanol poising, 2002, Clinical Toxicology, 40, 415-446.

pesar del gran número de métodos analíticos existentes para verificar la autenticidad del tequila [**18**].

Para contribuir al aseguramiento y control de calidad de manera global del aceite de oliva y tequila, y al contrarresto de las actuales problemáticas de adulteración con productos nocivos para la salud y de falsificación con productos de menor calidad, se desarrollaron dos MAM basados en huellas instrumentales cromatográficas agnostizadas para que puedan ser fácilmente transferibles entre laboratorios para el control de calidad alimentario de estos productos alimenticios. El primer MAM se utilizó para detectar adulteraciones de aceites de oliva virgen (AOV), virgen extra (AOVE) y aceite de oliva (AO) mezclados con aceites de orujo de oliva (AOO) y aceites de oliva refinados (AOR). El segundo MAM se empleó para la autentificación y diferenciación de las categorías '100 % agave' y 'mixto' del Tequila Blanco. Estos MAM fueron desarrollados por primera vez y expresamente para esta tesis doctoral.

A continuación, se presentan los artículos científicos derivados del desarrollo, aplicación y transferencia de los MAM previamente mencionados, en los cuales se detallada a profundidad cada uno de ellos, así como los resultados y conclusiones a las cuales se llegó.

[**18**] W.M. Warren-Vega, R. Fonseca-Aguiñaga, L.V. González-Gutiérrez, L.A. Romero-Cano, A critical review on the assessment of the quality and authenticity of tequila by different analytical techniques: recent advances and perspectives, 2022, Food Chemistry, 408, 135223.

3.2. Artículo científico II

Food Control 137 (2022) 108957

Contents lists available at ScienceDirect

Food Control

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

Instrument-agnostic multivariate models from normal phase liquid chromatographic fingerprinting. A case study: Authentication of olive oil

Christian H. Pérez-Beltrán, Ana M. Jiménez-Carvelo *, Sandra Martín-Torres, Fidel Ortega-Gavilán, Luis Cuadros-Rodríguez

Department of Analytical Chemistry, Faculty of Sciences, University of Granada, C/ Fuentenueva, s/n, E-18071, Granada, Spain

ARTICLE INFO

Keywords:
Max 6)
Instrument-agnostic chromatographic fingerprints
Instrument-independent multivariate models
Data mining and chemometrics
Olive oil authentication

ABSTRACT

The application of non-targeted analytical strategies such as instrumental chromatographic fingerprinting is commonly applied in the field of food authentication/food quality. Although the multivariate methods developed to date are able to solve any authenticity problem, they remain dependent on the instrument state where the signals were acquired, which difficult their transfer to other laboratories. The aim of this research is to develop multivariate models independent of both instrument state and time at which the signals were acquired. For this, chromatograms obtained from the polar fraction of different olive oil samples analysed by (NP)UHPLC-UV/Vis are transformed to instrument-agnostic fingerprints. Instrument independence is achieved by transferring the chromatographic behaviour of an 'ad-hoc' external standards mixture solution analysed throughout an analysis sequence to the remaining analysed samples.

The SIMCA models developed from the chromatographic fingerprint matrix before and after instrument-agnostizing showed significant differences in the number of samples classified as "inconclusive", with the after model showing the best results. Furthermore, the PLS-DA and SVM models obtained before and after signal instrument-agnostizing showed similar outcomes. The main conclusion of the work has been to verify that the instrument-agnostizing methodology could allow the building of multivariate classification models which could be transferred among different laboratories as they are not influenced by the signal acquisition time.

1. Introduction

The untargeted approach is an emergent approach which is increasingly used in the field of food authentication/food quality. Untargeted methodology is focused on the study of unspecific instrumental signals without taking on any previous knowledge of relevant/irrelevant food components and it is mainly represented by fingerprinting methodology (Muñoz Olivas, 2004; Creydt & Fischer, 2020). In this sense, the instrumental fingerprint of a foodstuff can be defined as a non-specific signal that contains sufficient information about the chemical composition or structure of a food product or a food commodity to be able to unequivocally characterise and/or differentiate it from others similar foodstuffs (Cuadros Rodríguez et al., 2016a).

The application of instrumental fingerprinting methodology involves resorting to advanced mathematical data processing methods to extract useful information which is not obvious and not explicitly shown, such as data mining/chemometric methods. Usually, the recorded analytical signal is subjected to a previous pre-processing in order to clean it before being used for the development of a multivariate model. The most commonly used pre-processing techniques are, autoscaling, mean centring, noise filtering, baseline correction and normalization (Jiménez Carvelo et al., 2020). In the case of chromatographic signals, it is also necessary to carry out a peak alignment. This last pre-processing step is probably the most important since retention times (RT) are often shifted among chromatographic analyses. There are different algorithms for peak alignment being COW (Tomasi et al., 2004) and icoshift (Tomasi et al., 2011) the most commonly used in chromatography.

A large amount of literature is available on analytical methods using different analytical techniques together with data mining/chemometric methods in the food science field, which are focused on solving almost any authentication or quality problems (Boccard & Rudaz, 2020; Jiménez Carvelo & Cuadros Rodríguez, 2021; Oliveri et al., 2020; Tahir et al., 2022). Despite of, there is an important challenge still to be solved: reposted multivariate methods are based on the instrumental

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

https://doi.org/10.1016/j.foodcont.2022.108957
Received 24 November 2021; Received in revised form 24 February 2022; Accepted 5 March 2022
Available online 8 March 2022

Instrument-agnostic multivariate models from normal phase liquid chromatographic fingerprinting. A case study: authentication of olive oil

Christian H. PÉREZ-BELTRÁN, Ana M. JIMÉNEZ-CARVELO✉, Sandra MARTÍN-TORRES, Fidel ORTEGA-GAVILÁN, Luis CUADROS-RODRÍGUEZ

Department of Analytical Chemistry, Faculty of Sciences, University of Granada, C/ Fuentenueva, s/n, E-18071 Granada (Spain).

Highlights

- Instrumental fingerprints of different olive oil classes were acquired by UHPLC-UV/Vis.
- Steps for instrument-agnostizing of chromatographic signals are described.
- Adulteration of olive oils was detected with 100% precision and sensibility.
- Suitable multivariate models were developed from instrument-agnostic chromatographic fingerprints.

Keywords

Instrument-agnostic chromatographic fingerprints; instrument-independent multivariate models; data mining and chemometrics; olive oil authentication

✉ Corresponding author (E-mail: amariajc@ugr.es)

1. Introduction

The untargeted approach is an emergent approach which is increasingly used in the field of food authentication/food quality. Untargeted methodology is focused on the study of unspecific instrumental signals without taking on any previous knowledge of relevant/ irrelevant food components and it is mainly represented by fingerprinting methodology (Muñoz Olivas, 2004; Creydt & Fischer, 2020). In this sense, the instrumental fingerprint of a foodstuff can be defined as a non-specific signal that contains sufficient information about the chemical composition or structure of a food product or a food commodity to be able to unequivocally characterise and/or differentiate it from others similar foodstuffs (Cuadros Rodríguez et al., 2016a).

The application of instrumental fingerprinting methodology involves resorting to advanced mathematical data processing methods to extract useful information which is not obvious and not explicitly shown, such as data mining/chemometric methods. Usually, the recorded analytical signal is subjected to a previous pre-processing in order to clean it before being used for the development of a multivariate model. The most commonly used pre-processing techniques are, autoscaling, mean centring, noise filtering, baseline correction and normalization (Jiménez Carvelo et al., 2020). In the case of chromatographic signals, it is also necessary to carry out a peak alignment. This last pre-processing step is probably the most important since retention times (RT) are often shifted among chromatographic analyses. There are different algorithms for peak alignment being COW (Tomasi et al., 2004) and icoshift (Tomasi et al., 2011) the most commonly used in chromatography.

A large amount of literature is available on analytical methods using different analytical techniques together with data mining/chemometric methods in the food science field, which are focused on solving almost any authentication or quality problems (Boccard & Rudaz, 2020; Jiménez Carvelo & Cuadros Rodríguez, 2021; Oliveri et al., 2020; Tahir et al., 2022). Despite of, there is an important challenge still to be solved: reposted multivariate methods are based on the instrumental fingerprints acquired by a single analytical instrument, at a specific and under particular conditions; this are instrument-sensitive fingerprints. This leads to multivariate models which are dependent of the analytical laboratory where the data have been acquired.

The performance of those models with samples analysed by different manufacturer instrument or by the same one but in different time periods is largely unknown and by experience unsuccessful. Focusing towards a global or universal model, more fundamental work is required.

Despite of different attempts to create universal linear retention indexes (LRI) in liquid chromatography (Rigano et al., 2021), such activity aims to identify the compounds of interest in the sample and not to standardize the instrumental fingerprint. In fact, a common occurrence is to find some non-negligible variations in retention times or even in peak intensities when carrying out replicate chromatographic analysis. These scrolling on the axes of the chromatographic intensity/time signals (or chromatograms) make it difficult to use the recorded signals to create a representative database capable of being used for reliable comparisons/identifications or for multivariate classification or quantitation model building.

Some proposal regarding the standardization of spectroscopic signals, usually NIR and Raman for quality control purposes in the medical and pharmaceutical fields, have been reported (Fornasaro et al., 2020; Gou et al., 2018; Zhang et al., 2019). This methodological practice has been called ’instrument cloning’ and is applied in order to obtain a ’transfer model’, mainly calibration models for analytical quantitation, was proposed by Wang et al. (1991). This procedure is based on the statement that the position coordinates of a spectrum are characteristic of each instrument and that they remain practically invariant over time. Generally, calibration transfer is implemented as follows: the spectrum of a sample obtained by the NIR or Raman equipment from which a particular multivariate calibration model has been developed (Master instrument) must have a similar spectral profile to the spectrum obtained by the equipment to which the model is to be transferred (Satellite instrument) (Folch Fortuny et al., 2017). Despite the recent industrial and technological progress, the spectra obtained by different NIRS instruments differ for various reasons, among which the instrument configuration and optics are the most common. Thus, each transfer model is only applicable to pairs of instruments.

However, no equivalent strategy was suggested specifically for chromatographic signals. In this context, Cuadros Rodríguez et al. (2021a, 2021b) have recently proposed an innovative methodology to be followed in order to obtain standardized instrumental fingerprints when the gas and liquid chromatography are employed; this methodology has been termed by the authors as instrument-agnostizing (Cuadros Rodríguez et al., 2021a, 2021b). It was proof to be able to standardize conventional chromatograms so that the new instrument-agnostic signal (fingerprint) is independent of the chromatographic state or the date of analysis, so that chromatographic fingerprints acquired from different instruments states (two or more) should have a high degree of similarity. For this purpose, both internal and external chemical standards series are used as instrumental references. Briefly, this methodology is summarised below: firstly, it is performed a stage for setting up an invariant set of standard retention scores (SRS) from the external standards, which is only applied once; then, the agnostizing step is performed in which both intensities and retention times of the signal is standardized using the previously established SRS. Note that this is the first methodology that attempt to obtain a database of EVOO instrumental fingerprints and, thus it can be employed as potential tool to achieve multivariate 'instrument-agnostic' models.

Olive oil is one of the main vegetable oils chosen by consumers due to its nutritional characteristics and health benefits, being one of most regulated and controlled foodstuffs in the European Union (EU). It should be noted that EU legislation allows the blending of olive oil with other vegetable oils, however, some European producer countries, such as Spain or Italy, have specific legislation which forbids the blending of olive oil with other vegetable oils. There are three different European marketing quality categories of edible olive oil: (i) extra virgin olive oil (EVOO), (ii) virgin olive oil (VOO) and (iii) olive oil (OO), the latter being a blend of chemically refined olive oil and EVOO/VOO. These oils vary in price and quality, due to their organoleptic and physico-chemical properties. In fact, EVOO and VOO achieve much higher prices on international markets than any other type of vegetable oil, which makes it potentially considered to be adulterated with lower quality edible vegetal oils, such as seed oils (e.g., sunflower oil), refined olive (ROO) oil and/or olive-pomace oil (OPO), in order to obtain a higher illicit profit. Currently, the European official method of analysis used to detect adulteration of EVOO/VOO with ROO or OPO involves carrying out several chemical analyses in order to determine specific analytical

parameters such as ECN42 and to quantify some particular compounds (chemical markers) or family of compounds such as triterpene dialcohols or waxes, among others, using different sample treatments and/ or analytical procedures for each one (Commission Regulation (EEC) No. 2568/91). As an example, the triterpene dialcohols such as erythrodiol and uvaol are separated from the unsaponifiable matter by thin-layer chromatography on a basic silica gel plate. The fractions recovered from the silica gel are derivatised into trimethylsilyl ethers and then analysed by gas chromatography. Thus, this method is highly time-consuming and entails a large consumption of chemicals.

Moreover, a wide number of different procedures for the adulteration detection of EVOO/VOO with different edible oils (sunflower, soybean, peanut, corn, rapeseed, hazelnut oils, among the most common) have been proposed (Zhang et al., 2021; Meenu et al., 2019). Basically, two methodologies outstand for this purpose: i) the use of nontargeted spectroscopic approaches, such as Raman (Duraipandian et al., 2019) and Fourier transform infrared (FTIR) (Abdallah et al., 2016; Karunathilaka et al., 2016); and ii) the use of high-performance liquid chromatography (HPLC) or gas chromatography (GC) for quantitative analysis of peculiar marker compounds (Mingchih et al., 2015; Jabeur et al., 2017).

Nonetheless, the adulteration of EVOO/VOO with refined olive oil and/or olive-pomace oil by means of HPLC and chemometrics has been addressed to a lesser extent. In fact, it was possible to find only five research studies involving one or both of these topics which mainly mass spectrometry (MS) as detection system and targeted approach are employed on minor polar compounds (Carranco et al., 2018; Drira et al., 2020; Li et al., 2021; Navratilova et al., 2022; Tata et al., 2022).It should be noted that all these studies were performed with non-standardized signals for the development of the multivariate models, what limit their implementation to routine analytical laboratory, being applicable only under the conditions of measurement under which they were carried out.

To date, the development of a single multivariate instrument-agnostic model has not been proposed in food authentication field using chromatographic signals. In this context, the current study proposes a multivariate analytical method for the detection of olive oil adulteration with ROO or OPO using agnostic-instrument chromatographic fingerprints for the first time.

In this sense, this study proposes the use of the chromatographic fingerprints from the polar compounds fraction of the olive oils of different quality categories, acquired using normal phase ultra-high-performance liquid chromatographic coupled to an ultraviolet–visible molecular absorption detector ((NP)UHPLC-UV/Vis), as a source of analytical information to set up instrument-agnostic multivariate classification models. The discrimination results from each method and strategy were compared and ranked using several classification performance metrics, such as sensitivity, specificity, precision, efficiency (or accuracy), area under the receiver operative curve (AUC), among others. More details on the specific features of the classification strategies and the meaning of the classification metrics can be found in the tutorial published by Cuadros Rodríguez et al. (2016b).

2. Materials and methods

2.1. Chemicals

HPLC-grade solvents, such as n-hexane, 2-propanol and ethanol were employed within the study. *N*-Hexane was purchased from Panreac Quimica S.L.U. (Barcelona, Spain), 2-propanol from Honeywell (Deutschland, Germany) and ethanol from VWR (Darmstadt, Germany). Deionized water was obtained using a Milli-Q system (Millipore, Bedford, MA).

Chemical standards, such as 1,2,3-trimethyl benzene (TMB) provided by Sigma-Aldrich (St. Louis, USA), propiophenone (PROP) provided by AlfaAesar (Kandel, Germany), 2,5-dimetylphenol (2,5-DP) provided by Sigma-Aldrich (St. Louis, USA), 2-naftol (2-NAF) provided by ACROS (Geel, Belgium) and ethyl paraben (EPB) provided by Fluka Chemika (Buch, Germany) were employed to create the external standard mix (ESM) solution. Each of the chemicals were added into the mix at 12, 100, 16, 4 and 40 mg/L, respectively, using n-hexane/2-propanol 99/1 (v/v) as solvent.

2.2. Samples

A total of 88 vegetables oils samples were analysed: 35 extra-virgin olive oil samples (EVOO) of different regions from Spain, 4 virgin olive oils (VOO), 4 olive oil (OO), 5 refined olive oil (ROO), 4 olive-pomace oil (OPO), and 36 blends (BLE) of EVOO or VOO with ROO or OPO. These blends represented adulterated olive oils with other olive oils of poorer quality in 20, 40 and 60%.

2.3. Sample preparation

1 g of oil was placed in a 10 ml tube and 4 ml of n-hexane were added into the tube for further agitation with vortex for 10 s. Then, 1 ml of ethanol/water 87/13 (v/v) mixture was added and vortexed for another 10 s. The polar fraction at the bottom of the tube was extracted and this step was repeated twice. Finally, the polar fraction was centrifuged for 3 min at 1500 g and further filtered with 0.22 μm nylon filters. The polar fraction solutions were frozen (-4 ºC) and kept in the dark until analysis.

2.4. Chromatography

(NP)UHPLC-UV/Vis analysis was performed with a Dionex Ultimate 3000 UHPLC + Focused chromatography system (Thermo Scientific, Waltham, MA, USA) equipped with a RS autosampler and column compartment. Detection was performed with an RS variable wavelength detector. Chromeleon™ version 7.0 software was used to visualize and export data. A silica stationary phase column (ZORBAX RX-SIL, 150 × 4.6 mm i.d, 5 μm) coupled to a pre-column with the same diameter (12.5 × 4.6 mm) were used through all the analysis. Both pre-column and column were kept at 35 ◦C during the experimental work.

The chromatographic analysis of the ESM was performed 32 times along 6 days in order to considerer in the calculation of the SRS as much variability as possible. Additionally, the ESM was analysed at the beginning and at the end of each chromatographic run for further calculation of the SRS, and as quality control of the behaviour of the equipment. For this purpose, 1.5 ml of the ESM were placed in a chromatographic glass vial for its corresponding analysis with a flow rate of 0.8 ml/min during the entire operation. The gradient mode of the mobile phase was the following: the ESM was injected at time 0 and was eluted with hexane for 15 min.

Then, solvent was changed to hexane-isopropanol 90/10 (v/v) for 2 min. Finally, from minute 17 to minute 21, the chromatographic system came back to the initial conditions of hexane 100%. Note that during the second day, ESM from the first day was analysed together in the same batch with ESM of day two; for day three, ESM from day two was analysed with ESM from day three; the same process was followed for the remaining days.

Just before the chromatographic analysis, 750 mL of the polar fraction solution, previously thawed, were added into a 2 ml chromatographic glass vial, and then 180 ml of TMB solution (100 mg/L in n-hexane/2-propanol 99/1, v/v) were added as an internal control standard. The vial was sealed and vortexed for 20 s and 5 mL of this solution were injected in the LC equipment. A flow rate of 1.2 ml/min was kept during the entire operation. The gradient mode of the mobile phase was the following: the samples were injected at time 0 and were eluted with hexane for 1 min, at a flow rate of 1.2 mL/min. Then, solvent was changed to hexane-isopropanol 80/20 (v/v) for 3 min. Afterwards, the solvent was changed again to hexane/isopropanol 60/40 (v/v) for 4 min, going back to 80/20 (v/v) after 2 min. Finally, from min 10 to minute 13, the system came back to the initial conditions of hexane 100%.

2.5. Methodology: development of a multivariate model from instrument-agnostic chromatographic fingerprints

In order to be able to have multivariate models for common use, the chromatographic signals must be standardized, and then the multivariable models are developed. In this regard, the following steps were needed to obtain the instrument-agnostic chromatograms. All data (88 samples × 1950 variables) used in this study were exported from the instrument software to an Excel environment (.csv, comma separated values), and then converted to Matlab environment (.mat). In this way, each chromatogram was firstly turned into a data vector.

The description of the process of standardization of signals as well as building of the multivariate model can be summarised in 6 major steps:

1. Application of the automatic Whittaker filter to correct the baseline of raw chromatograms in which values of $\lambda = 100$ and $p = 0.001$ were selected. λ indicates the baseline curvature to allow (the smaller this value, the more curved the baseline fit will be), whilst 'p' ($0 < p < 1$) indicates the asymmetry to use in the Whittaker filter (the smaller this value, the smaller the allowed negative proportion of the result that has been adjusted) (Wise et al., 2006).
2. Selection of the data to create the training and external validation data sets. The choice was performed through the Kennard-Stone algorithm already implemented in Matlab. The proportion of samples to include in the training and validation data sets was 70 and 30%, respectively. At the end, the training data set was composed by 61 samples, whilst the external validation set by 27.
3. Intensity normalization of the chromatogram from the 88 oil samples to create a homogeneous intensity scale. All the chromatographic intensities (height) were normalized taking as a reference the maximum peak of the internal standard TMB, assigning an intensity value = 1 on the y-axis. The peak of TMB was found among variables number 224 and 250.
4. Establishment of the standard retention scores (SRS) according to the protocol given by Cuadros et al. (2021b).
5. Replacement of the retention time values for the calculated SRS in each sample signal, in order to unify the scale on the x-axis.
6. Re-sampling to fix into a same number of variables all chromatograms. The specified range for the re-sampling function was from 1 to 5.8 considering the SRS scale, obtaining a variable reduction from 1256 to 575.

Once the chromatograms were standardized, different multivariate models were developed. The corresponding information can be found in the following section.

2.6. Multivariate analysis

After performing the six major steps outlined in subsection 2.5, exploratory analysis using principal component analysis (PCA) was performed to detect possible natural grouping or outliers in the data.

Furthermore, soft independent modelling of class analogies (SIMCA), partial least squares - discriminant analysis (PLS-DA) and support vector machine (SVM) were used to create alternative classification models capable to identify EVOO and VOO, and detect blends of these olive oils adulterated with ROO and OPO). For this, PLS_Toolbox (version 8.6.1, 2019, Eigenvector Research Inc., Manson, WA, USA) was used under Matlab (version R2013b, 8.2.0.701, The Mathworks Inc. MA, USA) environment.

3. Results and discussion

The raw chromatograms obtained can be observed in Fig. 1. The chromatograms from EVOO and VOO have some similarities around minutes 4.00 - 4.20 and 4.80 - 5.40 that are attributed to their chemical composition. However, these same two chromatograms also show differences of intensity between minutes 4.30 - 4.70 and a lack of a small peak around minute 6.80. Such differences could be attributed to the presence of small concentrations of defective compounds of the VOO that diminish its category from EVOO to VOO.

Additionally, in Fig. 1 B it can be appreciated that OO, ROO, OPO and BLE have similar fingerprints around times 4.70, 5.20 and 5.40 min. The pattern among the first three (OO, ROO and OPO) might be due to the similar chemical composition of these oils as they have all suffered a refining process to a greater or lesser extent. Moreover, the intense peak around at 1.40-1.60 min corresponds to the internal standard TMB added to each sample. In this regard, it can be observed that EVOO and VOO have a particular feature with respect to the chromatograms of the other oils, specifically, a bit more intense peaks around minutes 4.20–4.60.

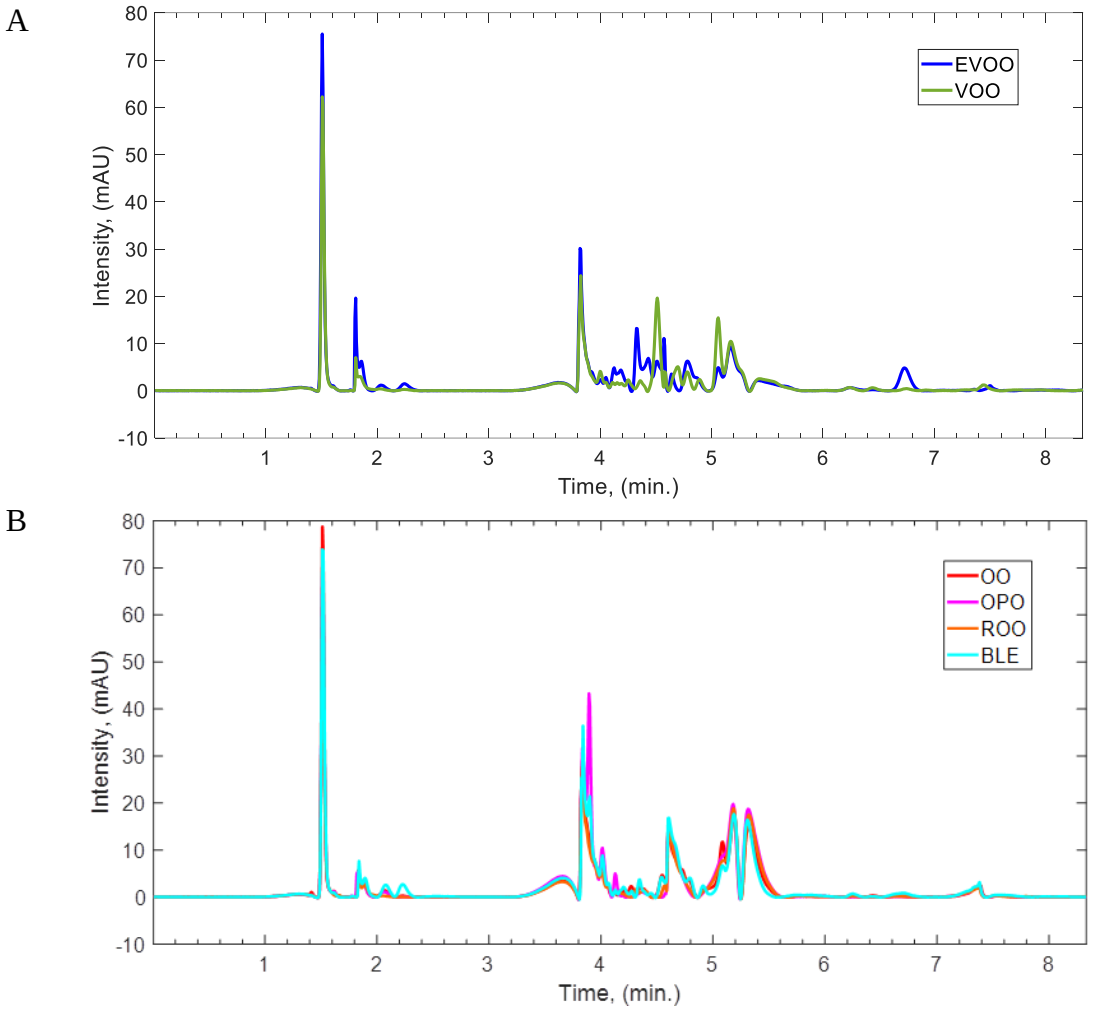


Figure 1. Overlapped raw chromatograms of six olive oils considering one representative sample per olive oil category: A) EVOO and VOO, and B) OO, ROO, OPO and BLE. EVOO: extra virgin olive oil, VOO: virgin olive oil, OO: olive oil, ROO: refined olive oil, olive-pomace oil (OPO) and BLE: blend of EVOO or VOO with ROO or OPO.

In order to be able of agnostizing the raw chromatograms, the ESM was analysed before and after each analytical run (the corresponding chromatogram is shown in Fig. 2). The first eluted compound was the internal standard (TMB) around 2.30 min, then PROP, 2,5-DP and 2-NAF at 4.37, 8.47 and 9.46 min respectively.

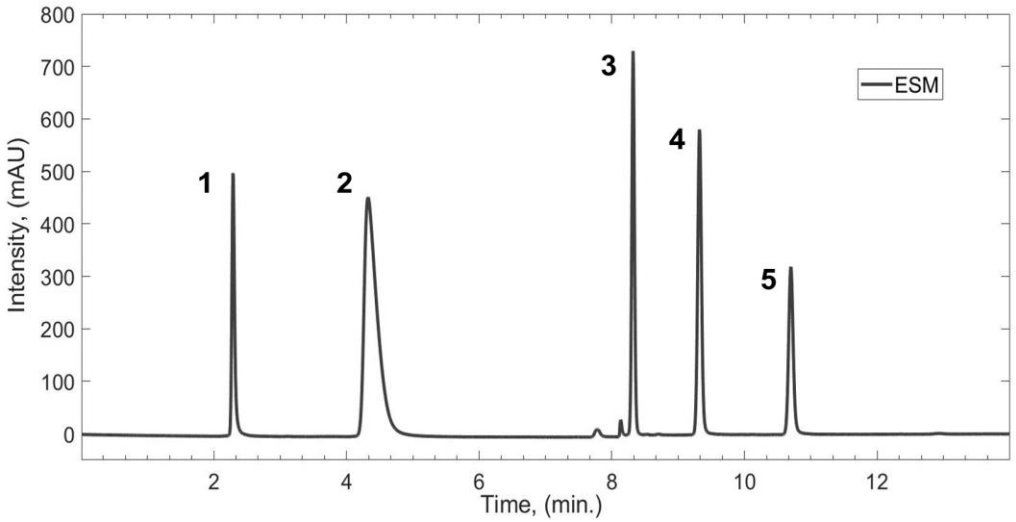


Figure 2. Chromatogram of the external standard mix (ESM) composed by five different chemical components: 1) 1,2,3-trimethyl benzene (TMB); 2) propiophenone (PROP); 3) 2,5-dimetylphenol (2,5-DP); 4) 2-naftol (2-NAF); and 5) ethyl paraben (EPB).

Then, the first step was to normalize the intensity of all chromatograms considering the most intense peak before minute 2, which was the internal standard TMB. Afterwards, an invariant reference chemical system for normalizing the retention values was established. For this purpose, the estimation of SRS values was performed (see subsection 2.5). The second step focused on the retention time normalization remains on the transference of retention standard scores from the ESM to any intensity-normalized chromatogram and involves the transformation of the chromatographic intensity-normalized vectors from the instrumental-dependent RT domain to an instrumental-agnostic SRS domain. Fig. 3 shows the overlapped chromatograms of the same six samples plotted in Fig. 1 after intensity normalization (time domain) and after agnostizing methodology on SRS domain.

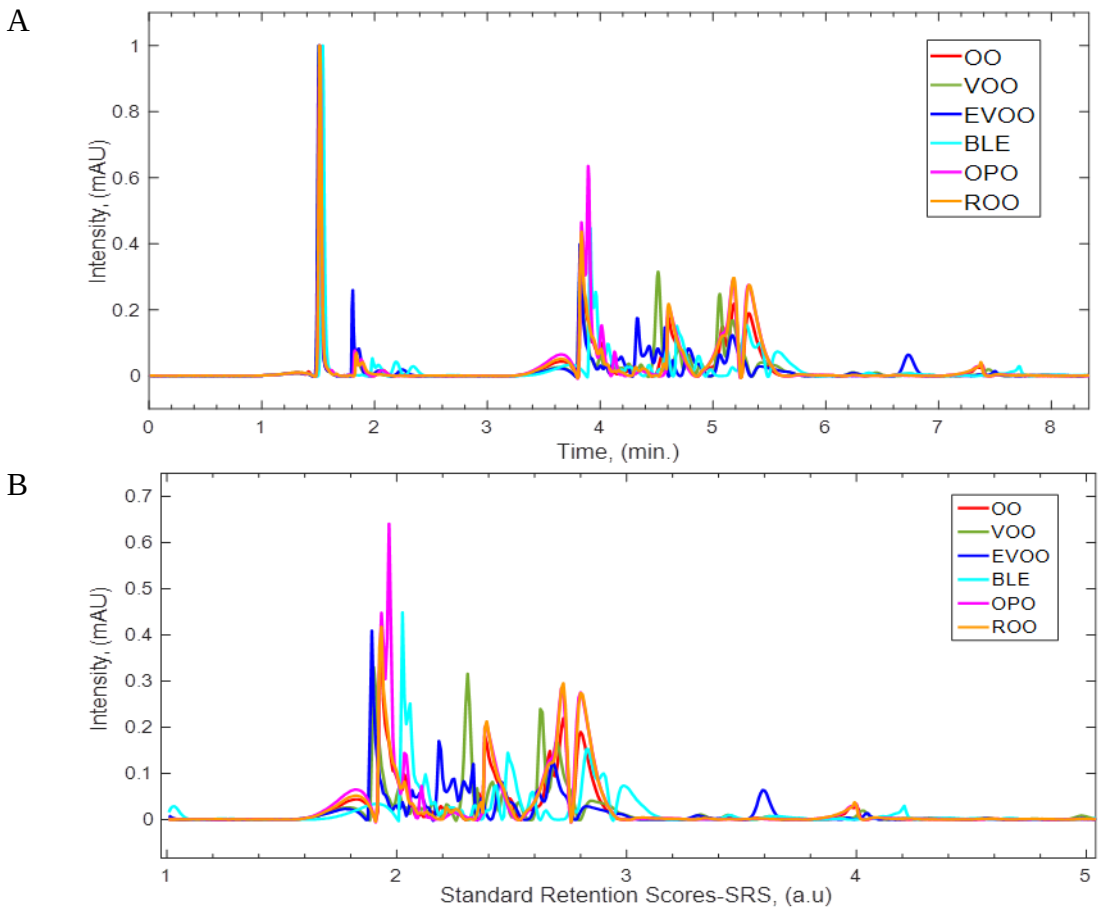


Figure 3. Overlapped chromatograms of the same six different olive oil samples as Figure 1 A) after intensity normalization (time domain) and B) after retention time normalization (SRS domain) EVOO: extra virgin olive oil, VOO: virgin olive oil, OO: olive oil, ROO: refined olive oil, olive-pomace oil (OPO) and BLE: blend of EVOO or VOO with ROO or OPO.

3.1. Exploratory analysis

The raw chromatograms (which were conventionally aligned using icoshift algorithm) and the instrument-agnostic fingerprints were used to perform the different steps of this multivariate study. Note that when aligned raw chromatograms are used as input data to develop multivariate models, they will be referred to as before-agnostizing models. Conversely, if instrument-agnostic fingerprints are used, the after-agnostizing model term is then employed.

First, an exploratory PCA was performed for both data sets and their corresponding score plots can be observed in Fig. 4 A and B, respectively. The before-agnostizing PCA model was built considering 3 principal components (PCs) which explained 73% of the total variance with a root mean square error for cross validation (RMSECV) = 1.87, whilst after-agnostizing PCA model was developed with 4 PCs explaining 67% of the total variance with RMSECV = 0.03. Only PC1 vs PC2 were used to perform the score plots in both cases, since they provided the best grouping overview. The EVOO and VOO were grouped together in both PCA score plots; the same ensued for the OO, ROO and OPO samples, which could be attributed to the similar chemical composition.

However, the different pattern in the two scores plots deserves further comment. Fig. 4A clearly shows the differentiation between the two groups mentioned above, but the same is not evident in the layout shown in Fig. 4B. This is because PCA is not a classification method, but outputs groups based on the variability observed in the corresponding input signals. These results in the BLE oils being further separated into three subgroups, possibly because the agnostizing of the signals enhances the dissimilarity amount blended oils. However, this fact does not hinder the aim of the exploratory analysis, which was to show that there is a difference between the two concerned groups, which is clearly evident. As a consequence, the application of appropriate classification methods should yield good results.

In the same regard as the results presented in this study, Drira et al. (2020) could identify grouping trends applying PCA among the EVOO and the EVOO/OPO adulterated samples using the profile from the phenolic compounds, sterolic composition and antioxidants. Nonetheless, authors used only nine samples in total within the study, which is a very low number of samples to ensure that the used information of the chromatographic profiles is sufficient enough to discriminate between the different vegetable oils. Navratilova et al. (2022) could not observe clear groups of EVOO and EVOO/ROO analysing the polar fingerprints with PCA. Finally, Carranco et al., 2018 also used chromatographic polar fingerprints of different EVOO samples, and EVOO samples adulterated with ROO and sunflower oil as analytical signal for PCA.

As a result, it was possible to find a tendency of the olive oils against the other vegetable oils samples, but authors do not mention if there was some pattern of EVOO against adulterated EVOO with ROO.

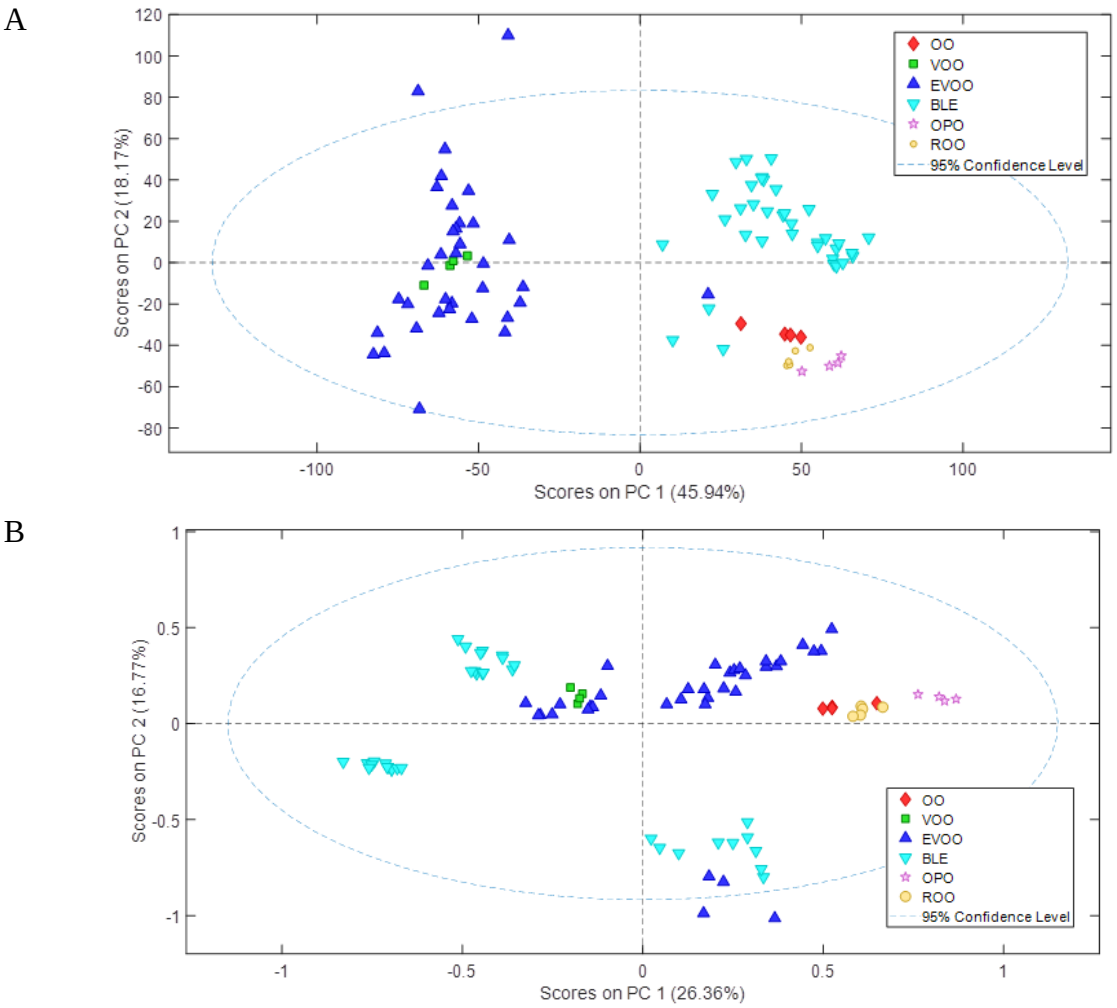


Figure 4. Exploratory PCA score plots of 88 olive oil samples belonging to six different quality categories: A) before agnostizing, and B) after agnostizing. EVOO: extra virgin olive oil, VOO: virgin olive oil, OO: olive oil, ROO: refined olive oil, olive-pomace oil (OPO) and BLE: blend of EVOO or VOO with ROO or OPO.

On the contrary to these studies, the current study included a wide number of samples, the PCA displayed a better grouping of all samples and it was possible to distinguish EVOO and VOO from OO, OPO, ROO and adulterated samples with PCA.

This demonstrates that the instrumental fingerprint of the polar fraction contains the information of interest to authenticate the olive oil as discussed above.

3.2. Authentication of olive oil – discrimination multivariate models

For the sake of clarity, only the characteristics of each after-agnostizing model developed with instrument-agnostic fingerprints, as well as the classification plots, classification contingencies and classification performance metrics tables for each one is presented here. In order to compare these outcomes with the results of the before-agnostizing multivariate models, classification plots, contingencies and metrics tables can be found in supplementary material.

Several multivariate classification models were built using both data sets (before and after agnostizing) employing SIMCA, PLS-DA and SVM as data mining methods to find the best multivariate method capable to differentiate among EVOO, VOO or OO from ROO, OPO or BLE. For this, two classes were considered to build the models: class 1 (EVOO/VOO/ OO) and class 2 (ROO/OPO/BLE). The training was set of 61 samples (24 EVOO, 2 VOO, 3 OO, 4 ROO, 3 OPO and 25 BLE) and further validated with a data set of 27 samples (11 EVOO, 2 VOO, 1 OO, 1 ROO, 1 OPO and 11 BLE), as outlined in subsection 2.5.

Firstly, SIMCA was employed. It is a multivariate classification method that builds models based on PCA and considers the classes independent from each other. For this particular case, the model was performed using 3 PCs for class 1 and 4 PCs for class 2, which explained 62.34% and 88.70% of the total variance, respectively. The classification outcomes can be evaluated using the Coomans' plot which is showed in Fig. 5. Coomans' plot is a visual representation of the separation between two classes, in which the two axes represent the normalized orthogonal distances of all the samples respect to each individual model. Optimally, the validation samples should be classified in the class 1 or class 2. In real conditions, some validation samples could be assigned to both classes simultaneously, in this case these samples are considered as inconclusive ones. In addition, some samples can be not recognized as belonging to any class.

It can be observed in Fig. 5 that there are three validation samples placed in the 'inconclusive' quadrant (bottom-left quadrant) which are samples of OO, ROO and blend of VOO (80%) with OPO (20%), respectively. The classification contingency is shown in Fig. 6.

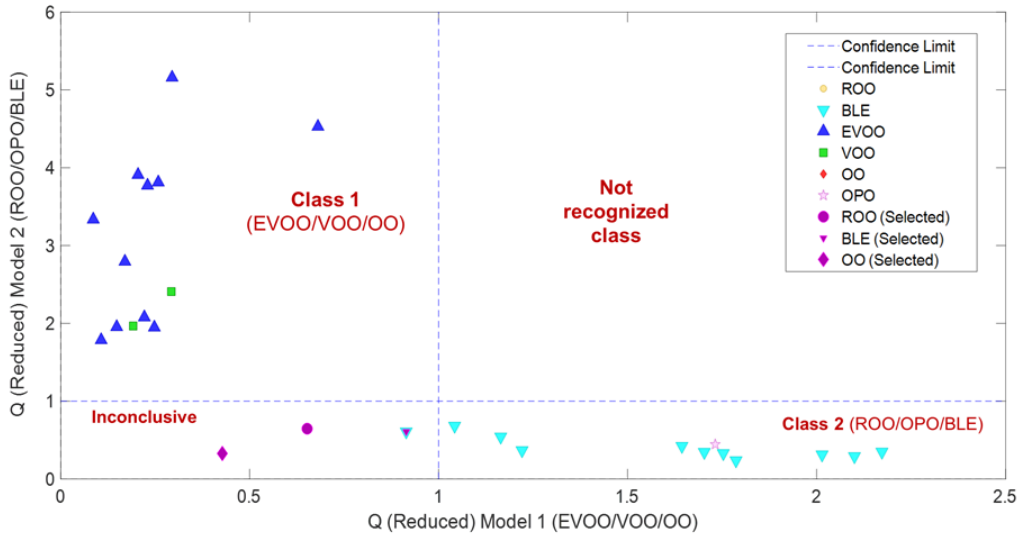


Figure 5. Cooman's classification plot of the validation set samples from the after-agnostizing SIMCA model. EVOO: extra virgin olive oil, VOO: virgin olive oil, OO: olive oil, ROO: refined olive oil, olive-pomace oil (OPO) and BLE: blend of EVOO or VOO with ROO or OPO. *(Left upper quadrant 'Class 1' includes EVOO, VOO, and OO samples; right upper quadrant is for not recognized samples; left bottom quadrant is for inconclusive samples; right bottom quadrant 'Class 2' includes ROO, OPO and BLE samples; 3 samples were classified as 'inconclusive').*

		14	13	27
Assignation	Not recognized (Nr)	0	0	0
	Inconclusive (I)	1 (3.70%)	2 (7.41%)	3
	Class 2 (ROO/OPO/BLE)	0	11 (40.74%)	11
	Class 1 (EVOO/VOO/OO)	13 (48.15%)	0	13
		Class 1	Class 2	
		Actual		

Figure 6. Validation contingencies from the after-agnostizing PLS-DA classification model. Class 1 (target class): EVOO: extra virgin olive oil, VOO: virgin olive oil, OO: olive oil) – Class 2 (non-target class): ROO: refined olive oil, OPO: olive-pomace oil, and BLE: blend of EVOO/VOO with ROO/OPO oils.

When comparing these classification outcomes with those obtained from the after-agnostizing SIMCA model, it is shown that the agnostizing step improved the results using the standard icoshift alignment, since the before-agnostizing SIMCA model placed in the 'inconclusive' quadrant 10 samples, providing worse classification results (see supplementary material, Figs. S1 and S2).

The next multivariate method was PLS-DA, which was performed with 3 latent variables (LVs) that could explain 77.99% of the total variance. The classification results can be observed in Fig. 7 in which only one OO sample belonging to class 1 was classified in class 2, since it did not trespass the threshold of 0.5, associating it to be more similar to ROO and OPO BLE samples. The classification contingency can be observed in Fig. 8. Note that the performance of both before-agnostizing and after-agnostizing PLS-DA models was the same (see supplementary material, Figs. S3 and S4).

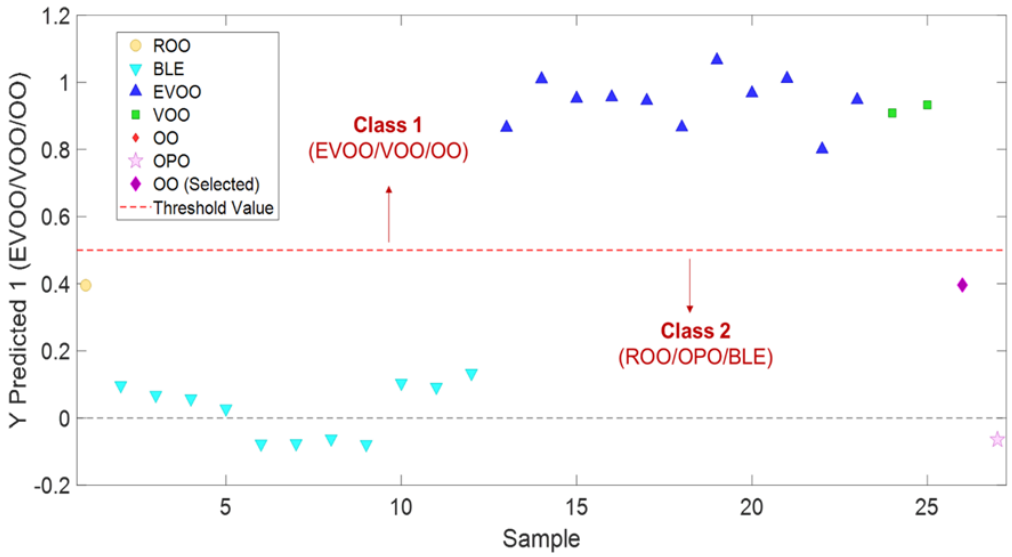


Figure 7. Classification plot of the validation set samples from the after-agnostizing PLS-DA model. EVOO: extra virgin olive oil; VOO: virgin olive oil, OO: olive oil; ROO: refined olive oils; OPO: olive-pomace oil and BLE: blend of EVOO/VOO with ROO/OPO oils. *(The solid line signifies the threshold decision of 0.5; the circled sample from class 1 is the only misclassified in class 2)*.

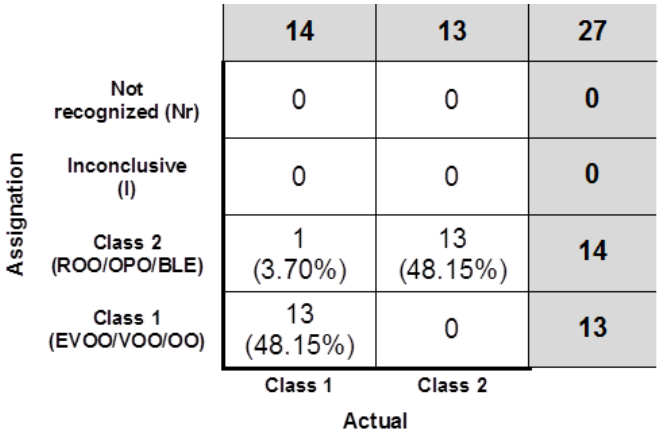


Figure 8. Validation contingencies from the after-agnostizing PLS-DA classification model. Class 1 (target class): EVOO: extra virgin olive oil, VOO: virgin olive oil, OO: olive oil) – Class 2 (non-target class): ROO: refined olive oil, OPO: olive-pomace oil, and BLE: blend of EVOO/VOO with ROO/OPO oils.

The after-agnostizing SVM classification model was performed considering the same two classes. The Kernel type algorithm radial basis function (RBF) with gamma and cost values, established by default in the PLS_Toolbox software, was applied. As observed in Fig. 9, all samples were classified within their corresponding classes. In this case, the OO sample belonging to the class 1, previously misclassified by SIMCA and PLS-DA classification models, was classified correctly. The classification 'contingency chart' can be observed in Fig. 10.

The same classification performance results were found using raw chromatograms (before-agnostizing) and instrument-agnostic fingerprints (after-agnostizing) (see supplementary material, Figs. S5 and S6). This finding suggests that SVM is a data mining/chemometric classification method suitable to be applied for the authentication of olive oil evidencing the huge difference of EVOO and VOO from ROO of any other kind. In addition, it is reaffirmed that the instrument-agnostizing methodology for the standardization of raw chromatograms yields equal or even better results than the conventional icoshift alignment but with the advantage that it results in single multivariate models without the need to repeat the data alignment step each time new samples are analysed.

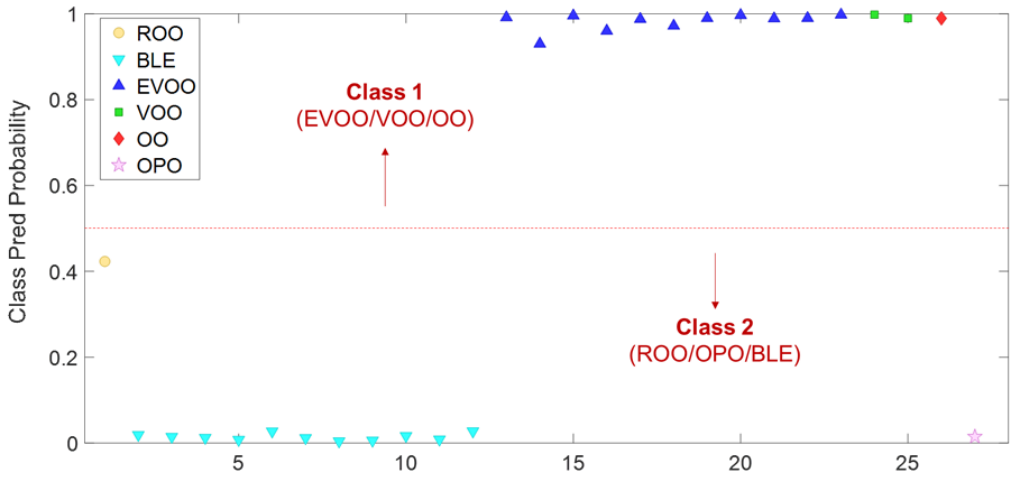


Figure 9. Classification plot of the validation set samples from the after-agnostizing SVM model. EVOO: extra virgin olive oil; VOO: virgin olive oil, OO: olive oil; ROO: refined olive oils; OPO: olive-pomace oil and BLE: blend of EVOO/VOO with ROO/OPO oils. (*The solid line signifies the threshold decision value of 0.5; all validation samples were rightly classified*).

Assignation	Class 1 (EVOO/VOO/OO)	Class 2 (ROO/OPO/BLE)	
	14	**13**	**27**
Not recognized (Nr)	0	0	**0**
Inconclusive (I)	0	0	**0**
Class 2 (ROO/OPO/BLE)	0	13 (48.15%)	**13**
Class 1 (EVOO/VOO/OO)	14 (51.85%)	0	**14**
	Class 1	**Class 2**	
	Actual		

Figure 10. Validation contingencies of the after-agnostizing SVM classification model. Class 1 (target class): EVOO: extra virgin olive oil, VOO: virgin olive oil, OO: olive oil) – Class 2 (non-target class): ROO: refined olive oil, OPO: olive-pomace oil, and BLE: blend of EVOO/VOO with ROO/OPO oils.

The contingency results were used to further calculate the classification performance metrics, presented in Table 1. The model SVM obtained the best results for sensitivity (SENS), specificity (SPEC), positive predictive value (PPV) and negative predictive value (NPV). For SIMCA and PLS-DA models, both of them obtained a SENS of 0.93, which indicates the ratio of agreement of the class 1; SIMCA model obtained a SPEC of 0.85 and

PLS-DA of 1.00 what indicates that the latter shows a better ratio of agreement of class 2. The PPV in both models was 1, indicating that the models are capable to correctly classify in all the cases the samples belonging to class 1, and that SIMCA model is better classifying the samples of class 2, since it obtained a NPV of 1.00 and PLS-DA of 0.93. In addition, Bayes' conditional probabilities 1/1 and 2/2, which report good quality of the model and/good probability of classification, are equal or close to one.

Note that, the parameters that indicate bad quality/probability of misclassification for the SVM model, like FPR, FNR, MR and Bayes' conditional probabilities 1/2 and 2/1, are equal to zero. In this regard, the SVM is capable to avoid wrong assignations with an MR value of 0, whilst PLS-DA and SIMCA classification models can perform wrong assignations with 0.04 and 0.11, respectively. Another example can be observed in PROB (1/2) which indicates that the SVM will not classify a sample from class 1 to class 2. On the contrary, PLS-DA and SIMCA models can make that misclassification with 0.07 and 0.08, respectively. Such results reveal that PLS-DA and SVM are better in classifying the samples used within this study, providing good results among data before and after agnosticism.

In this adulteration context, a similar study was performed by Tata et al. (2022) in which EVOO chromatographic polar fingerprints were analysed with PLS-DA and SVM to detect adulterations in EVOO with soft-refined olive oil. In this study, PLS-DA was used mainly as an exploratory technique, since authors declared to observe a good separation between the different kinds of vegetable oils. Afterwards, authors performed a SVM model with values reported on the training set of SENS, SPEC and EFFIC of 0.94, 0.93 and 0.95, respectively. However, such model was further validated with only six correctly classified samples. It is important to note that, the SVM model developed within the present study performs better than the model developed by Tata et al. (2022), since it provides improved quality performance metrics considering a larger number of samples. Additionally, it is worth noting that such results were obtained with instrument-agnostic fingerprints of EVOO what leads to an important interlaboratory application as well as to the expansion and implementation of multivariate methods for the control of olive oil authenticity.

Table 1. Summary of classification performance metrics of after-agnostizing SIMCA, PLS-DA and SVM models.

Classification performance metrics	SIMCA	PLS-DA	SVM
	Class 1 (EVOO/VOO/OO)		
Inconclusive rate (IR)	0.04	0.00	0.00
Sensitivity (SENS)	0.93	0.93	1.00
Specificity (SPEC)	0.85	1.00	1.00
False positive rate (FPR)	0.15	0.00	0.00
False negative rate (FNR)	0.07	0.07	0.00
Positive predictive value (precision) (PPV)	1.00	1.00	1.00
Negative predictive value (NPV)	1.00	0.93	1.00
Youden index (YOUD)	0.77	0.93	1.00
Positive likelihood rate (LR (+))	6.04	–	–
Negative likelihood rate (LR (-))	0.08	0.07	0.00
Classification odds ratio (COR)	71.50	–	–
F-measure (F)	0.96	0.96	1.00
Discriminant power (DP)	1.02	–	–
Efficiency (or accuracy) (EFFIC)	0.89	0.96	1.00
Misclassification rate (MR)	0.11	0.04	0.00
AUC (correctly classified rate)	0.89	0.96	1.00
Gini coefficient (Gini)	0.77	0.93	1.00
G-mean (GM)	0.89	0.96	1.00
Matthew’s correlation coefficient (MCC)	0.89	0.93	1.00
Chance agreement rate (CAR)	0.45	0.50	0.50
Chance error rate (CER)	0.50	0.50	0.50
Kappa coefficient (KAPPA)	0.80	0.93	1.00
PROB (1/1)	0.87	1.00	1.00
PROB (2/2)	0.92	0.93	1.00
PROB (1/2)	0.08	0.07	0.00
PROB (2/1)	0.13	0.00	0.00

The hyphen "–" signifies that the performance feature cannot be determined since it involves a division between zero.

4. Conclusions and future perspectives

The main problem with the expansion of multivariate models is that they are very dependent on the alignment of the chromatographic signals and the application of alignment algorithms such as icoshift implies repeating the process of applying this algorithm each time a new chromatographic signal is obtained. This involves that multivariate classification models must be retrained and validated with all chromatograms again. In this regard, due to the application of the instrument-agnostizing methodology this is not essential since the instrument dependence has been minimised and, once the model has been trained using instrument-agnostic fingerprints, it is not necessary to do it again. Therefore, this model could be transferred to another laboratory for its application. In fact, the authors are currently carrying out more experiments in collaboration with other laboratories in order to transfer and to implement a unique model.

In this study, the advantage of the instrument-agnostizing methodology over the application of a conventional chromatographic signal alignment procedure applying the icoshift algorithm has been demonstrated. Thus, we can conclude that the methodology for standardizing raw chromatograms has allowed to obtain instrument-agnostic fingerprints of olive oil independent from the chromatographic state or date of chromatographic analysis. This offers an important advance in knowledge as it provides the opportunity to establish the first universal database of olive oil chromatographic fingerprints, generating from these instrument-agnostic fingerprints single multivariate models that could be universally implemented in routine laboratories in order to easily authenticate olive oil.

CRediT authorship contribution statement

Christian H. Pérez-Beltrán: Investigation, Methodology, Software, Validation, Writing – original draft. **Ana M. Jiménez-Carvelo:** Investigation, Methodology, Resources, Writing – review & editing, Supervision, Project administration. **Sandra Martín-Torres:** Methodology, Software, Writing – review & editing. **Fidel Ortega-Gavilán:** Methodology, Software, Writing – review & editing. **Luis Cuadros-Rodríguez:** Resources, Writing – review & editing, Funding acquisition, Supervision, Project administration.

Declaration of interests

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

Acknowledgments

Funding for open access charge: University of Granada / CBUA.

Appendix A. Supplementary data

Table S1. Pre-processing details applied to the 88 olive oil chromatogram data.

Step	Pre-processing	Values
i	Grouping and overlapping of all chromatograms	–
ii	Selection of the range of interest of the chromatograms	1 – 7.6 min
iii	Decimation and filtering of chromatograms	1 and 0.040 a.u.
iv	Baseline correction	–
v	Peak alignment with the 'icoshift' function	1, 400, 700, 990 variables
vi	Mean center	–
vii	Additional baseline correction with the 'automatic Whittaker filter'	$\lambda = 100$ and $p = 0.001$ a.u.

Table S2. Summary of classification performance metrics of before-agnostizing SIMCA, PLS-DA and SVM models performed from the raw chromatograms.

Classification performance metrics	SIMCA	PLS-DA	SVM
	Class 1 (EVOO/VOO/OO) - Target		
Inconclusive rate (IR)	0.04	0.00	0.00
Sensitivity (SENS)	0.93	0.93	1.00
Specificity (SPEC)	0.23	1.00	1.00
False positive rate (FPR)	0.77	0.00	0.00
False negative rate (FNR)	0.07	0.07	0.00
Positive predictive value (precision) (PPV)	1.00	1.00	1.00
Negative predictive value (NPV)	1.00	0.93	1.00
Youden index (YOUD)	0.16	0.93	1.00
Positive likelihood rate (LR (+))	1.21	–	–
Negative likelihood rate (LR (-))	0.31	0.07	0.00
Classification odds ratio (COR)	3.90	–	–
F-measure (F)	0.96	0.96	1.00
Discriminant power (DP)	0.33	–	–
Efficiency (or accuracy) (EFFIC)	0.59	0.96	1.00
Misclassification rate (MR)	0.41	0.04	0.00
Correctly classified rate (AUC)	0.58	0.96	1.00
Gini coefficient (Gini)	0.16	0.93	1.00
G-mean (GM)	0.46	0.96	1.00
Matthew’s correlation coefficient (MCC)	0.46	0.93	1.00
Chance agreement rate (CAR)	0.30	0.50	0.50
Chance error rate (CER)	0.50	0.50	0.50
Kappa coefficient (KAPPA)	0.42	0.93	1.00
Bayes PROB (1/1)	0.57	1.00	1.00
Bayes PROB (2/2)	0.75	0.93	1.00
Bayes PROB (1/2)	0.25	0.07	0.00
Bayes PROB (2/1)	0.43	0.00	0.00

Figure S1. SIMCA Cooman's plot showing the classification of the validation set corresponding to the before-agnostizing model. EVOO: extra virgin olive oil; VOO: virgin olive oil, OO; olive oil, ROO: refined olive oils; OPO: pomace olive oil and BLE: blend of class 1 and class 2 olive oils. *(Left upper quadrant 'Class 1' includes EVOO, VOO, and OO samples; right upper quadrant is for not recognized samples; left bottom quadrant is for inconclusive samples; right bottom quadrant 'Class 2' includes ROO, OPO, and BLE samples; 11 samples were classified as 'inconclusive').*

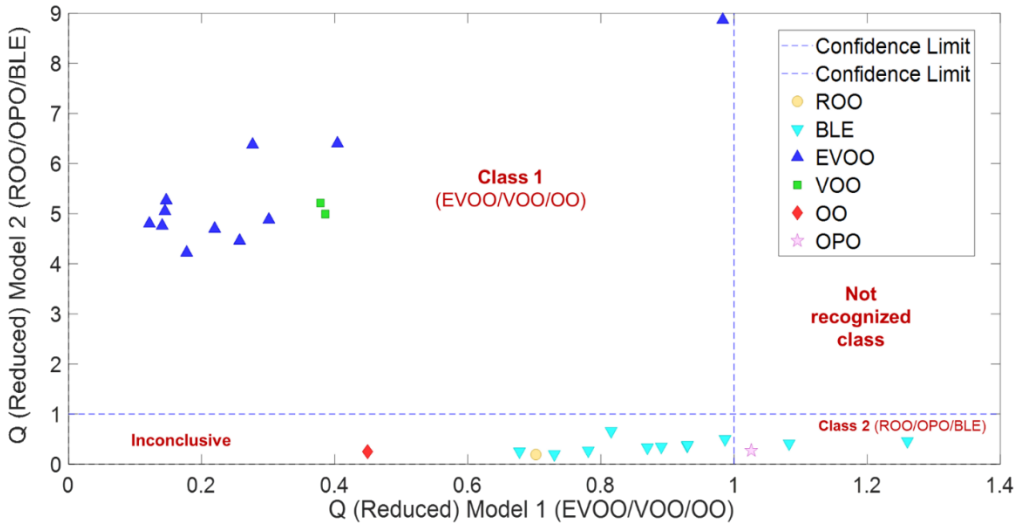


Figure S2. Validation contingency chart of the before-agnostizing SIMCA classification model (Class 1: target class (EVOO: extra virgin olive oil, VOO: virgin olive oil, OO: olive oil, ROO: refined olive oil, OPO: olive-pomace oil and BLE: blend of class 1 and class 2 olive oils).

Assignation	Class 1 (EVOO/VOO/OO)	Class 2 (ROO/OPO/BLE)	
	14	**13**	**27**
Not recognized (Nr)	0	0	**0**
Inconclusive (I)	1 (3.70%)	10 (37.00%)	**11**
Class 2 (ROO/OPO/BLE)	0	3 (11.15%)	**3**
Class 1 (EVOO/VOO/OO)	13 (48.15%)	0	**13**
	Class 1	**Class 2**	
	Actual		

Class 1: target class (EVOO: Extra Virgin olive oil, VOO: Virgin olive oil, OO: Olive oil); Class 2: non-target class (ROO: Refined olive oil, OPO: olive-pomace oil, BLE: blend)

Figure S3. PLS-DA plot showing the classification of the validation set corresponding to the before-agnostizing model. EVOO: extra virgin olive oil; VOO: virgin olive oil, OO: olive oil; ROO: refined olive oils; OPO: olive-pomace oil and BLE: blend of class 1 and class 2 olive oils (*The solid line represents the threshold decision value of 0.5. The circled sample from class 1 is the only misclassified in class 2*).

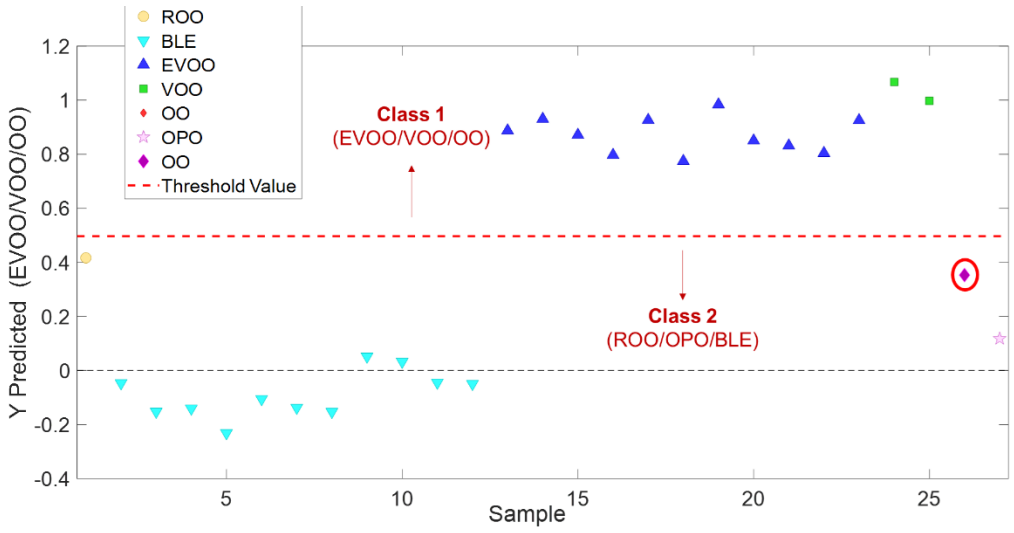


Figure S4. Validation contingency chart of the before-agnostizing PLS-DA classification model (Class 1: target class (EVOO: extra virgin olive oil, VOO: virgin olive oil, OO: olive oil, ROO: refined olive oil, OPO: olive-pomace oil and BLE: blend of class 1 and class 2 olive oils).

Assignation	Class 1 (EVOO/VOO/OO)	Class 2 (ROO/OPO/BLE)	
	14	**13**	**27**
Not recognized (Nr)	0	0	**0**
Inconclusive (I)	0	0	**0**
Class 2 (ROO/OPO/BLE)	1 (3.70%)	13 (48.15%)	**14**
Class 1 (EVOO/VOO/OO)	13 (48.15%)	0	**13**
	Class 1	**Class 2**	
	Actual		

Class 1: target class (EVOO: Extra Virgin olive oil, VOO: Virgin olive oil, OO: Olive oil); Class 2: non-target class (ROO: Refined olive oil, OPO: olive-pomace oil, BLE: blend)

Figure S5. SVM classification plot of the validation set corresponding to the before-agnostizing model. EVOO: extra virgin olive oil; VOO: virgin olive oil, OO: olive oil; ROO: refined olive oils; OPO: olive-pomace oil and BLE: blend of class 1 and class 2 olive oils (*The solid line represents the threshold decision value of 0.5; all validation samples were rightly classified*).

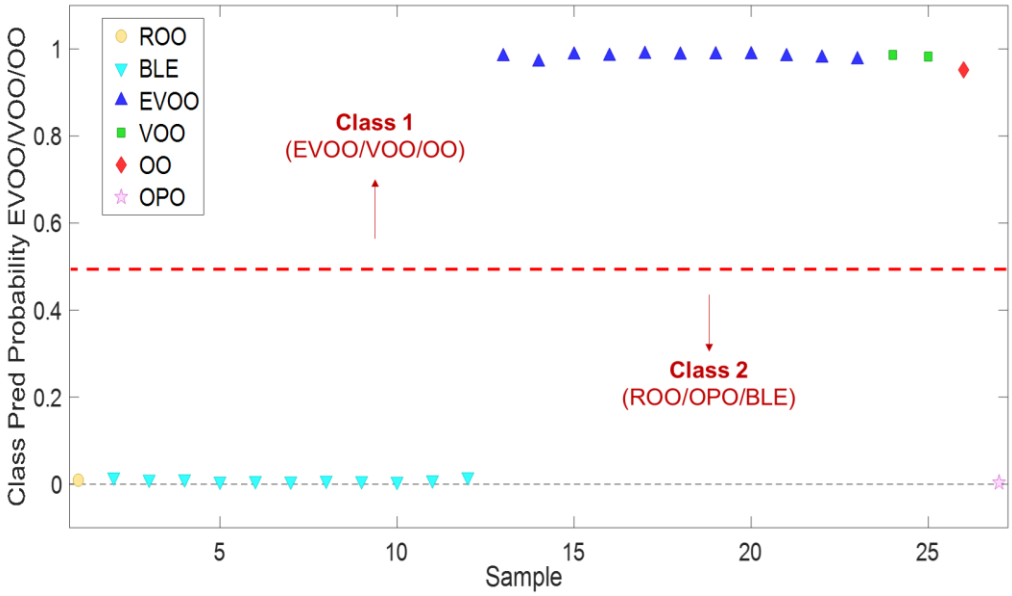


Figure S6. Validation contingency chart of the before-agnostizing SVM classification model (Class 1: target class (EVOO: extra virgin olive oil, VOO: virgin olive oil, OO: olive oil, ROO: refined olive oil, OPO: olive-pomace oil and BLE: blend of class 1 and class 2 olive oils).

		14	13	27
Assignation	Not recognized (Nr)	0	0	**0**
	Inconclusive (I)	0	0	**0**
	Class 2 (ROO/OPO/BLE)	0	13 (48.15%)	**13**
	Class 1 (EVOO/VOO/OO)	14 (51.85%)	0	**14**
		Class 1	**Class 2**	
		Actual		

Class 1: target class (EVOO: Extra Virgin olive oil, VOO: Virgin olive oil, OO: Olive oil); Class 2: non-target class (ROO: Refined olive oil, OPO: olive-pomace oil, BLE: blend)

References

Abdallah, M., Vegara-Barberán, M., Lerma-García, M. J., Herrero-Martínez, J. M., Simó-Alfonso, & E., & Guerfel, M. (2016). *European Journal of Lipid Science and Technology, 118*, 1236–1242. https://doi.org/10.1002/ejlt.201500041

Boccard, J., & Rudaz, S. (2020). Analysis of metabolomics data – a chemometric perspective. In S. Brown, T. Roma, & B. Walczak (Eds.), *Comprehensive chemometrics – chemical and biochemical data analysis* (pp. 483–505). Amsterdam: Elsevier.

Carranco, N., Farrés-Cebrián, M., Saurina, J., & Núñez, O. (2018). Authentication and quantification of fraud in extra virgin olive oils based on HPLC-UV fingerprinting and multivariate calibration. *Foods, 7*, 44. https://doi.org/10.3390/foods7040044

Commission Regulation (EEC) No 2568/91. Characteristics of olive oil and olive-residue oil and the relevant methods of analysis. Official Journal of European Union, 031.005, 1-128. https://www.fao.org/faolex/results/details/en/c/LEX-FA OC040621/

Creydt, M., & Fischer, M. (2020). Food authentication in real life: How to link nontargeted approaches with routine analytics? *Electrophoresis, 41*, 1665-1679. https://doi.org/10.1002/elps.202000030

Cuadros Rodríguez, L., Martín Torres, S., Ortega Gavilán, F., Jiménez Carvelo, A. M., López Ruíz, R., Garrido Frenich, A., Bagur González, M. G., & González Casado, A. (2021b). Standardization of chromatographic signals – Part II: Expanding instrument-agnostic fingerprints to reverse phase liquid chromatography. *Journal of Chromatography A, 1641*, 461973. https://doi.org/10.1016/j.chroma.2021.461973

Cuadros Rodríguez, L., Ortega Gavilán, F., Martín Torres, S., Medina Rodríguez, S., Jiménez Carvelo, A. M., González Casado, A., & Bagur González, M. G. (2021a). Standardization of chromatographic signals – Part I: Towards obtaining instrument-agnostic fingerprints in gas chromatography. *Journal of Chromatography A, 1641*, 461983. https://doi.org/10.1016/j.chroma.2021.461983

Cuadros Rodríguez, L., Pérez Castaño, E., & Ruiz Samblás, C. (2016b). Quality performance metrics in multivariate classification methods for qualitative analysis. *Trends in Analytical Chemistry, 80*, 612–624. https://doi.org/10.1016/j. trac.2016.04.021

Cuadros Rodríguez, L., Ruiz Samblás, C., Valverde Som, L., Pérez Castaño, E., & González Casado, A. (2016a). Chromatographic fingerprinting: An innovative approach for food 'identitation' and food authentication – a tutorial. *Analytica Chimica Acta, 909*, 9–23. https://doi.org/10.1016/j.aca.2015.12.042

Drira, M., Kelebek, H., Guclu, G., Jabeur, H., Selli, S., & Bouaziz, M. (2020). Targeted analysis for detection the adulteration in extra virgin olive oil's using LC-DAD/ESI-MS/MS and combined with chemometrics tools. *European Food Research and Technology, 246*, 1661–1677. https://doi.org/10.1007/s00217-020-03522-y

Duraipandian, S., Petersen, J. C., & Lassen, M. (2019). Authenticity and concentration analysis of extra virgin olive oil using spontaneous Raman spectroscopy and multivariate data analysis. *Applied Sciences, 9*, 2433. https://doi.org/10.3390/app9122433

Folch Fortuny, A., Vitale, R., de Noord, O. E., & Ferrer, A. (2017). Calibration transfer between NIR spectrometers: New proposals and a comparative study. *Journal of Chemometrics, 31*, Article e2874. https://doi.org/10.1002/cem.2874

Fornasaro, S., et al. (2020). Surface enhanced Raman spectroscopy for quantitative analysis: Results of a large-scale European multi-instrument interlaboratory study. *Analytical Chemistry, 92*, 4053–4064. https://doi.org/10.1021/acs. analchem.9b05658

Gou, S., Heinke, R., Stöckel, S., Rösch, P., Jürgen, P., & Bocklitz, T. (2018). Model transfer for Raman-spectroscopy - based bacterial classification. *Journal of Raman Spectroscopy, 49*, 627–637. https://doi.org/10.1002/jrs.5343

Jabeur, H., Drira, M., Rebai, A., & Bouaziz, M. (2017). Putative markers of adulteration of higher-grade olive oil with expensive pomace olive oil identified by gas chromatography combined with chemometrics. *Journal of Agricultural and Food Chemistry, 65*, 5375–5383. https://doi.org/10.1021/acs.jafc.7b00687

Jiménez Carvelo, A. M., & Cuadros Rodríguez, L. (2021). Data mining/machine learning methods in foodomics. *Current Opinion in Food Science, 37*, 76–82. https://doi.org/10.1016/j.cofs.2020.09.008

Jiménez Carvelo, A. M., Martín Torres, S., Cuadros Rodríguez, L., & González Casado, A. (2020). Nontargeted fingerprinting approaches. In C. M. Galanakis (Ed.), *Food traceability and authentication* (pp. 163–193). Oxford: Academic Press/Elsevier. https://doi.org/10.1016/B978-0-12-821104-5.00010-6.

Karunathilaka, S. R., Fardin Kia, A. R., 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

Li, Y., Wen, S., Sun, Y., Zhang, N., Gao, Y., & Yu, X. (2021). New method based on polarity reversal for detecting adulteration of extra virgin olive oil with refined olive pomace oil. *European Journal of Lipid Science and Technology*. https://doi.org/10.1002/ejlt.202100193

Meenu, M., Cai, Q., & Xu, B. (2019). A critical review on analytical techniques to detect adulteration of extra virgin olive oil. *Trends in Food Science & Technology, 91*, 391–408. https://doi.org/10.1016/j.tifs.2019.07.045

Mingchih, F., Chia-Fen, T., Guan-Yan, W., Su-Hsiang, T., Hwei-Fang, C., Ching-Hao, K., Che-Lun, H., Ya Min, K., Yang Chih, D., & Yu Mei, C. (2015). Identification and quantification of Cu-clorophyll adulteration of edible oils. *Food Additives and Contaminants: Part B, 8*, 157–162. https://doi.org/10.1080/19393210.2015.1025861

Muñoz Olivas, R. (2004). Screening analysis: An overview of methods applied to environmental, clinical and food analyses. *Trends in Analytical Chemistry, 23*, 203–216. https://doi.org/10.1016/S0165-9936(04)00318-8

Navratilova, K., Hurkova, K., Hrbek, V., Uttl, L., Tomaniova, M., Valli, E., & Hajslova, J. (2022). Metabolic fingerprinting strategy: Investigation of markers for the detection of

extra virgin olive oil adulteration with soft-deodorized olive oils. *Food Control, 134*, 108649. https://doi.org/10.1016/j.foodcont.2021.108649

Oliveri, P., Malegori, C., Mustorgi, E., & Casale, M. (2020). Application of chemometrics in the food sciences. In S. Brown, T. Roma, & B. Walczak (Eds.), *Comprehensive chemometrics – chemical and biochemical data analysis* (pp. 99–111). Amsterdam: Elsevier.

Rigano, F., Arigò, A., Oteri, M., La-Tella, R., Dugo, P., & Mondello, L. (2021). The retention index approach in liquid chromatography: An historical review and recent advances. *Journal of Chromatography A, 1640*, Article 461963. https://doi.org/10.1016/j.chroma.2021.461963

Tahir, H. E., Arslan, M., Mahunu, G. K., Mariod, A. A., Xiaobo, S. B. H. Z., Jiyong, S., El-Seedi, G. R., & Musa, T. H. (2022). The use of analytical techniques coupled with chemometrics for tracing the geographical origin of oils: A systematic review (2013–2020). *Food Chemistry, 366*, 130633. https://doi.org/10.1016/j.foodchem.2021.130633

Tata, A., Massaro, A., Damiani, T., Piro, R., Dall'Asta, C., & Suman, M. (2022). Detection of soft-refined oils in extra virgin olive oil using data fusion approaches for LC-MS, GC-IMS and FGC-Enose techniques: The winning synergy of GC-IMS and FGC-Enose. *Food Control, 133*, 108645. https://doi.org/10.1016/j.foodcont.2021.108645

Tomasi, G., Savorani, F., & Engelsen, S. B. (2011). Icoshift: An effective tool for the alignment of chromatographic data. *Journal of Chromatography A, 1248*, 7832–7840. https://doi.org/10.1016/j.chroma.2011.08.086

Tomasi, G., Van der Berg, F., & Andersson, C. (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics, 18*, 231–241. https://doi.org/10.1002/cem.859

Wang, Y., Veltkamp, D. J., & Kowalski, B. R. (1991). Multivariate instrument standardization. *Analytical Chemistry, 63*, 2750–2756. https://doi.org/10.1021/ac00023a016

Wise, B. M., Gallagher, N. B., Bro, R., Shaver, J. M., Winding, W., & Koch, R. S. (2006). *Chemometrics tutorial for PLS_toolbox and solo*. Wenatchee, WA, USA: Eigenvector Research, Inc.

Zhang, J., Guo, C., Cui, X., Cai, W., & Shao, X. (2019). A two-level strategy for standardization of near infrared spectra by multi-level simultaneous component analysis. *Analytica Chimica Acta, 1050*, 25–31. https://doi.org/10.1016/j. aca.2018.11.013

Zhang, J., Sun, H., & Lu, W. (2021). Recent advances in analytical detection of olive oil adulteration. *Food Science and Technology*, 1–10. https://doi.org/10.1021/acsfoodscitech.1c00254

3.3. Artículo científico III

Paper enviado a la revista científica Analytica Chimica Acta

Índice de impacto : (WoS) 6.911 (2021) **ISSN:** 1873-4324

Taking-up instrument-agnostic liquid chromatographic fingerprints: towards the creation of a global harmonised database and a single analytical multivariate method – Tequila authentication as a case study

Luis CUADROS-RODRÍGUEZ [a], Christian Hazael PÉREZ-BELTRÁN [a], José J. OLMOS-ESPEJEL [b], Guadalupe PÉREZ–CABALLERO [b], Ana M. JIMÉNEZ-CARVELO✉ [a]

[a] Department of Analytical Chemistry, Faculty of Sciences, University of Granada, C/ Fuentenueva, s/n, E-18071 Granada (Spain).

[b] Unidad de Investigación Multidisciplinaria, Facultad de Estudios Superiores Cuautitlán, campo 4, Universidad Nacional Autónoma de México, Cuautitlán Izcalli, (México).

Keywords

Reverse phase liquid chromatography; similarity analysis; global fingerprints database; tequila; single multivariate model; chemometrics

✉ Corresponding author (E-mail: amariajc@ugr.es; Phone: +34 958 24 07 97)

1. Introduction

The growing and complex food and beverages fraud problems require simple methodologies that can rigorously predict and detect potential vulnerabilities in products, cases and sectors. Over the years, the food and beverage industries have developed quality and safety management systems, which have made it possible to prevent food and beverages counterfeiting and/or poisoning. Safety systems typically use hazard analysis and critical control points (HACCP) schemes, which are recognised worldwide [**1**,**2**]. However, it should be noted that HACCP has demonstrated to be effective against accidental contamination, but has not been used to routinely detect or mitigate intentional fraudulent/criminal actions in the whole supply chain. In this context, it is necessary to perform a comprehensive fraud risk assessment. For this purpose, the vulnerability of occurrence (i.e., the probability of fraud occurring) and the severity of the fraud occurrence are the two main factors to be taken into account [**3**,**4**]. In this sense, different approaches to food fraud detection during different stages of the supply chain have been reported in the literature [**5**,**6**]. Nevertheless, the authenticity assurance of foodstuff when they enter the last step of the supply chain, i.e., when the food products are marketed at different levels (local, national or worldwide), remains to be solved.

In general terms, two main approaches could be considered for ensuring the authenticity of a food product, one of them is based on scientific analysis and the other one based on traceability verification [**7**,**8**,**9**]. Focusing on the first approach, the conventional analytical path of food and beverages authentication has focused on the qualification and/or quantification of specific components, which generates multi-parametric and/or compositional or profiling data of specific fractions obtained from different chemical, biomolecular or isotopic analyses, involving the performance of multiple measurements using a battery of analytical techniques and instrumental platforms, i.e., to verify that a product has not been adulterated or that it complies with current legislation.

Note that, since even the simplest food and beverage is also a complex multi-compositional matrix, the way to ensure its quality/authenticity should be carried out by applying a multivariate approach to evaluate the food as a whole. In the field of the food analytical chemistry, the application of the multivariate approach involves changing the conventional

schemes of the analytical methods one by one, focused on the identification and quantification of chemical compounds (targeted approach), and replacing them with new analytical methods based on the application of broad-based chemical information (untargeted approach) from which a sample of the product is analysed obtaining a non-specific but characteristic analytical signal of each one. In this context, the potential of the application of the instrumental fingerprinting methodology has been demonstrated, and proof of this are the studies found in literature in this regard [**10**,**11**]. Briefly, instrumental fingerprints are defined as non-specific analytical signals, which contain all the information of interest of the analysed product, allowing to unequivocally authenticating it.

Surprisingly, the application of instrumental fingerprinting-based analytical methods is still limited. In fact, to date there is no official analytical method based on the multivariate approach to assess the authenticity of food or beverage products as such [**12**]. But, in this line, a first approximation has been performed by the AOAC International, which has proposed the creation of a working group in order to promote the development of non-targeted analytical methods for the control of foodstuff, such as olive oil, milk (liquid and powder) and honey [**13**]. But even then, several challenges remain to be addressed when implementing multivariate analytical methods based on instrumental fingerprints in routine analysis, especially, when chromatographic techniques, such as gas and liquid chromatography, are employed. The central challenge for their implementation lies in minimizing or removing the dependence on the analytical instrument used to obtain the chromatographic signal, because it is well known that some non-negligible variations in retention times or even peak intensities may occur when repeated chromatographic analyses are performed.

In this regard, the solution to this issue would be to obtain an instrument-independent chromatographic signal which would lead to the creation of a global chromatographic fingerprints database. From this, it would be possible to generate a single multivariate model that would combine any chromatographic instrumental fingerprint obtained by applying the same analytical method in similar chromatographic equipment but in different laboratories or in the same equipment but on a different date. In order to achieve this objective, Cuadros et al. have recently proposed an innovative methodology to be followed

in order to obtain standardised instrumental fingerprints when the gas and liquid chromatography are employed, which has been named 'instrument-agnostizing'. Both intensity values and retention times of the chromatographic signal are standardised using first a reference intensity signal from a suitable internal standard (IS) and then an external standard mixture (ESM), respectively. Essentially, this last step is based on first establishing a set of system-independent constant values, named as 'Standard Retention Scores' (SRS), from the analysis of the ESM. A comprehensive description about the process of 'instrument-agnostizing' of chromatographic signals as well as, for instance, the comparison with traditional chromatographic signal alignment algorithms, such as 'icoshift' [**14**], can be read in the following papers published by the authors [**15**,**16**,**17**].

However, there are still gaps to be explored about the potential of the novel 'chromatographic-agnostizing' methodology, for example, its application for the creation of the first global chromatographic fingerprint database leading to the harmonised use of a single multivariate model. In this line, the innovation of this study is based on answering three analytical issues: (i) *is it possible to obtain a unique instrumental fingerprint for the same sample analysed in two different laboratories*? (ii) *Is it possible to achieve a single multivariate model built from instrument-agnostic chromatographic fingerprints?* And (iii) *does this method lead to successful results in the food and beverage authenticity field?*

The achievement of challenges involved in these matters is fully linked to two reports recently published by the European Union (EU) in which the need for a standardised database to ensure the quality of foodstuffs was remarked [**18**,**19**]. For this purpose, the assessment of the authenticity of tequila has been considered as a case study. Nowadays, the food fraud in beverages is increasing, in fact, the European Union Agency for Law Enforcement Cooperation (EUROPOL) has indicated in a report published in March 2022, that *the production of illicit food products, especially drinks, is increasingly professional and sophisticate*d [**20**]. It is, therefore, necessary to research and develop new analytical methods capable of detecting new frauds.

In this concern, a unique and characteristic instrument-agnostic chromatographic fingerprint per tequila sample was successfully obtained from two laboratories, one located in Spain and the other one in Mexico. First of all, a comprehensive similarity analysis

between the chromatographic signals acquired in the different laboratories was carried out. After that, a multivariate model for authenticating white tequila was built, being liquid chromatography coupled to a diode-array detector (HPLC-DAD) the analytical technique and partial least squares-discriminant analysis (PLS-DA) and support vector machine (SVM) the chemometric tools employed for the establishment of the single multivariate method able to authenticate the tequila.

2. Considerations prior to the signal agnostization stage

One important aspect before applying the 'instrument-agnostizing' methodology is that the chromatographic system should perform correctly and with consistency, obtaining similar shape and duration times for the analysed ESM and chromatographic fingerprints within and throughout the different sequences of analyses. When the results of the quality control analyses are favourable within the running sequence and the chromatographic system is in good condition, the 'instrument-agnostizing' methodology is carried out. However, the chromatographic system may present problems and provide different results among the ESM and the chromatographic fingerprints, thus, the chromatographic signal must be assessed to verify if it is appropriate to be agnostized or not.

On the one hand, to corroborate the correct functioning of the chromatographic system in terms of the ESM and fingerprint length, the ‘Runtime Ratio’ (RtR) has been developed 'ad hoc' to evaluate whether the instrumental fingerprint has shortened or expanded in reference to the previous one. Thus, the RtR should be calculated for each chromatographic fingerprint after each sequence of analysis, according to the following Equation:

$$RtR = \frac{t_{SF(last)} - t_{SF(first)}}{t_{ESM(last)} - t_{ESM(first)}} \quad (1)$$

where $t_{SF(last)}$ and $t_{SF(first)}$ are the last and first running time points selected from the sample fingerprint (SF), respectively, and $t_{ESM(last)}$ and $t_{ESM(first)}$ are the last and first running time points selected from the in-day ESM chromatogram, respectively.

On the other hand, it is well known that the retention times of the same signal in liquid chromatography may present shifts, either get advanced or get delayed, between analyses. To appraise these variations, the 'Starting-time Lag' (StL) has been introduced and should be also calculated for each chromatographic signal after each sequence of analysis as well. The StL is calculated as follows:

$$StL = t_{SF(first)} - RtR \cdot t_{ESM(first)} \quad (2)$$

where $t_{SF(first)}$ and $t_{ESM(first)}$ are the first running time points selected from the SF and in-day ESM chromatogram, respectively, and RtR is the previous calculated value for this parameter. Once the evaluation of both parameters, RtR and StL, for each chromatographic signal is completed and their results are kept the same through the different days of analyses, the 'instrument-agnostizing' methodology is then performed. Nonetheless, if the RtR and StL results are very diverse among chromatographic signals, it means that they could get shortened, expanded, advanced or delayed; therefore, the agnostization step cannot be performed. In this scenario, such effects can be minimised using the 'Equity Function', developed 'ad hoc' for this study.

The purpose of the Equity Function is to diminish and correct the possible variations of the chromatographic signals that might have occurred during the sequence analysis in order to resemble them to the acceptable ones, which have been obtained with the instrumental equipment in optimal conditions. For this, the Equity Function is described as:

$$\mathbf{X}_{SFcorrec} = f_{Eq}(\mathbf{X}_{SF}) \quad (3)$$

which can be rewritten as:

$$\mathbf{X}_{SFcorrec} = m \times (\mathbf{A} + \mathbf{X}_{SF}) \quad (4)$$

where m and $\mathbf{A}$ are multiplicative and additive corrective terms, respectively; $\mathbf{X}_{SF}$ is a one column type matrix (n × 1) from the defective sample fingerprint (SF) and $\mathbf{X}_{SFcorrec}$ is the one column type matrix (n × 1) that has been corrected. To obtain the value of the multiplicative corrective term, m, use Eq. (5):

$$m = \frac{RtR^{mg}}{RtR^{wf}} \quad (5)$$

where RtR^{mg} is the Runtime Ratio of the manager (Mg) system or the in-day reference signal, which is calculated as displayed in Eq. (1); and RtR^{wf} is the Runtime Ratio of the workforce (Wf) system or the in-day defective fingerprint. Please, note that the Mg and Wf terms will be used when instrumental fingerprints of different laboratories are compared, where the Mg term is for the reference laboratory and the Wf term is for the secondary laboratory.

To obtain the value of the additive corrective term, $\mathbf{A}$, use Eq. (6):

$$a = StL^{mg} - StL^{wf} \quad (6)$$

where StL^{mg} is the starting-time lag of the Mg system or the in-day reference fingerprint, which is calculated as displayed in Eq. (2); and StL^{wf} is the starting-time lag of the Wf system or the in-day defective fingerprint. Once the corrections are completed, the similarity analyses should be performed to corroborate the similitude of the corrected sample fingerprint with the reference fingerprint, as explained in subsection 3.5. Similarity analysis. Finally, if the outcome is successful, the 'instrument-agnostizing' methodology can be performed. The general scheme to be followed in order to implement the instrument-agnostizing methodology can be summarised in Figure 1.

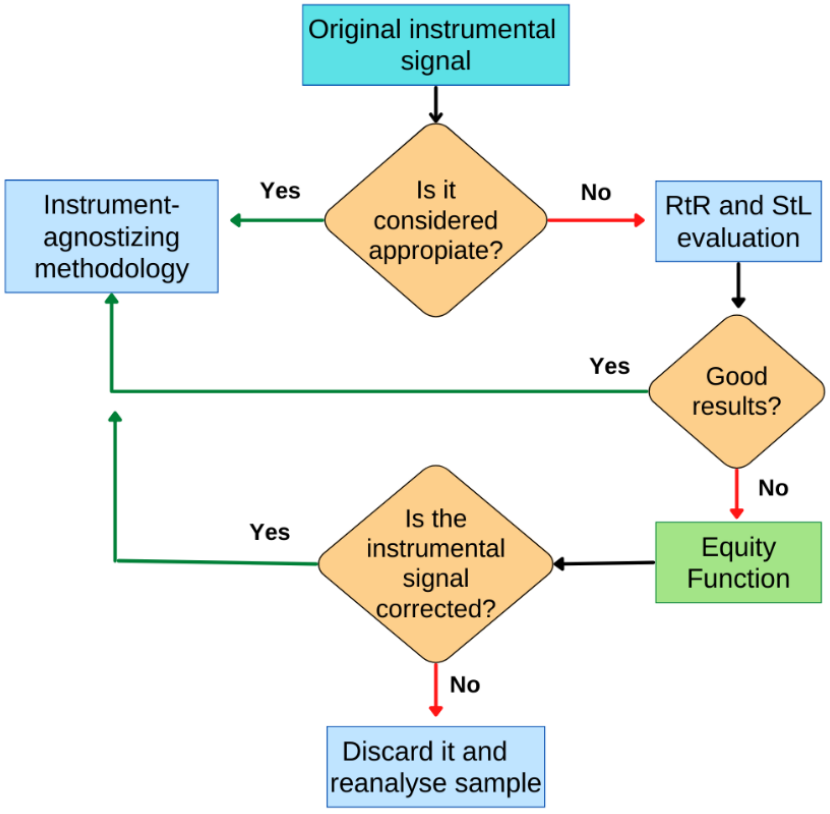


Figure 1. General scheme for the instrument-agnostizing methodology application. RtR: Runtime ratio, StL: Starting-time lag.

3. Materials and methods

3.1. Tequila samples

A total of 41 samples of White Tequila: 23 categorised as '*100 % agave*' and 18 as '*mixed*' were analysed and employed to perform the development (training and validation steps) of the different multivariate models. Additionally, 26 samples of White Tequila were analysed for determining their category through the developed multivariate models. All White Tequila samples were provided by the Regulatory Council of Tequila (CRT) in México.

3.2. Chemicals and reagents

Methanol (MeOH) with HPLC-grade was employed in the study. It was purchased from Panreac Química (Barcelona, Spain) and J.T. Baker (México City, México). The chemical standards such as 5-(hydroxymethyl)furfural (5HMF), furfural (FUR), 2-acetylfuran (2AF), 2-acetyl-5-methylfuran (2A5MF), shown in Table 1, were acquired from Dr. Ehrenstorfer - LGC Standards (Augsburg, Germany) and Sigma-Aldrich (Missouri, USA) to constitute the ESM.

Table 1. List of chemical compounds selected to constitute the external standard mixture (ESM), ordered by their experimental retention time.

Chemical compound	Molecular formula	Elution order
5-(Hydroxymethyl)furfural (5HMF)	$C_6H_6O_3$	(1)
Furfural (FUR)	$C_5H_4O_2$	(2)
2-Acetylfuran (2AF)	$C_6H_6O_2$	(3)
2-Acetyl-5-methylfuran (2A5MF)	$C_7H_8O_2$	(4)

3.3. (RP)LC method for analysis of Tequila and external standard mixture

All chromatographic analyses of White Tequila were performed under reproducibility conditions in Spain and Mexico. For this, 430 µL of tequila were mixed with 635 µL of deionised water, and 135 µL of 5HMF was added as internal standard (IS). Finally, the mixture was sealed and vortexed for 10 s.

The samples were analysed using reverse-phase liquid chromatography in two different HPLC-DAD pieces of equipment as described in Table 2. During the experimental work the columns were kept at 40°C and the chromatograms of the tequila samples and external standard mixture (EMS) were acquired at 220 nm and 280 nm, respectively.

Table 2. HPLC equipment used to analyse the different tequila samples.

Equipment	Specifications
HPLC1 *(reference system)*	Agilent 1260 series liquid chromatograph (Agilent Technologies, Santa Clara, CA, USA) equipped with a CH30 column thermostat (Eppendorf, Hamburg, Germany), a G1311A quaternary pump, a G1322A degasser and G1329A autosampler. Detection was performed with a G7115A Infinity II diode-array detector (DAD). Agilent ChemStation OpenLab CDS software (rev. C.01.09) for LC systems was used to export data to CSV format.
HPLC2	Agilent 1100 series liquid chromatograph (Agilent Technologies, Santa Clara, CA, USA) equipped with a G1316A Thermostatted Column Compartment (Eppendorf, Hamburg, Germany), a G1311A quaternary pump, a G1322A degasser and G1313A autosampler. Detection was performed with a G1315B Infinity II diode-array detector (DAD). Agilent ChemStation for LC 3D software (rev. A.10.02) was used to export data to CSV format.
Column1	Eclipse XDB-C18 (250 × 4.6 mm i.d, 5 µm)
Column2	Eclipse XDB-C18 (250 × 4.6 mm i.d, 5 µm)

The gradient mode of the mobile phase was performed in the same mode in both equipment, with a constant flow rate of 1.0 mL min^{-1}, as follows: 20 µL of each sample were injected at time 0 and were eluted with methanol-deionised water (MeOH-DW) 10/90 (v/v) for 3 min. Then, solvent was changed to MeOH-DW 70/30 (v/v) for 3 more min. Finally, the solvent was changed back to the initial conditions of MeOH-DW 10/90 (v/v) for 2 min. After each analysis, a cleaning step was performed with MeOH-DW 90/10 (v/v) for 5 min with the same flow rate to ensure the integrity of the column. Subsequently, the solvent was changed to initial conditions and kept for 2 min before the next analysis.

Additionally, an external standard mixture, consisting of 5HMF, FUR, 2AF and 2A5MF, was first analysed in the HPLC1 (considered as reference system, located in the laboratory in Spain) 27 times to obtain the standard retention scores (SRS). Moreover, it was analysed before and after each chromatographic run, in order to extrapolate the SRS values to the chromatograms of White Tequila and as a quality control step to monitor the proper behaviour of the equipment. The injected volume of ESM was 20 µL at time 0 and was

eluted with MeOH-DW 20/80 (v/v) for 2 min with a constant flow rate of 0.8 mL min^{-1} for the entire operation. Then, the solvent was changed to MeOH-DW 60/40 (v/v) for 1 min. Next, it was shifted to MeOH-DW 45/55 (v/v) for 5 min. Finally, the initial conditions of MeOH-DW 20/80 (v/v) were restored for 1 min. The cleaning step was also performed after each analysis of ESM.

3.4. Agnostizing of instrumental chromatographic signals

As mentioned in the introduction section, to reach a successful harmonised database of instrumental chromatographic fingerprints, which could be used to build a single multivariate model, it is necessary to standardise the signals (chromatograms) obtained by similar chromatographic pieces of equipment, with the aim of removing instrument variability effects as well as those due to time (date) of sample analysis. In this concern, to achieve standardised chromatographic signals from the two HPLC equipment, instrument-agnostizing methodology proposed by Cuadros et al. [**16**] was applied on raw chromatograms. Chromatographic raw data files, embedded in a data vector composed of 1650 intensity elements, were exported in 'comma separated value' (CSV) format, and then converted to MATLAB format (version R2017b, 9.3.0.713579, The Mathworks Inc. MA, USA).

Firstly, SRS values were established from the chromatographic signals obtained after the analysis of the EMS in the HPLC1 (see table 2). Subsequently, intensity normalization of the chromatograms using the peak of the internal standard 5HMF was carried out. Then, replacement of the retention time values by the calculated SRS in each sample chromatographic signal was performed in order to unify the scale on the x-axis. The last step was based on a re-sampling process to fix into a same number of variables all chromatograms. Thus, instrument-agnostic chromatographic fingerprints for each sample were achieved in data vectors, which were arranged into 629 variables.

3.5. Similarity analyses

In order to evaluate the potential of transfer of the agnostization methodology for obtaining instrument-agnostic chromatographic signals acquired on different laboratories and, thus, testing its applicability for the creation of a harmonised database of instrumental

fingerprints leading to a single multivariate model, a comprehensive similarity analysis was performed. This was based on comparing the chromatographic signals obtained on the Spain and Mexico laboratories before and after applying the instrument-agnostizing methodology. To this end, 10 tequila samples were randomly selected from the total samples, and their characteristic signals obtained in Mexico and Spain before and after applying the agnostizing process were compared.

The nearness similarity index (NEAR) [**21**] which is based on the closeness of two vectors in the space and is calculated from the normalised Euclidean distance, as depicted in Eq. (7) was applied:

$$NEAR(X_M X_S)=1 - \left[\sqrt{\frac{(X_M - X_S) \times (X_M - X_S)^T}{(X_M + X_S) \times (X_M + X_S)^T}} \right] \quad (7)$$

where X_M and X_S symbolised the data vectors which collect the Mexico- and Spain-acquired chromatographic signals, respectively, and the superscript T denotes the transposed matrix. Note that NEAR is calculated for each pair of chromatographic signals, thus, the term in square brackets represents the (0-1) normalised Euclidean distance between the two vectors.

Additionally, the cosine angle (COS) among the chromatographic signals (vectors) before and after the agnostizing process was also calculated to complement the similarity evaluation, as it has been done in previous studies [**21**,**22**,**23**]. The COS from X_M and X_S vectors is calculated as represented in Eq. (8):

$$COS(X_M X_S) = \frac{\sum_j X_{Mj} \times X_{Sj}}{\sqrt{\sum_j X_{Mj}^2 \times \sum_j X_{Sj}^2}} \quad (8)$$

where X_M and X_S symbolised the data vectors which collect the Mexico- and Spain-acquired chromatographic signals, respectively, and X_{Mj} and X_{Sj} symbolised each element of the considered (X_M) and reference (X_S)chromatographic signals, respectively.

Since only positive figures take place in the evaluated chromatographic signals, the COS results might be in the range from 0 (totally different vectors) to 1 (matching vectors).

A layout of experimental design of the similarity analysis is displayed in Figure 2.

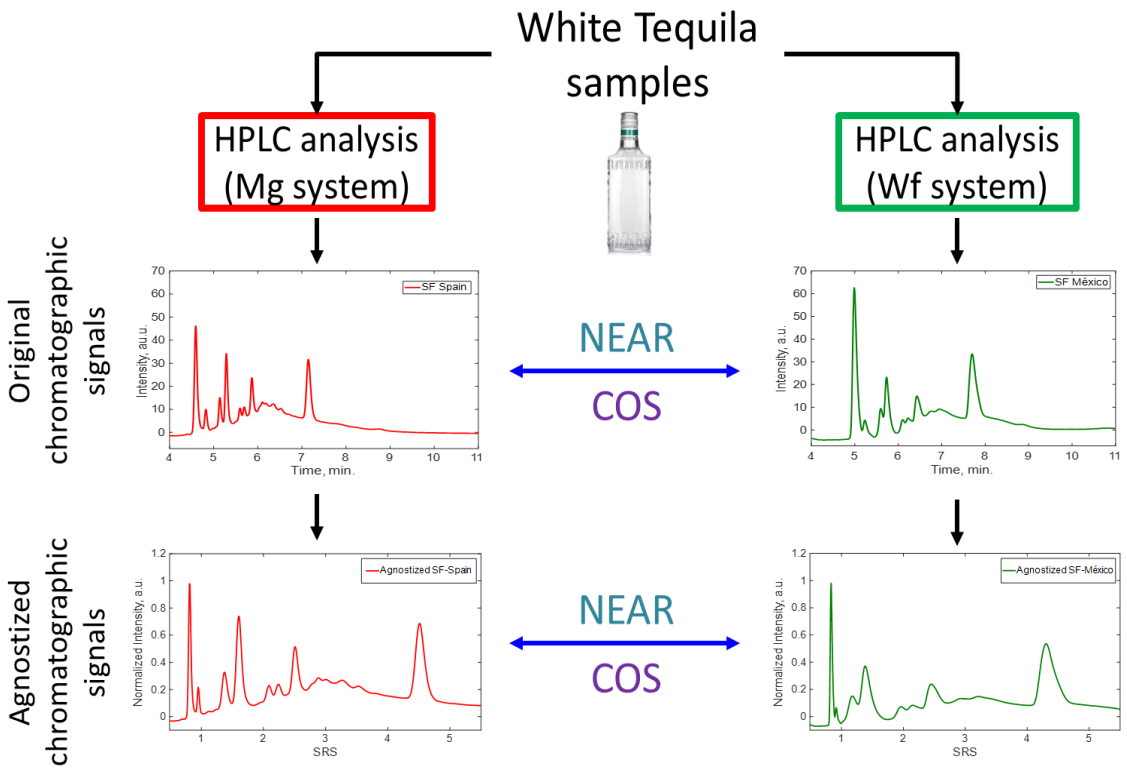


Figure 2. Experimental design layout of the similarity analyses performed among original and agnostized chromatographic fingerprints of White Tequila obtained from two different HPLC pieces of equipment, one considered as the manager (Mg) system and the other one as the workforce (Wf) system. The similarity parameters used in these studies were the nearness index (NEAR) and cosine angle (COS).

3.6. Multivariate analysis

All multivariate analysis was carried out using PLS_Toolbox (ver 8.6.1, Eigenvector Research Inc. MA, USA), working under the MATLAB environment. Prior to the development of the multivariate models, the next pre-processing was always applied over the data vectors: automatic weighted least squares (order: 2), autoscaling, and smoothing (filter width: 7, polynomial order: 1).

Partial least squares-discriminant analysis (PLS-DA) and support vector machine (SVM) were used as classification methods for the development of the multivariate models. The samples for the training and external validation sets were randomly designated, selecting them with the support of the Onion method [**24**]. The data obtained from the two different HPLC equipment were denominated as: 'TB' class from the Spanish term 'Tequila Blanco' (White Tequila) and 'TBM' class from the Spanish term 'Tequila Blanco Mixto' (Mixed White Tequila).

4. Results and discussion

In order to achieve a global harmonised database and a single analytical multivariate method, it is necessary to obtain standardised data with enough quality. Otherwise, the analytical multivariate method is influenced by the variations that the analytical instrument may present over the chromatographic fingerprints.

In this sense, a White Tequila database was created from an interlaboratory study, performed in Spain (SP) and México (MX), where the same Tequilas samples were analysed in both laboratories with different HPLC equipment. The chromatographic fingerprints of a ‘100 % agave’ White Tequila sample analysed in Spain and México can be observed in Figure 3 (a) and (b), respectively; as well as the ESM chromatograms obtained in Spain and México in Figure 4 (a) and (b), respectively.

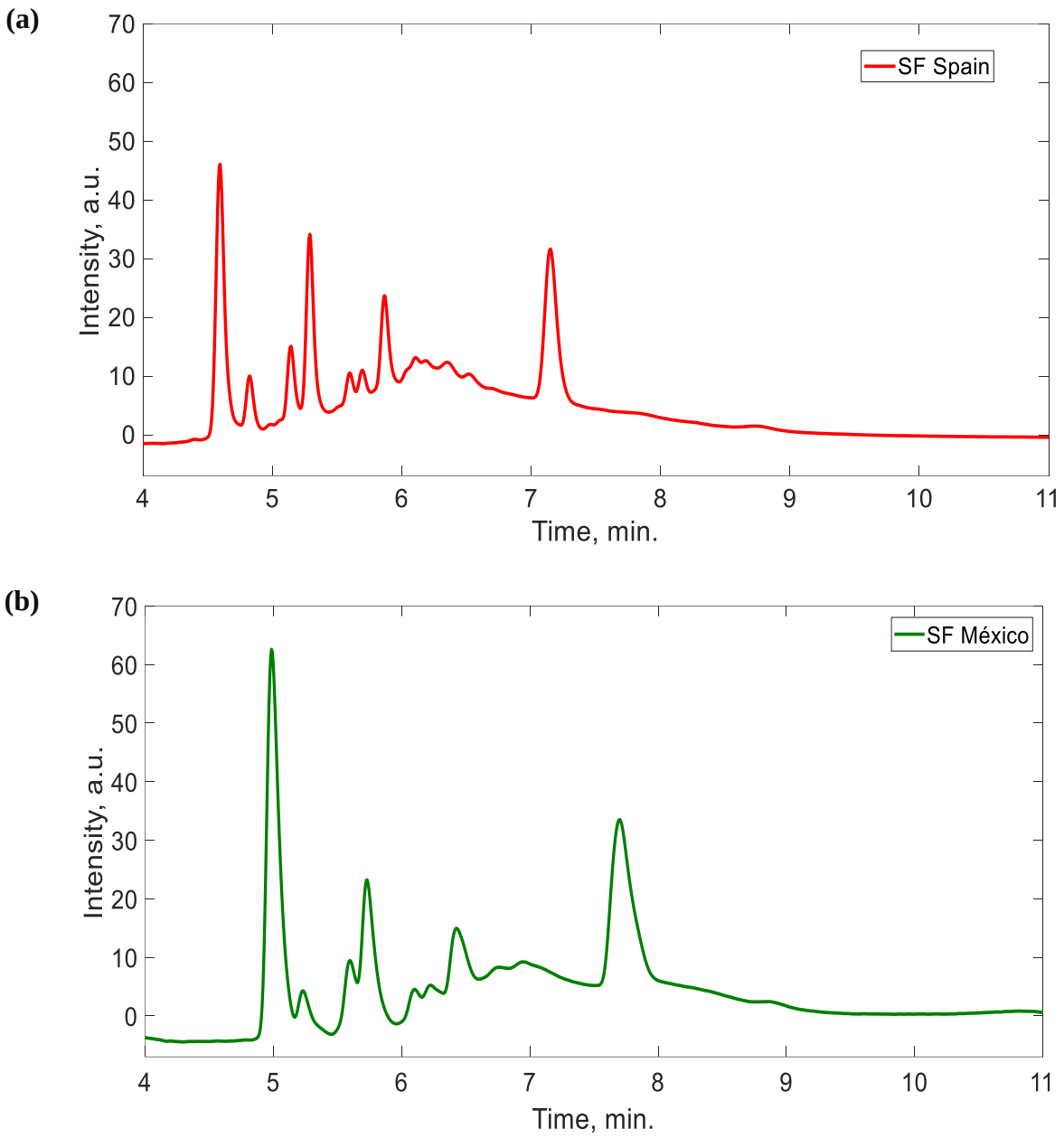


Figure 3. Sample fingerprints (SF) of the same '100 % agave' White Tequila sample obtained in (a) Spain and in (b) México.

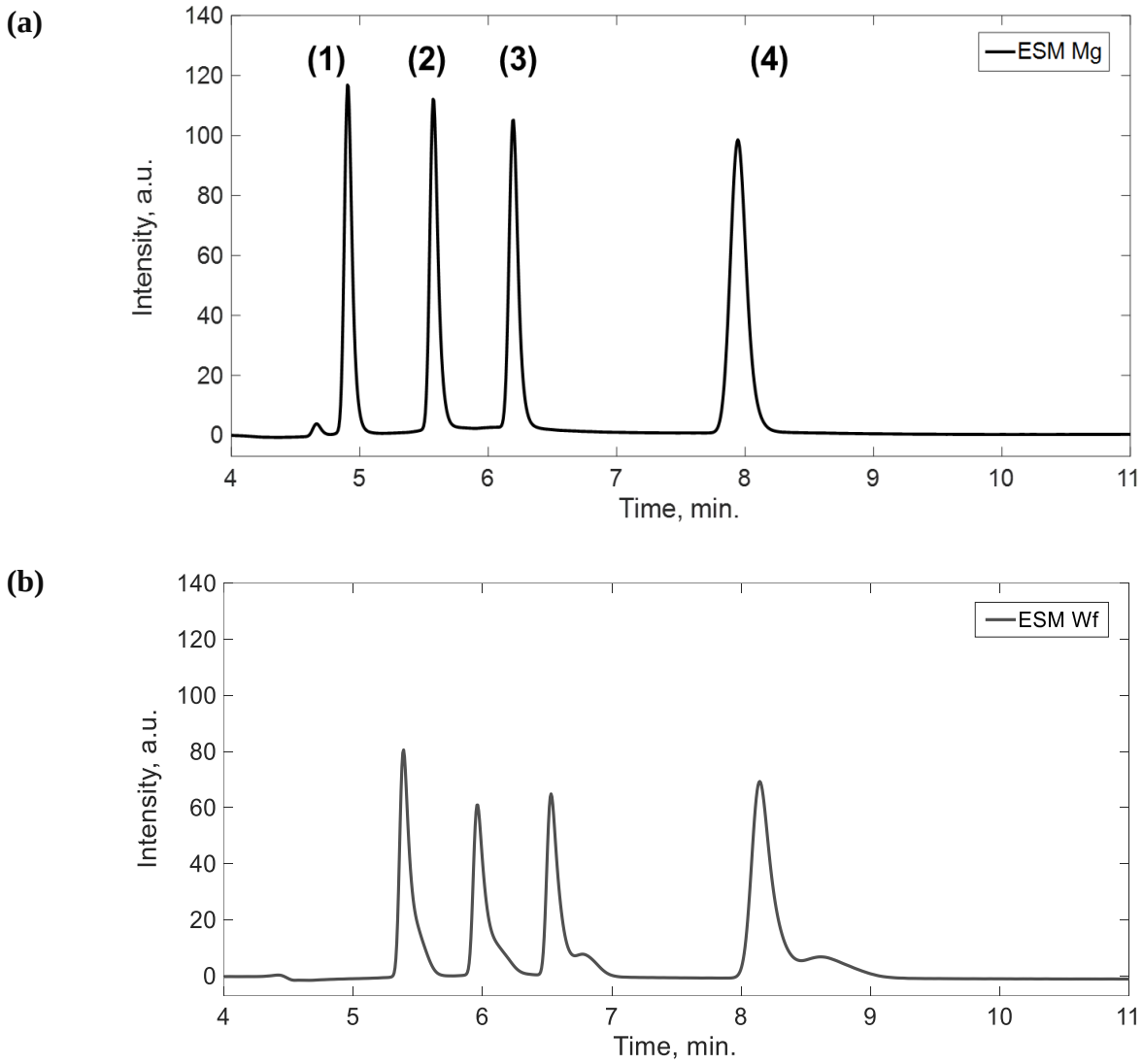


Figure 4. External standard mixture (ESM) chromatograms obtained from (a) the manager system in Spain and from (b) the workforce system in México. The ESM was composed by (1) 5-(hydroxymethyl)furfural, (2) furfural, (3) 2-acetylfuran and (4) 2-acetyl-5-methylfuran.

As observed from Figure 3 (a) and (b), the shape of both chromatographic fingerprints is similar, observing the characteristic peaks of furanic compounds present in tequilas [**25**], which are mostly generated by the Maillard reaction during the agave cooking step [**26**].

Nonetheless, the Mexican chromatographic fingerprint —considered from now on as the 'workforce (Wf) system'— presents less sensitivity and different runtime from the Spain chromatographic fingerprint —considered from now on as the 'manager (Mg) system'— observing an expansion effect of the chromatographic fingerprint. Surprisingly, the duration time of the ESM chromatograms of the Mg and Wf systems were similar, as observed in Figure 4, what led to identify that the Wf system was not working properly, since the chromatographic behaviour of the ESM and chromatographic fingerprints obtained with the Wf system was different. This situation was observed in most of the chromatographic fingerprints obtained with the Wf system, which was not consistent and caused different chromatographic behaviour among the ESM chromatograms and the chromatographic fingerprints.

In order to solve this situation, the Equity function was individually applied to each chromatographic fingerprint obtained with the Wf system, as explained in section 2. Considerations prior to the signal agnostization stage. An example of original and corrected chromatographic fingerprints can be observed in Figure 5 (a) and (b), respectively.

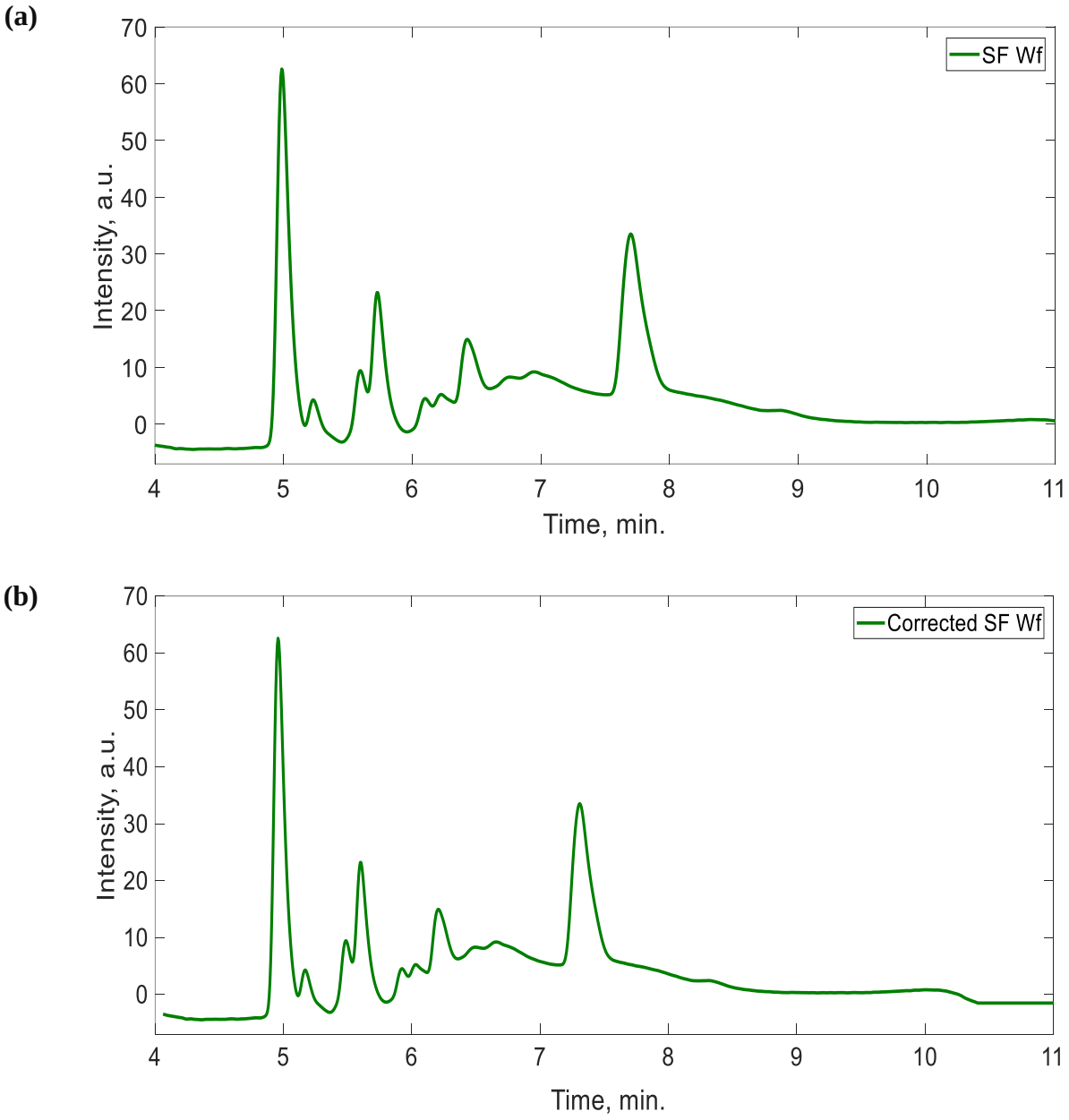


Figure 5. (a) Original sample fingerprint obtained with the workforce (Wf) system and (b) corrected sample fingerprint using the Equity function.

Once the chromatographic fingerprints of the Wf system were corrected, the 'instrument-agnostizing' methodology took place to harmonise them and to create a unique White Tequila database, as explained in subsection 3.4. Agnostizing of instrumental

chromatographic signals. Examples of agnostized chromatographic fingerprints can be observed in Figure 6 in which the time domain has been replaced by the new SRS domain and the intensities of the chromatographic fingerprints have been normalised. However, similarity analyses were performed to ensure that the chromatographic fingerprints were useful before the multivariate model building, as explained in the following subsection 4.1.

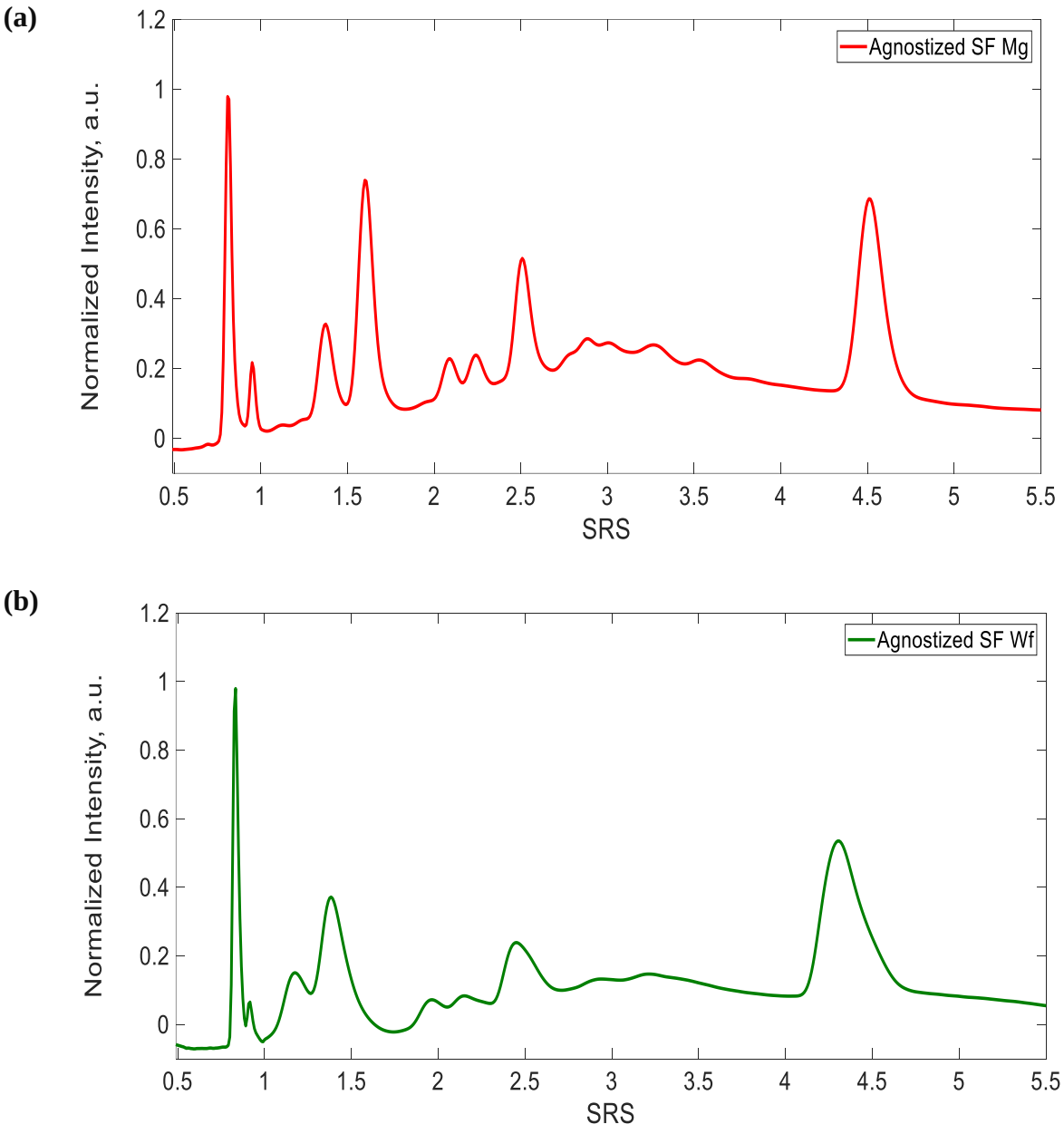


Figure 6. Agnostized sample fingerprints obtained from the (a) manager and (b) workforce systems.

4.1. Similarity analyses between chromatographic signals before and after agnostizing

For these analyses 10 chromatographic fingerprints of White Tequila obtained with the Mg system were compared with the same 10 corrected chromatographic fingerprints obtained with the Wf system. The first 5 samples belong to the '100 % agave' category, whilst the 5 remaining samples belong to the 'mixed' category. Such comparisons were performed before and after the agnostization of the chromatographic fingerprints using the NEAR index and cosine (COS) angle, whose results can be observed in Table 3.

Table 3. Similarity analysis results between chromatographic signals before and after their agnostization obtained from the manager and workforce system using the NEAR index and Cosine angle.

	NEAR		**COS**	
Mg *vs* Wf	***Before Agnost***	***After Agnost***	***Before Agnost***	***After Agnost***
Sample 1	0.2892	0.5534	0.3325	0.7049
Sample 2	0.3288	0.6245	0.3934	0.7649
Sample 3	0.3694	0.7335	0.4982	0.8674
Sample 4	0.4130	0.5956	0.5197	0.7264
Sample 5	0.2433	0.7404	0.2992	0.8882
Sample 6	0.4103	0.7588	0.5721	0.8901
Sample 7	0.3459	0.6957	0.4771	0.8327
Sample 8	0.4009	0.7647	0.5349	0.8952
Sample 9	0.3647	0.6612	0.526	0.8297
Sample 10	0.3711	0.7604	0.5423	0.8915

Mg: manager system ; Wf: workforce system

Both similarity indexes (NEAR and COS) present low values (less than 0.4) before the agnostization step, which indicate that both chromatographic fingerprints were dissimilar among them. After the agnostization step, both indexes improved their values (more than 0.7), with better outcomes for the COS angle, indicating that the chromatographic fingerprints are now able to be part of the White Tequila database and subsequent single analytical multivariate model.

4.2. Single analytical multivariate model: tequila authenticity

Once the chromatographic fingerprints of both the Mg and Wf systems were agnostized and verified to be similar, they were utilised to build the subsequent multivariate model, aimed at differentiating among the '100 % agave' (TB-target class) and 'mixed' (TBM-alternative class) categories of White Tequila.

- *Minimum validation requirements*

Since the ultimate goal of this single multivariate model is its implementation in a real analytical application, it is necessary to establish minimum validation requirements in order to evaluate if the method is fitted-for-purpose [**27**]. For this, the occurrence must be taken into consideration, which is a population-parameter that informs on the rate of samples of interest against the total sample population that are subjected to analysis in the laboratory. At the same time, this population-parameter has a direct influence on the quality performance metrics of sensitivity (SENS) and precision (PREC), which are calculated using the occurrence parameter and two other related indexes that should be established in advance: Index saving (I_{SAVING}) and assignation error index (I_{ERROR}) [**27**]. On the one hand, the I_{SAVING} informs on the economic savings the analytical laboratory may have, since the samples correctly classified would not be subjected to confirmatory analyses. On the other hand, I_{ERROR} indicates the risk of misclassifying a sample, either from the target or alternative class.

In this particular case, the samples of interest are the 25 samples that belong to the TB class, which have an occurrence of 0.56. Moreover, the I_{SAVING} and I_{ERROR} were set at 60 % and 15 %, respectively. Thus, the minimum expected SENS and PREC for the multivariate model are calculated according to the equations 9 and 10, obtaining values of 0.85 and 0.75 for each of them, respectively. Once the validation requirements were settled down, the multivariate model building was performed to confirm if it was capable to achieve such performances.

$$SENS = (I_{SAVING} - I_{ERROR}) \times \frac{1}{OCURR} \quad (9)$$

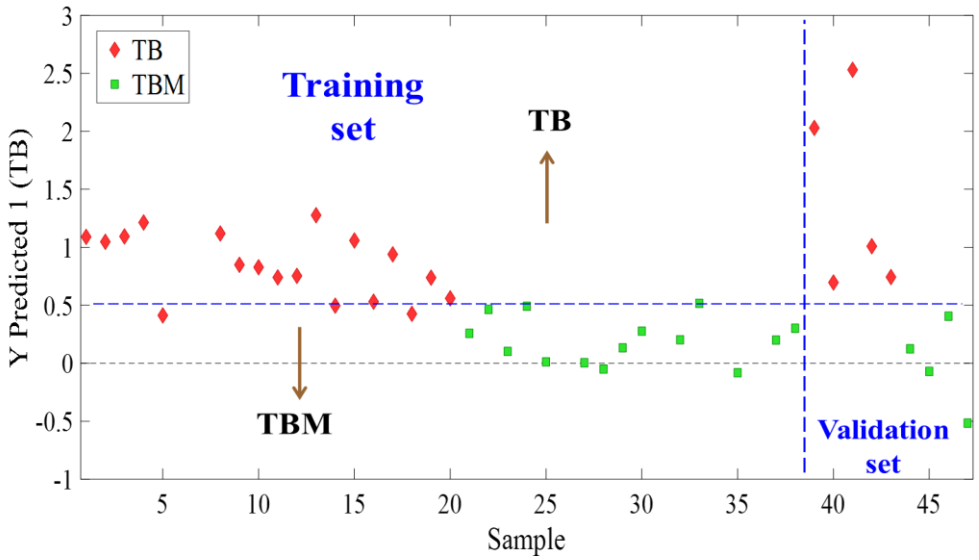
$$PREC = 1 - \frac{I_{ERROR}}{I_{SAVING}} \quad (10)$$

- *Analytical multivariate model building*

For this purpose, the initial multivariate models were performed with the chromatographic fingerprints obtained in Spain, which were considered as the reference ones obtained with the manager system. From the 41 total samples, 32 samples constituted the training set (18 TB and 14 TBM), 9 samples the external validation set (5 TB and 4 TBM) and 1 TBM sample was excluded from the training set, since it was an outlier detected from the initial exploratory analyses with principal component analyses.

The first chemometric tool to be used was PLS-DA, which was utilised to build a mathematic model using 5 LVs, which explained 59.8 % and 64.1 % of the cumulative variance in the X and Y blocks, respectively. The classification results for the training and external validation sets can be observed in Figure 7 (a) and the validation contingencies results in Figure 7 (b). A value = 0.5 was considered as a threshold value (dashed blue line), where samples with threshold values >0.5 were classified as TB and samples with threshold values <0.5 as TBM. From Figure 7 (a), it is evident the proper classification ability of the PLS-DA model, since all samples from the validation set were correctly classified. The validation results from Figure 7(b) were used to calculate the quality performance metrics SENS and PREC, which obtained values of 1 for both of them, indicating that the model achieved the minimum validation requirements to be implemented in real analytical applications.

(a)



(b)

Assignation	Class 1 (TB)	Class 2 (TBM)	
Inconclusive (I)	0	0	**0**
Class 2 (TBM)	0	4 (50 %)	**4**
Class 1 (TB)	5 (50 %)	0	**5**
	5	**4**	**9**

Actual: Class 1, Class 2

Figure 7. (a) Classification plot and (b) validation contingencies for the PLS-DA classification model using the chromatographic fingerprints of the manager system. Class 1: target class (TB: '100 % agave' White Tequila); class 2: alternative class (TBM: 'mixed' White Tequila). The dashed blue line in Figure 6(a) indicates the threshold limit of 0.5.

Once the mathematic model was built and its results validated, the next stage involved an extra evaluation of 26 additional samples analysed by the Mg system again. To perform this evaluation, the following activities were performed: first, the model was augmented with the samples of the external validation set; afterwards, the class of the 26 additional samples was predicted, obtaining that 4 TB samples were more similar to the TBM samples and 6 TBM samples more similar to the TB ones, as observed in Figure 8. Finally, the model was augmented once again including all samples.

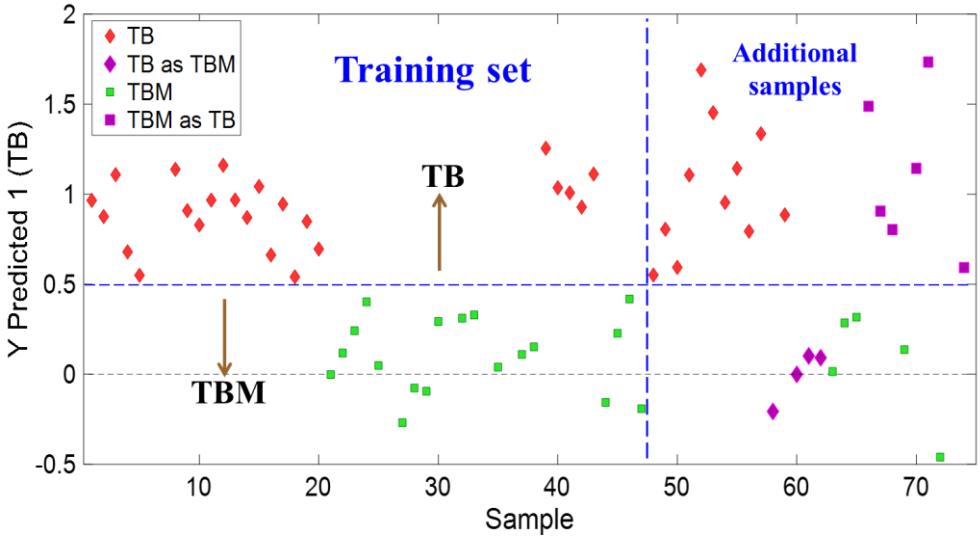


Figure 8. Classification plot for the PLS-DA classification model using 26 additional chromatographic fingerprints of White Tequila, obtained with the manager system in real conditions.

The second multivariate technique applied was SVM, which was performed with the radial basis function (RBF) kernel algorithm, gamma values between 10-6 – 10, cost values between 10-3 – 102, and PLS compression with 4 LVs. Please, note that the same development process of the PLS-DA model was followed for the building process of the SVM model: first, the model was trained and externally validated; afterwards, the samples were augmented onto the SVM model; then, the 26 additional samples were evaluated with the previous augmented SVM model. The initial classification results for both the training and external validation sets can be observed in Figure 9 (a), obtaining consistent results with the PLS-DA model in which all samples were correctly classified and the minimum validation requirements were surpassed. Afterwards, the samples were augmented onto the model and the additional 26 agnostized chromatographic fingerprints were evaluated, whose results can be observed in Figure 9 (b). In this case, the SVM model suggested that 4 TB samples (the same result provided by the PLS-DA model) have similar characteristics than the TBM samples, and that 7 TBM samples (1 more TBM sample than the PLS-DA model) have similar characteristics to the TB samples.

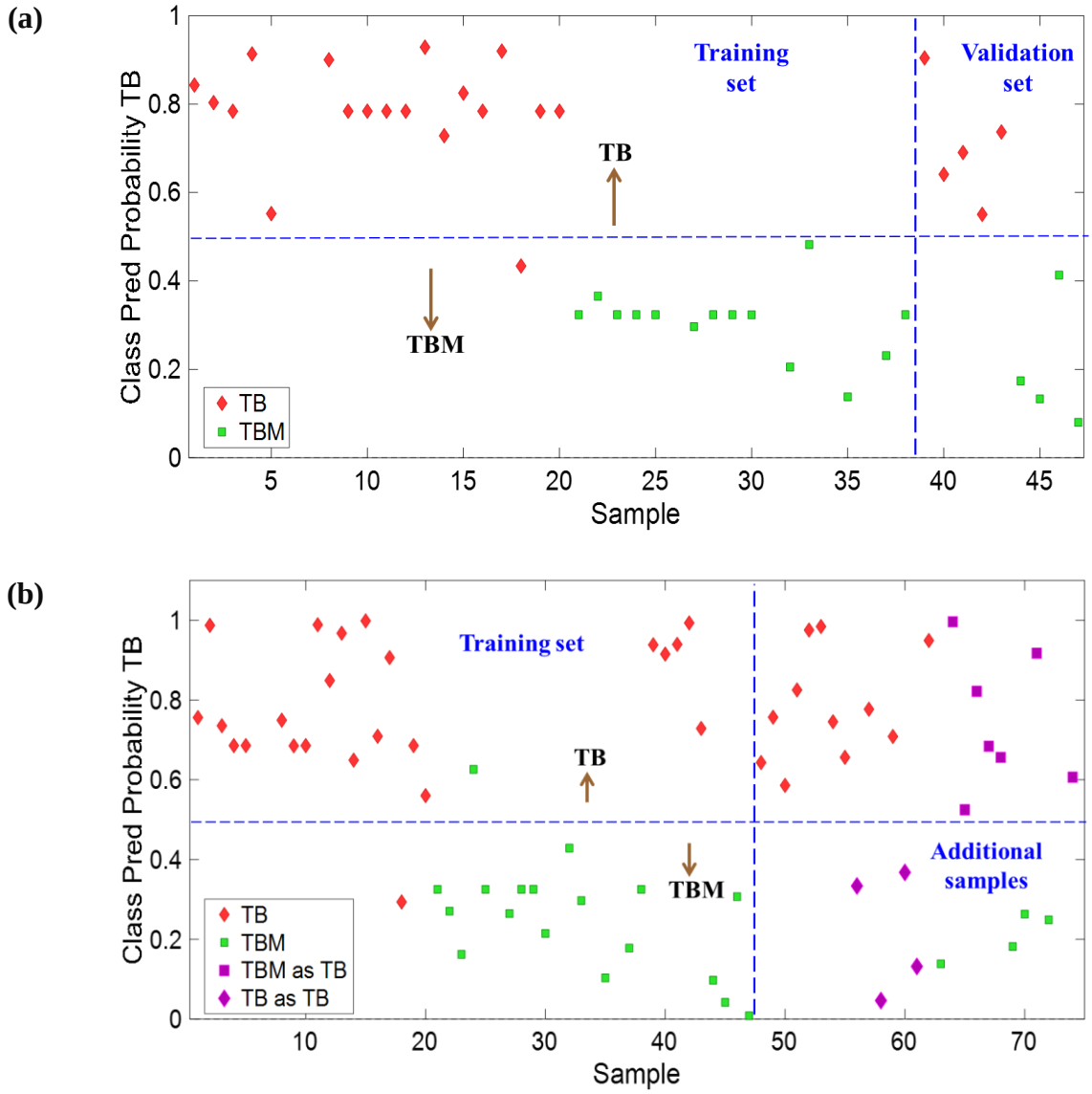


Figure 9. (a) Classification plot for the initial SVM classification model, and (b) classification plot for the SVM classification model in which 26 additional chromatographic fingerprints were assessed.

- *Analytical multivariate method in pilot conditions*

In a real quality control scenario of Tequila, both '100 % agave' and 'mixed' categories are analysed with different official and traditional methods before exporting them to more than 40 countries [**28**], where only the former can be exported in bottles and the later in large volumes [**29**].

The same quality control analyses are expected upon the arrival of the product in order to verify its original quality, since adulteration or falsification could take place during its transportation.

In this regard, the analytical multivariate method developed with the Mg system was applied over the agnostized fingerprints obtained in México, to probe the hypothesis that the same tequila samples analysed in different instrumental equipment at different periods of time allow to obtain the same results. This, would have two application strands: (i) the creation of harmonised and global databases, and (ii) the development of a single classification model.

On this subject, the 41 Mexican agnostized chromatographic fingerprints were evaluated with the previously built PLS-DA and SVM models. In this scenario, the PLS-DA multivariate model suggested that confirmatory analyses should be performed on 7 TB and 14 TBM samples, whilst the SVM model suggested 3 TB and 14 TBM samples, as it could be observed in Figure 10 (a) and (b), respectively, since the mathematic models detected in these samples the presence of different characteristics from the category displayed on the label.

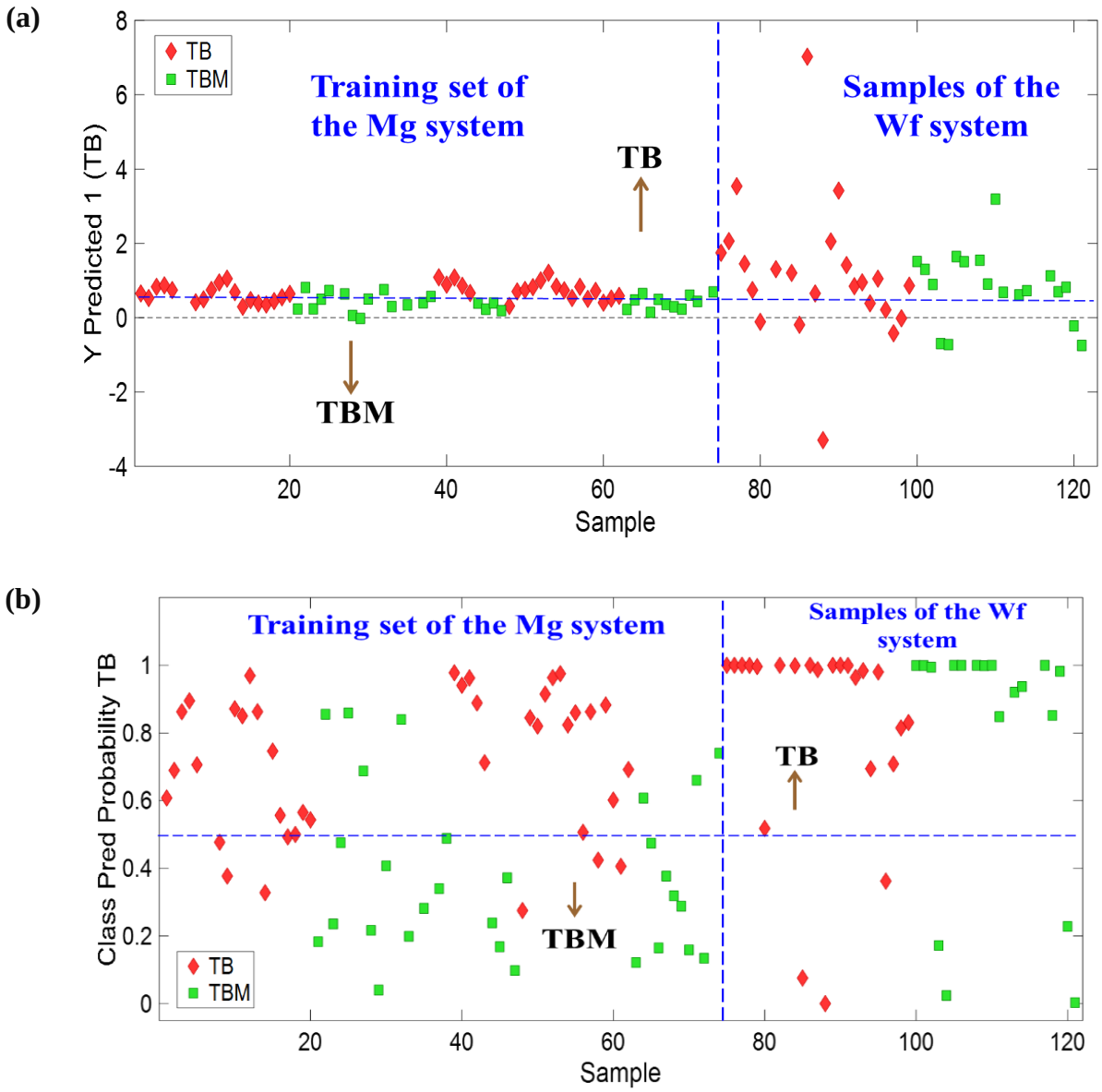


Figure 10. Classification plots in which 41 agnostized chromatographic fingerprints, analysed with the workforce (Wf) system, were evaluated with the (a) PLS-DA and (b) SVM classification models built with chromatographic fingerprints of the manager (Mg) system. Class 1: target class (TB: '100 % agave' White Tequila); class 2: alternative class (TBM: 'mixed' White Tequila). The dashed blue lines in Figure 10 (a) and (b) indicate the threshold limit of 0.5.

After the evaluation of the agnostized fingerprints obtained in México, the I_{SAVING} and I_{ERROR} were calculated in order to evaluate the performance of both mathematical models [**27**]. In this regard, the I_{SAVING} and I_{ERROR} values for the PLS-DA model were 48.8 % and

51.2 %, respectively, whilst for the SVM were 58.5 and 41.5, respectively. Therefore, the SVM multivariate model should be preferred, as it will allow the analytical laboratory to save up more costs and time of analysis with only 17 samples subjected to confirmatory analyses with traditional methods. It is worth mentioning that despite of the initial discrepancies among the chromatographic signals, the results presented here are of significant value, since it was possible to correct the signals through the Equity function and further use them for the creation of a single multivariate model through the instrument-agnostizing methodology, obtaining promising results.

5. Key findings and future implications

In view of the results attained within this study, it is concluded that the questions posed in the introduction section have been duly answered since it has been possible:

- to obtain similar *instrument-agnostic* chromatographic fingerprints of the same sample in different instrumental equipment located in two different countries, obtaining similarity index values of 0.7
- to create a global database of *instrument-agnostic* chromatographic fingerprints of tequila, which are independent from the equipment they were acquired with.
- to develop and validate according to previously established applicability requirements of the analytical method a single mathematical model with *instrument-agnostic* fingerprints with the aim of authenticating tequila.

Finally, it should be noted that this methodology represents a substantial advance in the field of food and beverage quality, since it is possible to generate databases of instrumental food fingerprints obtained in different laboratories, allowing the generation of single models with comparable results among laboratories and allowing to save economic and human resources, as it would only be necessary to analyse by traditional methods those samples whose results can be considered inconclusive.

Conflicts of interest

The authors declare that they have no conflict of interest.

Acknowledgements

The authors are deeply grateful to the Mexican 'Consejo Regulador del Tequila' (CRT) for providing the samples for this study. C.H. Pérez-Beltrán acknowledges also the scholarships from the Autonomous University of Sinaloa (México) and from the 'Asociación Universitaria Iberoamericana de Posgrado' (AUIP) and the 'Consejería de Transformación Económica, Industria, Conocimiento y Universidades' of the Regional Government of Andalusia (Spain) for a research stay. AMJC acknowledges the Grant (RYC2021-031993-I) funded by MCIN/AEI/501100011033 and "European Union NextGenerationEU/PRTR".

References

[1] F. Tian, A supply chain traceability system for food safety based on HACCP, blockchain & Internet of things, 2017 ICSSSM., (2017) 1-6. https://doi.org/10.1109/ICSSSM.2017.7996119

[2] S. Benito, The Management of Compounds that Influence Human Health in Modern Winemaking from an HACCP Point of View, Fermentation,5, (2019) 1-20. https://doi.org/10.3390/fermentation5020033

[3] V. Barrere, K. Everstine, J. Théolier, S. Godefroy, Food fraud vulnerability assessment: Towards a global consensus on procedures to manage and mitigate food fraud, Trends Food Sci. Technol. 100 (2020) 131-137. https://doi.org/10.1016/j.tifs.2020.04.002

[4] J.M. Soon, S.C. Kryzaniak, Z. Shuttlewood, M. Smith, L. Jack, Food fraud vulnerability assessment tools used in food industry, Food Control, 101 (2019) 225-232. https://doi.org/10.1016/j.foodcont.2019.03.002

[5] C. Brooks, L. Parr, J.M. Smith, D. Buchanan, D. Snioch, E. Hebishy, A review of food fraud and food authenticity across the food supply chain, with an examination of the impact of the COVID-19 pandemic and Brexit on food industry, Food Control 130 (2020) 108171. https://doi.org/10.1016/j.foodcont.2021.108171

[6] W.S. Alrobaish, L. Jacxsens, P. Spagnoli, P. Vlerick, Assessment of food integrity culture in food businesses through method triangulation, Food Control 141 (2022) 109168. https://doi.org/10.1016/j.foodcont.2022.109168

[7] A.F. El Sheikha (2020). Food authentication: Introduction, techniques, and prospects, in: C.M. Galanakis (Ed), Food Traceability and Authentication, Academic Press / Elsevier, Oxford, 2020, pp. 163-193. https://doi.org/10.1016/B978-0-12-821104-5.00006-4

[8] Y. Lu, P. Li, H. Xu, A Food anti-counterfeiting traceability system based on Blockchain and Internet of Things, Procedia Comput. Sci. 199 (2020) 629-636. https://doi.org/10.1016/j.procs.2022.01.077

[9] J. Q, L. Ruiz-Garcia, B. Fan, J.I. Robla Villalba, U. McCarthy, B. Zhang, Q. Yu, W. Wu, Food traceability system from governmental, corporate, and consumer perspectives in the European Union and China: A comparative review, Trends Food Sci. Technol. 99 (2020) 402-412. https://doi.org/10.1016/j.tifs.2020.03.025

[10] L. Cuadros-Rodríguez, C. Ruíz-Samblás, L. Valverde-Som, E. Pérez-Castaño, A. González-Casado, Chromatographic fingerprinting: An innovative approach for food 'identitation' and food authentication - A tutorial, Anal. Chim. Acta 909 (2016) 9-23. http://dx.doi.org/10.1016/j.aca.2015.12.042

[11] S. Medina, J.A. Pereira, P. Silva, R. Perestrelo, J.S. Camara, Food fingerprints – A valuable tool to monitor food authenticity and safety, Food Chem. 278 (2019) 144-162. https://doi.org/10.1016/j.foodchem.2018.11.046

[12] S. Esslinger, J. Riedl, C. Fauhl-Hassek, Potential and limitations of non-targeted fingerprinting for authentication of food in official control, Food Res. Int. 60 (2014) 189–204. https://doi.org/10. 1016/j.foodres.2013.10.015

[13] Food Authenticity, AOAC International (2019). https://www.aoac.org/wp-content/uploads/2019/08/FAM_Program_Update-_July_2019-1.pdf [Accessed on 27 September 2022]

[14] G. Tomasi, F. Savorani, S.B. Engelsen, Icoshift: An effective tool for the alignment of chromatographic data, J. Chromatogr. A 1248 (2011) 7832-7840. https://doi.org/10.1016/j.chroma.2011.08.086

[15] L. Cuadros Rodríguez, F. Ortega Gavilán, S. Martín Torres, S. Medina Rodríguez, A.M. Jiménez Carvelo, A. González Casado, M.G. Bagur González, Standardization of chromatographic signals – Part I: Towards obtaining instrument-agnostic fingerprints in gas chromatography, J. Chromatogr. A 1641 (2021) 461983. https://doi.org/10.1016/j.chroma.2021.461983

[16] L. Cuadros Rodríguez, S. Martín Torres, F. Ortega Gavilán, A.M. Jiménez Carvelo, R. López Ruíz, A. Garrido Frenich, M.G. Bagur González, A. González Casado, Standardization of chromatographic signals – Part II: Expanding instrument-agnostic fingerprints to reverse phase liquid chromatography, J. Chromatogr. A 1641, (2021) 461973. https://doi.org/10.1016/j.chroma.2021.461973

[17] C. H. Pérez-Beltrán, A.M. Jiménez-Carvelo, S. Martín-Torres, F. Ortega-Gavilán, L. Cuadros-Rodríguez, Icoshift: Instrument-agnostic multivariate models from normal phase liquid chromatographic fingerprinting. A case study: Authentication of olive oil, Food Control, 137 (2022) 108957. https://doi.org/10.1016/j.foodcont.2022.108957

[18] Framework for selecting and testing of food products to assess quality related characteristics: EU harmonised testing methodology. EC report, 25 April 2018.

[19] EUR 29779 EN, Results of an EU wide comparison of quality related characteristics of food products. European Commission report, June 2019.

[20] C. de Bolle, C. Archambeau, Intellectual property crime. Threat assessment 2022, EUIPO. (2022).
https://doi.org/10.2814/830719

[21] 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

[22] F. Ortega-Gavilán, L. Valverde-Som, F.P. Rodríguez-García. L. Cuadros-Rodríguez, M.G. Bagur-González, Homogeneity assessment of reference materials for sensory analysis of liquid foodstuffs. The virgin olive oil as case study, Food Chem.322 (2020) 126743

[23] K. Zhang, H. Wang, L. Zhong, L. Liu, R. Huang, H. Zhang, D. Xu, W. Yin, L. Li, H. Zang, Evaluation and monitoring of the API content of a portable near infrared instrument combined with chemometrics based on fluidized bed mixing process, J. Pharm. Innov. 17 (2022) 1136-1147.
https://doi.org/10.1007/s12247-021-09581-2

[24] N.B. Gallagher, D. O'Sullivan, Selection of representative learning and test sets using the Onion method, Eigenvector. [Accessed 08 September 2022]
https://eigenvector.com/wp-content/uploads/2020/01/Onion_SampleSelection.pdf

[25] A.C. Muñoz-Muñoz, A.C. Grenier, H. Gutiérrez-Pulido, J. Cervantes-Martínez, Development and validation of a high performance liquid chromatography-diode array detection method for the determination of aging markers in tequila, J. Chrom A. 1213 (2008) 218-223.
https://doi.org/10.1016/j.chroma.2008.10.018

[26] N.A. Mancilla-Margalli, M.G. López, Generation of Maillard compounds from inulin during the thermal processing of agave tequilana weber var. azul, J. Agric. Food Chem. 4 (2002) 806-812.
https://doi.org/10.1021/jf0110295

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

[28] Protection of tequila at the international level. Regulatory Council of Tequila. https://www.crt.org.mx/index.php/es/pages-3/proteccion-del-tequila-a-nivel-internacional

[29] Mexican Official Standard NOM-006-SCFI-2012, Alcoholic Beverages -Tequila- Specifications, National Advisory Committee on Standardization, User Safety, Commercial Information and Trade Practices (CCNNSUICPC), Mexican Government. https://www.crt.org.mx/images/documentos/Normas/NOM_006_SCFI_2012_Ingles.pdf (accessed January 14th 2023)

3.4. Comunicación a congresos

C.H. Pérez-Beltrán, A.M. Jiménez-Carvelo, S. Martín-Torres, F. Ortega-Gavilán, L. Cuadros-Rodríguez. Detección de adulteración en aceites de oliva a partir de la huella instrumental de la fracción polar y análisis multivariable. EXPOLIVA XX Simposio científico-técnico. Jaén, 2021. *Comunicación en formato Póster*.

C.H. Pérez-Beltrán, J.J. Olmos-Espejel, A.M. Jiménez-Carvelo, G. Pérez-Caballero, L. Cuadros-Rodríguez. Desarrollo de huellas instrumentales de tequilas blanco 100% agave y mixto mediante HPLC-DAD. Congreso Iberoamericano de Ciencia, Educación y Tecnología. Estado de México, 2021. *Comunicación en formato Póster*.

A.M. Jiménez-Carvelo, C.H. Pérez-Beltrán, J.J. Olmos-Espejel, G. Pérez-Caballero, L. Cuadros-Rodríguez, Towards the creation of a global harmonised database and a single analytical multivariate method - tequila authentication as a case study, Colloquium Chemometricum Mediterraneum, Padova, Italy, 2023, *Comunicación en formato Póster*.

Capítulo 4

MÉTODOS ANALÍTICOS MULTIVARIABLE BASADOS EN TÉCNICAS NO INVASIVAS

4. MÉTODOS ANALÍTICOS MULTIVARIABLE BASADOS EN TÉCNICAS NO INVASIVAS

4.1. Resumen

Los métodos analíticos oficiales de control de calidad alimentaria suelen estar basados en técnicas analíticas tradicionales, tales como la cromatografía de gases –usada para la determinación del contenido permitido de aldehídos, metanol y alcoholes superiores en tequilas [**1**] o para la determinación de la composición y contenido de esteroles y dialcoholes triterpénicos en aceites de oliva [**2**]–, o la cromatografía de líquidos –utilizada para la determinación de los niveles máximos de furfural en tequilas [**3**] o del contenido de triglicéridos en aceites de oliva [**2**]–.

La mayoría de las técnicas analíticas tradicionales aplican un enfoque dirigido y requieren ser llevadas a cabo en instalaciones especializadas de manera *off-line* con un tiempo prolongado de análisis, son de alta complejidad y, a menudo, requieren de un exhaustivo pretratamiento de la muestra. Una alternativa para disminuir estos inconvenientes y aumentar la eficiencia de los análisis de control de calidad alimentario es aplicar un enfoque no dirigido para desarrollar métodos analíticos multivariable (MAM), basados en el uso de técnicas espectroscópicas no invasivas y herramientas quimiométricas, que sean rápidos, confiables, sencillos y que el tratamiento de la muestra sea poco o nulo, para que puedan ser usados como métodos de análisis de cribado (métodos analíticos de vanguardia) para una detección más temprana de posibles fraudes alimentarios.

De esta manera, el uso de técnicas espectroscópicas propicia el desarrollo de estos métodos analíticos de vanguardia que permiten agilizar los análisis de control y aseguramiento de

[**1**] Norma Mexicana NMX-V-005-NORMEX-2018. Bebidas alcohólicas-Determinación de aldehídos, ésteres, metanol, y alcoholes superiores-Métodos en ensayo (prueba). Organismo Nacional de Normalización Sociedad Mexicana de Normalización y Certificación, S.C. Comité Técnico de Normalización Nacional para Bebidas Alcohólicas.

[**2**] Reglamento (CEE) Nº 2568/91 de la Comisión, de 11 de julio de 1991, relativo a las características de los aceites de oliva y de los aceites de orujo de oliva y sobre sus métodos de análisis.

[**3**] Norma Mexicana NMX-V-004-NORMEX-2018. Bebidas alcohólicas-Determinación de furfural-Métodos de ensayo (prueba). Organismo Nacional de Normalización Sociedad Mexicana de Normalización y Certificación, S.C. Comité Técnico de Normalización Nacional para Bebidas Alcohólicas.

calidad de una mayor cantidad de muestras, dejando las técnicas analíticas tradicionales para el desarrollo de métodos analíticos de retaguardia [**4**], como lo son los métodos cromatográficos, utilizados para confirmar ciertos resultados sospechosos de adulteración o falsificación provistos por los métodos de vanguardia.

Entre las técnicas espectroscópicas más exploradas para el control de calidad en la industria alimentaria se encuentran la espectroscopía de infrarrojo medio con transformada de Fourier (FTIR), de infrarrojo cercano (NIR) y/o Raman [**5**,**6**], cuyos espectros son habitualmente utilizados para desarrollar de manera individual métodos analíticos multivariable. Sin embargo, también es posible generar un MAM que combine las señales de diferentes técnicas analíticas con objeto de mejorar los resultados de clasificación/cuantificación obtenidos individualmente. A esta estrategia se le conoce como "fusión de datos" [**7**] o también "análisis multibloque" [**8**].

La fusión de datos puede realizarse a bajo, medio o alto nivel, cuya elección dependerá del objetivo del estudio. La fusión de datos de bajo nivel consiste en la unión de los datos de las diferentes técnicas analíticas en una sola matriz, la cual contendrá tantas filas como muestras analizadas y tantas columnas como variables. En la fusión de datos intermedia, o de nivel medio, primero, se extrae la información más relevante de los datos originales a través de las puntuaciones (*scores*) y/o los ponderales (*loadings*) referidas a las variables latentes obtenidas de PLS y/o PCA, y segundo, se concatenan en una única matriz. Por último, la fusión de datos de alto nivel, también llamada fusión de nivel de decisión, combina los resultados de cada modelo individual obtenido para cada fuente de datos. Dicha fusión de datos en análisis alimentario se ha enfocado mayoritariamente en

[**4**] M. Valcárcel, S. Cárdenas, Vanguard-rearguard analytical strategies, 2005, Trends in Analytical Chemistry, 24, 67–74.

[**5**] V.A. Tirado-Kulieva, E. Hernandez-Martinez, J.P. Suomela, Non-destructive assessment of vitamin C in food: a review of the main findings and limitations of vibrational spectroscopic techniques, 2022, European Food Research and Technology, 248, 2185 – 2195.

[**6**] E. Arendse, H. Nieuwoudt, L.S. Magwaza, J.F.I. Nturambirwe, O.A. Fawole, U.L. Opara, Recent advancements on vibrational spectroscopic techniques for the detection of authenticity and adulteration in horticultural products with specific focus on oils, juices and powders, 2021, Food and Bioprocess Technology, 14, 1 – 22.

[**7**] S.M. Azacarate, R. Ríos-Reina, J.M. Amigo, H.C. Goicoechea, Data handling in data fusion: Methodologies and applications, 2021, Trends in Analytical Chemistry, 143, 116355.

[**8**] P. Mishra, J.M. Roger, D.J. Rimbaud-Bouveresse, A. Biancolillo, F. Marini, A. Nordon, D.N. Rutledge, Recent trends in multi-block data analysis in chemometrics for multi-source data integration, 2021, Trends in Analytical Chemistry, 137, 116206.

problemas de clasificación, haciendo uso comúnmente de la inferencia Bayesiana basada en estimación de probabilidad [**9**,**10**].

Hoy en día, el uso de herramientas quimiométricas, como las descritas en la **Introducción** de esta tesis doctoral, hace posible manejar grandes volúmenes de datos, incluso aquellos fusionados a cualquier nivel y provenientes de distintas técnicas analíticas. No obstante, sigue siendo común el uso de señales instrumentales obtenidas de técnicas espectroscópicas no fusionadas para desarrollar MAM para el control de calidad alimentario.

Es por lo anterior, que esta sección aborda el desarrollo novedoso de tres MAM basados en técnicas espectroscópicas y herramientas quimiométricas. En el primero de ellos, se empleó FTIR con el cual se obtuvieron señales instrumentales del Tequila Blanco, las cuales fueron pre-tratadas mediante la corrección de la línea base de dos maneras distintas: (i) una corrección por categoría de tequila a la cual pertenecía cada muestra, y (ii) una corrección general de la línea base para todas las muestras por igual; posteriormente, dichas señales instrumentales fueron fusionadas para lograr una mejor diferenciación entre las categorías '100 % agave' y 'mixto'. El segundo MAM consistió en el análisis NIR de Tequilas Blanco, analizados previamente mediante FTIR, con la finalidad de comprobar la utilidad de esta nueva información adquirida en el rango del infrarrojo cercano para diferenciar entre las categorías '100 % agave' y 'mixto', y predecir su contenido alcohólico. El tercer y último estudio se basó en la técnica espectroscópica SORS, una técnica analítica emergente en el ámbito alimentario (véase **Introducción**, subsección **1.4**), la cual fue utilizada, en primera instancia, para realizar mediciones de Tequila Blanco a través de sus botellas originales y realizar análisis de similitud, y enseguida, para realizar mediciones a través de viales ámbar, obteniendo señales instrumentales utilizadas para diferenciar entre las dos mismas categorías de Tequila Blanco y predecir su contenido alcohólico. Véase el esquema de los tres estudios en la **Figura 11**.

[**9**] A. Biancolillo, R. Bucci, A.L. Magrì, A.D. Magrì, F. Marini, Data-fusion for multiplatform characterization of an Italian craft beer aimed at its authentication, 2014, Analytica Chimica Acta, 820, 23 – 31.

[**10**] L.I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, John Wiley & Sons, Inc., Hoboken, 2004.

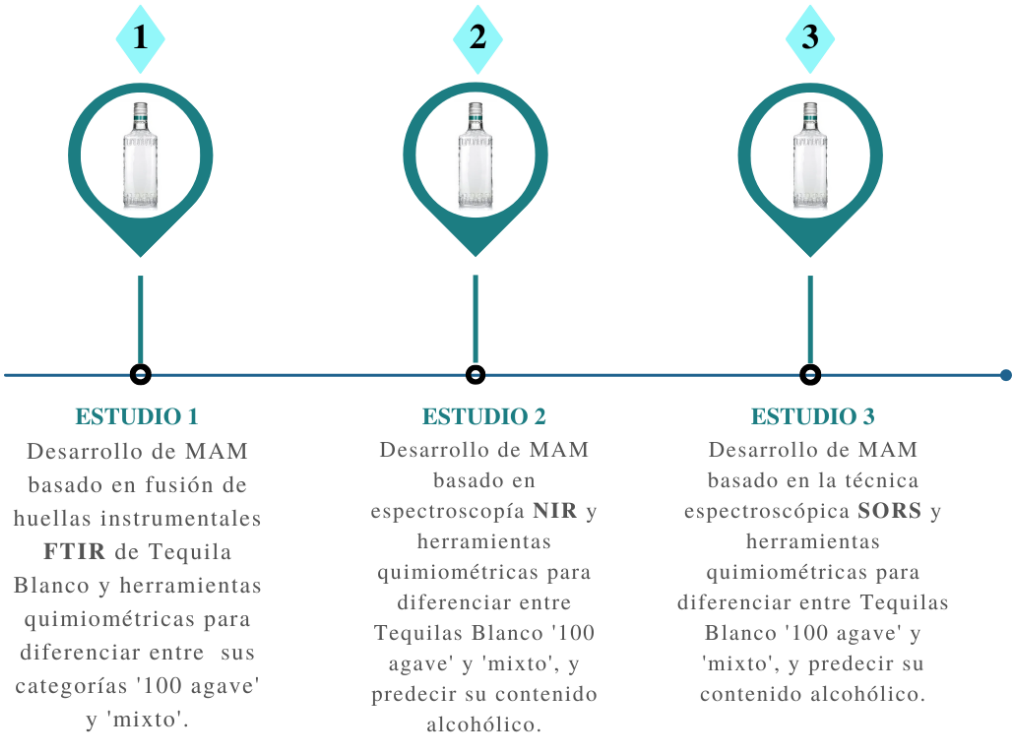


Figura 11. Esquema de los estudios llevados a cabo para el desarrollo de métodos analíticos multivariable (MAM) basados en técnicas no invasivas. *Estudio 1:* Desarrollo de un MAM en el cual se utilizó espectroscopía de infrarrojo medio con transformada de Fourier (FTIR), fusión de datos de bajo nivel y herramientas quimiométricas. *Estudio 2:* Desarrollo de un MAM basado en espectroscopía de infrarrojo en el rango cercano (NIR) y herramientas quimiométricas. *Estudio 3:* Desarrollo de un MAM *in-situ* basado en espectroscopía Raman con sistema de compensación espacial (SORS) y herramientas quimiométricas.

Estos MAM fueron desarrollados mediante estudios de investigación exclusivos para esta tesis doctoral, los cuales presentan un alto potencial de aplicación en la industria tequilera. Estos métodos analíticos multivariable podrían ser implementados en las empresas productoras de tequila, aunque no de manera exclusiva, para el control de calidad de sus productos y/o procesos de elaboración, así como también podrían ser utilizados por los organismos evaluadores de la conformidad para verificar y comprobar la calidad de los productos en puntos de venta y/o consumo, aumentando la eficiencia en el control de calidad y detección de tequilas adulterados o falsificados.

A continuación, se ponen de manifiesto de manera detallada los artículos científicos originados a partir de las investigaciones en las cuales se emplearon espectroscopía FTIR (*estudio 1*) y SORS (*estudio 3*), así como también se especifica la parte experimental y de resultados de la investigación con espectroscopía NIR (*estudio 2*), cuyos resultados no han sido enviados a publicación por las razones que posteriormente se aducen en el apartado correspondiente.

4.2. Artículo científico IV

chemosensors

MDPI

Article

A Sensor-Based Methodology to Differentiate Pure and Mixed White Tequilas Based on Fused Infrared Spectra and Multivariate Data Treatment

Christian Hazael Pérez-Beltrán [1], Víctor M. Zúñiga-Arroyo [2], José M. Andrade [3,*], Luis Cuadros-Rodríguez [1], Guadalupe Pérez-Caballero [2] and Ana M. Jiménez-Carvelo [1]

[1] Department of Analytical Chemistry, Faculty of Science, University of Granada, 18071 Granada, Spain; christianpb@correo.ugr.es (C.H.P.-B.); lcuadros@ugr.es (L.C.-R.); amariajc@ugr.es (A.M.J.-C.)
[2] Laboratorio de Fisicoquímica Analítica y Especiación Química, Unidad de Investigación Multidisciplinaria, Facultad de Estudios Superiores Cuautitlán, Campo 4, Universidad Nacional Autónoma de México, Cuautitlán-Izcalli 54714, Mexico; victor_0679_zvx@comunidad.unam.mx (V.M.Z.-A.); perezcg@unam.mx (G.P.-C.)
[3] Group of Applied Analytical Chemistry, Campus da Zapateira s/n, University of A Coruña, 15071 A Coruña, Spain
* Correspondence: jose.manuel.andrade@udc.es; Fax: +34-981-167065

Abstract: Mexican Tequila is one of the most demanded import spirits in Europe. Its fast-raising worldwide request makes counterfeiting a profitable activity affecting both consumers and legal distillers. In this paper, a sensor-based methodology based on a combination of infrared measurements (IR) and multivariate data analysis (MVA) is presented. The case study is about differentiating two categories of white Tequila: pure Tequila (or '100% agave') and mixed Tequila (or simply, Tequila). The IR spectra were treated and fused with a low-level approach. Exploratory data analysis was performed using PCA and partial least squares (PLS), whilst the authentication analyses were carried out with PLS-discriminant analysis (DA) and soft independent modeling for class analogy (SIMCA) models. Results demonstrated that data fusion of IR spectra enhanced the outcomes of the authentication models capable of differentiating pure from mixed Tequilas. In fact, PLS-DA presented the best results which correctly classified all fifteen commercial validation samples. The methodology thus presented is fast, cheap, and of simple application in the Tequila industry.

Keywords: white Tequila; authentication; IR spectroscopy; chemometrics; data fusion; beverage alcoholic industry

check for updates

Citation: Pérez-Beltrán, C.H.; Zúñiga-Arroyo, V.M.; Andrade, J.M.; Cuadros-Rodríguez, L.; Pérez-Caballero, G.; Jiménez-Carvelo, A.M. A Sensor-Based Methodology to Differentiate Pure and Mixed White Tequilas Based on Fused Infrared Spectra and Multivariate Data Treatment. *Chemosensors* **2021**, *9*, 47. https://doi.org/10.3390/chemosensors9030047

Academic Editor: Jose Vicente Ros Lis

Received: 1 February 2021
Accepted: 24 February 2021
Published: 27 February 2021

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

1. Introduction

Moderate and responsible consumption of spirits and wine is associated commonly with social relationships during leisure times and relaxing activities. People like tasting good products and experiencing new flavors and are prone to accept slightly higher expenses than usual. However, problems may arise when trust is broken by illicit practices like counterfeiting. At best 'only' economical losses are suffered, at worst health problems may arise.

Drinking a good spirit or wine is a hugely common social activity in many countries and thus an honest, historical, labor-demanding, and profitable industrial sector is devoted to those goods. Worldwide, the leading spirits markets in the 2010–2020 period according to value sales were (decreasing order) USA, Japan, Russia, China, UK, France, Germany, Mexico, Spain, and Canada [1]. According to the last European survey and statistics [2], during the period 2008–2013 EU annual production of spirits amounted to €22 billion and imports from third countries amounted to €1 billion (UK and France being the largest producers of spirits, with production valued at more than €5 and €4 billion, respectively) [3].

A sensor–based methodology to differentiate pure and mixed White Tequilas based on fused infrared spectra and multivariate data treatment

Christian Hazael Pérez–Beltrán [1], Víctor M. Zúñiga–Arroyo [2], José M. Andrade [3,*], Luis Cuadros–Rodríguez [1], Guadalupe Pérez–Caballero [2] and Ana M. Jiménez–Carvelo [1]

[1] *Department of Analytical Chemistry, Faculty of Science, University of Granada, 18071 Granada, Spain; christianpb@correo.ugr.es (C.H.P.-B.); lcuadros@ugr.es (L.C.-R.); amariajc@ugr.es (A.M.J.-C.)*

[2] *Laboratorio de Fisicoquímica Analítica y Especiación Química, Unidad de Investigación Multidisciplinaria, Facultad de Estudios Superiores Cuautitlán, Campo 4, Universidad Nacional Autónoma de México, Cuautitlán-Izcalli 54714, Mexico; victor_0679_zvx@comunidad.unam.mx (V.M.Z.-A.); perezcg@unam.mx (G.P.-C.)*

[3] *Group of Applied Analytical Chemistry, Campus da Zapateira s/n, University of A Coruña, 15071 A Coruña, Spain*

***** *Correspondence: jose.manuel.andrade@udc.es; Fax: +34-981-167065*

Keywords

White Tequila; authentication; IR spectroscopy; chemometrics; data fusion; beverage alcoholic industry

1. Introduction

Moderate and responsible consumption of spirits and wine is associated commonly with social relationships during leisure times and relaxing activities. People like tasting good products and experiencing new flavors and are prone to accept slightly higher expenses than usual. However, problems may arise when trust is broken by illicit practices like counterfeiting. At best 'only' economical losses are suffered, at worst health problems may arise.

Drinking a good spirit or wine is a hugely common social activity in many countries and thus an honest, historical, labor-demanding, and profitable industrial sector is devoted to those goods. Worldwide, the leading spirits markets in the 2010–2020 period according to value sales were (decreasing order) USA, Japan, Russia, China, UK, France, Germany, Mexico, Spain, and Canada [**1**]. According to the last European survey and statistics [**2**], during the period 2008-2013 EU annual production of spirits amounted to €22 billion and imports from third countries amounted to €1 billion (UK and France being the largest producers of spirits, with production valued at more than €5 and €4 billion, respectively) [**3**]. More recent figures indicate a European production of 25 million hectoliters of spirits and about €37 billion sales [**4**].

With such big figures in place, adulteration of alcoholic beverages has become a very important practice worldwide and a health-threating problem because methanol (a possible contaminant of ethanol) is very harmful for human health. A recent study presented by the EU Intellectual Property Office (EUIPO) compiled previous unreported studies in which around 5.3% of sales of spirits and wine correspond to counterfeited products (ca. €2.3 billion), which causes ca. €5.2 billion in total lost sales and affects around 31,850 lost employments [**5**]. Unfortunately, these figures are rising fast [**6**,**7**] as the €1.3 billion lost sales in 2016 demonstrated [**7**].

One of the most EU-demanded imported spirits is Mexican Tequila, which agrees with a worldwide rise in its demand during the last years and with ca. 3.6% increments until 2020 [**8**]. In 2020, a total of 345.2 million liters of Tequila (207.2 million '100% agave' and 137.9 million 'mixed') were produced within the denomination of origin of which 76.6% were exported [**7**].

Tequila is the spirit obtained after the double distillation of milled and cooked pinecones or stems of Mexican *Agave Tequilana Weber Blue* variety [**9**]. Tequila's denomination of origin was first established in 1974 [**10**] and recognized by the EU in 1997 [**11**]. The Tequila's Regulatory Council (in brief, CRT, Consejo Regulador del Tequila) was created in 1993 to certify the quality and authenticity of Tequila and to avoid its undue imitation outside Mexico. By 2017, five countries, including the EU, imported ca. 95% of exported Tequila (the US being by far the main importer).

In Mexico alone, Tequila distillers estimated that ca. 40-50% of Tequila sales were adulterated, yielding an economic loss of about US $0.55 billion (equivalent to 60 million liters) [**12**]. Various types of counterfeits are common, such as the addition of water, alcohol, colorants, and aromas to the original beverages, even under poor hygienic conditions, leading to potential risks associated to their consumption. In particular, the most frequent fraud is the increased amount of methanol in the beverage, well above the norm [**13**]. To complicate things further, it is almost impossible for the final consumer to visually identify counterfeited Tequilas because they are sold as 'original Tequila', sometimes even in bottles and boxes that are very good copies of the original ones.

White Tequila is the non–matured (in wood barrels) spirit of blue agave. It is also called 'Silver' Tequila and it is likely the most common type of Tequila sold worldwide. If the spirit is matured, it leads to Rested and Aged Tequila. A common form of counterfeiting White Tequila consists of an undue declaration at the label, stating that it is '100 % from blue agave' when, in fact, it was obtained from *Agave Tequilana Weber Blue* variety and other sources of fermentable sugars (sugar cane, other agaves, etc.). Please, note that the second possibility is legal as far as only 49% of the fermentable sugars employed to get Tequila proceed from sources different from blue agave, which should be called 'mixed Tequila' or simply 'Tequila' and its denomination must be declared on the label according to CRT guidelines. Despite health might not be affected in this case, the denomination of origin and/or the label declaration is violated yielding deceptive products for consumers (including economic deception) and leading to economic losses to distilleries.

In previous works, reviews were made to ascertain the analytical techniques most commonly applied to authenticate Tequilas and it was concluded that from a pragmatic point of view usual infrared (IR) and ultraviolet-visible (UV-Vis) spectrometric techniques would be favored for routine quality control [**14**,**15**]. In fact, a study was presented in order to show that common UV-Vis spectrometry can be a powerful tool to assess different qualities of Tequila (with significantly different prices: White, Gold, Rested, and Aged Tequilas) and that they can be easily differentiated from Mezcal, another Mexican spirit [**16**].

The use of these spectrometric nondestructive techniques is usually potentiated by hybridizing them with multivariate chemical-data analysis (MVA) algorithms (in brief, chemometrics). That may lead to off-line or in-line sensors of potential interest for quality control and authentication analyses in the modern alcoholic beverage industry and related organizations. In addition, they are powerful, potential tools to evaluate the denominations of origin and to perform geographical discrimination of products. In addition, chemometric methods play an important role in the recent popularity of analytical screening methodologies. The essential objective of chemometrics is to extract useful information and get meaningful conclusions through the most appropriate data treatment, modeling, and characterization of samples, and the study of related variables [**17**–**20**].

Data fusion consists of combining results of different measurement techniques in hopes that they all can lead to more powerful and reliable authentication models, and to a better understanding of the chemistry behind the samples [**21**–**23**]. Data fusion can be performed at low, mid, and high levels. The former consists simply of concatenating data from all sources into a single 'raw' data matrix. The mid-level approach first extracts separately important information from each data set and then concatenates it into one matrix. Finally, the high-level approach concatenates the individual results derived from classification or regression models to obtain a final data matrix [**23**]. Despite several researches aimed at studying different properties of Tequila with chemometrics [**14**,**24**,**25**], up to date, to the best of the authors' knowledge, this is the first work which considers data fusion to handle spectroscopic data from Tequila.

The aim of this paper is to develop a sensor-based methodology that combines IR spectrometry, chemometric tools, and data fusion to address a complex problem: to evaluate whether White Tequila is 100% from Blue Agave or whether it comes from a mixture of blue agave plus other sources of sugars (recall that to be considered Tequila, at least 51% of them should be from *Agave Tequilana Weber Blue* variety). This is an insidious problem because the physicochemical and organoleptic properties of both categories of White Tequila are very similar (but for gross adulterations) and despite health might not be affected, deceptive products are released to consumers.

2. Materials and Methods

2.1. Samples

A total of sixty-five samples of White Tequila from different Mexican brands were used in this study. Thirty-six samples of White Tequila were of the '100% agave' category (TB, from Spanish 'Tequila Blanco') and twenty-nine were of the 'mixed' category (TBM, from Spanish 'Tequila Blanco Mixto'). Samples were obtained in specialized Mexican liquor stores from well-known Tequila producers whose labels displayed the corresponding CRT quality seals, some bottles were provided by CRT itself. This collection was selected to include different factories, brands, raw materials, processing units, etc. Those details cannot be disclosed due to confidentiality agreements.

2.2. Instrumentation

A PerkinElmer Frontier FTIR Spectrometer, equipped with a single-reflection diamond PerkinElmer UATR device, was employed. Samples were withdrawn from the bottles and poured (50 μL) directly on the surface of the UATR to perform the measurements and capped with a lid. Then, the crystal was gently dried with cotton, washed with ethanol and distilled water, and dried with lens cleaning tissues. MIR (medium-range infrared) spectra were obtained in the 4000-400 cm^{-1} range, using 16 scans, with a nominal resolution of 4 cm^{-1}, and background corrected (a background per sample). They were transformed to absorbance, ATR corrected (using the built-in PerkinElmer proprietary function) to take

account of radiation penetration and baseline corrected; finally, they were digitized and exported (1 datum per cm^{-1}).

Baseline correction was made setting the correction points manually for each type of Tequila before performing the MVA. The spectra of all Tequilas showed the same spectral bands, with minor differences in the spectral band centered around 1045 cm^{-1}. In particular, the band was a bit broader for '100% blue agave' Tequilas as it extended to 954.5 cm^{-1}, whereas the band for the mixed Tequilas finished at 957.8 cm^{-1}.

A practical problem arises here when the methodology is to be implemented in routine quality control laboratories because although selecting the baseline correction is simple for the authenticated samples used in modeling (e.g., a set of wavenumbers per category of White Tequila) this is not the case for real samples, for which the class is unknown (or suspicious) and so selecting the "correct" baseline is not possible. A possible solution is to base the models on fused data gathered from different baseline corrections, as studied in next section.

2.3. Multivariate Analysis and Data Fusion

The spectral data treatments, digitation, and exportation were made using the PerkinElmer proprietary software, Spectrum. Unsupervised models used to visually explore data –such as principal component analysis (PCA)– and supervised models employed for classification purposes –like partial least squares (PLS), soft independent modeling for class analogy (SIMCA), and partial least squares discriminant analysis (PLS-DA)– were built using PLS_Toolbox 8.6.1 (Eigenvector Research, Manson,WA USA) for MATLAB environment (MathWorks Inc., Massachusetts, USA, versions R2015a and R2017b). The classification results of each model were evaluated with several quality performance metrics (QPM), mainly: (i) sensitivity (SENS), (ii) specificity (SPEC), (iii) positive predictive value (PPV) or precision, and (iv) negative predictive value (NPV). Twenty-two complementary QPMs for each classification model –calculated from the four main QPMs previously mentioned– are also presented. However, detailed explanations for each of them are out of the scope of this work. The reader is kindly referred to the literature cited herein where detailed information can be found [**26**,**27**].

After preliminary studies focused on detecting outliers and searching for the spectral regions less affected by the spectral bands of the OH group (water plus alcohols), the spectral working region was established at the fingerprint region, i.e., 1800-451 cm^{-1}.

At the same time, the baseline correction had to be considered in a pragmatic way. As samples to be investigated might be in reality of any of the two categories (classes), and so without a sound reason to select the baseline correction of the 'mixed' or of the '100% agave' Tequilas, a way to consider this into the models is to develop them with two baseline corrections:

1. A unique correction for all kinds of White Tequilas (4000, 957, and 450 cm^{-1}); and
2. A baseline correction specific for each class (4000, 1854, 954.5, and 450 cm^{-1} for Tequilas '100% agave'; and 4000, 1854, 957.8, and 450 cm^{-1} for the mixed ones).

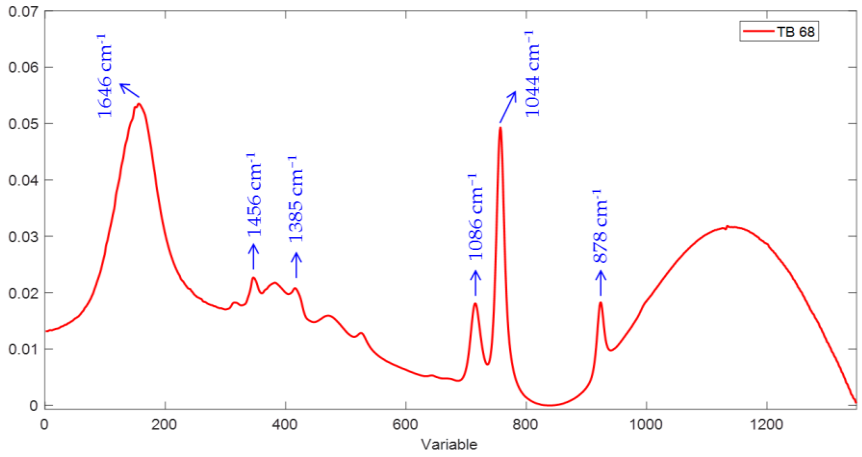
For simplicity, let us call blocks (or matrices) **A** and **B** the options in points 1 and 2, respectively. Hence, the low-level data fusion [**21**] consisted of concatenating block **B** to block **A** after its 1350th variable, forming a new fused-data matrix [**A B**], of the size 65 × 2700 (samples × variables). In this way, the suspicious sample will be corrected at the common points and at the '100% agave' ones (it will be absolutely rare to sell '100% agave' Tequila as 'mixed').

The training set was made up of fifty samples randomly selected, twenty-seven for TB class and twenty-three for TBM class. The fused-data matrix [**A B**] was auto-scaled and used for the exploratory analysis with PCA and PLS, and for the authentication analysis with SIMCA and PLS-DA. Venetian blinds was employed as cross-validation technique, using ten PCs (PCA and SIMCA) or LVs (PLS and PLS-DA), ten splits, and one sample per split.

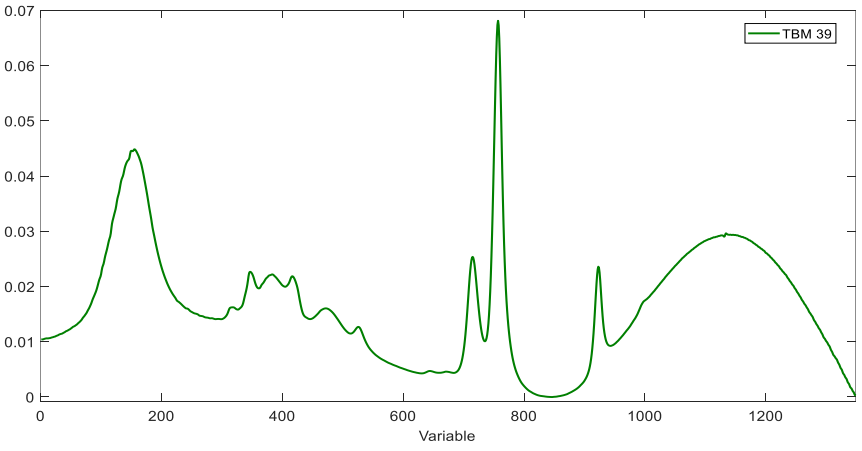
Different classification models were studied, each of them with different setups, and once the most satisfactory model for each chemometric approach was selected, external validation was performed using a validation set composed of 15 samples, 9 for TB and 6 for TBM classes, which was auto-scaled using the parameters derived from the training set.

3. Results and Discussion

The IR fingerprints of two representative samples of White Tequila –TB ('100% agave') and TBM ('mixed')– are displayed in Figure 1. Strong molecular vibrations are observed in this characteristic region of White Tequila. In essence, the pronounced peak from variable #155 (1646 cm^{-1}) is assigned to the bending vibration of OH in water and to the O–H deformation, which is important in the identification of alcohol compounds.



(**a**)



(**b**)

Figure 1. IR spectra of two samples of White Tequila. (a) TB: '100% agave' Tequila and (b) TBM: 'mixed' Tequila.

The small peak at variable #345 (1456 cm^{-1}) corresponds to C–OH bending deformation. The small band at variable #416 (1385 cm^{-1}) may be due to the stretching vibrations of C–O of tartaric acid whilst the band at variable #715 (1086 cm^{-1}) may be due to the stretching vibration of C–O of sucrose (C–O stretching of secondary alcohols). These two latter bands overlap with the C–H stretching of CH_2 and CH_3 moieties. The sharp peak represented by variable #757 (1044 cm^{-1}) is attributed to the stretching vibration of the C–O bond in primary alcohols, and the small peak at variable #923 (878 cm^{-1}) is assigned to some meta–disubstituted aromatic compounds in White Tequila [**28-30**].

The fused-data matrix is shown in Figure 2. It is worth noting that independent data analysis of blocks **A** and **B** showed poor results, but when data were fused, the exploratory analysis and the authentication analyses improved notably (i.e., the number of misclassifications decreased).

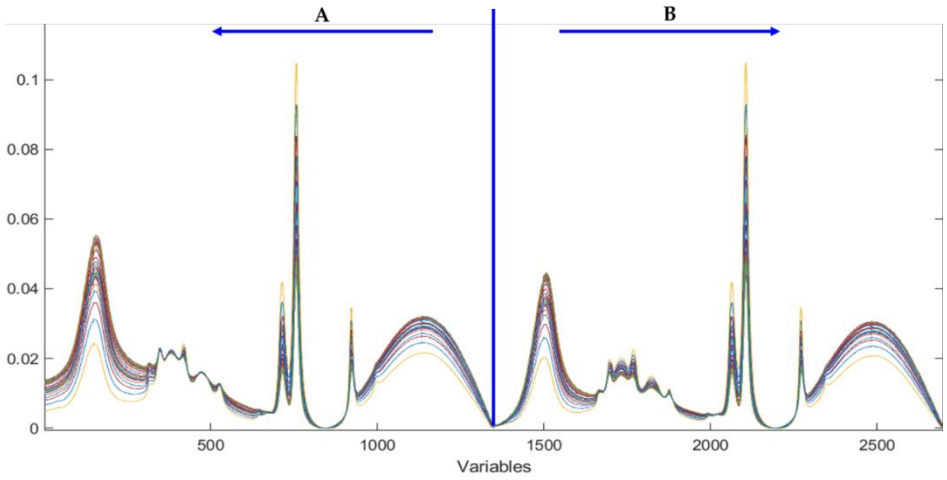


Figure 2. IR spectroscopic fused data of White Tequilas. Variables #1-1350 represent block A in which the base line was corrected in the same points for both TB and TBM classes. Variables #1351-2700 represent block B in which the baseline was corrected in different points according to each class (see section 2.3 for details).

3.1. Exploratory Analysis

The main goal of exploratory data analysis is to visually identify possible groups of samples [**31**], being the most common technique for this PCA. Despite the typical predictive use of PLS, it can also be used for exploration purposes since its algorithm pursuits the calculation of factors that capture both variance and correlation [**32**], thus resulting more powerful than PCA.

For the [**A B**] fused dataset, the PCA model required three principal components (PCs), which were selected according to the explained variance (97.8%) and visualization of the information of interest (results in this respect where very poor); whilst for the PLS model, three latent variables (LVs) were considered, which explained 97.3% of the variance, with root mean square errors for calibration (RMSEC) and cross-validation (RMSECV) of 0.50 and 0.55, respectively (they are very similar, so overfitting is not expected). Note that PC1 explained ca. 92.0% of the information but it is not displayed here because it 'merely' ordered the samples according to the percentage of ethanol they had. However, no useful groups of samples could be visualized. Indeed, the only vision that allowed some differentiation (very poor) is presented in Figure 3. If different color codes or symbols would not have been used no groups would have been guessed. On the contrary, PLS displays quite clear groups, with the two categories easily differentiated (Figure 4).

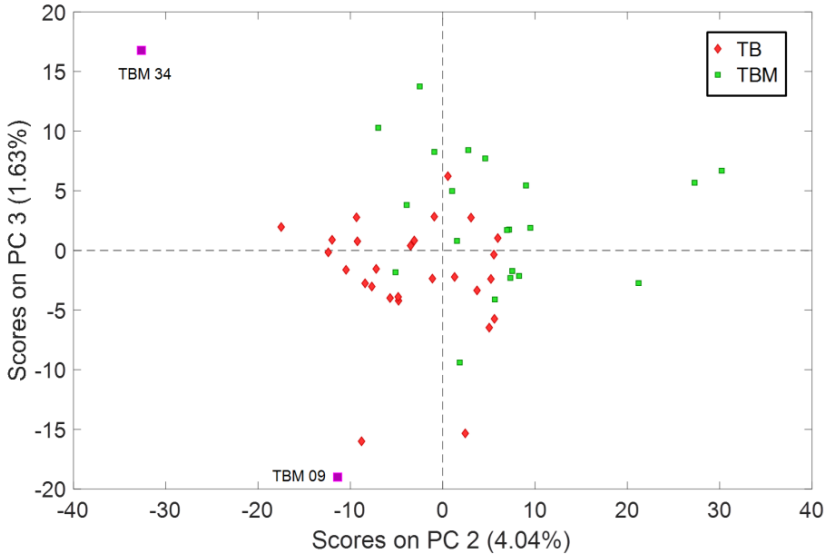


Figure 3. Exploratory analysis – PCA score plots of White Tequilas fused data (TB: '100% agave, and TBM: 'mixed').

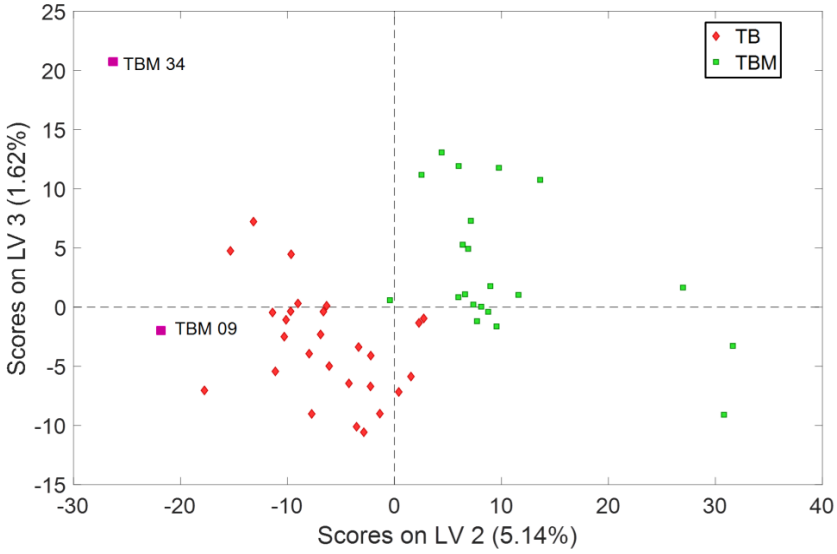


Figure 4. Exploratory analysis – PLS score plots of White Tequilas fused data (TB: '100% agave', and TBM: 'mixed').

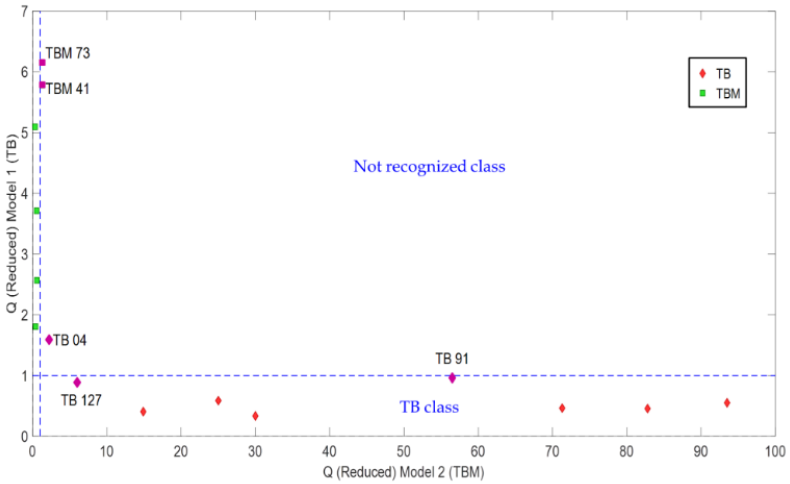
Additionally, samples TBM 09 and TBM 34 were misplaced in the score plots, but only TBM 34 was considered as an outlier since it affected the classification models and, thus, it was excluded from the models. The reason for such different behavior among the two techniques is, precisely, what makes PLS so powerful: the extraction of spectral information maximally related to the property of interest, while PCA only decomposes the overall variance in its major components (regardless of being useful to see groups of samples).

3.2. Authentication Analysis

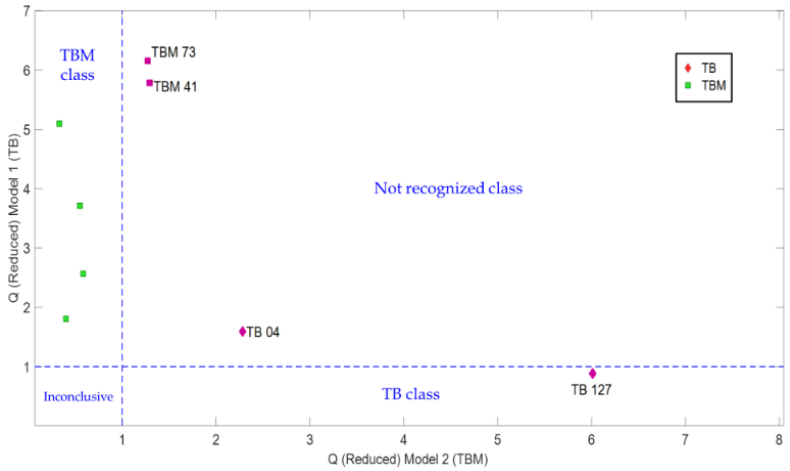
After data exploration, SIMCA and PLS-DA models were employed to authenticate White Tequila considering the '100% agave' category as target class. The results are summarized below.

3.2.1. SIMCA

This chemometric method builds a classification model with classes independently modelled (hence, its name) from the training set. First, an individual PC model is built per class. In this case, since previous PCA exploratory analysis suggested three PCs, they were studied and kept for both classes. They explained 98.6% and 98.1% of the variance for TB and TBM, respectively. Then, spatial regions of probability around each Tequila class are constructed considering the average distances of the samples to the PCs subspace and a critical distance is calculated by means of an F statistic [**32**].



(a)



(b)

Figure 5. Coomans' plot for SIMCA validation samples. (a) Complete Cooman's plot and (b) magnified view of the bottom, left–hand part of (a).

The classification results can be evaluated with the Coomans' plot (Figure 5). Samples coded as TB 04, TBM 41 and TBM 73 are placed in the upper–right quadrant, which are considered as not recognized and are not classified in any of the two classes. Moreover, samples TB 91 and TB 127 are correctly classified, but close to the boundary of the not recognized quadrant, which may indicate that these samples could present some authenticity (i.e., composition) problem.

Table 1 presents the contingency results that were used to calculate the QPM figures (Table 2). The sensitivity (SENS) also known as 'recall' or 'true positive rate' indicates the ratio of agreement of TB class, and specificity (SPEC) or 'true negative rate' shows the ratio of agreements of TBM class. For the SIMCA model, SENS and SPEC were 0.89 and 0.67, respectively. Though these values can be considered low, the result for the predictive capability (both positive predictive value (PPV) or precision, and negative predictive value (NPV)) is 1.00 for each of them. These results indicate that the model is good in assigning samples to classes, since precision represents the proportion of target samples correctly allocated in TB class and NPV the proportion of target samples correctly assigned to TBM in relation to the total samples assigned to those classes, respectively.

Table 1. Classification of the validation samples using SIMCA.

		9	6	15
Assignation	**Not recognized (Nr)**	1	2	**3**
	Inconclusive (I)	0	0	**0**
	TBM	0	4 (66.6%)	**4**
	TB	8 (88.8%)	0	**8**
		TB	**TBM**	
		Actual		

TB: target class (White Tequila '100% agave'); TBM: non-target class (White Tequila 'mixed').

Table 2. Main quality performance metrics (QPM) for PLS-DA and SIMCA models.

Quality Metrics	PLS–DA	SIMCA
Sensitivity (SENS)	1.00	0.89
Specificity (SPEC)	1.00	0.67
Positive predictive value (precision) (PPV)	1.00	1.00
Negative predictive value (NPV)	1.00	1.00

3.2.2. PLS-DA

PLS-DA is a latent variable-based method that, first, builds a PLS regression model using a set of latent variables (LV) that are employed to establish limits for the classes and, then, it carries out a discriminant analysis (DA) to classify the samples into a particular class. The boundaries among the different classes are calculated using the Tequilas training set only.

PLS-DA model from the IR fused data. For this, '100% agave' and 'mixed' Tequilas were coded as 1 and 0, respectively. A threshold value of 0.5 was established as a decision criterion for the classification of the samples. In this sense, samples with scores above the decision criterion (values around one) were classified as '100% agave' whereas samples with scores below the decision criterion (values around zero) were considered as 'mixed'. Moreover, an uncertainty region of plus/minus 0.1 the threshold value was established to improve the reliability of the validation results.

Figure 6 displays the PLS-DA classification plot. The solid line represents the threshold value whereas the dotted lines indicate the uncertainty region, where the sample is labeled as inconclusive. For the current model, all samples were correctly classified in their corresponding categories in validation. However, validation samples TB 04, TB 127, and TBM 54 are close to the uncertainty region.

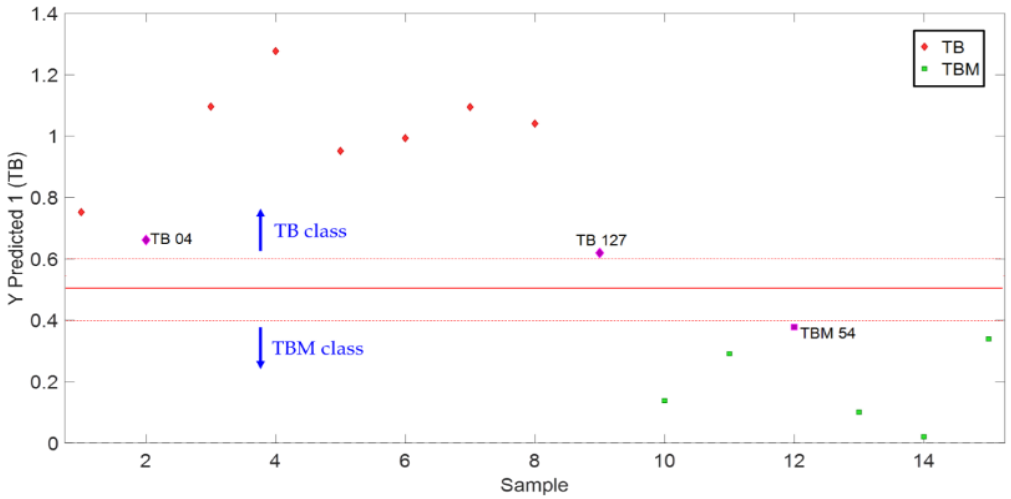


Figure 6. Classification plot for PLS–discriminant analysis (DA) validation results. The solid red line represents the threshold decision value (0.5) and the dotted red lines the uncertainty region.

In fact, special attention must be given to samples TB 04 and TB 127 because slightly problematic results were obtained for them also with SIMCA, where sample TB 04 was not recognized by the model and TB 127 was allocated near the limit of 'not recognized' quadrant.The contingency results for validation are shown in Table 3, which were employed to obtain the main four QPMs (Table 2). According to these results, it becomes clear that the PLS-DA model is able to correctly classify new samples in the corresponding class with 100% confidence.

Table 3. Classification of the validation samples using PLS-DA.

		9	6	15
Assignation	**Inconclusive (I)**	0	0	**0**
	TBM	0	6 (100%)	**4**
	TB	9 (100%)	0	**8**
		TB	**TBM**	
		Actual		

TB: target class (White Tequila '100% agave'); TBM: non-target class (White Tequila 'mixed').

In addition, twenty-two other QPMs reported in literature were calculated for a more comprehensive evaluation of the models (Table 4). The conclusions derived from the QPMs results shown in Table 4 are in agreement with the typical four main QPMs (Table 3). They all indicate that the PLS-DA model is capable to classify unknowns better than SIMCA since the parameters that report good quality of the model and/or good probability of classification (i.e., SENS, SPEC, PPV, NPV, YOUD, EFFIC, AUC, MCC, and Bayes' conditional probabilities 1/1 and 2/2) are equal or close to one. For instance, YOUD indicates that PLS-DA model is more reliable for classifying both classes than SIMCA model; EFFIC shows the ratio between agreements (i.e., correct assignations) and the total number of objects, which is better for PLS-DA than for SIMCA; and AUC demonstrates that PLS-DA is better avoiding errors during classification than SIMCA.

Table 4. Quality performance metrics (QPM) summary chart of the selected models.

Quality Metrics	PLS-DA		SIMCA	
	TB	TBM	TB	TBM
Inconclusive rate (IR)	0.00	0.00	0.00	0.00
Sensitivity (SENS)	1.00	1.00	0.89	0.67
Specificity (SPEC)	1.00	1.00	0.67	0.89
False positive rate (FPR)	0.00	0.00	0.33	0.11
False negative rate (FNR)	0.00	0.00	0.11	0.33
Positive predictive value (precision) (PPV)	1.00	1.00	1.00	1.00
Negative predictive value (NPV)	1.00	1.00	1.00	1.00
Youden index (YOUD)	1.00	1.00	0.56	0.56
Positive likelihood rate (LR (+))	-	-	2.67	6.00
Negative likelihood rate (LR (−))	0.00	0.00	0.17	0.38
Classification odds ratio (COR)	-	-	16.00	16.00
F-measure (F)	1.00	1.00	0.94	0.80
Discriminant power (DP)	-	-	0.66	0.66
Efficiency (or accuracy) (EFFIC)	1.00	1.00	0.80	0.80
Misclassification rate (MR)	0.00	0.00	0.20	0.20
AUC (Area under the receiver operating curve)	1.00	1.00	0.78	0.78
Gini coefficient (Gini)	1.00	1.00	0.56	0.56
G-mean (GM)	1.00	1.00	0.77	0.77
Matthews correlation coefficient (MCC)	1.00	1.00	0.77	0.77
Change agreement rate (CAR)	0.52	0.52	0.43	0.43
Change error rate (CER)	0.48	0.48	0.48	0.48
Kappa coefficient (KAPPA)	1.00	1.00	0.65	0.65
PROB (1/1)	1.00	1.00	0.80	0.80
PROB (2/2)	1.00	1.00	0.80	0.80
PROB (1/2)	0.00	0.00	0.20	0.20
PROB (2/1)	0.00	0.00	0.20	0.20

The hyphen "-" indicates that the performance feature cannot be determined since involves a division between zero.

Moreover, the parameters that indicate bad quality/probability of misclassification (i.e., FPR, FNR, MR, and Bayes' conditional probabilities 1/2 and 2/1) are low or equal to zero. In this sense, MR displays the ratio between errors (i.e., wrong assignations) and the total number of objects, which is lower for PLS-DA than for SIMCA. PROB (1/2) refers to the probability of an object being assigned to TBM class when it belongs to TB, which is better for PLS-DA than for SIMCA; and analogously for PROB (2/1).

Furthermore, SIMCA was not able to classify three samples in any class (TB 04, TBM 41, and TBM 73), considering them as not recognized. All QPMs for PLS-DA were equal to 1.00 with no inconclusive samples. Every Tequila sample of the '100% agave' (TB) category was correctly classified (probability = 1), as well as the 'mixed' (TBM) ones (probability = 0).

Noteworthy, no 'mixed' Tequila samples were included in the '100% agave' category, which is a very positive and relevant result. Additional studies in which particularly conflictive products (sold as low-price Tequilas in urban and rural marketplaces) and potential fraudulent samples are considered, are under development.

4. Conclusions

The current research presented the development of an analytical methodology based on the combination of a common but powerful screening sensor, an FTIR instrument, and multivariate classification tools to authenticate two commercial categories of White Tequila ('100% agave' and 'mixed'). The mid-IR spectra of a broad collection of commercial samples were collected and baseline corrected according to their classes. However, this posed a problem when studying truly unknown samples because it is not obvious how to baseline correct their spectra.

Therefore, a low-level data fusion approach was considered (to the best of our knowledge, this is the first study which involves data fusion to authenticate Tequilas). In this context, the solution was to develop multivariate models based on fused data gathered from diverse baseline corrections since differences between both classes are taken into consideration. This avoided subjective decisions when trying to select the 'correct' baseline correction for unknowns. In this sense, the results presented here demonstrated that multivariate discrimination models addressed the authentication of the two Tequila categories. Noteworthy, no 'mixed' Tequila sample was included in the '100% agave' category, which is a very positive and relevant result. The model finally selected was even able to cope with two complex samples declared as '100% agave' which were not obviously classified as such.

This sensor-based methodology is fast, cheap, non-destructive, and it allows for potential on-line/at-line implementations, and it can be straightforwardly implemented in routine quality control laboratories. The use of data fusion approaches to develop the multivariate authentication models for Tequila is highly recommended not only for the industry but for regulatory organizations.

Author Contributions: Conceptualization and methodology: all authors; formal analysis: V.M.Z.-A. and C.H.P.-B.; validation: G.P.-C., L.C.-R., J.M.A., and A.M.J.-C.; investigation: all authors; resources: G.P.-C.; data curation: all authors; writing—original draft preparation: C.H.P.-B., L.C.-R., and A.M.J.-C.; writing—review and editing: all authors; funding acquisition: G.P.-C. All authors have read and agreed to the published version of the manuscript.

Funding: G.P.-C. acknowledges a research grant from UNAM, Universidad Nacional Autónoma de México (Grants: PIAPI 2042 and PAPIIT IT200918).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: C.H.P.-B. acknowledges Universidad Autónoma de Sinaloa (México) for a PhD scholarship and further support.

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

References

1. Statista. Available online: https://www.statista.com/statistics/496130/leading-spirits-markets-worldwide-based-on-valuesales/ (accessed on 8 January 2021).
2. Eurostats, Manufacture of Beverages Statistics-NACE Rev. 2, 2013. Archive: Manufacture_of_beverages_statistics_-_NACE_Rev._2. Available online: https://ec.europa.eu/eurostat/statistics-explained/ (accessed on 8 January 2021).
3. Wajsman, N.; Arias Burgos, C.; Davies, C. (European Union Intellectual Property Office) The Economic Cost of IPR Infringement in Spirits and Wine; European Union Intellectual Property Office: Alicante, Spain, 2016.
4. Spirits Europe. Available online: https://spirits.eu/issues/internal-market/introduction-3 (accessed on 9 January 2021).
5. *2020 Status Report on IPR Infringement, Why IP Rights are Important, IPR Infringement, and the Fight against Counterfeiting and Piracy*; European Union Intellectual Property Office: Alicante, Spain, 2020. [CrossRef]
6. Trends in Trade in Counterfeit and Pirated Goods, Illicit Trade, OECD Publishing, Paris/European Union Intellectual Property Office- OECD/EUIPO (2019). Available online: https://doi.org/10.1787/g2g9f533-en (accessed on 9 January 2021).
7. Consejo Regulador del Tequila. Available online: https://www.crt.org.mx/EstadisticasCRTweb/ (accessed on 10 January 2021).
8. Drinks International. Available online: http://drinksint.com/news/fullstory.php/aid/6772/DI_Annual_Bar_Report:_Tequila_html (accessed on 10 January 2021).
9. Norma Oficial Mexicana NOM-006-SCFI-2012, Bebidas alcohólicas-Tequila-Especificaciones. Mexican National Official Bulletin (Mexican Governent). Available online: http://www.dof.gob.mx/nota_detalle.php?codigo=5282165&fecha=13/12/2012 (accessed on 26 February 2021).

10. Secretaría de Patrimonio y Fomento Industrial—Dirección General de Invenciones y Marcas. Declaración General de Protección a la Denominación de origen 'Tequila', Mexican National Official Bulletin (Mexican Governent). Número del oficio: 16-I.-57348 (9 December 1974). Available online: https://www.crt.org.mx/index.php/es/pages-3/declaratoria (accessed on 26 February 2021).

11. Agreement between the European Community and the United Mexican States on the mutual recognition and protection of designations for spirit drinks. European Communities. Off. J. Eur. Comm. **1997**, L152, 16–26.

12. Ruiz-Pérez, A.; Pérez-Castañeda, J.I.; Castañeda-Guzmán, R.; Pérez-Ruiz, S.J. Determination of Tequila quality by photoacoustic analysis. Int. J. Thermophys. **2013**, 34, 1695–1702.

13. Wang, L.; Sun, D.-W.; Pu, H.; Cheng, J.-H. Quality analysis, classification, and authentication of liquid foods by near-infrared spectroscopy: A review of recent research developments. Crit. Rev. Food Sci. Nutr. **2016**, 57, 1524–1538.

14. Fernández-Lozano, C.; Gestal-Pose, M.; Pérez-Caballero, G.; Revilla-Vázquez, A.L.; Andrade-Garda, J.M. Multivariate classification techniques to authenticate Mexican commercial spirits. In Quality Control in the Beverage Industry; Grumezescu, A.M., Holban, A.M., Eds.; Elsevier: Amsterdam, The Netherlands, 2019; Volume 17, Chapter 8, pp. 259–288.

15. Pérez-Caballero, G.; Andrade, J.M.; Olmos, P.; Molina, Y.; Jiménez, I.; Durán, J.J.; Fernandez-Lozano, C.; Miguel-Cruz, F. Authentication of tequilas using pattern recognition and supervised classification. Trends Anal. Chem. **2017**, 94, 117–129.

16. Andrade, J.M.; Bellabio, D.; Gómez-Carracedo, M.P.; Pérez-Caballero, G. Nonlinear classification of comercial Mexican tequilas. Chemometrics **2017**, 2939.

17. Jiménez-Carvelo, A.M.; González-Casado, A.; Bagur-González, M.G.; Cuadros-Rodríguez, L. Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity—A review. Food Res. Int. **2019**, 122, 25–39.

18. Kamiloglu, S. Authenticity and traceability in beverages. Food Chem. **2019**, 277, 12–24.

19. Jiménez-Carvelo, A.M.; Osorio, M.T.; Koidis, A.; González-Casado, A.; Cuadros-Rodríguez, L. Chemometric classification and quantification of olive oil in blends with any edible vegetable oils using IR-ATR and Raman spectroscopy. LWT–Food Sci. Technol. **2017**, 86, 174–184.

20. Jiménez-Carvelo, A.M.; Pérez-Castaño, E.; González-Casado, A.; Cuadros-Rodríguez, L. 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. **2017**, 221, 1784–1791.

21. Ríos-Reina, R.; Callejón, R.M.; Savorani, F.; Amigo, J.M.; Cocchi, M. Data fusion approaches in spectroscopic characterization and classification of PDO wine vinegars. Talanta **2019**, 560–572.

22. Buratti, S.; Malegori, C.; Benedetti, S.; Oliveri, P.; Giovanelli, G. E-nose, e-tongue and e-eye for edible olive oil characterization and shelf life assessment: A powerful data fusion approach. Talanta **2018**, 131–141.

23. Borràs, E.; Ferré, J.; Boqué, R.; Mestres, M.; Aceña, L.; Busto, O. Data fusion methodologies for food and beverage authentication and quality assessment—A review. Anal. Chim. Acta **2015**, 1–14.

24. Ceballos-Magaña, S.G.; De Pablos, F.; Jurado, J.M.; Martín, M.J.; Alcázar, A.; Muñiz-Valencia, R.; Gonzalo-Lumbreras, R.; Izquierdo-Hornillos, R. Characterisation of tequila according to their major volatile composition using multilayer perceptrón neural networks. Food Chem. **2013**, 136, 1309–1315.

25. Muñoz-Muñoz, A.C.; Pichardo-Molina, J.L.; Ramos-Ortiz, G.; Barbosa-García, O.; Maldonado, J.L.; Meneses-Nava, M.A.; Ornelas-Soto, N.E.; Escobedo, A.; López-de-Alba, P.L. Identification and quantification of furanic compounds in tequila and mezcal using spectroscopy and chemometric methods. J. Braz. Chem. Soc. **2010**, 21, 1077–1087.

26. 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.

27. Ballabio, D.; Grisoni, F.; Todeschini, R. Multivariate comparison of classification performance measures. Chemom. Intell. Lab. Syst. **2018**, 174, 33–44.

28. Anjos, O.; Santos, A.; Estevinho, L.M.; Caldeira, I. FTIR–ATR spectroscopy applied to quality control of grape-derived spirits. Food Chem. **2016**, 205, 28–35.

29. Nagarajan, R.; Gupta, A.; Mehrotra, R.; Bajaj, M.M. Quantitative analysis of alcohol, sugar, and tartaric acid in alcoholic beverages using attenuated total reflectance spectroscopy. J Autom. Methods Manag. Chem. **2006**, 2006, 45102.

30. Griffiths, P.R. Introduction to the theory and instrumentation for vibrational spectroscopy. In Handbook of Vibrational Spectroscopy; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2006; Volume 1, pp. 33–43.

31. Martín-Torres, S.; Jiménez-Carvelo, A.M.; González-Casado, A.; Cuadros-Rodríguez, L. Authentication of the geographical origin and the botanical variety of avocados using liquid chromatography fingerprinting and deep learning methods. Chemom. Intell. Lab. Syst. **2020**, 103960.

32. Wise, B.M.; Gallagher, N.B.; Bro, R.; Shaver, J.M.; Winding, W.; Koch, R.S. Chemometrics Tutorial for PLS_Toolbox and Solo; Eigenvector Research, Inc.: Wenatchee, WA, USA, 2006.

4.3. Estudio 2

Método analítico multivariable basado en espectroscopía NIR y herramientas quimiométricas para diferenciar entre Tequila Blanco '100 % agave' y 'mixto'

NOTA: El contenido de la presente investigación científica forma parte de un estudio que no se publicó, debido a que los resultados no ofrecieron ninguna mejora sobre lo inicialmente planteado, lo cual era una fusión de bajo nivel de datos del Tequila Blanco obtenidos con la técnica espectroscópica de infrarrojo medio con transformada de Fourier (FTIR) del estudio # 1 con los datos de la técnica espectroscópica de infrarrojo cercano (NIR) utilizada en este estudio #2.

Por tanto, en este apartado se describe solamente la metodología empleada y los resultados obtenidos derivados de utilizar la técnica espectroscópica NIR y herramientas quimiométricas, empleadas para la diferenciación entre las categorías '100 % agave' y 'mixto' del Tequila Blanco y para la predicción del contenido alcohólico de las muestras analizadas.

1. Materiales y Métodos

1.1. Muestras

Un total de 55 muestras de Tequila Blanco de distintas marcas fueron empleadas en este estudio, de las cuales 30 pertenecían a la categoría '100 % agave' (TB - Tequila Blanco) y 25 a la categoría 'mixto' (TBM - Tequila Blanco Mixto). Dichas muestras fueron otorgadas por el Consejo Regulador del Tequila (CRT) [**1**] de México.

1.2. Instrumentación

Se utilizó un espectrómetro Frontier IR Dual-Range System, PerkinElmer, equipado con una esfera de integración NIR IntegratIR™ de 50.8 mm (2") de diámetro, PIKE Technologies. Se analizaron 2 mL de cada muestra de Tequila Blanco, sin ningún tipo de tratamiento, los cuales se vertieron en una celda de cuarzo y fueron cubiertos con un émbolo de oro de 1.0 mm de paso óptico para evitar la dispersión de la radiación electromagnética. Las huellas instrumentales NIR se obtuvieron en el rango 10000-4000 cm^{-1}, usando 16 escaneos, con una resolución nominal de 4 cm^{-1} y correcciones de fondo entre el análisis de una muestra y otra. Las huellas instrumentales obtenidas fueron transformadas de transmitancia a absorbancia y corregidas manualmente en su línea base en los puntos 10000, 9400.67, 8868.69, 7784.51, 6006.13, 5461.28 y 4000 cm^{-1}, mediante el software Spectrum™ 10 STD, PerkinElmer. Finalmente, se digitalizaron y exportaron en formato ASC (Action Script).

1.3. Análisis de datos multivariable

Las huellas instrumentales originales NIR, formadas por 6000 variables cada una (1 dato por cm^{-1}), fueron exportadas a formato MATLAB (Mathworks, Massachusetts, USA, v. R2017b). El conjunto de datos fue divido en un subconjunto de entrenamiento conformado por 42 muestras (24 TB / 18 TBM) y un subconjunto de validación externa formado por 11 muestras (6 TB / 5 TBM). Dicha selección de muestras se realizó de manera aleatorizada en una proporción 80-20% para cada uno de los subconjuntos previamente mencionados.

Los análisis de datos multivariable fueron llevados a cabo con el software PLS_Toolbox (v. 8.6.1, 2019, Eigenvector Research In., Manson, WA, USA). Las herramientas quimiométricas empleadas para el análisis exploratorio fueron análisis de componentes

principales (PCA) y regresión parcial lineal mediante mínimos cuadrados (PLSR), mientras que para los estudios de clasificación se emplearon k-vecinos cercanos (kNN), modelado flexible e independiente por analogía de clases (SIMCA), análisis discriminante mediante regresión parcial de mínimos cuadrados (PLS-DA) y sistema de aprendizaje automático mediante vectores soporte (SVMs). Para la predicción del contenido alcohólico se utilizó PLSR y regresión mediante SVMR. Asimismo, para los análisis exploratorios se aplicaron conjuntamente los siguientes pre-procesados: filtro Whitaker y 1ª derivada; para los estudios de clasificación se seleccionó el rango 6000-4000 cm^{-1}, se normalizó (área = 1) y se realizó un centrado en la media; mientras que para la predicción del contenido alcohólico se utilizaron las 6000 variables a las cuales se les aplicó el filtro Whitaker y 1ª derivada como pre-procesado.

2. Resultados y Discusión

El Tequila contiene una gran variedad de compuestos volátiles [**2**], como alcoholes superiores, aldehídos, ácidos grasos, ésteres, compuestos azufrados, algunos compuestos fenólicos y hasta antioxidantes, los cuales le otorgan su aroma y sabor, los cuales han sido identificados a través de cromatografía de gases [**3**,**4**], cromatografía de líquidos [**5**], entre otros métodos analíticos [**6**]. No obstante, el Tequila está constituido mayoritariamente por agua y etanol en proporciones aproximadas de 60 y 40 % [**7**], respectivamente.

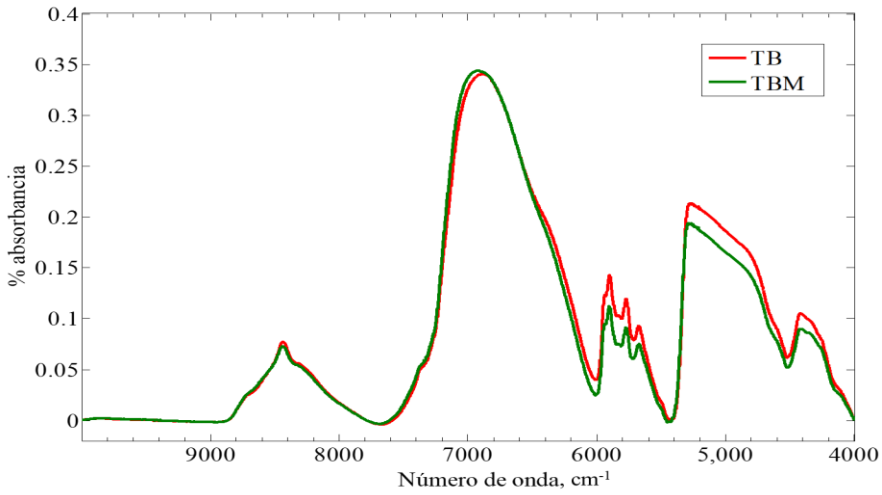


Figura 1. Espectros NIR de Tequila Blanco. TB: '100 % agave' y TBM: 'mixto'.

La presencia de ambos componentes mayoritarios se puede corroborar en la Figura 1 en donde se muestran las vibraciones características de agua y etanol de dos huellas instrumentales de Tequila Blanco, categorías TB (espectro color rojo) y TBM (espectro color verde), obtenidas mediante espectroscopía NIR.

Las intensas bandas alrededor de 7000 cm^{-1} y 5200 cm^{-1} se atribuyen a las bandas características de combinación de $\nu_1 + \nu_3$ y $\nu_1 + \nu_3$, respectivamente, del agua, las cuales están superpuestas con las bandas de combinación del etanol $2\times\nu(OH)$ y $\nu(OH) + \delta(OH)$; mientras que las regiones 6000-5500 cm^{-1} e inferior a 4500 cm^{-1} a las vibraciones $2\times\nu(CH)$ y $\nu(CH) + \delta(CH)$ del etanol [**8**,**9**,**10**].

Tal como se puede apreciar en la Figura 1, las señales instrumentales de ambas categorías de tequila son muy similares, lo que imposibilita su diferenciación y autentificación a simple vista. Es por ello, que se requiere la aplicación de herramientas quimiométricas que permitan extraer información relevante de dichas señales para lograr su adecuada autentificación, tal como se detalla en los siguientes subapartados.

2.1. Análisis exploratorio

El análisis exploratorio se llevó a cabo para estudiar el agrupamiento y/o comportamiento natural de este conjunto de datos obtenido de muestras de Tequila Blanco de categorías '100 % agave' y 'mixto', y para detectar posibles muestras anómalas. Se realizaron diversos pre-procesados, encontrando que los mejores resultados para estos análisis exploratorios eran obtenidos al aplicar conjuntamente el filtro Whitaker y 1ª derivada.

El primer análisis exploratorio fue llevado a cabo mediante PCA, el cual se construyó con 6 componentes principales (PCs, *principal components*), mismas que explicaban el 99.2 % de la varianza total. Al realizar el análisis de la PC1 *vs* PC2 y graficar sus puntuaciones (véase Figura 2 (a)), se observó que las muestras se ordenaban en el plano según su contenido alcohólico, siendo las de menor graduación alcohólica las situadas en la parte inferior y las de mayor graduación alcohólica las situadas en la parte superior. Asimismo, dicha gráfica de puntuaciones propició la detección de dos muestras anómalas (círculos marcados en magenta), las cuales presentaban una señal instrumental distinta en el intervalo 6000-5500 cm^{-1} (véase Figura 2 (b)), debido a que el análisis espectroscópico de dichas

muestras se realizó con una cantidad inferior a la requerida, por lo que dichas señales instrumentales fueron excluidas de los siguientes análisis de datos multivariable.

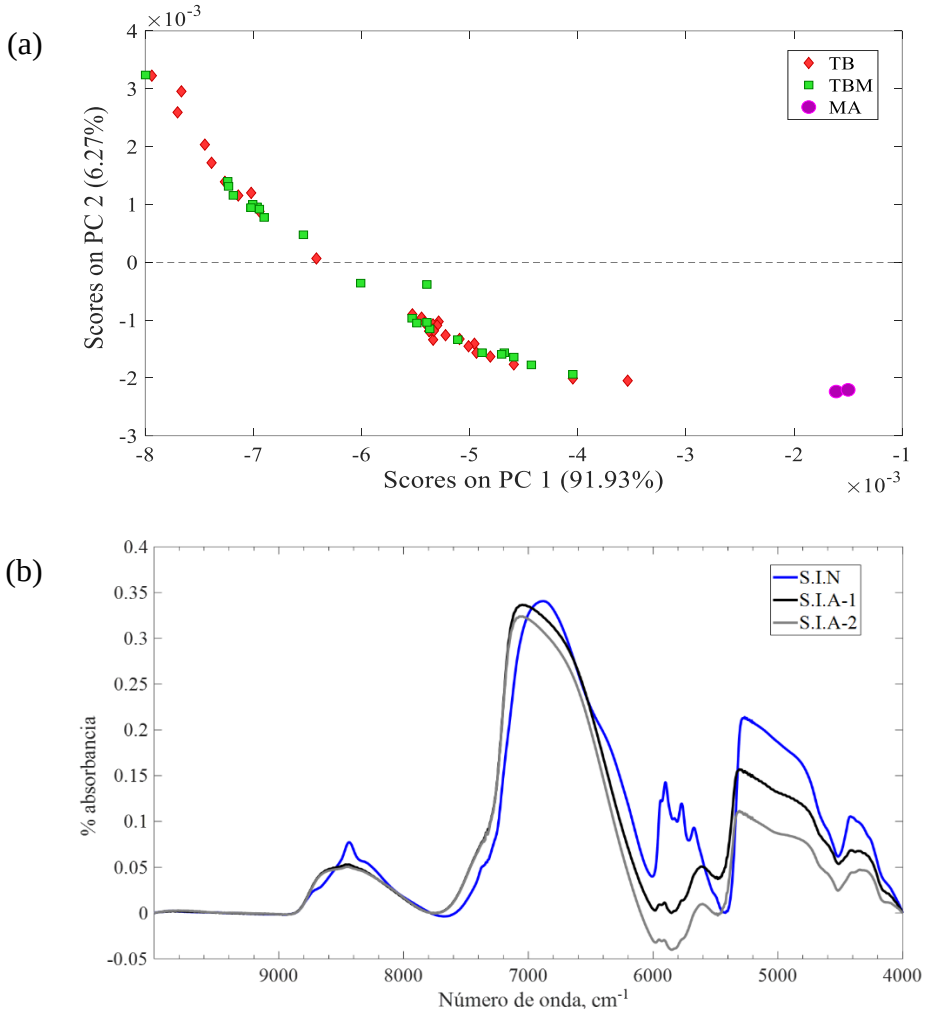


Figura 2. Análisis exploratorio a conjunto de datos de Tequila Blanco. (a) Gráfico de puntuaciones de PCA, PC1 *vs* PC2, elaborado a partir de datos de las categorías 100 % agave (TB) y mixto (TBM), en el cual se observa el comportamiento natural de las muestras y dos muestras anómalas (MA) en color magenta. (b) Señales instrumentales normales (SIN) y anómalas (SIA) de Tequila Blanco, identificadas a partir del análisis exploratorio PCA.

Posteriormente, se realizó un segundo análisis exploratorio mediante PLSR, el cual fue construido con 6 variables latentes (LVs, *latent variables*), que explicaban el 99.1 % y 92.7 % de la varianza total en los bloques X e Y, respectivamente, y cuyo gráfico de

puntuaciones (LV1 *vs* LV2) se observa en la Figura 3. En este caso y a diferencia del gráfico de puntuaciones del PCA, se pueden distinguir claramente las agrupaciones de las muestras pertenecientes a ambas categorías de tequila, lo cual demuestra la ventaja del PLSR sobre el PCA debido a su capacidad para capturar tanto la varianza y correlación entre los datos [**11**]. Una vez finalizado el análisis exploratorio, dichos datos fueron empleados en los siguientes análisis de datos multivariable.

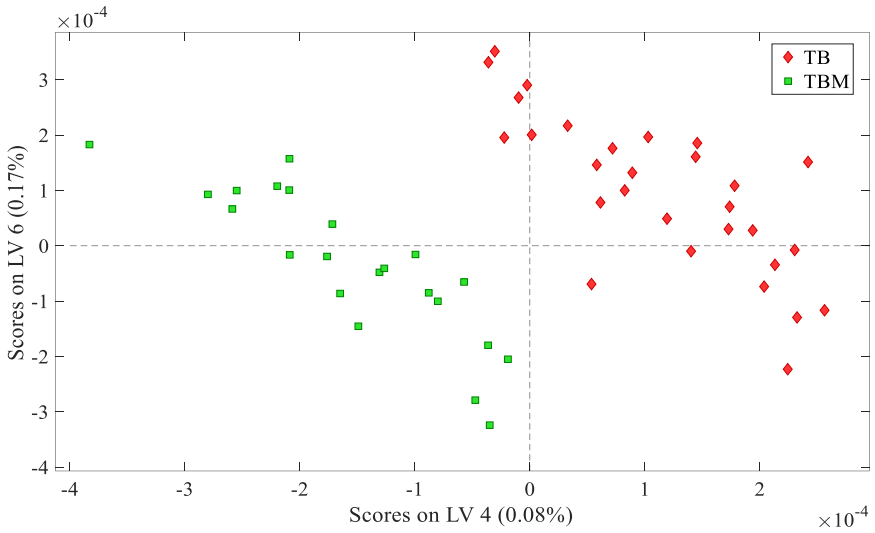


Figura 3. Gráfico de puntuaciones (LV4 *vs* LV6) de análisis exploratorio mediante PLSR a conjunto de datos de Tequila Blanco '100 % agave' (TB) y Tequila Blanco 'Mixto' (TBM).

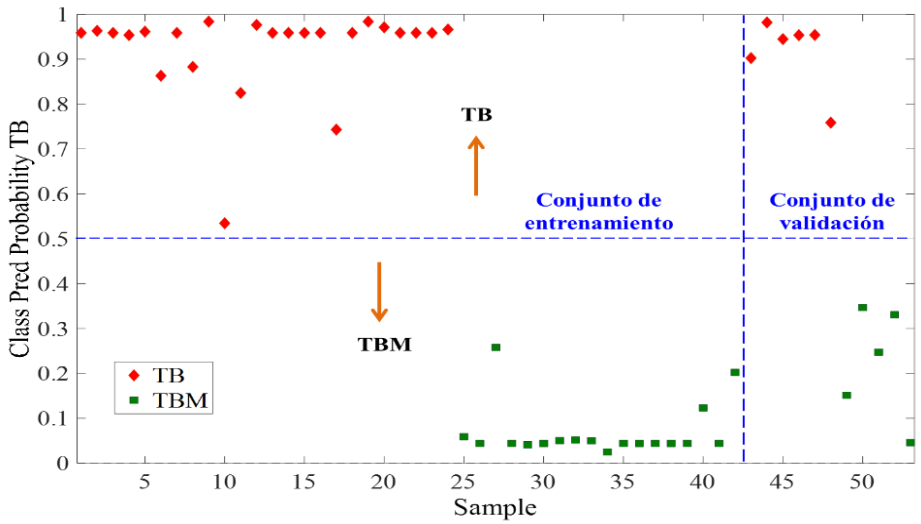
2.2. Análisis de clasificación

Para la realización de este tipo análisis se emplearon distintas herramientas quimiométricas de análisis supervisado, en las cuales se consideró a la categoría '100 % agave' (TB) como clase diana, debido a que es la categoría con mayor valor económico y, por tanto, más propensa a sufrir adulteraciones y/o falsificaciones. Después de un estudio exhaustivo de distintos pre-procesados, se seleccionaron únicamente 2000 variables (6000-4000 cm^{-1}) de las 6000 variables originales para eliminar la mayor cantidad posible de interferencia debida a las características bandas de combinación ($\nu_1 + \nu_3$) del agua. Seguidamente, dichos datos fueron normalizados (área = 1) y centrados en la media. A continuación, se describen los resultados obtenidos para cada uno de los modelos matemáticos.

❒ SVM

Para la elaboración de este modelo matemático se utilizó la función Kernel de 'base radial' con valores 'gama' y 'costo' comprendidos entre los rangos 10^{-6}-10 y 10^{-3}-10^{2}, respectivamente; así como una compresión adicional mediante PLS con 16 LVs. Se estableció un valor de 0.5 como límite de decisión para la clasificación de las muestras, donde un valor > 0.5 correspondía a las muestras pertenecientes a la clase TB, mientras que valores < 0.5 a muestras pertenecientes a la clase TBM. La gráfica de clasificación, representada por la Figura 4 (a), muestra los resultados de los conjuntos de entrenamiento y validación externa en donde todas las muestras fueron correctamente diferenciadas en sus clases correspondientes. Los resultados de este conjunto de validación se ponen de manifiesto en la tabla de contingencias de validación, expuesta a través de la Figura 4 (b).

(a)



(b)

Clase asignada	TB	TBM	T
	6	5	11
Inconcluso (I)	0	0	0
TBM	0	5 (50%)	5
TB	6 (50%)	0	6

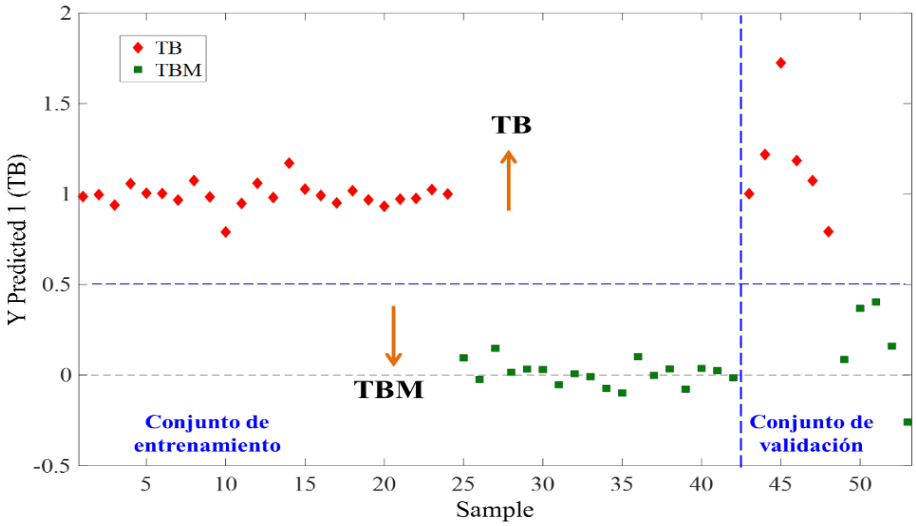
Clase Referencia

Figura 4. (a) Gráfica de clasificación y (b) tabla de contingencias de validación para el modelo de clasificación SVM. Clase diana: TB-*Tequila Blanco '100 % agave'*; Clase alternativa: TBM-*Tequila Blanco 'Mixto'*.

❒ PLS-DA

El modelo PLS-DA fue construido utilizando 18 LVs, las cuales explicaban el 100 % y 98.3 % de la varianza acumulada en los bloques X e Y, respectivamente, con un error promedio de clasificación de validación cruzada (RMSECV) de 0.604. Al igual que en el modelo SVM, se estableció un límite de 0.5 como criterio de decisión, asociando los valores > 0.5 a muestras de la clase TB y valores < 0.5 a muestras de la clase TBM, tal como se muestra en su gráfica de clasificación, representada por la Figura 5 (a). Asimismo, en ella se aprecia la correcta clasificación de todas las muestras de los conjuntos de entrenamiento y validación externa, cuyos resultados de validación se recopilan en su correspondiente tabla de contingencias en la Figura 5 (b).

(a)



(b)

Clase asignada	TB	TBM	T
	6	5	11
Inconcluso (I)	0	0	0
TBM	0	5 (50%)	5
TB	6 (50%)	0	6

Clase Referencia

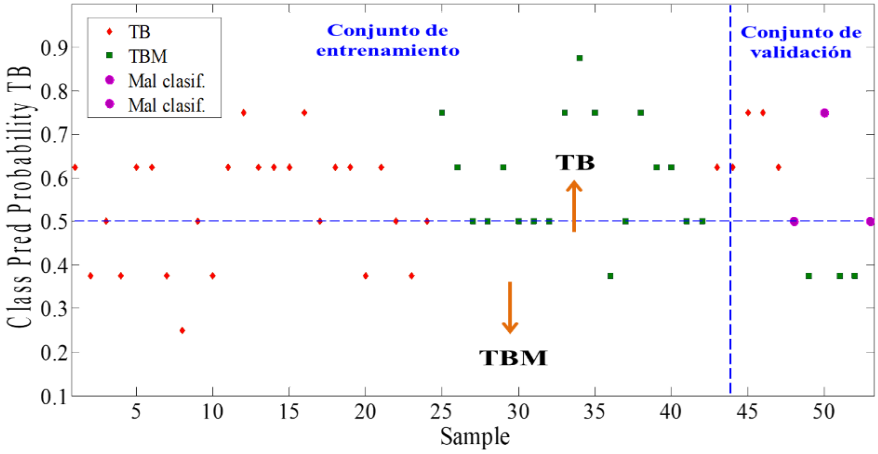
Figura 5. (a) Gráfica de clasificación y (b) tabla de contingencias de validación para el modelo de clasificación PLS-DA. Clase diana: TB-*Tequila Blanco '100 % agave'*; Clase alternativa: TBM-*Tequila Blanco 'Mixto'*.

❒ kNN

Para la construcción de este modelo matemático se consideró como número óptimo de vecino un valor de 8 (k = 8), con el cual fue posible encontrar una pequeña diferenciación entre ambas clases bajo estudio. El límite de decisión se mantuvo igual que en los modelos SVM y PLS-DA, siendo los valores > 0.5 asociados a las muestras de la clase TB y valores < 0.5 a la clase TBM, tal como se muestra en la gráfica de clasificación de la Figura 6 (a).

En este caso, el modelo matemático presentó dificultades para diferenciar las muestras del conjunto de entrenamiento, lo cual se vio reflejado en la validación del modelo con 3 muestras mal clasificadas (círculos marcados en color magenta). Los resultados que pueden apreciarse en la tabla de contingencias de validación en la Figura 6 (b).

(a)



(b)

Clase asignada	TB	TBM	T
	6	5	11
Inconcluso (I)	1 (8.33%)	1 (10%)	2
TBM	0	3 (30%)	3
TB	5 (41.67%)	1 (10%)	6
	TB	TBM	T

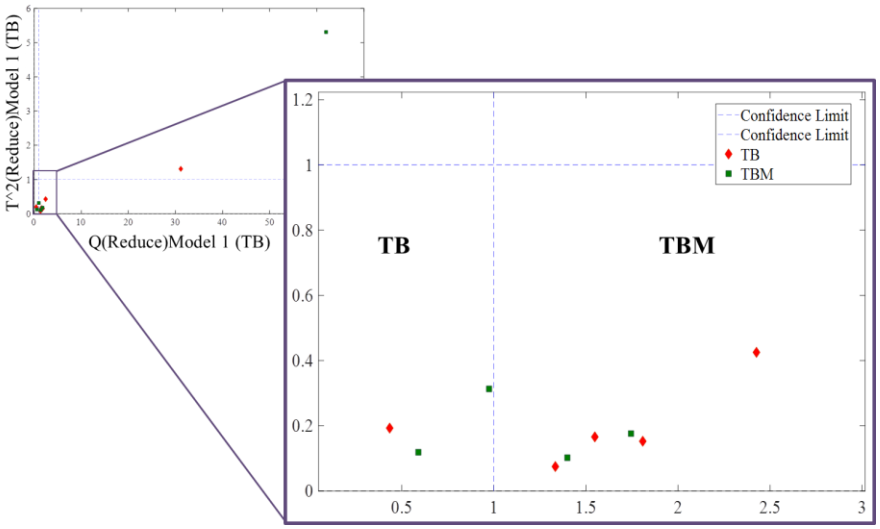
Clase Referencia

Figura 6. (a) Gráfica de clasificación y (b) tabla de contingencias de validación para el modelo de clasificación kNN. Clase diana: TB-*Tequila Blanco '100 % agave'*; Clase alternativa: TBM-*Tequila Blanco 'Mixto'*.

❒ SIMCA

Para encontrar los mejores resultados se estudiaron 2 estrategias en la construcción del modelo matemático SIMCA: (i) modelo de clasificación con dos clases de entrada (2iC-SIMCA), en el cual se utilizaron tanto la clase diana (TB) y la clase alternativa (TBM) para entrenar el modelo, y (ii) modelo de clasificación con una clase de entrada (1iC-SIMCA), en el cual sólo la clase diana fue utilizada para la etapa de entrenamiento del modelo.

(a)



(b)

Clase asignada	Clase Referencia: TB	Clase Referencia: TBM	T
	6	5	11
Inconcluso (I)	1 (8.33%)	1 (10%)	2
TBM	4 (33.33%)	2 (20%)	6
TB	1 (8.33%)	2 (20%)	3
	TB	TBM	T

Clase Referencia

Figura 7. (a) Gráfica de clasificación y (b) tabla de contingencias de validación para el modelo de clasificación 1iC-SIMCA. Clase diana: TB-*Tequila Blanco '100% agave'*; Clase alternativa: TBM-*Tequila Blanco 'Mixto'*.

A pesar del gran potencial de esta herramienta quimiométrica y de las dos estrategias estudiadas, no fue posible obtener una clara diferenciación de ambas clases de tequila, tal como se aprecia en la gráfica de clasificación de la Figura 7 (a), en donde se muestran los estadísticos T^2 de Hotelling y Q para la clase diana con un nivel de confianza del 95 %. Las muestras con valores de T^2 y $Q < 1$ son asociadas a la clase TB, mientras que las muestras con valores de T^2 y $Q > 1$ a la clase TBM. Los resultados para el conjunto de validación se pueden apreciar en la tabla de contingencias en la Figura 7 (b).

De manera complementaria, se calcularon las 4 principales métricas de calidad en el desempeño de cada uno de los modelos de clasificación mencionados –sensibilidad (SENS), especificidad (ESPEC), valor predictivo positivo o precisión (PREC) y el valor predictivo negativo (VPN) [**12**]– las cuales se muestran en la Tabla 1. Los mejores resultados, con valores = 1, se obtuvieron con los modelos matemáticos SVM y PLS-DA, seguidos del modelo kNN y por último 1iC-SIMCA.

En este sentido, los modelos SVM y PLS-DA son capaces de clasificar correctamente todas las muestras TB y TBM, lo cual se demuestra a través de las métricas SENS y ESPEC = 1, respectivamente. Asimismo, ambos modelos poseen una excelente precisión (PREC = 1), lo cual indica la proporción de muestras TB bien clasificadas en relación a todas las muestras totales asignadas a la clase TB; así como también un excelente VPN = 1, lo cual indica la habilidad de ambos modelos para proveer la proporción de muestras TBM bien clasificadas en relación a las muestras totales asignadas a la clase TBM.

Cabe mencionar que, a pesar que el modelo kNN clasificó mal solamente una muestra de TB, obteniendo buenos resultados de SENS y PREC = 0.83 en ambas métricas de calidad, sus resultados de ESPEC = 0.60 indican su falta de habilidad para clasificar muestras de la clase TBM. Por tanto, los modelos matemáticos SVM y PLS-DA arrojan mayor seguridad en los resultados y, por tanto, para su implementación en análisis de rutina y cribado para el control de calidad del Tequila Blanco, ya que son capaces de clasificar ambas categorías correctamente y de "aprender" de los datos que se van incluyendo para aumentar el conjunto de entrenamiento, por lo que la clasificación de nuevas muestras desconocidas será cada vez más confiable.

Tabla 1. Principales métricas de calidad en el desempeño para los modelos de clasificación SVM, PLS-DA, 1iC-SIMCA y kNN.

	SVM	**PLS-DA**	**kNN**	**1iC-SIMCA**
Métricas	*Clase diana (TB, Tequila Blanco 100 % agave)*			
Sensibilidad (SENS)	1.00	1.00	0.83	0.17
Especificidad (ESPEC)	1.00	1.00	0.60	0.40
Valor predictivo positivo o precisión (PREC)	1.00	1.00	0.83	0.33
Valor predictivo negative (VPN)	1.00	1.00	1.00	0.33

2.3. Predicción del contenido alcohólico

El conjunto de entrenamiento estuvo conformado por 38 muestras (22 TB / 16 TBM) y el conjunto de validación externa por 9 muestras (6 TB / 3 TBM), los cuales se utilizaron para construir los modelos matemáticos de regresión PLSR y SVMR para predecir su contenido alcohólico. El primero de ellos se construyó con 6 LVs, las cuales explicaban el 99.24 % y 96.90 % de la varianza acumulada en los bloques X e Y, respectivamente, con un RMSEC de 1.481 y R^2 = 0.969 (ver Figura 8 (a)). El segundo modelo de regresión, SVMR, fue construido utilizando 22 vectores soporte (SVs, *support vectors)*, función Kernel de base radial, y valores óptimos de gama y costo de 0.00032 y 100, respectivamente. Dicho modelo matemático proporcionó un RMSEC de 0.091 y un R^2 = 1.0 (véase Figura 8 (b)).

Además de las métricas de calidad en el desempeño de la predicción del contenido alcohólico comentadas anteriormente, se calcularon otras 5 métricas, sugeridas en las prácticas estandarizadas para la validación de calibraciones multivariable realizadas empíricamente (ASTM E2617 [**13**]).

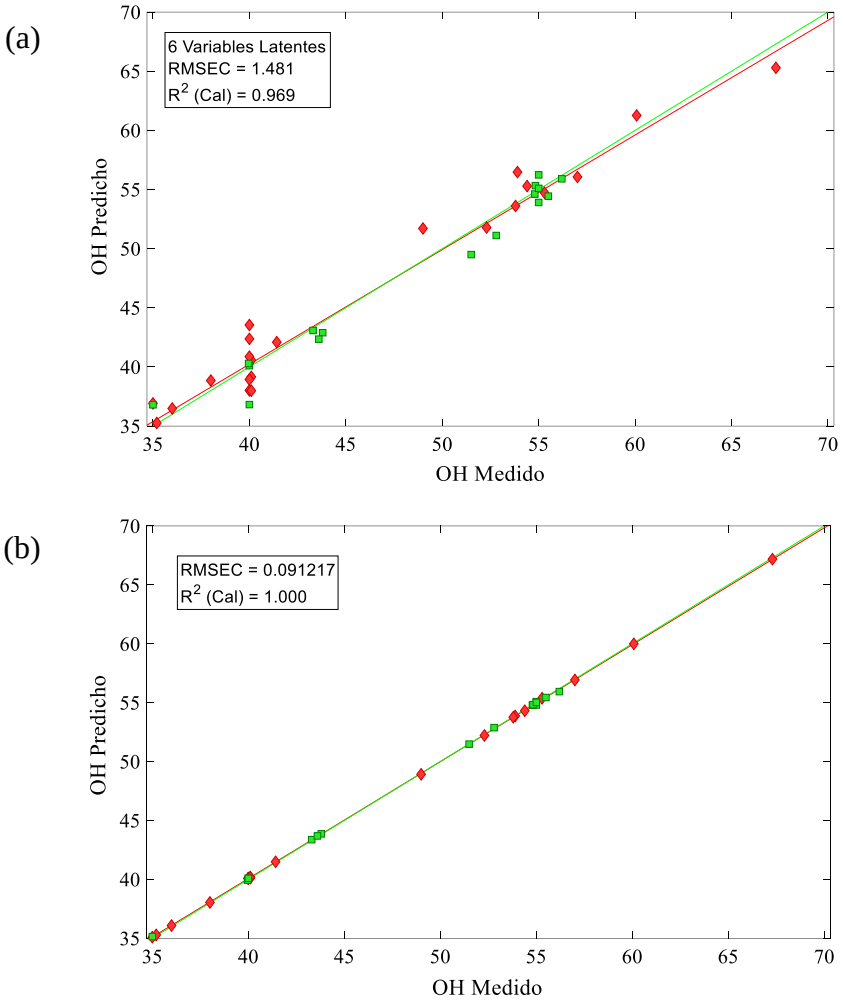


Figura 8. Curvas de calibración para la cuantificación del contenido alcohólico de muestras de Tequila Blanco mediante los modelos matemáticos (a) PLSR y (b) SVMR.

Como se puede observar a partir de las métricas de calidad de la Tabla 2, el modelo de regresión SVMR obtuvo mejores resultados que el modelo PLSR en cuanto a los errores que estos modelos pueden presentar al momento de realizar las predicciones. No obstante, el porcentaje de la desviación estándar de los residuos de validación (SDV), el cual hace referencia a la concordancia entre el valor predicho y el valor de referencia, de ambos modelos es muy similar, indicando que las predicciones del contenido alcohólico de ambos modelos son muy próximas entre sí.

Tabla 2. Métricas de calidad en el desempeño de la predicción del contenido alcohólico de muestras de Tequila Blanco mediante SVMR y PLSR.

	PLSR	**SVMR**
Métricas	Valor (%)	
Coeficiente de determinación (R^2)	0.969	1.00
Error cuadrático medio (RMSE)	7.21	6.22
Error absolute medio (MAE)	5.17	3.79
Error absoluto de la mediana (MdAE)	6.16	5.02
Error estándar de validación (SEV)	6.86	6.47
Desviación estándar de los residuos de validación (SDV)	6.47	6.21

3. Conclusiones

La adulteración y falsificación de bebidas alcohólicas espirituosas, como el tequila, ha ido en aumento a través de los años, con lo cual se requieren de métodos analíticos auxiliares para agilizar y maximizar la detección de dichas bebidas apócrifas. En este trabajo, se ha desarrollado un método analítico multivariable basado en la espectroscopía de infrarrojo cercano (NIR) y herramientas quimiométricas para el control de calidad de las categorías '100 % agave' y 'mixto' del Tequila Blanco. Las herramientas quimiométricas con las cuales fue posible obtener información de relevancia de las huellas instrumentales y, por tanto, mejor desempeño fueron SVM y PLS-DA, con las cuales ha sido posible diferenciar entre ambas categorías de Tequila Blanco, logrando una sensibilidad, especificidad, precisión y valor predictivo negativo = 1.

Del mismo modo, se desarrolló un modelo matemático para predecir el contenido alcohólico de las muestras, con el cual fue posible obtener un $R^2 = 1.0$ al utilizar regresión mediante SVMR. Dichos métodos analíticos multivariables podrían ser implementados en laboratorios analíticos de rutina para agilizar el análisis de muestras y aumentar la detección de bebidas alcohólicas adulteradas o falsificadas, así como también podría ser utilizado por la industria tequilera para agilizar el control de calidad de los productos elaborados en sus instalaciones.

Referencias

1. Consejo Regulador del Tequila. https://www.crt.org.mx.

2. L. Soto-Romero, L.J. Gutiérrez-Osnaya, L. Trejo-Fragoso, Revisión de los compuestos responsables del olor y sabor del Tequila, 2016, Investigación y Desarrollo en Ciencia y Tecnología de Alimentos, 1,910-915.

3. B. Vallejo-Córdoba, A.F. González-Córdova, M.C. Estrada-Montoya, Tequila volatile characterization and ethyl ester determination by solid pase microextraction gas chromatography/mass spectrometry analysis, 2004, Journal of Agricultural and Food Chemistry, 52, 5567-5571.

4. A. Peña-Alvares, S. Capella, R. Juárez, C. Labastida, Determination of terpens in tequila by solid pase microextraction-gas chromatography-mass spectrometry, 2006, Journal of Chromatography A, 1134, 291-297.

5. S.E.R Bukovsky-Reyes, L.E. Lowe, W.M. Brandon, J.E. Owens, Measurement of antioxidants in distilled spirits by a silver nanoparticle assay, 2018, Journal of the Institute of Brewing, 124, 291-299.

6. W.M. Warren-Vega, R. Fonseca-Aguiñaga, L.V. González-Gutiérrez, L.A. Romero-Cano, A critical review on the assessment of the quality and authenticity of tequila by different analytical techniques: Recent advances and perspectives, 2023, Food Chemistry, 408, 135223.

7. Norma Oficial Mexicana NOM-006-SCFI-2012, Bebidas alcohólicas-Tequila-Especificaciones. Comité Consultivo Nacional de Normalización de Seguridad al Usuario, Información Comercial y Prácticas de Comercio (CCNNSUICPC). https://www.crt.org.mx/images/documentos/Normas/NOM-006-SCFI-2012.pdf. Acceso: 20-01-2023

8. O.A. Kolomiets, D.W. Lachenmeier, U. Hoffmann, H.W. Siesler, Quantitative determination of quality parameters and authentication of vodka using near infrared spectroscopy, 2010, Journal of Near Infrared Spectroscopy, 18, 59-67.

9. D. Livermore, Q. Wang, R.S. Jackson, Understanding near infrared spectroscopy and its applications in the distillery, in K.A. Jacques, T.P. Lyons and D.R. Kelsall (Eds.), The alcohol textbook, a reference for the beverage, fuel and industrial alcohol industries, Nottingham University Press, pp. 145-170.

10. M. Pontes, S. Santos, M. Araujo, L. Almeida, R. Lima, E. Gaiao, U. Souto, Classification of distilled alcoholic beverages and verification of adulteration by near infrared spectrometry, 2006, Food Research International, 39, 182–189.

11. B.M. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Winding, R.S. Koch, Chemometrics Tutorial for PLS_Toolbox and Solo. Eigenvector Research, Inc. Wenatchee, WA, USA, 2006.

12. L. Cuadros Rodríguez, E. Pérez Castaño, C. Ruiz Samblás, Quality performance metrics in multivariate classification methods for qualitative analysis, 2016, Trends in Analytical Chemistry, 80, 612–624.

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

4.4. Artículo científico V

Microchemical Journal 183 (2022) 108126

Contents lists available at ScienceDirect

Microchemical Journal

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

Non-targeted spatially offset Raman spectroscopy-based vanguard analytical method to authenticate spirits: White Tequilas as a case study

Christian Hazael PÉREZ-BELTRÁN [a,*], Guadalupe PÉREZ–CABALLERO [b], José M. ANDRADE [c], Luis CUADROS-RODRÍGUEZ [a], Ana M. JIMÉNEZ-CARVELO [a,*]

[a] Department of Analytical Chemistry, Faculty of Sciences, University of Granada, C/ Fuentenueva, s/n, E-18071 Granada, Spain
[b] Laboratorio de Fisicoquímica Analítica y Especiación Química, Unidad de Investigación Multidisciplinaria. Facultad de Estudios Superiores Cuautitlán, Campo 4. Universidad Nacional Autónoma de México, México
[c] Group of Applied Analytical Chemistry, University of A Coruña, Campus da Zapateira s/n, E-15071, A Coruña, Spain

ARTICLE INFO

Keywords:
Chemometrics
Spirits fraud
Spatially offset Raman spectroscopy
Tequila authentication

ABSTRACT

Adulteration and counterfeiting are ongoing problems for alcoholic drinks, including beers, wines, and spirits. To fight against them, official analytical methods need to be complemented with faster, trustworthy, non-invasive and *in-situ* ones, which have been named as vanguard methods, to increase the efficiency in the detection probability of truly adulterated alcoholic drinks. The analytical methodology proposed here synergistically combines a novel measurement analytical technique (spatially offset Raman spectroscopy, SORS) with chemometrics methods, i.e., principal component analysis (PCA), soft independent modeling of class analogies (SIMCA), partial least squares regression-discriminant analysis (PLS-DA), support vectors machine, (SVM) and quantitative partial least squares regression (PLSR). The applicability of the proposal is tested with Tequila to (i) differentiate among *100 % agave* and *mixed* white packaged Tequilas, and (ii) to predict the alcoholic content. SORS spectra of 51 samples were obtained in the 300–2000 cm^{-1} range, from which classification and regression models were developed. The best classification performances were obtained with PLS-DA and SVM with 100 % sensitivity, specificity and overall classification rate. PLSR exposed a better trend of the samples than PCA in the exploratory analysis; and yielded predictive models capable of foreseeing alcoholic contents with average errors lower than 4 %. These results demonstrate the potential of this fast, *in-situ* analytical approach to be used as a vanguard analytical method to screen adulterated or counterfeited Tequilas and to assess the conformity of the alcoholic stated in the label.

1. Introduction

Criminal activity against consumers continues unabated, in fact, European Union Intellectual Property Office (EUIPO) and European Union Agency for Law Enforcement Cooperation (EUROPOL) have indicated in a last report published in March 2022 that *the production of illicit food products, especially drinks, is increasingly professional and sophisticated* [1]. However, in terms of health and food safety, the weightiness of food and drink fraud will depend on the type of fraud. In some cases, the consequences are limited to consumer deception, since offenders pass off lower value products as higher value foods or drinks for illicit financial profit. Specifically in drinks, the most frequent fraud is that committed in alcoholic beverages, so-called spirits. In fact, in the last two years, adulteration of this type of product has been detected, such as the case of the Whiskey fraud in Spain in 2020 [2] or the adulteration of alcoholic beverages in Santo Domingo in April 2022, which resulted in the death of several people [3].

There is a battery of recognized and well-described analytical methods for detecting different types of adulteration for each particular alcoholic beverage, most of them based on the identification and quantification of specific chemical markers. Despite traditional analytical methods proved to be reliable, accurate and are suitable tools for production control, they often do not comply with the principles of green chemistry, since they involve the use of environmentally unfriendly reagents, are time-consuming and frequently expensive, considering them as rearguard methods [4]. This gives opportunity for the development and application of alternative analytical methods, which are characterized by being miniaturized, transportable, simple,

* Corresponding authors.
E-mail addresses: christianpb@correo.ugr.es (C.H. PÉREZ-BELTRÁN), amariajc@ugr.es (A.M. JIMÉNEZ-CARVELO).

https://doi.org/10.1016/j.microc.2022.108126
Received 12 October 2022; Received in revised form 28 October 2022; Accepted 29 October 2022
Available online 5 November 2022

Research Paper

Non-targeted Spatially Offset Raman Spectroscopy-based vanguard analytical method to authenticate spirits: White Tequilas as a case study

Christian Hazael PÉREZ-BELTRÁN✉ [a], Guadalupe PÉREZ–CABALLERO [b], José M. ANDRADE [c], Luis CUADROS-RODRÍGUEZ [a], Ana M. JIMÉNEZ-CARVELO✉ [a]

[a] Department of Analytical Chemistry, Faculty of Sciences, University of Granada, C/ Fuentenueva, s/n, E-18071 Granada (Spain).

[b] Laboratorio de Fisicoquímica Analítica y Especiación Química, Unidad de Investigación Multidisciplinaria. Facultad de Estudios Superiores Cuautitlán, Campo 4. Universidad Nacional Autónoma de México (México).

[c] Group of Applied Analytical Chemistry, University of A Coruña, Campus da Zapateira s/n, E-15071, A Coruña (Spain).

Keywords

Chemometrics; Spirits fraud; Spatially offset Raman spectroscopy; Tequila authentication.

✉ Corresponding author (E-mail: christianpb@correo.ugr.es; Phone: +34 958 24 07 97)

✉ Corresponding author (E-mail: amariajc@ugr.es; Phone: +34 958 24 07 97)

1. Introduction

Criminal activity against consumers continues unabated, in fact, European Union Intellectual Property Office (EUIPO) and European Union Agency for Law Enforcement Cooperation (EUROPOL) have indicated in a last report published in March 2022 that *the production of illicit food products, especially drinks, is increasingly professional and sophisticated* [**1**]. However, in terms of health and food safety, the weightiness of food and drink fraud will depend on the type of fraud. In some cases, the consequences are limited to consumer deception, since offenders pass off lower value products as higher value foods or drinks for illicit financial profit. Specifically in drinks, the most frequent fraud is that committed in alcoholic beverages, so-called spirits. In fact, in the last two years, adulteration of this type of product has been detected, such as the case of the Whiskey fraud in Spain in 2020 [**2**] or the adulteration of alcoholic beverages in Dominican Republic in April 2022, which resulted in the death of several people [**3**].

There is a battery of recognized and well-described analytical methods for detecting different types of adulteration for each particular alcoholic beverage, most of them based on the identification and quantification of specific chemical markers. Despite traditional analytical methods proved to be reliable, accurate and are suitable tools for production control, they often do not comply with the principles of green chemistry, since they involve the use of environmentally unfriendly reagents, are time-consuming and frequently expensive, considering them as rearguard methods [**4**]. This gives opportunity for the development and application of alternative analytical methods, which are characterized by being miniaturized, transportable, simple, rapid, low-cost and capable of providing overall analytical information that is reliable and representative. The application of these type of alternative analytical methods, which have been named as vanguard methods, increase the efficiency of control laboratories since they make possible the analysis of only suspicious samples by rearguard methods [**4**]. The term vanguard method does not refer to the fact that the methodology presented in this study is highly recent and innovative, as might be inferred at first. It suggests that such a methodology could be applied as a first analytical approach to quickly process laboratory samples.

In this sense, a vanguard analytical method is often a forward screening method that allows the selection of suspect samples that will subsequently be subjected to a full backward analytical method, *i.e.*, a rearguard method.

In this sense, the use of non-targeted spectroscopic analytical techniques, such as conventional Raman or medium and near infrared spectroscopies, constitute well-established methodologies that fit most requirements to get vanguard analytical methods as they require minimum or null sample preparation. Despite of providing unspecific signals (spectroscopic instrumental fingerprints), they became popular to determine the composition/adulteration of food and beverages to ensure the authenticity and traceability [**5**]. One essential and inherent subsequent step after the application of spectroscopic techniques is the use of multivariate chemical data analysis or chemometrics, which together have created a synergistic and powerful analytical methodology that is regularly applied in the food industry to extract important and non-evident (or hidden) information from the raw spectra by developing mathematical models [**6**,**7**,**8**]. Quite recently, a new and more advanced Raman spectroscopy modality, termed spatially offset Raman spectroscopy (SORS), appeared and it shows highly promising capabilities for spirit quality and authenticity control. The fundamentals of SORS are like the conventional Raman spectroscopy, although in SORS the Raman signal is obtained at certain millimeters off the laser spot, making it possible to collect photons emitted from samples contained within opaque packaging materials [**9**]. This means that it is possible to carry out the analysis directly on the product within the container, without the need to alter the original package/sample, making SORS one of the few truly non-invasive analytical techniques. Even though this novel approach was first developed for the pharmaceutical industry, it expanded rapidly to the food industry to analyze packaged beverages in a fast and non-destructive manner [**9**]; for instance: Vodka, Gin and Whisky through their containers [**10**]. However, no applications have been found to authenticate Tequilas.

Tequila is a representative spirit from México that holds an Official Designation of Origin (DOT –*from the Spanish term 'Denominación de Origen Tequila'*–), which is regulated by the Mexican Government and the Regulatory Council of Tequila (CRT) through the official Mexican standard NOM-006-SCFI-2012 [**11**].

Additionally, two categories of Tequila can be distinguished: (i) '100% agave' Tequila if only sugars from the juice of the *Agave Tequilana Weber Blue* variety are used for the fermentation process, and (ii) 'mixed' Tequilas if any combination with other sources of reducing sugars (never more than 49%) are added to the process. The commonest commercial product is White Tequila, so, this paper focused on it [**11**]. Tequila can be classified in five classes according to their aging process in oak or holm oak containers: '*Silver or White*', *'Aged'*, *'Extra-aged'* and *'Ultra-aged'* according to whether maturation lasts for < 2 months, ≥ 2 months, ≥ 2 years or ≥ 3 years, respectively. *'Gold Tequila'* corresponds to commercial mixtures of White Tequila with Aged, Extra-aged or Ultra-aged Tequilas [**11**].

Currently, many adulteration and counterfeiting cases are still reported, not only at Mexico but in other countries. The main adulteration practice is to substitute ethanol with methanol or, less frequently, with propanol, ethylene glycol, aldehydes and others [**12**]. In 2021, a production of 527 million of liters of tequila was reported by the CRT, whose quality and authenticity were evaluated using representative samples extracted from the distilleries and analyzed independently at the CRT. All the aforementioned classes of Tequila are inspected by the CRT using standardized analytical techniques, such as liquid and gas chromatography or atomic absorption spectroscopy, to adhere to current official analytical methods. Several quality parameters are determined, *e.g.*, furfural, esters, aldehydes, methanol, higher alcohols, reducing and total sugars. An exemplary routine verification is whether the alcoholic content, using a digital densimeter method at 20 ºC, which is established in the Mexican standard NMX-V-013-NORMEX-2019 [**13**], is between 35 and 55% (v/v).

The studies found in the literature concerning the assessment of tequila authenticity are focused on (i) some chemical markers, (ii) a specific spectral region of interest (ROI), or (iii) Red, Green and Blue (RGB) color coordinates obtained after the Tequila analysis by chromatographic and spectroscopic analytical techniques [**14**,**15**,**16**,**17**,**18**,**19**]. For example, Contreras et al. [**20**] applied UV-Vis spectroscopy to identify adulterated and fake Tequilas (between White and Rested Tequilas) or Perez-Beltran et al. [**21**] employed FTIR and data fusion approach for distinguishing between pure and mixed White Tequilas.

However, surprisingly no studies have been found where the full RAMAN spectrum is used as an unspecific instrumental fingerprint but characteristic of each tequila together with chemometric tools for tequila authentication.

In this regard, the innovation of this work lies in developing a fast and non-invasive vanguard analytical method for the *in-situ* screening quality control of spirits using SORS. Its applicability is demonstrated to ensure Tequila from Mexico in the following terms: (i) discriminate White Tequilas ('100% agave' vs 'mixed'), and to (ii) predict and verify the alcoholic content. For this, SORS spectra were used together chemometric tools to develop suitable classification and quantitation multivariate analytical methods. Classification methods were validated in terms of sensitivity, specificity, precision, negative prediction value, among other 21 classification performance metrics and estimated following the study published by Cuadros-Rodríguez et al. (2016) [**22**]. In addition, the quantitative method for determining the alcohol content was validated according to the ASTM E2617 standard [**23**].

2. Materials and methods

2.1. Tequila samples

A total of 51 White Tequila samples were provided by the CRT in México, and analyzed in Spain, as described in the 'spatially offset Raman spectroscopy (SORS) measurements' section. Thirty White Tequilas belonged to the '100% agave' White Tequila category (TB - from the Spanish term 'Tequila Blanco') and twenty-one to the 'mixed' White Tequilas (TBM - from the Spanish term 'Tequila Blanco Mixto'). The alcoholic content of all these samples was determined by the CRT using a digital densimeter at 20ºC [**13**].

2.2. Spatially offset Raman spectroscopy (SORS) measurements

Vaya Raman SORS equipment (Agilent Technologies, Santa Clara, CA, USA) was used. The excitation radiation was 830 nm with a maximum power laser of 450 mW, obtaining Raman spectra in the low frequencies range, from 350 to 2000 cm^{-1}, with 12 to 20 cm^{-1}

spectroscopic resolution. The SORS measurements of the 51 White Tequila samples were performed directly through amber vials lasting 30 s, approximately.

2.3. Similarity analyses

In order to make sure that this methodology can be transferable to any other situation, similarity analyses were performed. SORS measurements were directly performed on four original bottled tequilas marketed in Spain (2 mixed White Tequilas, 1 mixed Rested Tequila and 1 mixed Tamarind flavored White Tequila). Afterwards, 2 mL of each of them were transferred to amber glass vials, similar to those used to transport the Mexican tequila samples, and measured. Once both spectra for each sample were acquired, the similarity among them was assessed calculating the corresponding nearness similarity index [**24**], which is based on the proximity of two vectors in space and is calculated from the standardized Euclidean distance, as depicted in Eq. (1).

$$\mathrm{NEAR}(X_{SORS}, X_{CRS}) = 1 - \left[\sqrt{\frac{(X_{SORS} - X_{CRS})\ \mathrm{x}\ (X_{SORS} - X_{CRS})^{T}}{(X_{SORS} + X_{CRS})\ \mathrm{x}\ (X_{SORS} + X_{CRS})^{T}}}\right] \qquad (1)$$

where X_{SORS} and X_{CRS} symbolize both SORS and conventional Raman spectra, respectively, and the superscript T denotes the transposed matrix [**25**].

2.4. Multivariate data analyses

SORS raw data were exported from .csv format (comma-separated values) to MATLAB environment (Mathworks, Massachusetts, USA, v. R2013b). The exported spectra contained 1651 variables, each. The training set was constituted by 41 samples (24 of TB type and 17 of TBM type) whilst the external validation set contained 10 different samples (6 TB and 4 TBM). Splitting was performed applying the Kennard-Stone selection method (so-called CADEX algorithm), which was deployed on the TB and TBM classes independently in order to select the samples of the validation set.

The multivariate data analyses were carried out using the PLS_Toolbox software (v. 8.6.1, 2019, Eigenvector Research In., Manson, WA, USA). The applied chemometric tools were principal component analysis (PCA) and partial least squares regression (PLSR) for exploratory analysis, soft independent modeling of class analogy (SIMCA), partial least squares-discriminant analysis (PLS-DA) and support vector machines (SVM) for classification, and PLSR was also used to quantify the alcoholic content of the samples. Mean centering and smoothing were used as pre-processing techniques depending on the multivariate method, as described in 'exploratory analyses' and 'classification analyses'. The proper number selection of the PCs and LVs of the models was based on the study of their root mean square error for calibration (RMSEC), or for prediction (RMSEP) and for cross-validation (RMSECV) plots, and the total explained variance, avoiding overfitting in each case.

3. Results and discussion

3.1. SORS analyses and characterization

When SORS analyses are performed, two measurements are acquired: one at zero offset and another one with a laterally spatial offset of 0.7 mm from the point of incidence of the laser to the collection point [**9**]. This separation favors the photons from the lower layers to be radiated from a spot laterally shifted from the incidence zone while the photons on the upper package are radiated from the same incidence zone [**26**]. Afterwards, internal pre-processing and normalization are performed by the equipment, obtaining a final Raman spectrum with no contribution of the container. The Raman spectra of the two categories of white tequilas can be observed in Fig. 1.

The intense peaks located at 882 cm^{-1} and 1053 cm^{-1} are attributed to the stretching and deformation modes of the skeletal C-C-O moieties, whilst the peak at 1090 cm^{-1} is associated to the stretching mode of the C-O bond. The peaks at 1279 cm^{-1} and 1455 cm^{-1} are assigned to the deformation wagging mode and to the wagging mode of CH_2, respectively [**15**,**27**].

Additionally, the two small peaks around 1610 cm^{-1} and 1728 cm^{-1} are associated to the cyclic ketone structure, which is the basis of furanic compounds in tequila. Noteworthy, those peaks are more intense for the TB category than for the TBM one, as TB proceeds only from fermentable sugars of the *Agave Tequilana Weber Blue* variety (through the Maillard reaction [**28**] when cooked). On the contrary, TBM might or might not present these spectral Raman peaks because this category of tequilas can be produced from mixtures of fermentable sugars, so that the production of furanic compounds might not occur [**29**].

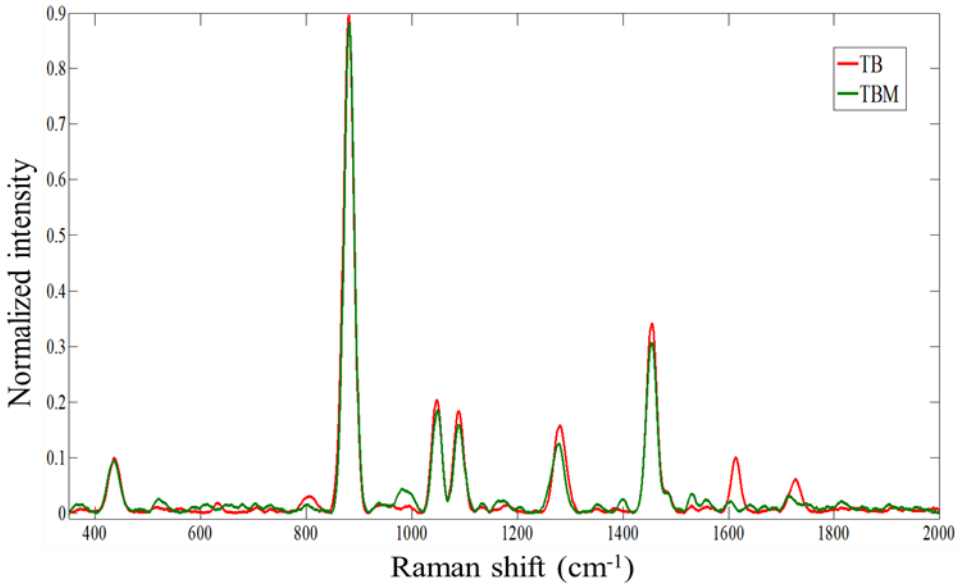


Figure 1. Raman spectra of a '100% agave' White Tequila sample (TB) and a 'mixed' White Tequila (TBM) one.

These acquired signals (Raman spectra), which are here used to evaluate the authenticity and quality of White Tequilas, are non-specific instrumental fingerprints and make it necessary the application of multivariate data analyses, as described in the following subsections.

3.2. SORS and conventional Raman spectra similarity analyses

A point-by-point comparison, using the nearness similarity index (NEAR), among the four pairs of spectra (data vectors) corresponding to the tequila samples marketed in Spain was performed to assess their similarity when the spectroscopic measurements are performed through the original tequila glass bottle or through amber glass vials (used as reference). The expected NEAR results of the standardized Euclidean distance range from 0 to 1, being 1 the maximum similarity among the spectra. Fig. 2 displays the spectra of the four analyzed samples within their original glass bottles and the spectra of the samples transferred to the vial.

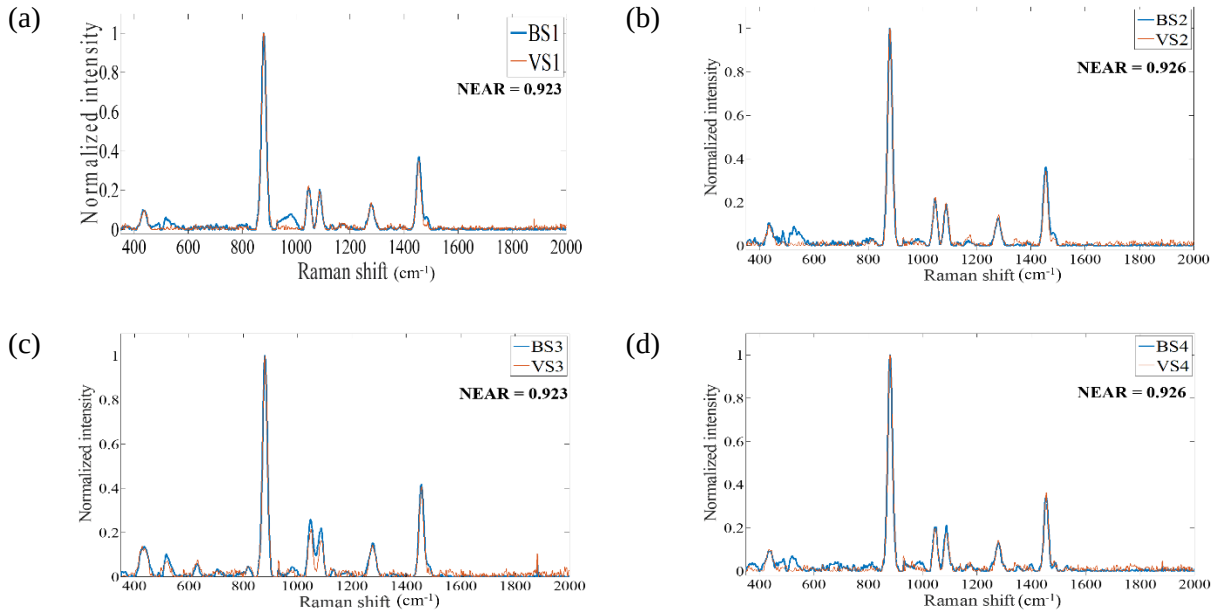


Figure 2. Similarity plots of four sample pairs of White Tequila (S1-S4) measured through the original bottle (BS) and amber vial (VS), considered as the reference spectrum.

As it can be observed in Fig. 2, each pair of overlapping spectra are similar at first glance and this fact is further confirmed when the 'nearness similarity' index (NEAR) is calculated, obtaining NEAR values > 0.92, which indicates that both spectra are largely similar with almost null influence of the original glass bottles over the measurements (the remaining ca. 0.08% can be considered as random noise).

According to these results, it is evident that the methodology presented here has potential application to the *in-situ* quality control and authentication analysis of tequila.

3.3. Exploratory analyses

Exploratory analyses were performed to screen the natural grouping of the 51 tequilas samples. For these studies, the spectral data was previously mean centered. First, a PCA was built considering 5 principal components (PCs), which explained 75.9% of the cumulative variance, whose main scores plot is displayed in Fig. 3. Nonetheless, it can be observed that the samples do not follow any specific trend among categories.

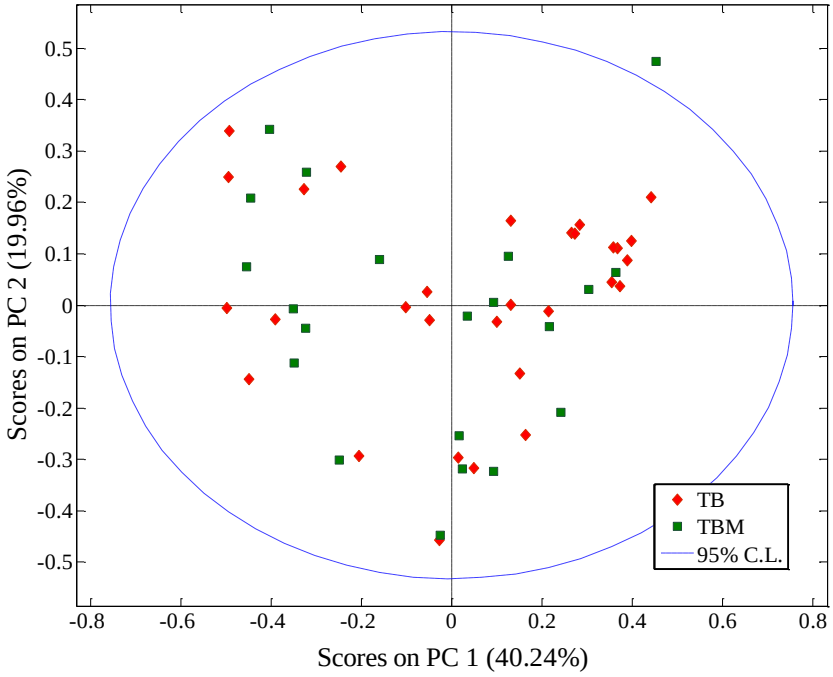


Figure 3. Exploratory PC1 vs PC2 scores plot from the 51 samples PCA model showing two different categories of White Tequilas. TB: '100% agave' White Tequila (n = 30) and TBM: 'mixed' White Tequilas (n = 21).

Furthermore, PLSR was used to explore these samples. The model was built with 5 latent variables (LVs) explaining 71.1% of the cumulative variance in the X-block and 85.8% in the Y-block. Fig. 4 shows the LV2 vs LV3 scores plot, where the TB category concentrates

(although not unequivocally) in the upper-right region of the plot and the TBM category to the left. The different results among PCA and PLSR lies basically in the very nature of the PLS latent variables that capture both variance and correlation [**30**], yielding best results when PLSR is applied, as it was also found when looking for groups among FTIR fused data of 100% agave and mixed White Tequilas [**21**]. Additionally, there are some samples placed out of the 95% confidence limit that might be considered as outliers (see Figures 3 and 4), however, it was noticed through the normalized (or reduced) Hotelling-T^2 leverages *vs* Q residuals plot that those samples had a normal behavior, discarding the existence of outliers. Thus, all samples were included in the following data analyses.

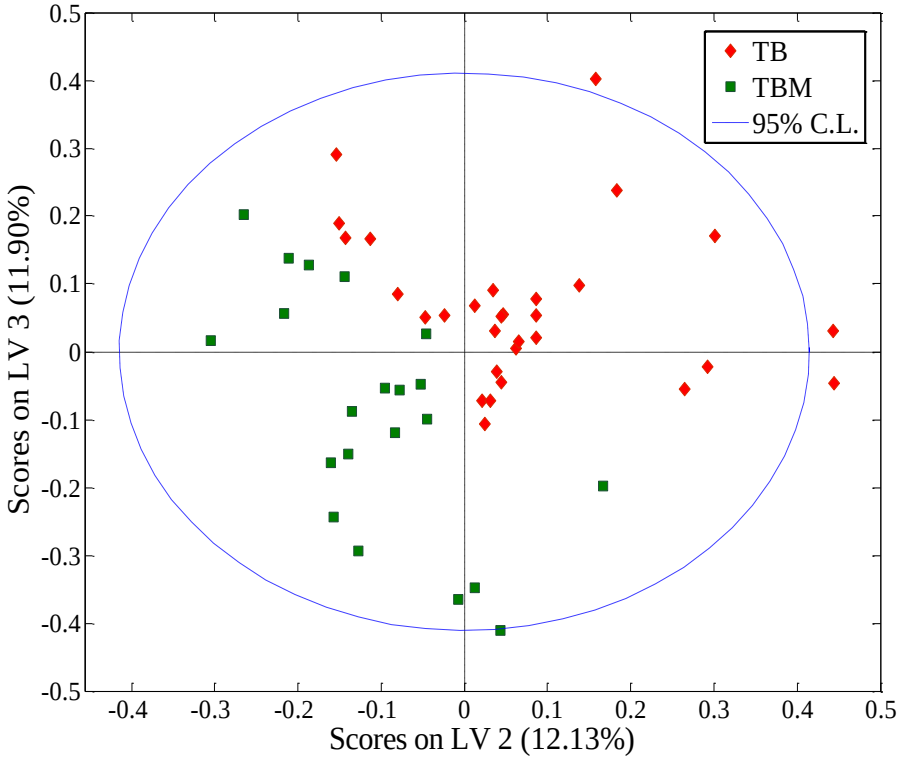


Figure 4. Exploratory LV2 vs LV3 scores plot from the 51 samples PLS model showing two different categories of White Tequilas. TB: '100% agave' White Tequila (n = 30) and TBM: 'mixed' White Tequilas (n = 21).

3.4. Classification analyses

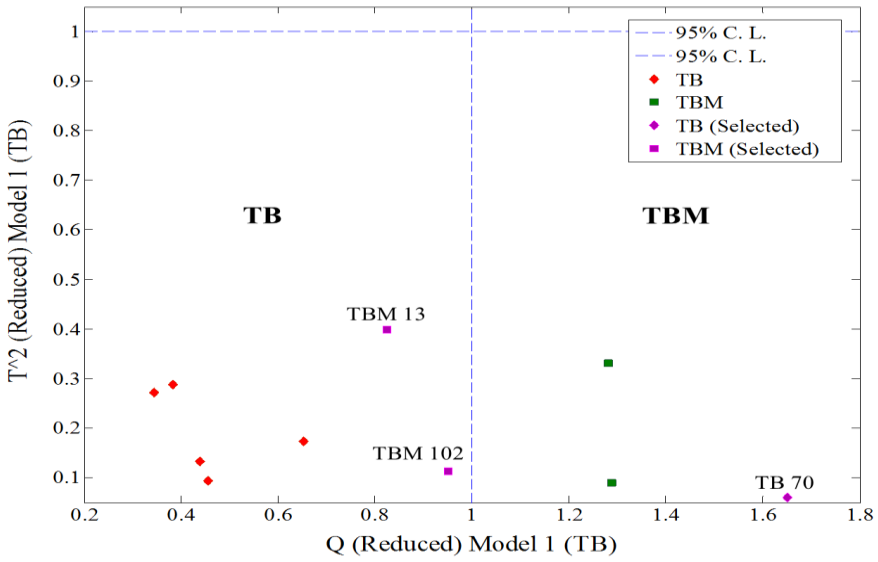
The next step after the exploratory analysis was the development of non-targeted multivariate analytical methods to discriminate among TB and TBM. For all classification models, mean centering and smoothing (Savitski-Golay, 15 points for filter width and 1st order polynomial) were used as pre-processing techniques. Smoothing is a low-pass filter that removes high-frequency noise [**30**]. The target class is TB as it is the category with more probability to be adulterated due to its economic profit. The results of the final classification models are discussed next.

❒ One Class-SIMCA

The developed SIMCA models were generated using two strategies: (i) two input-class classification (2iC-SIMCA) model, in which the model is trained using two classes (TB and TBM), and (ii) one input-class classification (1iC-SIMCA) model, in which the model is trained only with the 'target class' (TB). Within the 1iC-SIMCA strategy, two options were evaluated: (a) using the aforementioned calibration and validation data sets and (b) augmenting the validation set using all the 21 TBM and the previous 6 TB samples. It was found that the 1iC-SIMCA approach presented the best results using 5 PCs.

The 1iC-SIMCA classification plot (Fig. 5a) depicts the normalized (or reduced) Hotelling's T^2 and Q statistics of the target class, at a 95% confidence level. Samples from the validation set with normalized T^2 and Q values < 1 (left-bottom quadrant) are those considered as the target class (TB), whereas samples with T^2 and Q values > 1 (right-bottom quadrant) are considered as non-TB (or TBM). In this sense, samples TBM 13 and TBM 102 are misclassified as TB and sample TB 70 as TBM, indicating that further confirmatory analyses should be performed. These results are used to create the corresponding validation contingencies of the classification model, as shown in Fig. 5b.

(a)



(b)

Assignation	Class 1 (TB)	Class 2 (TBM)	
Inconclusive (I)	0	0	0
Class 2 (TBM)	1 (8.33%)	2 (25%)	3
Class 1 (TB)	5 (41.67%)	2 (25%)	7
	6	4	10

Actual

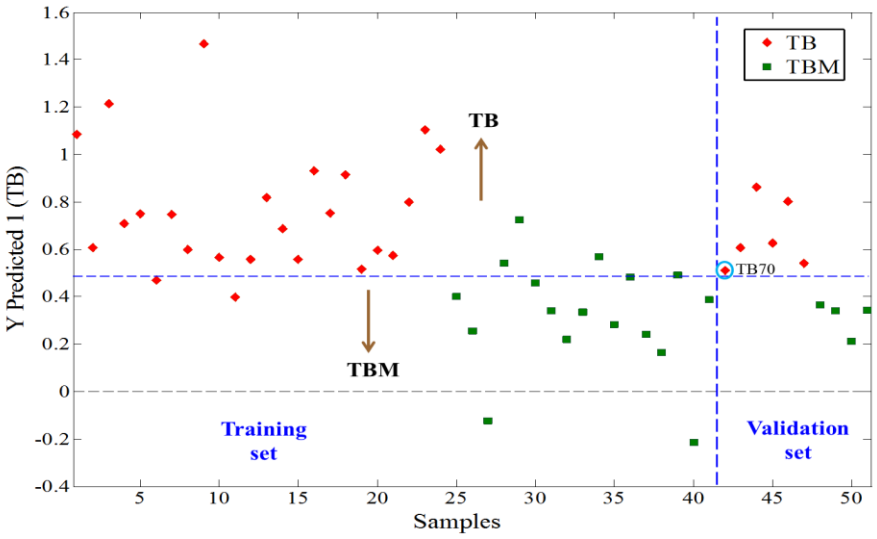
Figure 5. (a) Classification plot and (b) validation contingencies for the one input-class SIMCA classification model. Class 1: target class (TB: '100% agave' White Tequila); class 2: non-target class (TBM: 'mixed' White Tequila) (The magenta-marked samples in figure 5a are the misclassified samples).

❒ PLS-DA

The PLS-DA model was built using 4 latent variables, which explained 78.3% and 44.1% of the cumulative variance of both X- and Y-variable blocks, respectively. A threshold value of 0.5 was established as a decision criterion for the classification of the samples; scores (weights) > 0.5 correspond to TB and < 0.5 to TBM, as can be observed in the classification plot represented by Fig. 6a.

The validation contingencies of the PLS-DA classification model are shown in Fig. 6b. Note that all validation samples were correctly classified, even though some samples from the training set were misclassified. This demonstrates the powerful generalization capabilities of the PLS-DA model.

(a)



(b)

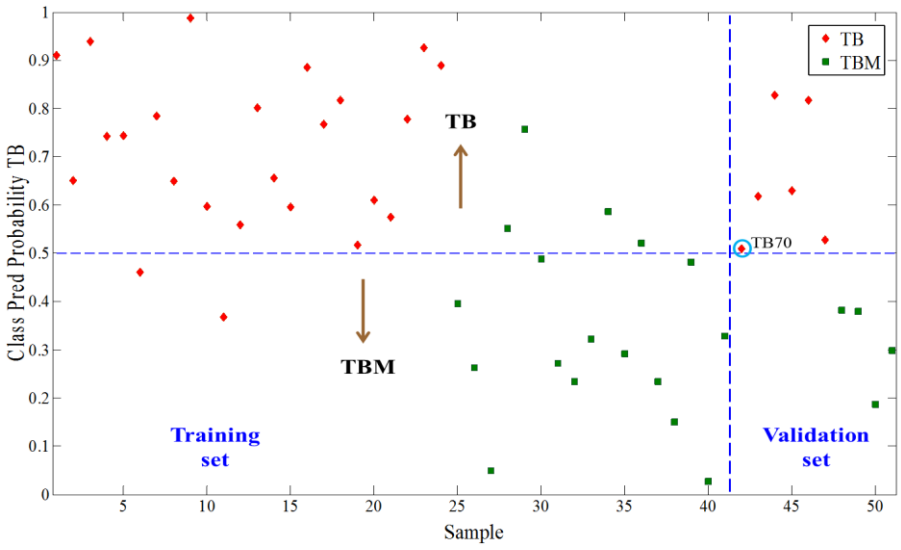
Assignation	6	4	10
Inconclusive (I)	0	0	**0**
Class 2 (TBM)	0	4 (50%)	**4**
Class 1 (TB)	6 (50%)	0	**6**
	Class 1	**Class 2**	
	Actual		

Figure 6. (a) Classification plot and (b) validation contingencies for the PLS-DA classification model. Class 1: target class (TB: '100% agave' White Tequila); class 2: non-target class (TBM: 'mixed' White Tequila). (The dashed line in figure 6a indicates the 0.5 threshold level).

❒ SVM

Support vectors machine (SVM) was performed using the radial basis function (RBF) kernel algorithm with the gamma and cost values studied in the 10^{-6}-10 and 10^{-3}-10^{2} ranges, respectively, and PLS compression with 4 LVs. The classification results for both the training and validation samples are displayed in Fig. 7a. The results are almost the same as the PLS-DA ones, suggesting that sample TB 70 should undergo further confirmatory analyses, since it is very close to the threshold value. The SVM validation contingencies are displayed in Fig. 7b.

(a)



(b)

Assignation	6	4	10
Inconclusive (I)	0	0	0
Class 2 (TBM)	0	4 (50%)	4
Class 1 (TB)	6 (50%)	0	6
	Class 1	Class 2	

Actual

Figure 7. (a) Classification plot and (b) validation contingencies for the SVM classification model. Class 1: target class (TB: '100% agave' White Tequila); class 2: non-target class (TBM: 'mixed' White Tequila). (The dashed line in figure 7a marks the 0.5 threshold level).

As a matter of comparison, the classification performance metrics for the classification models were calculated from the results of the validation contingencies (see Table 1) [**22**], considering TB as the target class. The most popular metrics are discussed here; however, the detailed explanation of each of them is out of the scope of this work and interested readers are kindly forwarded to ref [**22**] for specific details on this topic. In principle, satisfactory classifications lead to classification performance metrics close to 1 and bad models to 0. For instance, Table 1 shows that PLS-DA and SVM models have a sensitivity (SENS) = 1, whilst 1iC-SIMCA a and b yields SENS = 0.83, which indicates that PLS-DA and SVM models classify better the TB samples than 1iC-SIMCA. Specificity (SPEC) indicates that the TBM samples are correctly classified, being better for PLS-DA and SVM models with a value = 1 than for 1iC-SIMCA a and b with SPEC = 0.50 and 0.33, respectively. In fact, the 1iC-SIMCA b model, validated with all the TBM samples, provided worse classification results than 1iC-SIMCA a, validated with fewer TBM samples.

Additionally, the positive predictive value (PPV) (so-called precision) informs on the proportion of agreements in relation to all assigned values of TB class whilst the negative prediction value (NPV) takes into account the ratio between agreements and the total number of TBM samples. For PLS-DA and SVM those metrics were = 1, whereas for the 1iC-SIMCA a and b models PPV were = 0.71 and 0.26, and NPV = 0.67 and 0.88, respectively. The overall classification rate (OCR) was 100%, 100% and 83% for PLS-DA, SVM and 1iC-SIMCA, respectively, and the Matthews correlation coefficient (MCC) – which might be considered a compendium of the overall classification ability of the models– was 1.0, 1.0 and 0.36 for the same classification models.

When the validation set 'a' is applied on the 1iC-SIMCA model, the validation results are relatively good; however, the results are fictitious as this set does not represent the reality of the sample population. The good results are due to the fact that in the validation set 'a' only 4 TBM samples (non-target class) are considered, but when the number of TBM samples is increased (validation set 'b'), the model does not classify well. That is, the model classifies almost all TBM samples as belonging to the TB class, which is related to the results shown in the exploratory analysis and the no clustering tendency of the classes, so it

is not possible to establish regions for each of them. Therefore, the SIMCA class modeling method is not suitable for the purpose of this study.

Table 1. Summary of classification performance metrics for 1iC-SIMCA, PLS-DA and SVM models.

	1iC-SIMCA		PLS-DA	SVM
	a	b		
Metrics	*Target class (100% agave White Tequila, TB)*			
Sensitivity (SENS)	0.83	0.83	1.00	1.00
Specificity (SPEC)	0.50	0.33	1.00	1.00
False positive rate (FPR)	0.50	0.67	0.00	0.00
False negative rate (FNR)	0.17	0.17	0.00	0.00
Positive predictive value (PPV) (precision)	0.71	0.26	1.00	1.00
Negative predictive value (NPV)	0.67	0.88	1.00	1.00
Youden index (YOUD)	0.33	0.17	1.00	1.00
Positive likelihood rate (LR(+))	1.67	1.25	–	–
Negative likelihood rate (LR(-))	0.33	0.50	0.00	0.00
Classification odds ratio (COR)	5.00	2.50	–	–
F-measure (F)	0.77	0.40	1.00	1.00
Discriminant power (DP)	0.39	0.22	–	–
Efficiency (or accuracy) (EFFIC)	0.70	0.44	1.00	1.00
Misclassification rate (MR)	0.30	0.56	0.00	0.00
AUC (correctly classified rate) (CCR)	0.67	0.58	1.00	1.00
Gini coefficient (Gini)	0.33	0.17	1.00	1.00
G-mean (GM)	0.65	0.53	1.00	1.00
Matthews' correlation coefficient (MCC)	0.36	0.15	1.00	1.00
Chance agreement rate (CAR)	0.54	0.39	0.52	0.52
Chance error rate (CER)	0.48	0.35	0.48	0.48
Kappa coefficient (KAPPA)	0.35	0.09	1.00	1.00
PROB (TB/TB)	0.71	0.26	1.00	1.00
PROB (nTB/nTB)	0.67	0.88	1.00	1.00
PROB (TB/nTB)	0.33	0.13	0.00	0.00
PROB (nTB/TB)	0.29	0.74	0.00	0.00

The hyphen "–" signifies that the performance feature cannot be determined since it involves a division between zero.

a and b: models validated using 10 (6 TB and 4 TBM) and 27 (6 TB and 21 TBM) samples as external validation sets, respectively.

The classification ability of the models obtained in this study (PLS-DA and SVM models) are better than others previously reported for different purposes (despite a direct, straightforward comparison is not possible) applying PCA-linear discriminant analysis (LDA), with an overall classification rate (OCR) = 0.90, SENS = 0.90 and SPEC = 0.96 [**17**].

Furthermore, in a previous study [**18**] in which nine models were built using mean-centered UV-Vis spectroscopic data to differentiate various classes of Tequila, it was found that nonlinear models behaved better than linear ones (EFFIC > 0.94).

In this context, it is worth noting that class modeling methods, such as 1iC-SIMCA, are particularly suitable for real-world authentication problems where the target class is always defined from the authentic or genuine product and is modeled with a large number of samples, since it is less common to find adulterated samples. This approach has a great potential when sufficient number of authentic samples (target class) are available, *i.e.*, ideal scenario, being capable to properly identify new samples obtained from non-authentic products and differentiate them from those specimens of genuine ones. However, for this particular study, the available samples to build a more reliable 1iC-SIMCA model were limited, since '100% agave' White Tequila is only produced in certain regions of México and the accessibility of a variety of samples is rather narrow. A good alternative to address this situation are discriminant methods, such as PLS-DA and SVM because it aimed at classifying two mutually excluding classes ('100% agave' and 'mixed') of the same quality sort of tequila ('White Tequila'). In fact, it was evidenced that the validation results of the 1iC-SIMCA model depend on the number and type of samples included in the test set, but PLS-DA and SVM models provided better ability to correctly classify samples from both classes. However, this discriminant strategy is not free from the drawback of misclassifying new samples coming from non-genuine products with some different composition from those already used in the training step, which is a risk that practitioners must evaluate and take into account when extending the application of the method.

3.5. Alcoholic content quantitation

A PLSR-based quantitation analytical method was calibrated to predict the alcoholic content of the tequila samples. As detailed above, the reference values were obtained by the CRT following the official method. The PLSR model was built using mean centering to preprocess the spectra and including 5 LVs in the model which explained 73.6% and 97.1% of the cumulative variance for the X- and Y-variable blocks, respectively.

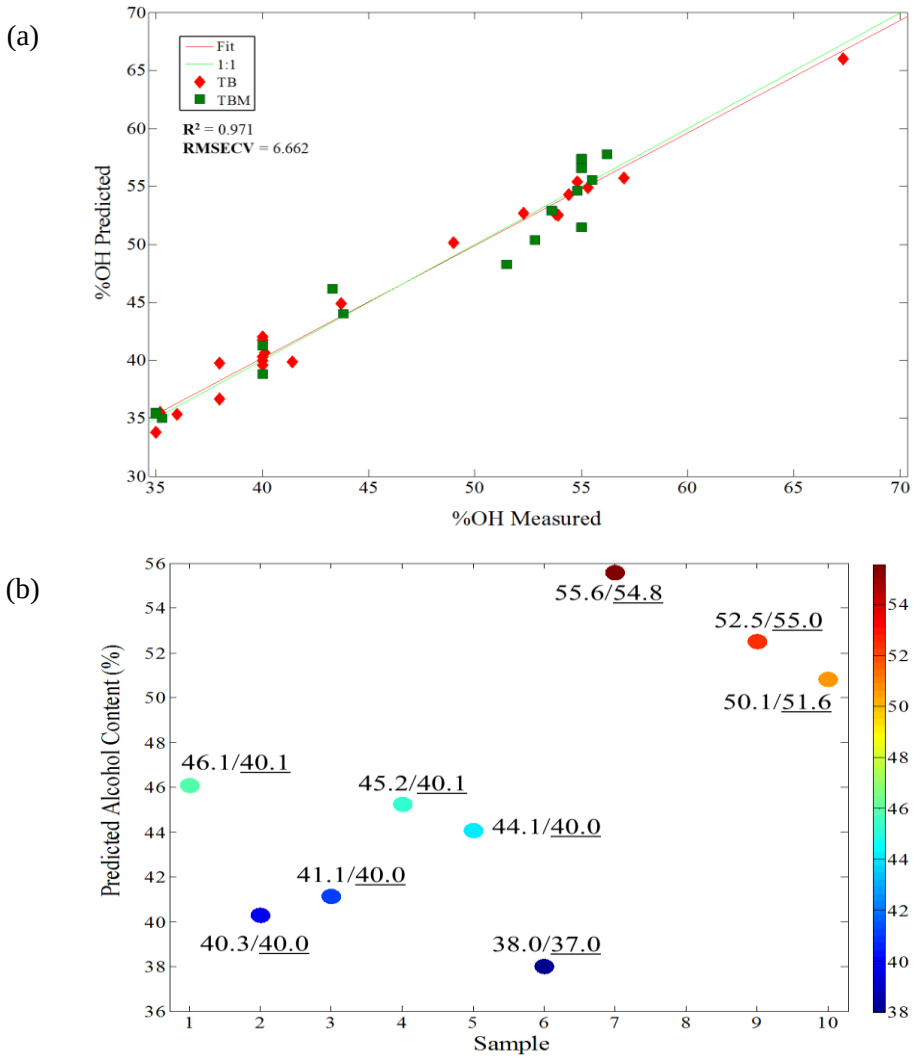


Figure 8. PLSR alcoholic predictions (% v/v) for White Tequila samples. (a) Calibration curve, and (b) alcoholic content plot of the validation set samples. The circles are colored according to the predicted alcoholic content from the vertical color scale. Each sample displays the predicted value against the real value of alcoholic content, which is underlined.

Fig. 8 compares the PLSR predicted alcoholic contents against the total alcoholic content reported by the CRT. The evaluation of this model was performed with the quantitation performance metrics, as observed in Table 2.

Table 2. Performance metrics in the quantitation of the alcoholic content of the Tequila samples that constitute the validation set.

Metrics	**Value (%)**
Coefficient of determination (R^2)	0.971
Root mean square error (RMSE)	3.32
Mean absolute error (MAE)	1.82
Median absolute error (MdAE)	2.61
Standard error of validation (SEV)	3.14
Standard deviation of validation residuals (SDV)	2.65

The first quantitation performance metric is the coefficient of determination (R^2) with a value = 0.971, evidencing a good fitting. The following four metrics are related to different sorts of errors the model might present (root mean square error, mean absolute error, median absolute error and standard error of validation), all of them with values less than 4%; the sixth metric is the standard deviation of validation residuals (SDV = 2.7%), indicating that the agreement of the predictions of the empirical model with the reference values is high, which results in a quite good predictive ability.

Note that PLSR has been previously applied to predict the alcoholic content of different Tequilas using FTIR, obtaining very good results [**19**] Moreover, a vector network analyzer with an open-ended coaxial probe kit was used for the same purpose [**31**]

PLSR has also been applied to quantitate the furfural, 2-acetylfuran and 5-methylfurfural content in White Tequilas and Mezcals samples with acceptable results [**29**]. It would have been interesting to compare the results obtained here with those of another report in which SORS was applied to study the adulteration of Vodka, Gin and Whisky with methanol, but prediction of the alcoholic content was not considered [**10**].

4. Conclusions

Economic losses for the industry of alcoholic beverages and societal health problems are two relevant consequences of the adulteration and counterfeiting of commercialized spirits, which have not ceased over the years. To streamline the authentication surveillance of these products, current official rearguard methods need to be complemented with vanguard, faster and reliable *in-situ* screening analytical methods. In this regard, the present study reports for the first time the combination of the SORS analytical technique and chemometrics to discriminate between '100% agave' and 'mixed' White Tequilas and to predict their alcoholic content. It should be noted that the potential of the *in-situ* non-invasive SORS measurement implemented here has been verified by means of a similarity analysis. This demonstrated that the spectra obtained after analyzing Tequilas through the original bottle and through amber vials are almost the same, obtaining nearness indexes close to 1. Afterwards, models were developed and assessed with several classification performance metrics, which indicated that satisfactory classifications and predictions were achieved. PLS-DA and SVM presented the best OCR = 100%, evidencing that the combination of SORS and some chemometric methods is able to discern among '100% agave' and 'mixed' White Tequilas. Finally, a PLSR quantitation model demonstrated an excellent ability to predict the alcoholic content of the samples.

The approach presented here offers an alternative analytical method for routine authentication tasks undergone by official regulatory bodies. It is reliable and fast for *in-situ* screening purposes and, can complement and accelerate the quality control and authentication processes of commercial spirits, such as Tequila.

Conflicts of interest

The authors declare that they have no conflict of interest.

Acknowledgements

The authors are deeply grateful to the Mexican "Consejo Regulador del Tequila" (CRT) for providing the samples for this study, as well as to the 'Sánchez Baldiviezo' Association for facilitating the samples for the similarity study.

One author, C.H. Pérez-Beltrán acknowledges also the scholarships from the Autonomous University of Sinaloa (México) and from the "Asociación Universitaria Iberoamericana de Posgrado" (AUIP) and the "Consejería de Transformación Económica, Industria, Conocimiento y Universidades" of the Regional Government of Andalucía (Spain) for a research stay.

References

[1] C. de Bolle, C. Archambeau. Intellectual property crime. Threat assessment 2022. EUIPO. (2022). https://doi.org/10.2814/830719.

[2] Fourteen arrested and 300,000 bottles of counterfeit whisky seized. (2020). https://www.elconfidencial.com/espana/2020-12-10/catorce-detenidos-intervenidas-300-000-botellas-whisky-falso-guardia-civil_2866480/ (*Spanish version*). Accessed 27 July 2022.

[3] Six people have died as a result of adulterated alcohol. (2022). https://www.elcaribe.com.do/destacado/seis-personas-han-fallecido-a-causa-del-alcohol-adulterado-en-2022/ (*Spanish version*). Accessed 27 July 2022.

[4] M. Valcárcel, S. Cárdenas, Vanguard-rearguard analytical strategies, Trends. Anal. Chem. 24 (2005) 67-74. https://doi.org/10.1016/j.trac.2004.07.016.

[5] A.M. Jiménez Carvelo, S. Martin Torres, L. Cuadros Rodríguez, A. González Casado, Food Authentication and Traceability, in: C.M. Galanakis (Ed.), Nontargeted fingerprinting approaches, Academic Press, 2021, pp. 163-194. https://doi.org/10.1016/B978-0-12-821104-5.00010-6.

[6] P. Oliveri, C. Malegori, E. Mustorgi, M. Casale, Comprehensive Chemometrics – Chemical and Biochemical Data Analysis, in: S. Brown, R. Tauler, B. Walczak (Eds.), Application of chemometrics in the food sciences, Elsevier, 2020, pp. 99-111. https://doi.org/10.1016/B978-0-12-409547-2.14748-1.

[7] A.M. Jiménez Carvelo, L. Cuadros Rodríguez, Data mining/machine learning methods in foodomics, Curr. Opin. Food Sci. 37 (2021) 76-82. https://doi.org/10.1016/j.cofs.2020.09.008.

[8] A.M. Jiménez Carvelo, A. González Casado, M.A. 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. https://doi.org/10.1016/j.foodres.2019.03.063.

[9] 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. https://doi.org/10.1016/j.lwt.2021.111822.

[10] D.I. Ellis, R. Ecccles, Y. Xu, J. Griffen, H. Muhamadali, P. Matousek, I. Goodall, R. Goodacre, Through-container, extremely low concentration detection of multiple chemical markers of counterfeit alcohol using a handheld SORS device, Sci. Rep. 7 (2017) 12082. https://doi.org/10.1038/s41598-017-12263-0.

[11] Mexican Official Standard NOM-006-SCFI-2012, Alcoholic Beverages -Tequila-Specifications, National Advisory Committee on Standardization, User Safety, Commercial Information and Trade Practices (CCNNSUICPC), Mexican Government. https://www.crt.org.mx/images/documentos/Normas/NOM_006_SCFI_2012_Ingles.pdf (accessed 13 June 2022).

[12] D.G. Barceloux, R. Bond, E.P. Krenzelok, H. Cooper, J.A. Vale, American academy of clinical toxicology practice guidelines on the treatment of methanol poising, J. Toxicol. Clin. Toxicol. 40 (2002) 415-446. https://doi.org/10.1081/CLT-120006745.

[13] Mexican Standard NMX-V-013-NORMEX-2019, Bebidas alcohólicas-determinación del contenido alcohólico (por ciento de alcohol en volumen a 20ºC) (% Alc. Vol.) -Métodos de ensayo (prueba) (*in Spanish*). National Advisory Committee on Standardization of the Economy Secretariat (CCONNSE), Mexican Government.

[14] L.I. Espinosa Vega, A. Belio Manzano, C.A. Mercado Ornelas, I.E. Cortes Mestizo, V.H. Méndez García. Aging spectral markers of tequila observed by Raman

spectroscopy, Eur, Food Res. Technol. 245 (2019) 1031-1036. https://doi.org/10.1007/s00217-018-3203-4.

[15] C. Frausto Reyes, C. Medina Gutiérrez, R. Sato Berrú, L.R. Sahagún, Qualitative study of ethanol content in tequilas by Raman spectroscopy and principal component analysis, Spectrochim. Acta A Mol. Biomol. Spectrosc. 61 (2005) 2657-2662. https://doi.org/10.1016/j.saa.2004.10.008.

[16] C. Fernández Lozano, M. Gestal Pose, G. Pérez Caballero, A.L. Revilla Vázquez, J.M. Andrade Garda, Quality Control in the Beverage Industry, in: A. Grumezescu, A.M. Holban (Eds.), Multivariate classification techniques to authenticate Mexican commercial spirits, Academic Press, 2019, pp. 259-287. https://doi.org/10.1016/B978-0-12-816681-9.00008-4.

[17] A. Gómez, D. Bueno, J.M. Gutiérrez, Electronic eye based on RGB analysis for the identification of tequilas, Biosensors 11 (2021) 68-83. https://doi.org/10.3390/bios11030068.

[18] G. Pérez-Caballero, J.M. Andrade, P. Olmos, Y. Molina, I. Jiménez, J.J. Durán, C. Fernández-Lozano, F. Miguel-Cruz, Authentication of tequilas using pattern recognition and supervised classification, Trends Anal. Chem. 94 (2017) 117-129. https://doi.org/10.1016/j.trac.2017.07.008.

[19] D.W. Lachenmeier, E. Richling, M.G. López, W. Frank, P. Schreier, Multivariate analysis of FTIR and ion chromatographic data for the quality control of tequila, J. Agric. Food Chem. 53 (2005) 2151-2157. https://doi.org/10.1021/jf048637f.

[20] U. Contreras, O. Barbosa García, J.L. Pichardo Molina, G. Ramos Ortíz, J.L. Maldonado, M.A. Meneses Nava, N.E. Ornelas-Soto, P.L. López-de-Alba, Screening method for identification of adulterate and fake tequilas by using UV–VIS spectroscopy and chemometrics, Food Res. Int. 43 (2010) 2356-2362. https://doi.org/10.1016/j.foodres.2010.09.001.

[21] C.H. Pérez Beltrán, V.M. Zuñiga Arroyo, J.M. Andrade, L. Cuadros Rodríguez, G. Pérez Caballero, A.M. Jiménez Carvelo, A sensor-based methodology to differentiate

pure and mixed White Tequilas based on fused infrared spectra and multivariate data treatment, Chemosensors 9 (2021) 47-59. https://doi.org/10.3390/chemosensors9030047.

[22] 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.

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

[24] 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.

[25] F. Stilo, A.M. Jiménez Carvelo, E. Liberto, C. Bicchi, S.E. Reichenbach, L. Cuadros Rodríguez, C. Cordero, Chromatographic fingerprinting enables effective discrimination and identitation of high-quality Italian Extra-virgin olive oils, J. Agric. Food Chem. 69 (2021) 8874-8889. https://doi.org/10.1021/acs.jafc.1c02981.

[26] 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. https://doi.org/10.1366/0003702053641450.

[27] 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.

[28] N.A. Mancilla Margalli, M.G. López, Generation of Maillard compounds from inulin during the thermal processing of Agave tequilana Weber var. azul, J. Agric. Food Chem. 50 (2002) 806-812. https://doi.org/10.1021/jf0110295

[29] A.C. Muñoz Muñoz, J.L Pichardo Molina, G. Ramos Ortíz, O. Barbosa García, J.L. Maldonado, M.A. Meneses Nava, N.E. Ornelas Soto, A. Escobedo, P.L. López de

Alba, Identification and quantification of furanic compounds in tequila and mezcal using spectroscopy and chemometric methods, J. Braz. Chem. Soc. 21 (2010) 1077-1087. https://doi.org/10.1590/S0103-50532010000600018.

[30] B.M. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Winding, R.S. Koch, Chemometrics Tutorial for PLS_Toolbox and Solo, Eigenvector Research, Inc. Wenatchee, WA, USA, 2006.

[31] T.K. Kataria, M.E. Sosa Morales, J.L. Olvera Cervantes, A. Corona Chavez, Dielectric properties of tequila in the microwave frequency range (0.5-20 GHz) using coaxial probe, Int. J. Food Prop. 20 (2017) S377-S384. https://doi.org/10.1080/10942912.2017.1297949.

4.5. Comunicación a congresos

C.H. Pérez-Beltrán, G. Pérez-Caballero, J.M. Andrade, L. Cuadros-Rodríguez, A.M. Jiménez-Carvelo. Discriminación de Tequilas Blancos '100% agave' y 'mixtos' mediante espectroscopia Raman con compensación espacial y análisis multivariable. XXII Reunión de la Sociedad Española de Química Analítica (SEQA), Oviedo, España, 2022. *Comunicación en formato Póster.*

A.M. Jiménez-Carvelo, C.H. Pérez-Beltrán, G. Pérez-Caballero, J.M. Andrade, L. Cuadros-Rodríguez. Comparación de técnicas analíticas no invasivas (FTIR, NIR & SORS) para la evaluación de calidad del tequila. XVII Reunión del Grupo Regional Andaluz de la Sociedad Española de Química Analítica (GRASEQA), Sevilla, España, 2022. *Comunicación en formato Póster.*

Capítulo 5

DISCUSIÓN INTEGRADA

5. DISCUSIÓN INTEGRADA

5.1. Visión general

El rápido aumento de la población mundial ha incrementado la demanda de alimentos, así como también la necesidad de asegurar su calidad, seguridad y sostenibilidad. Muchos productos alimenticios tienen gran importancia sobre la salud y sobre la economía del consumidor y del resto de la cadena alimentaria, lo cual hace que sean apreciados como productos alimenticios de alto valor, ya que pueden ayudar a prevenir enfermedades y/o por las ganancias económicas que reportan. Históricamente, productores y/o vendedores deshonestos han llevado a cabo actividades ilegales, como adulteración, falsificación (no conformidad) o contaminación que constituyen fraude alimentario, para obtener ganancias económicas ilícitas, las cuales pueden dañar la salud e intereses del consumidor cuando los productos alimenticios son alterados en cualquier medida.

Dos productos alimenticios de alto valor y con grandes volúmenes de exportación que han sufrido fraude ampliamente a través de los años hasta la actualidad, son el aceite de oliva y el tequila. De hecho, se ha detectado que la falsificación es la actividad ilícita más común de fraude alimentario para el aceite de oliva, en la cual se mezclan aceites de oliva de máxima calidad, como lo son el aceite de oliva virgen (AOV) y virgen extra (AOVE), con aceites de oliva de menor calidad y se reporta su contenido en la etiqueta como si fuesen exclusivamente AOV o AOVE. Por su parte, el tequila también sufre de problemas de falsificación, encontrando comúnmente casos de dilución con agua y casos de mal etiquetado en los cuales tequilas de clases inferiores (Tequila Oro o Reposado) se reportan como tequilas de clases superiores (Tequila Añejo o Extra-Añejo) debido a la semejanza de colores ocasionada por la adición de compuestos edulcorantes coloreados; adicionalmente, el tequila también sufre de problemas de adulteración, siendo el metanol el adulterante más habitual, lo cual ha resultado en problemas de salud letales para los consumidores.

Hoy en día, la detección de este tipo de adulteraciones y/o falsificaciones suele ser tardía, ya que se descubren una vez que los productos se encuentran en el mercado, así como también suele ser larga y laboriosa debido a que esta detección se realiza mediante métodos analíticos convencionales en laboratorios analíticos especializados.

Ante esta situación, hay una necesidad urgente de apoyar a los métodos analíticos convencionales (métodos de retaguardia) con métodos analíticos auxiliares de detección y cribado (métodos de vanguardia) que sean rápidos, eficientes y exactos para agilizar y mejorar los actuales procesos de aseguramiento y control de calidad del aceite de oliva y el tequila.

Es por lo anterior, que en esta tesis doctoral se han desarrollado diferentes métodos analíticos multivariable para realizar de manera ágil el control de la calidad alimentaria, los cuales puedan ser implementados de manera auxiliar tanto en puntos de venta y/o consumo como en los distintos laboratorios. Primeramente, en el **capítulo 1** se ponen de manifiesto los principios y fundamentos sobre los cuales están basados los métodos analíticos multivariable que se han desarrollado a lo largo de esta investigación. Seguidamente, en el **capítulo 2** se describen los inicios de la implementación del control de procesos en la industria alimentaria, así como el avance propiciado por la implementación del enfoque multivariable, el cual se demuestra a través de una recopilación bibliográfica de diferentes estudios científicos que siguen los lineamientos de la Calidad mediante el Diseño (QbD) y utilizan las herramientas quimiométricas reconocidas por la Tecnología Analítica de Procesos (PAT). Posteriormente, en el **capítulo 3** se describen las diferentes actividades llevadas a cabo para lograr el desarrollo de dos métodos analíticos multivariable (MAM) cualitativos basados en la aplicación de cromatografía de líquidos de altas prestaciones (HPLC) y herramientas quimiométricas de clasificación, con los cuales ha sido posible generar señales instrumentales independientes del instrumento analítico (denominadas, señales *agnostizadas*), mismas que pueden ser transferidas entre laboratorios para realizar pruebas comparativas de control de calidad alimentario. Por su parte, en el **capítulo 4** se proponen tres nuevos MAM basados en las técnicas espectroscópicas de infrarrojo medio con transformada de Fourier (FTIR), de infrarrojo cercano (NIR) y Raman con sistema de compensación espacial (SORS), conjuntamente con herramientas quimiométricas, para que

puedan ser implementados durante el proceso de producción o puntos de venta y/o consumo de estos productos alimenticios. Por último, en el presente **capítulo 5** se realiza una discusión de manera integrada de cada uno de los capítulos anteriores.

El objetivo final de desarrollar estos MAM es el de proporcionar a la sociedad y a la industria alimentaria con nuevas herramientas de menor complejidad experimental que permitan agilizar y realizar con mayor eficiencia el aseguramiento y control de calidad de productos alimentarios, ayudando a cuidar y mantener la salud del consumidor.

5.2. Nuevos métodos analíticos multivariable para el aseguramiento y control de calidad alimentaria

Como se constata en el apartado anterior, es necesario contar con nuevos métodos analíticos para combatir los constantes problemas de adulteración y falsificación de diferentes productos alimenticios. Es por ello que en esta tesis doctoral se han desarrollado dos tipos de métodos analíticos multivariable para controlar y verificar la calidad del aceite de oliva y del tequila: unos basados en cromatografía de líquidos y otros en técnicas espectroscópicas, y ambos conjuntamente con herramientas quimiométricas.

A continuación, se brinda una discusión integral sobre su desarrollo, optimización y comparación de resultados entre ellos, así como también la aplicabilidad que podrían tener en las industrias correspondientes, siguiendo los lineamientos establecidos por QbD y PAT.

5.2.1. Desarrollo y aplicación de nuevos métodos analíticos multivariable basados en cromatografía de líquidos y su transferencia mediante agnostización instrumental

Tanto el aceite de oliva como el tequila son productos altamente demandados en todo el mundo, por consecuencia, enormes volúmenes de estos productos se transportan tanto al interior como al exterior de España y México, respectivamente. Esta situación, hace necesaria la existencia de métodos analíticos en cada lugar donde se da la comercialización del aceite de oliva y tequila, con el objetivo de evaluar su calidad final y, así, asegurar que

no han sufrido alteración alguna durante su producción, transporte, almacenaje o cualquier otra etapa de la cadena alimentaria. No obstante, y a pesar de que existen métodos analíticos oficiales para el análisis y control de calidad de estos productos alimenticios, los resultados obtenidos son altamente dependientes de los instrumentos analíticos utilizados, cabiendo la posibilidad de obtener resultados distintos para los mismos productos.

Lo mismo sucede para los métodos analíticos multivariable (MAM) que están basados en la metodología de señales instrumentales (huellas) y modelos matemáticos quimiométricos; por lo que, es conveniente desarrollar MAM que generen señales instrumentales independientes del instrumento analítico con el cual han sido obtenidas, minimizando y eliminando los efectos que puedan ejercer sobre ellas. De esta manera, dichas señales instrumentales podrían ser transferidas entre laboratorios analíticos y/o ser transferidas a una base de datos única para construir un modelo matemático global para facilitar y agilizar la comparación de resultados de control de calidad de los productos alimenticios bajo estudio en distintos puntos de análisis.

Bajo esta premisa, se inició una **primera etapa** para lograr generar señales instrumentales transferibles a partir de un MAM basado en la aplicación de cromatografía de líquidos y herramientas quimiométricas. La técnica analítica utilizada fue cromatografía liquida de ultra altas prestaciones en fase normal acoplado a un detector ultravioleta visible ((NP)UHPLC-UV/Vis), con la cual fue posible desarrollar un método analítico capaz de analizar y obtener las señales cromatográficas de muestras de aceite de oliva en tan solo diez minutos. Los especímenes estuvieron conformados por muestras auténticas de aceite de oliva virgen (AOV), virgen extra (AOVE) y aceite de oliva (AO) (mezcla comercial de aceite de oliva refinado con una pequeña proporción de aceite de oliva virgen o virgen extra), así como también por muestras de aceite de oliva refinado (AOR), aceite de orujo de oliva (AOO) y por muestras conformadas por distintas mezclas y en distintas proporciones de AOV o AOVE con AOR y/o AOO. La selección de muestras se realizó de esta manera ya que las adulteraciones y falsificaciones más comunes del aceite de oliva suelen estar enfocadas hacia el AOV y AOVE con AOR y/o AOO; así, se pudo simular los casos más comunes de fraude alimentario de la vida real.

Una vez obtenidas las señales cromatográficas, el estudio siguió dos rutas para realizar una comparación de sus resultados finales. En la primera, se utilizaron las señales cromatográficas originales, las cuales fueron internamente alineadas aplicando únicamente el algoritmo '*icoshift*' [**1**]; en la segunda, las señales cromatográficas originales fueron normalizadas en intensidad y tiempo mediante la metodología de *agnostización instrumental*, con la finalidad, en ambos casos, de generar modelos matemáticos capaces de detectar las adulteraciones del aceite de oliva anteriormente descritas.

La selección del algoritmo '*icoshift*' fue debida a que es el algoritmo empleado por defecto en aplicaciones cromatográficas para eliminar los desplazamientos observados en los tiempos de retención entre diferentes análisis cromatográficos. Este alineamiento de tiempos de retención es útil para la identificación y cuantificación de los picos, pero es especialmente importante como un paso en el tratamiento de las señales instrumentales para mantenerlas invariantes entre muestras, permitiendo desarrollar modelos matemáticos con capacidades predictivas adecuadas de clasificación y/o regresión, en los cuales la interpretación química de los compuestos no se ve comprometida.

No obstante, el alineamiento de estas señales instrumentales y los modelos matemáticos que se generen a partir de ellas están limitados al instrumento analítico y lugar de trabajo en donde fueron obtenidos, por lo que, no es posible realizar la transferencia de estas señales instrumentales a otros laboratorios analíticos para comparar resultados ni para utilizarlas con modelos matemáticos construidos con señales instrumentales adquiridas con instrumentos analíticos diferentes. Con la finalidad de dar solución a esta problemática, se aplicó la metodología de *agnostización instrumental* a las señales cromatográficas originales de las diferentes muestras de aceite de oliva antes descritas con el objetivo de comprobar, posteriormente, si esta metodología generaría los mismos resultados.

Durante la fase experimental fue importante la selección de una sustancia química que sería agregada a cada una de las muestras a analizar, ya que sería considerada como patrón interno (PI) a partir del cual la intensidad de cada una de las señales cromatográficas sería normalizada. Así mismo, fue primordial una selección cuidadosa y adecuada de un

[1] G. Tomasi, F. Savorani, S.B. Englensen, *i*coshift: An effective tool for the alignment of chromatographic data, 2011, Journal of Chromatography A, 1218, 7832-7840.

conjunto de sustancias químicas que conformarían la mezcla patrón externa (MPE) –la cual debía de contener compuestos químicos con características similares a la muestra y cubrir todo el rango de tiempos de retención de la señal cromatográfica original– que es utilizada para realizar la traslación del dominio de tiempo de retención a un dominio de valores estandarizados de tiempo, constantes independientes del sistema (SRSs, '*standard retention scores*').

Del mismo modo, fue necesario asegurar el buen comportamiento del sistema cromatográfico mediante el análisis de la MPE al inicio y al final de cada tanda de análisis, ya que de esta manera se aseguró que las señales instrumentales no variasen a lo largo de cada tanda de análisis.

Una vez hecha la selección de las diferentes sustancias químicas de interés y realizados los análisis, cada una de las señales cromatográficas fueron *agnostizadas* de manera individual y, a partir de ellas, se llevó a cabo la elaboración de modelos multivariable, tal como se detalla en el **capítulo 3** (subsección **3.1**, apartado **2.5**). El establecimiento de los modelos matemáticos desarrollados, tanto con señales cromatográficas originales como con las señales cromatográficas *agnostizadas*, se inició con análisis exploratorio mediante análisis de componentes principales (PCA) para constatar la estructura del conjunto de señales y estudiar el comportamiento de cada una de las muestras, y se continuó con análisis de clasificación, haciendo uso de las herramientas quimiométricas SIMCA, PLS-DA y SVM, para distinguir aceites de oliva auténticos de aquellos que fueron adulterados con aceites de oliva de menor calidad. Un hallazgo importante de este estudio fue que los resultados obtenidos para el análisis exploratorio y de clasificación con los datos derivados de las señales originales y *agnostizadas* fueron muy similares. Mediante el PCA se observó que las muestras de cada una de las clases presentaron una distribución ligeramente distinta en la gráfica de '*scores*', pero mantuvieron un comportamiento y agrupamiento similar. Por su parte, los modelos matemáticos de clasificación PLS-DA y SVM presentaron resultados idénticos al utilizar los datos de las señales cromatográficas originales o *agnostizadas*, pero con SIMCA mejoraron substancialmente, ya que al desarrollar dicho modelo matemático con las señales originales se obtuvo una especificidad de 0.23, mientras que con las señales *agnostizadas* de 0.85.

Con los resultados descritos anteriormente, se comprobó que la metodología de *agnostización instrumental* es útil para generar señales cromatográficas independientes del instrumento analítico con las cuales se pueden desarrollar modelos matemáticos de clasificación para identificar muestras de aceite de oliva que han sido adulteradas, proporcionando resultados similares o mejores que al utilizar los datos cromatográficos originales. Esto presenta un hecho sumamente importante, ya que se establece la posibilidad de generar bases de datos globales de distintos productos alimenticios, con lo cual se podrían desarrollar métodos analíticos multivariable universales, agilizando el control de calidad en distintos laboratorios.

Una vez comprobada la funcionalidad de la metodología de *agnostización instrumental*, se continuó con una **segunda etapa** para comprobar la hipótesis de transferencia de señales cromatográficas entre laboratorios. Este segundo estudio consistió en el análisis mediante cromatografía de líquidos de altas prestaciones en fase reversa acoplado a un detector ultravioleta-visible (HPLC-UV/Vis) de muestras de Tequila Blanco para obtener sus señales cromatográficas, las cuales se obtuvieron en dos laboratorios distintos, ubicados en España (ES) y México (MX), respectivamente.

El objetivo de este estudio requería la obtención de señales cromatográficas de la misma muestra, pero de diferentes laboratorios en condiciones de reproducibilidad para aplicar sobre ellas la metodología de *agnostización instrumental* y, con ellas, generar una base de datos global a partir de la cual construir modelos matemáticos capaces de agilizar los procesos de control de calidad intra e interlaboratorio. En este caso se seleccionó al tequila como matriz de estudio, específicamente, la clase Tequila Blanco en sus dos categorías '100 % agave' y 'mixto', ya que, por un lado, es la clase que más se comercializa y, por otro, estas dos categorías difieren significativamente en precio una de la otra y la única manera de distinguir entre ellas es a través del etiquetado de la botella, puesto que en apariencia son exactamente iguales.

Del mismo modo que en el estudio previo del aceite de oliva, este estudio comenzó con el desarrollo del método de análisis de los tequilas y con la detección del compuesto químico que sería utilizado como PI en cada una de las muestras a analizar. Para ello, se seleccionaron los compuestos químicos que conformarían la MPE y se desarrolló su

método de análisis, teniendo especial cuidado que la duración de su cromatograma incluyera la totalidad de la señal cromatográfica de los tequilas. Posteriormente, se dio inicio a las tandas de análisis en las cuales se comenzó con el análisis de la MPE, seguido del análisis de las muestras de tequila, y finalizó con el análisis de la MPE, nuevamente. Ejemplos del cromatograma de la MPE y de la señal cromatográfica original de una muestra de Tequila Blanco '100 % agave' se pueden apreciar en la **Figura 12 (a)** y **(b)**, respectivamente.

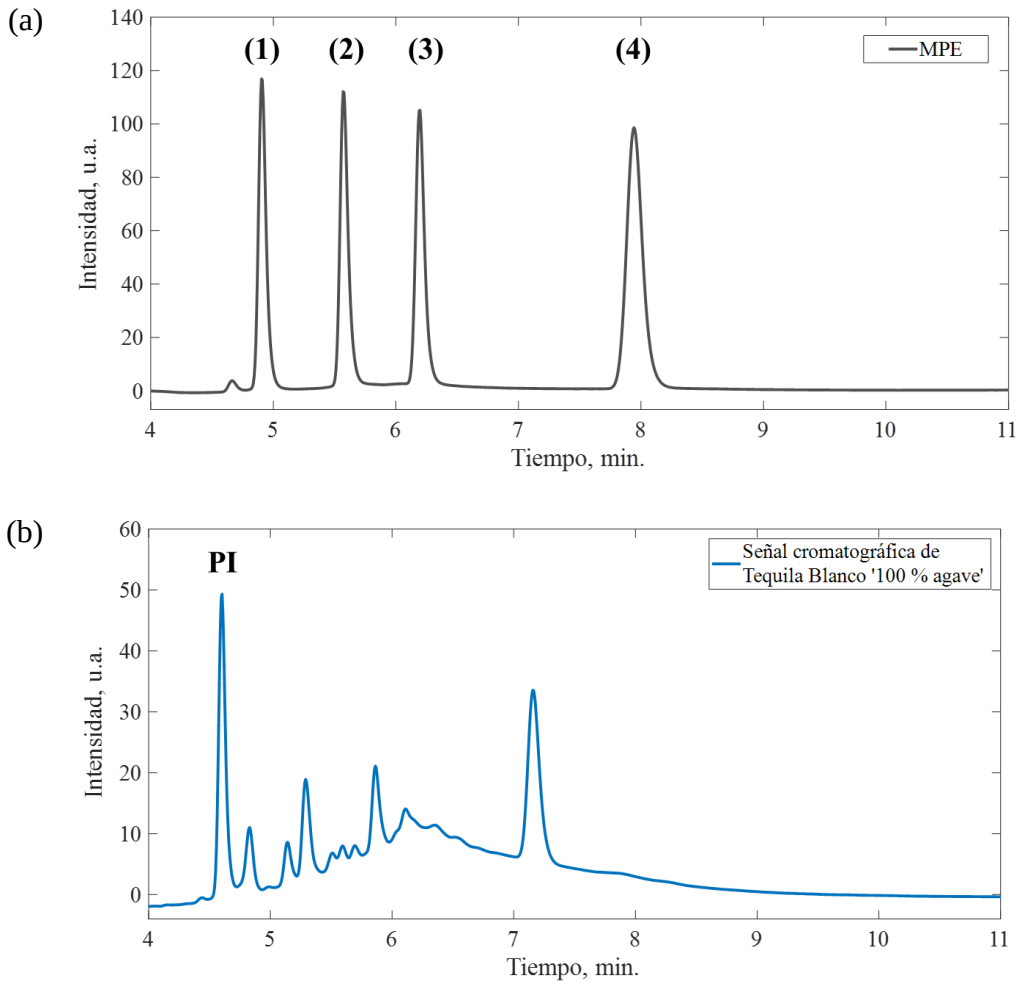


Figura 12. Ejemplos de (a) cromatograma de la mezcla patrón externa (MPE) empleada en el desarrollo del método analítico multivariable del tequila y (b) señal cromatográfica original de una muestra de Tequila Blanco '100 % agave'.
La MPE estuvo conformada por los compuestos (1) 5-(hidroximetil)furfural, (2) furfural, (3) 2-acetilfurano y (4) 2-acetil-5-metil furano. La señal cromatográfica contó con la presencia del patrón interno (PI) 5-(hidroximetil)furfural).

El análisis de la MPE al inicio y al final de cada tanda de análisis fue de utilidad para monitorear el control de calidad de los análisis y el funcionamiento del cromatógrafo utilizado en cada caso.

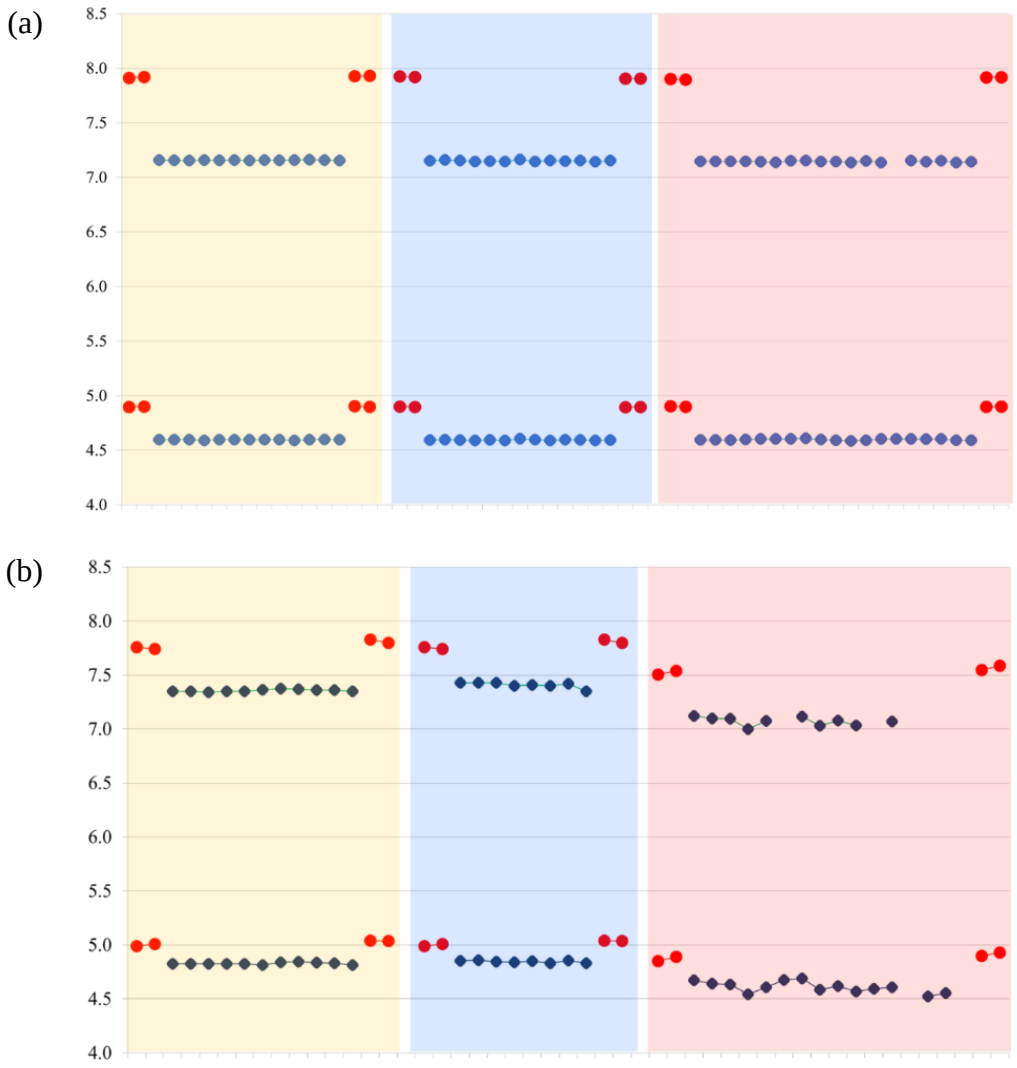


Figura 13. Ejemplos de cuadros control del (a) sistema de referencia y del (b) sistema subordinado creados a partir de tres tandas de análisis diferentes.
Los puntos rojos alrededor de los minutos 5 y 7.5 – 7.9 representan al 1er compuesto (5-(hidroximetil)furfural) y 4to compuesto (2-acetil-5-metil furano) de la mezcla patrón externa, respectivamente; mientras que los puntos azules alrededor de los minutos 4.5 – 4.7 y 7.0 – 7.49 representan al patrón interno (5-(hidroximetil)furfural) presente en la señal cromatográfica y al último compuesto químico de la misma señal cromatográfica.
*Ver **Figura 12** para obtener una referencia visual de los compuestos mencionados.*

Conforme se obtenían las señales cromatográficas, se verificaba la calidad de los datos obtenidos a través de los cuadros control elaborados para cada uno de los laboratorios en España (Mg, *'manager system'* o sistema de referencia) y México (Wf, *'workforce system'* o sistema subordinado), los cuales se pueden apreciar en las **Figuras 13 (a)** y **(b)**, respectivamente.

Tanto el cuadro control del sistema Mg como el del Wf fueron elaborados considerando los tiempos de retención de un compuesto químico inicial y otro final, tanto de la MPE como de la señal cromatográfica. A partir de estos cuadros control se puede apreciar que el sistema Mg tuvo un comportamiento constante y adecuado cada día de análisis tanto para la MPE como para las señales cromatográficas, mientras que el sistema Wf presentó ligeras diferencias en el inicio y duración de las señales cromatográficas de los tequilas, pero sí mantuvo un perfil más similar para los cromatogramas de la MPE.

Debido a esta situación, no fue posible implementar de primera instancia la metodología de *agnostización instrumental* sobre las señales cromatográficas del sistema Wf, ya que la integridad de los datos debió ser asegurada, primeramente. Para ello, se desarrolló la **función de equiparación** ('*equity function*'), tal como se describe en el **capítulo 3** de esta tesis doctoral (subsección **3.2**, apartado **2.0**). Al aplicar dicha función sobre las señales cromatográficas del sistema Wf fue posible, en primer lugar, comprobar cuantitativamente a través de los parámetros de **relación de tiempo de análisis** (**RtR**, *'runtime ratio'*) y **desfase de inicio de análisis** (**StL**, *'starting-time lag'*) que dichas señales cromatográficas sufrieron un tipo de efecto de alargamiento debido a una posible falla aleatoria del sistema de bombeo del instrumento analítico y, en segundo lugar, disminuir dicho defecto de cada una de las señales cromatográficas para que pudieran ser consideradas en la siguiente etapa del tratamiento de datos.

Después de la corrección de las señales cromatográficas del sistema Wf, se procedió a realizar la agnostización de las señales cromatográficas de ambos sistemas Wf y Mg –véase **capítulo 3** (subsección **3.2**, apartado **3.4**)–. Posteriormente, se realizaron distintos estudios de similitud entre 10 señales cromatográficas de tequila (5 TB y 5 TBM) utilizando el índice de proximidad (NEAR) y el ángulo coseno (COS), para verificar y confirmar la integridad de las nuevas señales corregidas y *agnostizadas* antes de proceder con la

construcción de la base de datos única de Tequila Blanco. En este sentido, se comprobó que la similitud entre las señales instrumentales antes de la corrección era muy pobre, con un promedio de NEAR = 0.354 y de COS = 0.470, así como también se demostró que después de su corrección y *agnostización* la similitud entre las señales cromatográficas se vio aumentada considerablemente, con un promedio de NEAR = 0.689 y COS = 0.828, lo cual se consideró satisfactorio para el siguiente desarrollo de los modelos matemáticos.

Finalmente, se construyeron modelos matemáticos para comprobar la autenticidad entre Tequilas Blanco '100 % agave' y 'mixto' a partir de la base de datos de señales cromatográficas *agnostizadas* del sistema Mg. Posteriormente, se evaluó la autenticidad de Tequilas Blancos utilizando sus señales cromatográficas obtenidas con el sistema de Wf, las cuales fueron introducidas a los modelos matemáticos desarrollados con los datos del sistema Mg. Con ello, se buscó poner a prueba la funcionalidad de la metodología de *agnostización instrumental* para la creación de una base de datos global y para la transferencia de señales instrumentales *agnostizadas* a aplicar en métodos analíticos multivariable, construido sobre la base de un modelo único y universal.

Las herramientas quimiométricas utilizadas para desarrollar los modelos matemáticos de clasificación en este estudio fueron PLS-DA y SVM. En una primera instancia, la construcción de todos los modelos matemáticos se realizó con la base de datos del sistema Mg, dividida en conjunto de entrenamiento (32 muestras) y conjunto de validación (9 muestras). Dichos modelos fueron evaluados y retroalimentados con otros conjuntos de validación del sistema Mg, siguiendo los pasos del diagrama de flujo de la **Figura 14**.

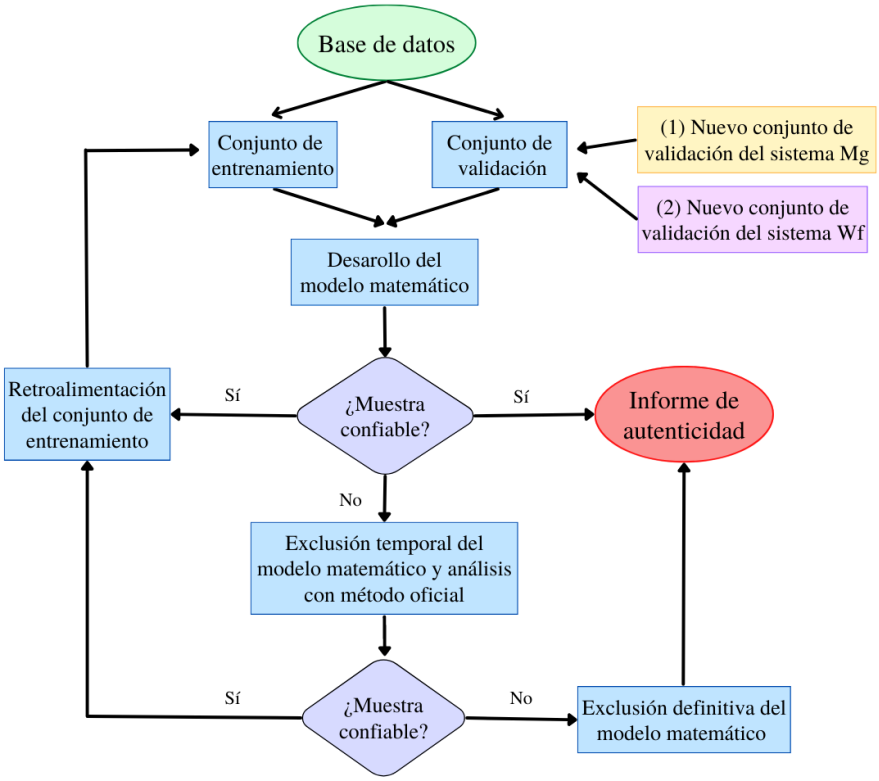


Figura 14. Diagrama de flujo del proceso de desarrollo y retroalimentación de modelos matemáticos a partir de señales cromatográficas *agnostizadas* del sistema de referencia (Mg, manager) y del sistema subordinado (Wf, worforce).

Debido a que se propone el uso de este método multivariable en aplicaciones analíticas reales, es necesario establecer *a priori* unos requisitos mínimos de validación para asegurar que dicho método será apto para su uso previsto. Para ello, se debe considerar un parámetro poblacional, llamado ocurrencia [**2**], el cual índica la relación del número de muestras de interés contra el número total de muestras. Con este parámetro, se debe hacer el cálculo de los parámetros de calidad en el desempeño de sensibilidad (SENS) y precisión (PREC), los cuales deben ser igualados o superados por el modelo matemático bajo desarrollo. A su vez, también se deben de establecer el índice de ahorro (I_{SAVING}) y el índice de error de asignación (I_{ERROR}), los cuales informan sobre los ahorros económicos que tendrían lugar

[2] A.M. Jiménez-Carvelo, L. Cuadros-Rodríguez, The occurrence: a meaningful parameter to be considered in the validation of multivariate classification-based screening methods – application for authenticating virgin olive oil, 2020, Talanta, 208, 120467.

en el laboratorio analítico al no realizar análisis confirmatorios a muestras correctamente clasificadas y sobre el riesgo de clasificar incorrectamente una muestra de cualquiera de las dos clases, respectivamente.

Así pues, con la ocurrencia de las muestras TB de 0.56 se establecieron unos requisitos mínimos de validación de I_{SAVING} de 60 % y de I_{ERROR} de 15 %, con los cuales se obtuvieron unos valores esperados de SENS y PREC de 0.85 y 0.75, respectivamente.

Una vez establecidos los requisitos mínimos de validación, se realizó la validación de los modelos matemáticos PLS-DA y SVM, obteniendo resultados excelentes de 1.00 para la SENS y PREC, lo que indicó que ambos modelos podían seguir siendo utilizados para las siguientes evaluaciones. En este sentido, las nueve muestras del conjunto de validación pasaron a formar parte del conjunto de entrenamiento, retroalimentando el modelo matemático con dichas señales cromatográficas de muestras de tequilas. Con la finalidad de mejorar la habilidad predictiva de clasificación del modelo matemático, se evaluaron 26 señales cromatográficas *agnostizadas* adicionales de tequila, las cuales fueron analizadas con el mismo sistema Mg. En esta ocasión, el modelo matemático construido con PLS-DA clasificó 4 muestras TB como TBM y 6 TBM como TB, mientras que el modelo matemático con SVM clasificó 4 muestras TB como TBM y 7 TB como TBM. Estos resultados indican que dichas muestras deben pasar por un proceso de confirmación mediante métodos analíticos tradicionales, ya que presentan características más similares a la clase contraria.

Por último, se procedió a evaluar la autenticidad de las mismas 41 muestras de Tequila Blanco analizadas con el sistema Wf, cuyas señales cromatográficas fueron previamente corregidas y *agnostizadas*. Una vez hecha la evaluación correspondiente, se encontró que el modelo matemático construido con PLS-DA identificó a 7 muestras de TB como TBM y a 14 TBM como TB, mientras que el modelo matemático basado en SVM identificó 3 muestras TB como TBM y 14 TBM como TB –véase **capítulo 3** (subsección **3.2**, apartado **4.2**)–. Debido a estos resultados, se decidió optar por el modelo matemático construido con SVM como mejor opción para la clasificación de nuevas muestras, cuyas señales cromatográficas fueron obtenidas con un instrumento analítico distinto en un laboratorio diferente.

Si bien es cierto que hay muestras que parecen estar mal clasificadas, estos resultados pueden interpretarse en la vida real en que solamente se requiere el análisis confirmatorio de 17 muestras mediante métodos analíticos oficiales, reduciendo la carga de trabajo en un 60 %, lo cual coincide con el I_{SAVING} inicialmente establecido. Esto se traduce en ahorro de tiempo y recurso económico que puede implementarse para el análisis de más muestras, incrementando la posibilidad de detectar muestras adulteradas y/o falsificadas.

Un aspecto sumamente importante a tener en cuenta previo a la aplicación de la metodología de *agnostización instrumental*, que ha quedado constatado en este segundo estudio, es la calidad e integridad de las señales cromatográficas, las cuales se obtienen con un correcto funcionamiento del instrumento analítico, ya que a partir de ellas se genera la base de datos global con la cual se elaboran los modelos matemáticos únicos capaces de comparar la calidad de una misma muestra analizada en diversos laboratorios.

En este sentido, se pone de manifiesto una nueva aportación a la ciencia con la cual realizar transferencia de señales cromatográficas (cromatogramas) que alimentan un único método analítico multivariable basado en cromatografía de líquidos haciendo uso de la metodología de *agnostización instrumental* y, cuando sea necesario, de la función de equiparación. Con esto, se presentan nuevas oportunidades y mejoras para asegurar el control de calidad de los productos a nivel mundial, lo cual permitirá combatir el fraude de productos alimenticios de una manera más ágil y eficiente.

5.2.2. Desarrollo de nuevos métodos analíticos multivariable complementarios para el control de calidad alimentario basados en técnicas no destructivas

Las adulteraciones y falsificaciones de los productos alimenticios de alto valor pueden producirse de distintas maneras, en distintos lugares y con diferentes niveles de sofisticación. Por ello, adicionalmente al desarrollo de los dos métodos analíticos multivariable descritos en el apartado anterior, también se crearon otros tres métodos analíticos multivariable basados en las técnicas no destructivas derivadas de las espectroscopías FTIR, NIR y SORS.

El **primer MAM** se basó en la aplicación de la espectroscopía FTIR y herramientas quimiométricas de clasificación, con el cual fue posible diferenciar entre las categorías '100 % agave' (TB) y 'mixto' (TBM) del Tequila Blanco. Inicialmente, dicha tarea se realizó con los espectros originales transformados de transmitancia a absorbancia y, aún con la ayuda de la quimiometria, no se obtenían resultados favorables, ya que los espectros de ambas categorías eran idénticos –véase **capítulo 4** (subsección **4.1**, apartado **3**)–.

Posteriormente, se advirtió que los espectros de ambas categorías de Tequila Blanco presentaban una ligera diferencia en la banda ubicada alrededor de 1045 cm^{-1}, la cual era un poco más ancha para la categoría '100 % agave', ya que se extendía hasta 954.5 cm^{-1}, mientras que para la categoría 'mixto' terminaba en 957.8 cm^{-1}.

Con esto, se procedió a realizar manualmente dos tipos de corrección de línea base de los espectros, manipulando lo menos posible cada espectro para evitar introducir variaciones en los datos y evitar perder de información valiosa. Estas correcciones fueron: (1) exactamente en los mismos puntos independientemente de la categoría de tequila (4000, 957, and 450 cm^{-1}), y (2) en función de la categoría (TB: 4000, 1854, 954.5, and 450 cm^{-1}; TBM: 4000, 1854, 957.8, and 450 cm^{-1}). Dichos datos fueron utilizados por separado para la elaboración de distintos modelos matemáticos, obteniendo mejores resultados con los datos cuya línea base fue corregida según la categoría de tequila.

No obstante, dicha solución no resultaría practica al implementar el método analítico en los laboratorios de rutina, pues la categoría de la muestra problema suele ser desconocida y, por tanto, no existe una selección idónea para corregir la línea base. Ante esta situación, se optó por realizar la fusión a bajo nivel de los espectros resultantes de aplicar ambas correcciones de la línea base, en el rango 1800-450 cm^{-1}, puesto que fue la región espectral que mejores resultados presentó en estudios previos, con lo cual se obtuvo una matriz fusionada más pequeña usada para desarrollar los siguientes modelos matemáticos.

En este estudio se evaluaron las herramientas quimiométricas SIMCA y PLS-DA, así como también la utilización de distintos métodos de pre-procesamiento. La mejor solución fue obtenida aplicando autoescalado a los datos, ya que se le otorgó la misma importancia a todas las variables que conformaban el espectro. Asimismo, mediante el uso de PLS-DA se logró clasificar correctamente todas las muestras del conjunto de validación con resultados

para las métricas de calidad en el desempeño de SENS, SPEC, PREC y NPV de 1.00 para todos ellas, mientras que los valores encontrados para SIMCA fueron 0.89, 0.67, 1.00 y 1.00, respectivamente.

Una vez establecido el MAM basado en datos FTIR fusionados, se procedió a la segunda etapa inicialmente planteada, la cual consistía en realizar una segunda fusión, también de bajo nivel, de los espectros obtenidos con la técnica NIR para complementar el rango espectral del infrarrojo y realizar modelos matemáticos con dichos datos. Un ejemplo de un vector resultante de la fusión de los datos de las técnicas FTIR y NIR se puede apreciar en la **Figura 15**. No obstante, los resultados de los modelos matemáticos no presentaban mejoría alguna sobre los resultados comentados anteriormente.

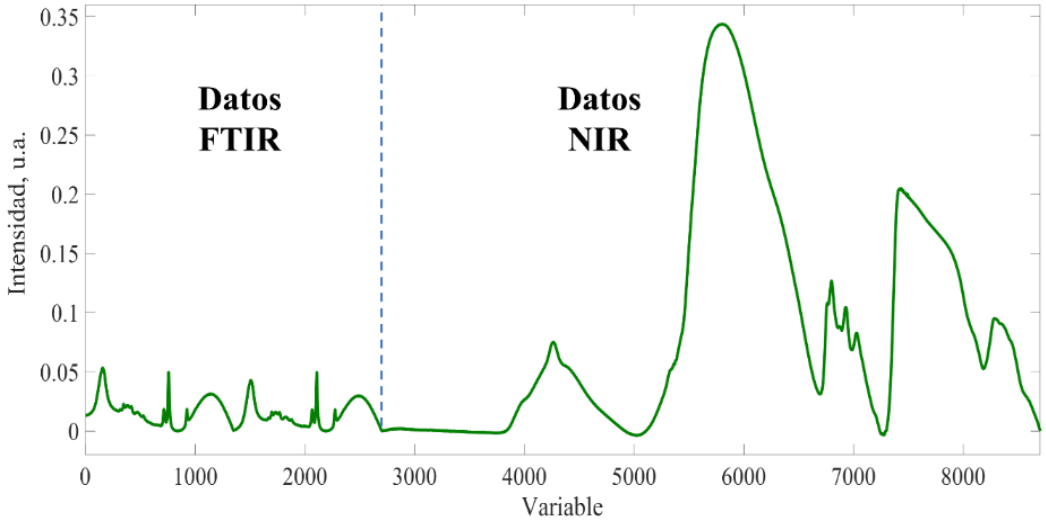


Figura 15. Ejemplo de un vector con datos fusionados de las técnicas espectroscópicas de infrarrojo medio con transformada de Fourier (FTIR) e infrarrojo cercano (NIR).

Fue por lo anterior, que se planteó desarrollar un **segundo MAM** basado únicamente en la técnica NIR con la finalidad de distinguir entre las mismas dos categorías de Tequila Blanco y para predecir cuantitativamente su contenido alcohólico, tal como se describe en el **capítulo 4** (subsección **4.2**, apartado **1**). Para ello, las señales instrumentales NIR, al igual que las señales instrumentales FTIR, fueron transformadas a absorbancia y corregidas en su línea base, pero en esta ocasión se realizó la corrección en los mismos puntos independientemente de su categoría. Dicho proceso de corrección de línea base fue

identificado como esencial al hacer uso de estas técnicas espectroscópicas para el análisis de los tequilas, ya que durante las etapas iniciales de estos estudios se trabajó con los datos originales, obteniendo resultados deficientes. Este es el motivo por el cual se recomienda altamente el uso de este método de pre-procesamiento a las señales instrumentales FTIR y/o NIR de tequilas después de realizar los respectivos análisis espectroscópicos.

Posteriormente, se realizaron los diversos análisis de datos multivariable en los cuales se incluyeron análisis exploratorios con PCA y PLS, análisis de clasificación con SVM, PLS-DA, kNN y SIMCA, y análisis de cuantificación del contenido alcohólico con PLSR y SVMR. Inicialmente, el análisis exploratorio con PCA permitió identificar muestras anómalas y una clara tendencia en el agrupamiento de las muestras, el cual atendía al contenido alcohólico de cada una de ellas; mientras que al realizar la exploración con PLS se distinguieron claramente dos grupos, uno para cada categoría de Tequila Blanco. Enseguida, se realizaron los modelos matemáticos de clasificación aplicando SVM y PLS-DA que presentaron valores de 1.00 en cada una de las métricas de calidad en el desempeño (SENS, SPEC, PREC y NPV). Por último, en este estudio fue posible realizar las cuantificaciones del contenido alcohólico de las muestras de tequila, ya que dicha información fue facilitada por el Consejo Regulador del Tequila (CRT) a través de la Universidad Nacional Autónoma de México (UNAM). Para estos análisis, se utilizaron 6 muestras menos que en los análisis de clasificación, debido a que se desconocía la información de su contenido alcohólico, por lo que se excluyeron del modelo matemático. Una vez realizadas las predicciones, se encontró que el modelo matemático desarrollado con SVMR era capaz de predecir un poco mejor y confiable el contenido alcohólico de las muestras desconocidas que el modelo matemático desarrollado con PLSR, ya que obtuvieron coeficientes de determinación (R^2) de 1.00 y 0.97, respectivamente. Lo mismo se pudo comprobar a través de otras 5 métricas de calidad en el desempeño, sugeridas en las prácticas estandarizadas para la validación de calibraciones multivariable realizadas empíricamente (ASTM E2617 [**3**]).

Un aspecto destacable de las etapas de análisis mencionadas anteriormente, es que fue necesaria la utilización de diferentes métodos de pre-procesado en una de ellas, ya que los

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

métodos que funcionaron en los análisis exploratorios y de predicción del contenido alcohólico no funcionaron durante los análisis de clasificación. Por esta razón, se realizó un estudio exhaustivo de diferentes pre-procesados que podrían funcionar en cada caso, concluyendo que, por un lado, los métodos de filtro de Whitaker y 1ª derivada fueron útiles en los análisis exploratorios y de predicción del contenido alcohólico debido a que el filtro Whitaker contribuyó a una corrección adicional de la línea base de los espectros, mientras que la 1ª derivada promovió la eliminación de señales sin importancia de la misma línea base [**4**,**5**].

Por otro lado, los métodos de normalización (área = 1) y centrado en la media en conjunto con la selección del rango espectral 6000 - 4000 cm^{-1} fueron los que permitieron obtener los mejores resultados durante los análisis de clasificación. En primer lugar, la reducción de variables (4000 variables) permitió eliminar grandes cantidades de interferencias asociadas mayoritariamente al agua presente en cada una de las muestras de tequila; la normalización de estos espectros reducidos permitió corregir los efectos de escala que se presentaban en esta región, ocasionados por efectos comúnmente presentes en técnicas espectroscópicas, como el efecto de longitud de trayectoria (*'pathlength'*), de dispersión (*'scattering'*) o variaciones en la sensibilidad del detector [**6**]; mientras que con el centrado en la media se hicieron evidentes las diferencias entre las distintas señales instrumentales de las muestras analizadas [**4**], lo cual permitió a los modelos matemáticos obtener buen desempeño al distinguir entre Tequilas Blanco '100 % agave' y 'mixto'.

Dados estos resultados, se puede concluir que el desarrollo de este segundo MAM basado en NIR fue menos eficiente, ya que las señales instrumentales obtenidas con dicha técnica analítica requirieron de un mayor y complejo tratamiento estadístico para lograr acceder a la información verdaderamente importante. No obstante, sigue siendo una buena opción como método de análisis auxiliar de cribado para el control y aseguramiento de calidad del tequila.

[4] B.M. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Winding, R.S. Koch, Chemometrics Tutorial for PLS_Toolbox and Solo, Eigenvector Research, Inc., WA, USA, 2006.

[5] P.H.C Eilers, A perfect smoother, 2003, Analytical Chemistry, 75, 3631-3636.

[6] K. Wang, G. Chi, R. Lau, T. Chen, Multivariate calibration of near infrared spectroscopy in the presence of light scattering effect: a comparative study, 2011, Analytical Letters, 44, 824-836.

Una vez desarrollados los dos MAM basados en espectroscopia de infrarrojo, se decidió estudiar la posibilidad de elaborar un **tercer MAM** que pudiera ser implementado para evaluar la calidad del tequila y aceite de oliva en puntos de venta y/o consumo. La técnica espectroscópica elegida para su elaboración fue la técnica analítica denominada espectroscopía Raman con sistema de compensación espacial (SORS), con la cual se presentan nuevas oportunidades en el ámbito de la calidad alimentaria [**7**,**8**], ya que se podría aumentar la incidencia de detección de bebidas alcohólicas adulteradas o falsificadas, no solo en el laboratorio analítico, sino justo en el lugar que pueden llegar a ser adquiridas o ingeridas por el consumidor.

Este estudio se inició con un análisis de similitud, utilizando el índice NEAR, entre espectros Raman de muestras de tequila obtenidos a través de sus botellas originales y espectros Raman obtenidos a través de viales ámbar –véase **capítulo 4** (subsección **4.3**, apartado **3)**–. Dicho análisis se llevó a cabo previamente al desarrollo del MAM para asegurar que esta metodología pudiera ser transferible a cualquier otra situación. De hecho, se comprobó, con valores NEAR > 0.92, que la huella instrumental obtenida a través de las botellas originales era prácticamente la misma que la obtenida a través de los viales ámbar, por lo que se decidió continuar con el desarrollo del correspondiente MAM.

Primero, se realizaron las mediciones espectroscópicas de las muestras de tequila, las cuales se encontraban en viales ámbar para asimilar los contenedores de vidrio opaco en los cuales se encuentran algunos tequilas en el mercado. Enseguida, se probaron diferentes métodos de pre-procesado, encontrando que los métodos de suavizado y centrado en la media ayudaban a remover el ruido de las señales instrumentales y a acentuar las diferencias entre ellas, respectivamente [**4**], que dieron lugar a una buena clasificación de ambas categorías de tequila. Los modelos matemáticos de clasificación fueron desarrollados con SIMCA siguiendo dos modalidades (según se consideraba una clase de entrada ‹1iC-SIMCA› o dos clases de entrada ‹2iC-SIMCA›), PLS-DA y SVM, siendo estos últimos dos métodos

[7] 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, 2023, Food Control, 146, 109522.

[8] A.M. Jiménez Carvelo, P. Li, S.W. Erasmus, H. Wang, S.M. Van Ruth, Spatial-temporal event analysis as a prospective approach for signaling emerging food fraud-related anomalies in supply chains, 2023, Foods, 12, 61.

discriminantes los que clasificaron correctamente la totalidad de las muestras del conjunto de validación.

Por último, se realizó la predicción del contenido alcohólico únicamente con PLSR, debido a que se buscó la opción más sencilla sin dejar de ser confiable. Con ello, se obtuvieron predicciones bastantes cercanas a los valores reales otorgados por el CRT, lo cual se evidenció a través del parámetro de desviación estándar de los residuos de validación, SDV = 2.65, así como también se obtuvo un buen ajuste de los datos con un R^2 de 0.97. De esta manera, quedó cumplimentado el tercer MAM mediante el cual se constata la facilidad para realizar mediciones espectroscópicas en los distintos puntos de venta y/o consumo del tequila u otros productos, ya que dichos análisis pueden realizarse a través de botellas de vidrio opacas en las cuales se comercializa esta bebida alcohólica.

Tal como se ha mostrado en este apartado, se desarrollaron tres nuevos métodos analíticos multivariable que podrían ser utilizados para el aseguramiento y control de calidad de las categorías del Tequila Blanco. Los tres MAM estuvieron basados en técnicas analíticas no invasivas y en la aplicación de herramientas quimiométricas con diferentes métodos de pre-procesamiento para cada uno de ellos.

De todas las herramientas quimiométricas utilizadas para la diferenciación del Tequila Blanco '100 % agave' y 'mixto', PLS-DA y SVM otorgaron los mejores resultados. Sin embargo, de estas dos técnicas de reconocimiento de pautas supervisado, se recomienda el uso de PLS-DA para su aplicación en la vida real debido a la sencillez de sus cálculos matemáticos, lo cual hace que el desarrollo de su modelo matemático sea más rápido y que pueda ser realizado con ordenadores de características más comunes, como lo puede ser un ordenador portátil. Por su parte, el uso de SVM sigue siendo aconsejable cuando se requiera verificar los resultados de PLS-DA.

De igual manera, dichos resultados no hubieran sido favorables sin el adecuado tratamiento de los datos obtenidos de las tres técnicas espectroscópicas. Esto quedó demostrado en cada uno de los estudios, puesto que cada tipo de datos requirió un método de pre-procesamiento diferente.

De esta manera, se podría concluir que un MAM es mejor o peor que otro, sin embargo, no debe de olvidarse que dichos MAM fueron desarrollados para brindar alternativas a la industria alimentaria y sus partes involucradas, para realizar un aseguramiento y control de calidad más eficiente y para que puedan ser empleados por organismos evaluadores de la conformidad en sus propios laboratorios analíticos o en los puntos de venta y/o consumo, así como también para que puedan ser implementados por la industria como parte de sus procesos de aseguramiento y control calidad siguiendo los lineamientos de QbD y PAT.

En este sentido, en el siguiente apartado se comentan algunos beneficios y propuestas de implementación de dichos MAM a lo largo de las cadenas alimentarias del tequila y aceite de oliva.

5.3. Implicación en la industria alimentaria

Habiéndose puesto de manifiesto el aumento de actividades ilícitas de fraude alimentario a lo largo de esta tesis doctoral, como la adulteración y falsificación de productos alimenticios de alto valor, es evidente la necesidad de contar con nuevos métodos analíticos auxiliares para aumentar y fortalecer el aseguramiento y control de la calidad de los productos alimenticios. Por ello, los MAM aquí planteados ofrecen nuevas oportunidades para la industria alimentaria, en especial, para las industrias tequilera y olivarera, y sectores relacionados, como lo son los laboratorios analíticos de rutina y los organismos oficiales de control de calidad, ya que dichos métodos pueden ser implementados por cada uno de ellos adaptándolos a los objetivos a perseguir.

En este sentido, las empresas productoras de tequila y aceite de oliva podrían ver mejorado sus rendimientos económicos, la seguridad de sus procesos y la calidad de sus productos al implementar estos MAM a lo largo de sus procesos de producción dentro del marco establecido por la QbD y haciendo uso de herramientas establecidas por el **PAT**. Lo anterior, se sustenta en la optimización de la calidad final del producto, la cual se logra diseñando y prediciendo sus características finales deseables mediante las herramientas multivariables del QbD y PAT, implementadas en puntos críticos de control (PCC). Dichas herramientas, comentadas en el **capítulo 2**, han sido empleadas para el desarrollo de cada

uno de los MAM de esta tesis doctoral, cuyo uso es propuesto en distintos PCC de las cadenas alimentarias del tequila y aceite de oliva mediante matraces Erlenmeyer en el esquema general de la **Figura 16**.

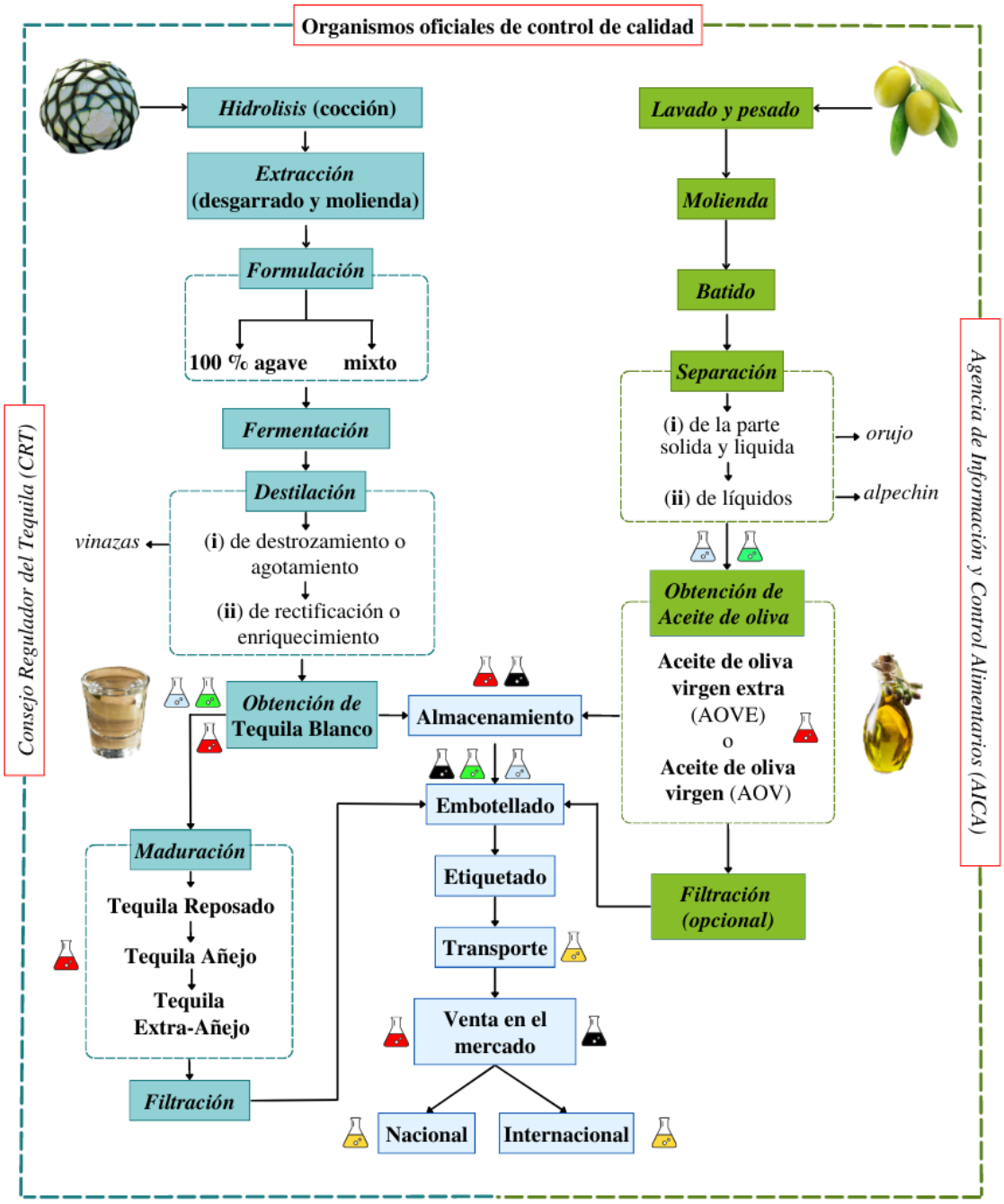


Figura 16. Esquema general con las etapas principales de las cadenas alimentarias del tequila y aceite de oliva en las cuales se propone el uso de distintos métodos analíticos multivariable (MAM) desarrollados en esta tesis doctoral. (*La descripción continúa en la siguiente página).*

Los matraces de Erlenmeyer representan a los distintos MAM y los puntos de los procesos en los cuales se sugiere su uso. El matraz con contenido: *verde* representa al MAM basado en espectroscopía de infrarrojo medio con transformada de Fourier (FTIR), *azul* al MAM basado en espectroscopía de infrarrojo cercano (NIR), *negro* al MAM basado en espectroscopía Raman con sistema de compensación espacial (SORS), *amarillo* al MAM basado en señales cromatográficas *agnostizadas*, y los matraces con contenido *rojo* representan los puntos en los cuales los organismos oficiales de control de calidad suelen tomar alicuotas de muestras para su posterior análisis con métodos analíticos oficiales.

Tal como se aprecia en la **Figura 16**, se propone el uso de los MAM una vez que ambos productos son obtenidos, ya que dichos MAM han sido desarrollados con productos elaborados, por lo que su implementación en las etapas anteriores no tendría cabida. No obstante, al considerar los procesos de producción dentro del marco de QbD, dichos procesos deberían ser de flujo continuo y monitoreados en cada una de las etapas, tal como se realizó en diversos estudios llevados a cabo a nivel industrial o planta piloto –véase **capítulo 2** (subsección **2.1**, apartado **2**)–.

De esta manera, el MAM basado en la técnica FTIR (representado por el matraz con contenido verde) y el MAM basado en la técnica NIR (representado por el matraz con contenido azul) podrían ser implementados primeramente durante la etapa de obtención de cada producto de manera: (i) *at*-line, tomando una pequeña muestra del producto alimenticio y realizando su respectivo análisis dentro de las mismas instalaciones; (ii) *on-line*, desviando la muestra de la línea de producción para realizar su análisis correspondiente, y regresándola posteriormente a la línea de producción; e (iii) *in-line*, realizando las mediciones directamente sobre las muestras en la línea de producción. Asimismo, ambos MAM podrían ser implementados de la misma manera entre las etapas de almacenamiento y embotellado para asegurar que la calidad del producto sigue siendo la misma de una etapa a la otra.

En la misma de imagen la **Figura 16**, se puede apreciar que dicho esquema se encuentra rodeado por un recuadro que representa a los organismos oficiales de control de calidad para el tequila y el aceite de oliva, como son el Consejo Regulador del Tequila (CRT) de México, o la Agencia de Información y Control Alimentarios (AICA) de España, respectivamente.

Dichos organismos tienen la capacidad de realizar inspecciones y verificaciones a lo largo de las cadenas alimentarias, de solicitar documentación necesaria para comprobar el cumplimiento de los requisitos de calidad, así como también la libertad de tomar alícuotas de muestras para su posterior análisis de control de calidad [**9**,**10**].

De hecho, los matraces de Erlenmeyer con contenido rojo representan los puntos en los cuales el CRT y la AICA suelen tomar estas alícuotas (etapas de obtención del producto, almacenamiento y venta en el mercado), las cuales son analizadas de manera *off-line*, lo que implica su traslado a laboratorios analíticos oficiales y de rutina para verificar el cumplimiento de los parámetros de calidad [**9**,**11**].

5

En este sentido, el MAM basado en SORS (representado por el matraz con contenido negro) es propuesto para su aplicación de manera *in-situ* por el CRT y AICA durante la inspección a la empresa, haciendo uso de este método para analizar los productos que ya se encuentren embotellados, lo cual permitiría realizar una comprobación preliminar de su calidad. Del mismo modo, también se propone su empleo cuando los productos estén en el mercado, ya sea en puntos de venta y/o consumo, permitiendo la detección oportuna de productos posiblemente adulterados y/o falsificados, lo cual impactaría directa y favorablemente en la salud y economía de los consumidores.

Del mismo modo, también podrían verse beneficiados al aplicar estos MAM los laboratorios analíticos oficiales y de rutina a los cuales se envían las alícuotas de muestras tomadas por los organismos evaluadores durante sus inspecciones. Dichos beneficios podrían ser (i) la reducción de tiempo y costos al analizar, en una primera instancia, las muestras desconocidas con alguno de los MAM basados en técnicas espectroscópicas, lo cual disminuiría los análisis con técnicas tradicionales, y (ii) la agilización para autentificar la calidad de productos alimenticios entre distintos laboratorios, mediante la aplicación de un MAM global. Esto último, implicaría la implementación del MAM basado en señales cromatográficas *agnostizadas* (representado por el matraz con contenido amarillo en el

[9] Norma Oficial Mexicana NOM-006-SCFI-2012, Bebidas alcohólicas-Tequila-Especificaciones, Comité Consultivo Nacional de Normalización de Seguridad al Usuario, Información Comercial y Prácticas de Comercio (CCNNSUICPC), Gobierno de México.

[10] Ley 12/2013, de 2 de agosto, de medidas para mejorar el funcionamiento de la cadena alimentaria.

[11] Real Decreto 760/2021, de 31 de agosto, por el que se aprueba la norma de calidad de los aceites de oliva y de orujo de oliva.

esquema de la **Figura 16**), por lo que se propone su utilización al final de la cadena alimentaria antes del transporte (nacional o internacional) de las muestras y después de su arribo en destino, lo cual permitiría una comparación de su calidad entre laboratorios más ágil, debido a la creación y uso de una base de datos global con la cual se desarrollarían los modelos matemáticos para dichas comparaciones.

Por último, es importante mencionar que a pesar de la utilidad demostrada y las ventajas que traería consigo la implementación de estos nuevos MAM, aún queda trabajo por llevar a cabo para que su aplicación sea una realidad en la industrias tequilera y olivarera.

Un reto importante es la disposición de las empresas a adoptar dichos MAM en sus líneas de producción, ya que implicaría una inversión económica y humana que probablemente las empresas no estén dispuestas a asumir debido a que podrían ver reducida su productividad por un periodo corto de tiempo; no obstante, la rentabilidad que les reportaría sería mucho mayor a mediano y/o largo plazo que le inversión realizada. Del mismo modo, también podrían presentarse retos dentro de los organismos evaluadores de control de calidad, ya que se requeriría cierto grado de concientización para aceptar y aplicar procedimientos adicionales, así como también se necesitarían cambios a nivel legislativo para autorizar estos MAM como parte de sus actividades de control y aseguramiento de calidad.

No obstante, se seguirá trabajando en esta línea de investigación para lograr la implementación de estos MAM en sus respectivas industrias, así como su divulgación para que sean ampliamente conocidos y puedan llegar a ser adoptados por otros sectores de la industria alimentaria. De igual manera, se continuará trabajando para consolidar la metodología de *agnostización instrumental* en cromatografía de líquidos utilizando distintos detectores, así como también se planea la demostración de su aplicabilidad en cromatografía de gases mediante la utilización de distintos instrumentos analíticos.

Capítulo 6

CONCLUSIONES GENERALES

6. CONCLUSIONES GENERALES

- Se han desarrollado nuevos métodos analíticos multivariable rápidos, confiables y eficientes con aplicabilidad demostrada en la autentificación de la calidad del tequila y aceite de oliva, por lo que podrían ser implementados en la industria alimentaria para aumentar la eficiencia en el aseguramiento y control de calidad de estos productos alimenticios.

- Se ha descrito un método analítico multivariable rápido, económico y no invasivo, basado en la técnica espectroscópica FTIR y aplicando fusión de datos y herramientas quimiométricas, para autentificar la calidad de las categorías '100 % agave' y 'mixto' del Tequila Blanco, logrando una diferenciación excelente entre ambas con PLS-DA.

- La técnica de fusión de datos permitió desarrollar modelos matemáticos de clasificación considerando conjuntamente las distintas correcciones de línea base hechas a los diferentes espectros FTIR de ambas categorías de Tequila Blanco, lo cual evita tener que aplicar en concreto un tipo de corrección de línea base a nuevas muestras cuya categoría se desconoce.

- Se ha desarrollado un método analítico multivariable no destructivo, basado en la técnica espectroscópica NIR y aplicando herramientas quimiométricas de clasificación mediante SVM y PLS-DA, para autentificar la calidad entre las categorías '100 % agave' y 'mixto' del Tequila Blanco y predecir su contenido alcohólico, mediante SVMR y PLSR. La adecuada selección de los métodos de pre-procesamiento durante el tratamiento de los datos influye en su totalidad para obtener modelos matemáticos con excelente o deficiente habilidad de clasificación y/o predicción. En este caso, SVM y PLS-DA presentaron los mejores resultados.

- La técnica analítica SORS, conjuntamente con herramientas quimiométricas, han sido empleadas para desarrollar un método analítico multivariable capaz de discriminar inequívocamente entre las categorías de Tequila Blanco (PLS-DA y SVM) y de predecir su contenido alcohólico (PLSR) a partir de espectros Raman obtenidos midiendo a través

de sus envases, lo cual le otorga un alto potencial para ser aplicado en puntos de venta y consumo.

- Queda constatada la aplicabilidad de la metodología de *agnostización instrumental* a través del desarrollo de dos métodos analíticos multivariable basados en HPLC, con los cuales fue posible obtener huellas cromatográficas independientes (huellas *agnostizadas*) de las condiciones particulares de cada instrumento analítico.

- Se logró contribuir al aseguramiento y control de calidad del aceite de oliva mediante la realización de un método analítico multivariable basado en huellas cromatográficas *agnostizadas* y herramientas quimiométricas, con el cual fue posible detectar adecuadamente adulteraciones realizadas a aceites de oliva virgen y virgen extra.

- Mediante la aplicación de la metodología de *agnostización instrumental* ha sido posible:

  (i) transferir las señales analíticas (cromatogramas) obtenidas mediante la aplicación de un método analíticomultivariable basado en HPLC, a través de huellas cromatográficas *agnostizadas* de muestras de Tequila Blanco obtenidas en cromatógrafos y laboratorios diferentes,

  (ii) crear una base de datos global de huellas cromatográficas *agnostizadas* de las categorías '100 % agave' y 'mixto' de Tequila Blanco, y

  (iii) desarrollar un modelo matemático único para diferenciar entre ambas categorías de muestras de Tequila Blanco analizadas en laboratorios ubicados en España y México.

- La metodología de *agnostización instrumental*, inicialmente propuesta en el propio grupo de investigación, ha sido ampliada con nuevo conocimiento, el cual es necesario aplicar previamente para obtener señales instrumentales correctamente *agnostizadas*. Con el fin de aplicar la metodología de *agnostización instrumental* es necesario asegurar previamente la integridad de las señales instrumentales a *agnostizar*, las cuales deben haber sido obtenidas libres de efectos que alteran el perfil de la señal (deformaciones) mediante un sistema cromatográfico bajo control y correcto funcionamiento.

- Se ha implementado la función de equiparación (*'equity function'*) para corregir los posibles efectos deformantes de alargamiento o estrechamiento que hayan podido sufrir las señales instrumentales al ser obtenidas mediante un sistema cromatográfico fuera de control.

- Los parámetros relación de tiempo de análisis (RtR, *'runtime ratio'*) y el desfase de inicio de análisis (StL, *'starting-time lag'*) han sido desarrollados, como parte de la función de equiparación, para comprobar cuantitativamente las diferencias entre las señales cromatográficas del mismo o diferentes laboratorios analíticos.

- Se ha demostrado la aplicabilidad y funcionalidad de cada uno de los métodos analíticos multivariable desarrollados durante esta tesis doctoral, y se ha propuesto y argumentado su uso en las cadenas alimentarias del tequila y del aceite de oliva.

Mural de las 7 virtudes del Tequila en espacios de realidad y leyenda

Foto tomada por el autor en Tequila, Jalisco, México

Cuenta la leyenda que en el cielo de un lugar llamado Tequitlan vivía Mayahuel, una joven de apariencia muy hermosa que poseía una planta mágica la cual podía proveer de alimento y bebida, ideal para los primeros habitantes de la zona. Sin embargo, se encontraba prisionera entre el maleficio de su abuela, quien era uno de los demonios conocidos como Tzitzimime.

Quetzalcóatl, cuando se transformaba en el Viento Cósmico, era nombrado Ehécatl. En uno de sus viajes por tierras jaliscienses se enamoró de la bella Mayahuel. Tras varios días pensando en ella, una noche con el cielo despejado, Ehécatl decidió escabullirse entre las estrellas para convencerla de escapar a la Tierra, repartir sus bondades entre los mortales y vivir su propia historia de amor. Ella aceptó y esa misma noche se escapó con él.

Para amarse y mantenerse a salvo de la malvada abuela, se convirtieron en un árbol de dos ramas. Así entrelazados se juraron amor eterno.

Sin embargo, la abuela de Mayahuel al notar su ausencia, reunió a otros demonios para bajar a buscarla. Tras encontrar el árbol, destrozó la rama en la que se escondía la bella joven y los pedazos los repartió entre las Tzitzimimes, quienes la devoraron en un instante.

Ehécatl devastado, recogió los restos de su amada para enterrarlos. Al poco tiempo, en ese preciso lugar nació la primera planta de agave que extendió sus raíces y llenó los campos del tan característico verde azulado.

Una tarde lluviosa, los dioses desataron su enojo con rayos que iluminaban los caminos, quemando el corazón de varios agaves que desprendieron el dulce aroma de una especie de miel. Los indígenas no dudaron dos veces en probarlo y quedaron maravillados con el regalo que les heredó Mayahuel: el Tequila.

UNIVERSIDAD
DE GRANADA