XLIV JORNADAS DE VITICULTURA Y ENOLOGÍA TIERRA DE BARROS

IV Congreso Agroalimentario de Extremadura

CENTRO UNIVERSITARIO SANTA ANA ALMENDRALEJO

Del 3 al 6 de Mayo 2022

XLIV JORNADAS DE VITICULTURA Y ENOLOGÍA
DE LA TIERRA DE BARROS

IV CONGRESO AGROALIMENTARIO DE EXTREMADURA

Edita:
Centro Universitario Santa Ana
C/ IX Marqués de la Encomienda, nº 2
Almendralejo
Tel. 924 661 689
http//www.univsantana.com

Colabora: Cajalmendralejo

Ilustración de portada:
© ALBERTO CATILLO

Diseño original:

Tecnigraf S.A.

Maquetación: Virginia Pedrero

ISBN: 978-84-7930-112-0

D.L.:

Imprime: Impresal

Caracterización y discriminación de parafinas de uso agroalimentario mediante espectroscopia Vis-NIR y aprendizaje automático

Barea-Sepúlveda, M.

Ferreiro-González, M.

Calle, J.L.P.

Palma, M.

Departamento de Química Analítica, Facultad de Ciencias, Universidad de Cádiz, Campus Internacional de Excelencia Agroalimentaria (ceiA3); IVAGRO, 11510 Puerto Real, Cádiz.

RESUMEN

Las parafinas son productos derivados del petróleo (PDPs) con un amplio espectro de aplicaciones de consumo e industriales, incluida la agroalimentaria, que varían según su composición química. Este estudio presenta un método basado en la espectroscopia visible e infrarroja cercana en combinación con el aprendizaje automático para la correcta caracterización y discriminación de los dos tipos de parafinas más comúnmente comercializadas. Por otra parte, los datos espectroscópicos combinados con algoritmos de aprendizaje automático no supervisados, como el análisis jerárquico de conglomerados (HCA), y con algoritmos de aprendizaje automático supervisados no pa-

ramétricos, como las máquinas de vectores de soporte (SVM) y los bosques aleatorios (RF), permitieron caracterizar y discriminar las muestras en función de la composición molecular. Los resultados obtenidos demostraron la idoneidad de esta técnica analítica rápida, ecológica y económica como alternativa a los métodos actuales para el control de calidad automático de las parafinas.

Palabras clave: Parafinas; Productos Derivados del Petróleo; Espectroscopia Visible-Infrarrojo Cercano; Aprendizaje Automático; Máquinas de vectores de soporte; Árboles Aleatorios; Huella Espectral.

ABSTRACT

Waxes are petroleum-derived products (PDPs) with a wide spectrum of industrial and consumer applications that vary according to their chemical composition. This study presents a method based on visible and near-infrared spectroscopy in combination with machine learning for the correct characterization and discrimination of the two most marketed types of petroleum waxes. Moreover, the spectroscopic data combined with unsupervised machine learning algorithms, such as hierarchical cluster analysis (HCA), and with nonparametric supervised machine learning algorithms, such as support vector machines (SVM) and random forests (RF), allowed to characterize, and discriminate, the samples based on their molecular composition. The results obtained demonstrated the suitability of this fast, environmentally friendly, and cost-effective analytical technique as an alternative to current methods for automatic quality control of petroleum waxes.

Key words: Petroleum Waxes; Petroleum-Derived Products; Visible-Near Infrared Spectroscopy; Machine learning; Support vector machine; Random Forest; *Spectralprint*.

1. INTRODUCCIÓN

Las parafinas son un producto derivado del petróleo (PDP) que se obtienen a partir del refino de los aceites lubricantes y que presentan un amplio espectro de aplicaciones industriales (*incluida la agroalimentaria)* y de consumo. Desde un punto de vista químico, las parafinas son mezclas complejas compuestas principalmente por cadenas largas de hidrocarburos saturados

(*n*-parafinas, isoparafinas y cicloparafinas) y, por tanto, son poco reactivas [1]. Además, también pueden incluir en su composición otros componentes menores como hidrocarburos aromáticos y compuestos de azufre y nitrógeno. Sin embargo, debido a sus aplicaciones en la industria agroalimentaria y cosmético-farmacéutica, las parafinas que van destinadas a estos fines han de ser sometidas a un proceso de hidrotratamiento [2] para eliminar estos compuestos minoritarios con el fin de alcanzar un alto grado de pureza, mejorar el color, eliminar el olor y satisfacer así los requisitos establecidos por la Administración de Alimentos y Medicamentos (FDA) [3,4] y la Farmacopea de la Unión Europea (Ph. Eur.) [5]. Las parafinas pueden clasificarse en macrocristalinas y microcristalinas en función del tipo de hidrocarburo saturado predominante en su composición, así como del patrón de empaquetamiento de sus cadenas en la red formada. En este sentido, las parafinas macrocristalinas están compuestas principalmente por *n*-parafinas (alcanos lineales), mientras que las microcristalinas por isoparafinas (alcanos ramificados) [6,7]. Estas diferencias en su composición molecular permiten que las propiedades fisicoquímicas de cada tipo de parafina sean ligeramente diferentes y, por tanto, sus aplicaciones. Es por ello que, las parafinas macrocristalinas se utilizan principalmente en la fabricación de velas, lápices de colores y papel parafinado. A su vez, también se aplican en la industria alimentaria como aditivo en los chicles [6,8,9] siempre que adquieran el grado alimentario requerido para este fin. Por su parte, las parafinas microcristalinas son más elásticas y flexibles, utilizándose su versión de calidad alimentaria y cosmético-farmacéutica para estabilizar la estructura de barras de labios, para recubrir frutas y quesos. Además de ser empleadas para modificar las propiedades cristalinas de las ceras macrocristalinas [9,10].

Hasta el momento, el control de calidad de las parafinas en la industria petroquímica está regulado internacionalmente por las normas establecidas por la Sociedad Americana de Pruebas y Materiales (ASTM), que se basan principalmente en métodos para medir y evaluar las propiedades fisicoquímicas, como el punto de fusión [11], el punto de congelación (ASTM D938) [12], la penetración de la aguja (ASTM D1321) [13], el contenido de aceite (ASTM D721) [14], el olor (ASTM D1833) [15], y el color (ASTM D156) [16]. Sin embargo, la clasificación más aceptada para las ceras de petróleo es la definida por la ASTM-TAPPI (1963) [17], que las divide en función de su punto de congelación y su índice de refracción a 212 ^{o}F. A su vez, la

cromatografía de gases (GC), generalmente acoplada a la espectrometría de masas (MS), también se ha utilizado como técnica de referencia para llevar a cabo la identificación individual de los hidrocarburos en las ceras y otras PDP [1,18,19]. Sin embargo, el procedimiento de identificación en esta técnica analítica puede llevar mucho tiempo y ser difícil de reproducir. Así pues, teniendo en cuenta que las aplicaciones de las parafinas varían según sus propiedades, disponer de metodologías analíticas rápidas y no destructivas para discriminar el tipo de parafina resulta de gran interés tanto para la industria del petróleo como para la industria agroalimentaria y cosmético-farmacéutica en términos de automatización del proceso de control de calidad de este producto. En este marco, las técnicas espectroscópicas como la espectroscopia visible e infrarroja cercana (Vis-NIR), la espectroscopia infrarroja (IR) o la espectroscopia ultravioleta y visible (UV-Vis), utilizadas como métodos de cribado y perfilado global, constituyen tecnologías no destructivas, rápidas, respetuosas con el medio ambiente y de operación *in situ*, que pueden suponer una alternativa a los métodos de referencia en el análisis de las ceras del petróleo con mayor precisión y robustez.

En particular, la espectroscopia Vis-NIR ha demostrado su eficacia en aplicaciones industriales y de laboratorio a lo largo de los años [20,21]. Sin embargo, el uso de esta técnica espectroscópica genera una gran cantidad de información en un periodo limitado, por lo que el manejo de este volumen de datos requiere la aplicación de algoritmos de aprendizaje automático para transformar los datos en información interpretable, así como para generar modelos predictivos que permitan construir aplicaciones interactivas para automatizar los procesos de control de calidad a nivel industrial [22,23]. En cuanto a los algoritmos de aprendizaje automático, existen numerosas *técnicas de aprendizaje automático no supervisado, como el análisis jerárquico de conglomerados (HCA)*, que se utilizan principalmente para el reconocimiento de patrones dentro del conjunto de datos, y supervisadas, como la *máquina* de vectores de apoyo (SVM) y el bosque aleatorio (RF), que se emplean para generar modelos predictivos de clasificación y/o regresión [23,26]. De este modo, la espectroscopia Vis-NIR se ha aplicado con éxito en el sector de la investigación petroquímica en combinación con técnicas de aprendizaje automático supervisado, como PLS, para cuantificar el contenido de aceite extraíble mediante el proceso de desparafinado utilizando tolueno y metil etil cetona (MEK) [1]. Además, esta técnica espectroscópica también se ha aplicado con éxito para la discriminación de la

gasolina según su octanaje en combinación con algoritmos de aprendizaje automático no supervisado, como HCA, y algoritmos supervisados como SVM y RF [28-30].

En base a lo anterior, este estudio tiene como objetivo evaluar la aplicabilidad de la espectroscopia Vis-NIR en combinación con algoritmos de aprendizaje automático no supervisados (HCA) y supervisados (SVM y RF) para la discriminación de parafinas según su composición molecular (macrocristalina y microcristalina), las cuales son de alto interés en el sector agroalimentario dada su aplicación, principalmente, como materiales de contacto con alimentos.

2. MATERIALES Y MÉTODOS

2.1. Muestras

Se han emplearon un total de 60 muestras de parafinas, 36 macrocristalinas y 24 microcristalinas, suministradas por la Compañía Española de Petróleos, S.A.U., (CEPSA) refinería de San Roque (Cádiz, España). Las muestras se tomaron en diferentes años e incluyeron parafinas macrocristalinas y microcristalinas hidrogenadas (calidad alimentaria) y no hidrogenadas (calidad no alimentaria) para obtener un conjunto de datos heterogéneo que permitiera una adecuada generalización de los modelos de aprendizaje automático supervisados. Antes de los análisis, las ceras (0,4 g) se almacenaron en viales sellados de 10 mL (Agilent Crosslab) y se fundieron en un horno a 80 °C durante 10 min. Posteriormente, las muestras se solidificaron a temperatura ambiente (25 ^{o}C) dentro del mismo vial con el fin de obtener una superficie sólida plana y homogénea que cubriera completamente el fondo del vial.

2.2. Obtención de espectros Vis-NIR

Los espectros Vis-NIR se registraron en un analizador FOSS XDS Rapid Content™ con tecnología de infrarrojo cercano XDS (FOSS Analytical, Hilleroed, Dinamarca), utilizando el software de análisis de rutina ISIscan (FOSS Analytical, Hilleroed, Dinamarca). Las medidas se realizaron con las muestras almacenadas en los viales sellados de 10 mL, las cuales se analizaron en el rango de 400 — 2500 nm con una resolución espectral de 0,5 nm.

Se recogieron un total de 32 scans por muestra, utilizando después el espectro medio. Todas las muestras se analizaron por duplicado. Finalmente, el espectro medio Vis-NIR obtenido para cada muestra de cera se colocó en una matriz de datos D_{mxn} donde *n* es el número de muestras de cera (*n* = 60), y *m* es el número de valores de absorbancia (*m* = 4200).

2.3. Análisis de datos y software

Todos los análisis de datos se realizaron con RStudio (R versión 4.1.2, Boston, MA, USA). Previamente a la aplicación de las técnicas de aprendizaje automático se llevó a cabo un pretratamiento espectral con el fin de minimizar el ruido instrumental y eliminar las contribuciones no deseadas debido tanto a las propiedades físicas de la muestra como a las derivadas por las ligeras variaciones en las condiciones de registro. Para ello, se aplicó la primera derivada al espectro Vis-NIR de cada muestra mediante el método de Savitzky-Golay (ventana móvil de 11 puntos y polinomio de segundo orden). Por otro lado, las técnicas de aprendizaje automáticas empleadas inculyeron la aplicación del HCA para realizar un estudio exploratorio con el fin de encontrar patrones y agrupar tendencias en el conjunto de datos, y la aplicación de los algoritmos de SVM y RF para el desarrollo de modelos predictivos de clasificación.

3. RESULTADOS Y DISCUSIÓN

3.1. Análisis espectral

Los espectros Vis-NIR típicos (datos brutos; $D_{4200x60}$) para los dos tipos de parafinas, macrocristalina y microcristalina, se muestran en la Fig. 1 (A). Se detectó una curva de perfil similar para ambos tipos. Sin embargo, tras una inspección visual de los espectros, se pudieron observar discrepancias en la intensidad de la absorbancia en algunas regiones espectrales, que pueden ser relevantes para discriminarlas. Las diferencias en la intensidad de la absorbancia se observan en la región visible entre 400 — 700 nm. Especialmente en la región de la longitud de onda de 400 nm, donde las parafinas microcristalinas muestran una mayor intensidad de absorbancia con respecto a las macrocristalinas. La absorción a 400 nm está relacionada con compuestos que absorben en el rango de luz violeta-azul, lo que indica que

las parafinas microcristalinas muestran una coloración ligeramente más amarillenta con respecto a las macrocristalinas. Por otro lado, se observan diferencias en la intensidad de absorción entre 800 — 1680 nm en la región NIR. Concretamente, las parafinas macrocristalinas muestran una mayor intensidad de absorción en comparación con las ceras microcristalinas. Estas regiones NIR corresponden a las bandas del primer, segundo y tercer sobretono. Cabe destacar las bandas entre 1100 — 1275 nm y 1350 — 1500 nm, que están asociadas al segundo sobretono del C-H y al primer sobretono de las combinaciones de C-H. Esto indicaría que existen diferencias químicas entre los dos tipos de parafinas en términos de su composición en hidrocarburos. Los resultados espectrales obtenidos que aquí se presentan fueron comparados con el estudio realizado por Palou et. al (2014) [1] en el que se utilizó la espectroscopia NIR para analizar muestras de parafinas refinadas para evaluar el aceite removible mediante MEK. Esta comparación mostró que, en la región NIR, los espectros de las muestras aquí estudiadas presentaban un perfil similar al reportado por estos autores y, por tanto, estaban de acuerdo con la bibliografía. En la Fig. 1 (B) se muestran los espectros de la primera derivada utilizando el método de Savitzky-Golay. La aplicación del método de filtrado Savtizky-Golay implica una ligera pérdida de datos al principio y al final de la matriz de datos. Por ello, se redujo a $D_{4190x60}$. La descripción de los espectros de absorción Vis-NIR de la primera derivada muestra algunos picos prominentes en la región NIR. En concreto, la banda a 1200 nm, asociado al segundo sobretono del C-H, la banda a 1600 nm, relacionado con sobretono de la banda de -ArCH, las bandas a 1700 nm, que están asociados al primer sobretono del C-H, y, finalmente, y la banda a 2200 nm que está relacionado con las bandas de combinaciones del C-H. Cabe destacar que la banda a 1600 de longitud de onda estarían relacionados con los compuestos menores presentes tanto en parafinas macrocristalinas como en las microcristalinas que no han sido sometidas al proceso de hidrotratamiento.

3.2. Estudio exploratorio

Para corroborar la tendencia de las muestras de cera a agruparse según su composición molecular, se llevó a cabo un HCA. Esta técnica de aprendizaje automático no supervisado se aplicó al conjunto de datos preprocesados ($D_{4190x60}$). Para este análisis, se seleccionó la distancia euclidea como medida de distancia y el método de Ward como método de vinculación. La elec-

ción del método de vinculación se determinó a partir de la comparación del coeficiente de aglomeración de distintos métodos (Promedio, Completo, Simple y Ward). Los coeficientes de aglomeración más cercanos a 1 indicarían una estructura de agrupación más fuerte. En este caso, el método Ward presentó el mayor coeficiente de aglomeración (0,95) entre los métodos de vinculación evaluados. Los resultados obtenidos mediante el HCA se representaron en el dendrograma mostrado en la Fig. 2. De acuerdo con los resultados se puede observar que las muestras tienden a agruparse en dos clusters principales. Por un lado, el clúster de color negro está completamente formado por todas las muestras de cera microcristalina. Por otro lado, el cluster de color gris está formado por todas las muestras de cera macrocristalina. Así pues, los resultados indican que existe una fuerte tendencia a agrupar las ceras según su composición molecular.

A pesar de los buenos resultados obtenidos a partir del HCA, esta técnica no supervisada tiene la desventaja de no poder realizar predicciones futuras. Por lo que la aplicación de técnicas no supervisadas de inteligencia artifical se hacen de gran interés para contar con modelos predictivos que puedan ser implementados en el control de calidad de este producto.

3.3. Modelos de clasificación

Se generaron y compararon un total de dos modelos basados en los algoritmos de clasificación SVM y RF para la discriminación de las parafinas según su composición molecular (ceras macrocristalinas y microcristalinas). Para construir ambos clasificadores, el conjunto de datos preprocesados se dividió aleatoriamente en un conjunto de entrenamiento (división = 0,7) y un conjunto de prueba (división = 0,3). El conjunto de entrenamiento se utilizó durante la optimización de los hiperparámetros y el proceso de entrenamiento y el conjunto de prueba para la validación de los modelos generados. Por otro lado, se utilizó la validación cruzada (CV) de *5-folds* durante el proceso de optimización de hiperparámetros y entrenamiento para minimizar el sobreajuste de los modelos. El rendimiento de los modelos se llevó a cabo utilizando el *accuracy* y *kappa* como métricas. El *accuracy* se calculó como el porcentaje de muestras clasificadas correctamente dividido por el número total de muestras clasificadas, mientras que el parámetro *kappa* se calculó como la diferencia entre el accuracy observada menos el *accuracy* esperado dividida por 1 menos el *accuracy* esperado.

En primer lugar, para la construcción del clasificador SVM con la implementación de *kernel* gausiano, se realizó la optimización de los hiperparámetros (C y σ) utilizando el conjunto de entrenamiento. Para ello, se seleccionó el método de búsqueda en cuadrícula con el crecimiento exponencial de C y σ. En este caso, se evaluaron valores de $\log_2$C y $\log_2\sigma$ estaban en el rango de -10 a 10 en intervalos de 0,5. Cada combinación de opciones de parámetros se probó utilizando un CV de *5-folds* y aquellos que presentaron el mejor *accuracy* fueron seleccionados. La Fig. 3 muestra el gráfico de contorno para la búsqueda de los valores de C y σ que proporcionaron el mejor accuracy del CV de *5-folds*. Se puede observar que a medida que el $\log_2$C, y por tanto C, aumentaba, la precisión del CV a *5-folds* era mayor. El valor óptimo de C se fijó en 0,7071 ($\log_2$C = -0,5), ya que fue el valor mínimo que permitió obtener la máxima precisión del CV de 5-*folds*. Por otro lado, como se puede observar en la Fig. 3, la precisión del CV 5-*folds* aumenta a medida que disminuye el valor de $\log_2\gamma$ *y, por consiguiente, de* γ. *El valor de* γ *controla el comportamiento del kernel* y, a medida que su valor aumenta, también lo hace la flexibilidad del modelo. En este caso, el valor óptimo de σ se fijó en $9{,}766 \cdot 10^{-4}$ ($\log_2\sigma$ = -10). Cabe destacar que los mejores resultados se obtuvieron con los valores más bajos de σ, lo que sugiere que los grupos son prácticamente linealmente separables. Tras el ajuste de los hiperparámetros, se entrenó el modelo con los valores óptimos de C y σ obtenidos utilizando el conjunto de entrenamiento y aplicando un CV de *5-folds*, obteniendo un *accuracy* del 97,8% y un valor de *kappa* del 0,95. Por último, se evaluó el rendimiento del clasificador de SVM generado utilizando el conjunto de prueba, obteniendose un *accuracy* del 100% y una *kappa* de 1, confirmando el excelente rendimiento del modelo para discriminar las parafinas según su composición molecular. Para construir el modelo de RF, se establecieron los valores óptimos de los hiperparámetros *mtry* y número de árboles de decisión. La raíz cuadrada del número total de predictores se utilizó como valor óptimo de *mtry* y, concretamente, fue igual a 64,73 (4.200 predictores). Por otra parte, en RF el número de árboles de decisión no es un hiperparámetro crítico, ya que añadir más árboles de decisión no implica riesgos de sobreajuste y mejora el rendimiento del modelo. Sin embargo, su valor debe ser determinado previamente por el analista para estabilizar el error y minimizar la pérdida de recursos computacionales. En este sentido, el número de árboles de decisión se fijó en 100 ya que es un número lo suficientemente alto como para estabilizar el error sin que suponga un coste computacional significativo. Los valores óptimos estable-

cidos para *mtry* y el número de árboles de decisión se utilizaron entonces para entrenar el modelo RF con el conjunto de entrenamiento aplicando un CV de *5-folds*. Se obtuvieron excelentes resultados para el CV de *5-folds* con un *accuracy* del 97,8% y un *kappa* de 0,95. Posteriormente, se evaluó el rendimiento del modelo utilizando el conjunto de pruebas, obteniendo un *accuracy* del 100% y un *kappa* de 1. Esto confirmó que se obtuvo un modelo RF fiable y preciso para la discriminación de las parafinas de petróleo en términos de su composición molecular.

Además de disponer de modelos predictivos fiables, uno de los objetivos que se persiguen al aplicar técnicas de perfilado global para la automatización del proceso de control de calidad es encontrar un conjunto reducido de señales que caractericen las muestras y permitan diferenciarlas fácilmente. El SVM, por la propia naturaleza del algoritmo, no permite establecer las longitudes de onda más relevantes directamente relacionadas con la discriminación de las parafinas macrocristalinas y microcristalinas. Sin embargo, la RF permite realizar esta tarea. Así, según el criterio de disminución media de la impureza del nodo, se establecieron las longitudes de onda más relevantes. En concreto, las variables que presentaron una importancia relativa superior al 50%, constituyendo un total de 7 longitudes de onda (1174,50 nm, 1751,00 nm, 1810,50 nm, 1928,5 nm, 2041,00 nm, 2117,00 nm y 2189,50 nm). Por lo tanto, las 7 longitudes de onda seleccionadas se utilizaron para la construcción de la huella espectral característica de ambos tipos de ceras. Los diagramas de araña de las huellas espectrales se muestran en la Fig. 4. De acuerdo con los resultados obtenidos, se puede observar una forma similar de la huella espectral para ambas parafinas. Sin embargo, se pueden detectar diferencias en términos de absorbancia. Por un lado, las parafinas macrocristalinas mostraron su máxima absorbancia a una longitud de onda de 2189,50 nm, mientras que el porcentaje de absorbancia para esta longitud de onda en la parafina microcristalina es de 0,7 (70% del máximo de absorbancia). A su vez, se pudo apreciar que había otras longitudes de onda para la parafina macrocristalina por encima del 50% del máximo de absorbancia, concretamente, a 1174,50 nm, 1751,00 nm y 2041,00 nm. Las demás longitudes de onda estaban por debajo de 0,5 (50% de la absorbancia máxima). Por otra parte, para las parafinas microcristalinas, la absorbancia máxima se alcanzó en λ = 1751,00 nm, mientras que el porcentaje de absorbancia para esta longitud de onda en las parafinas macrocristalinas es de 0,7. Además, en la parafina microcristalina se ob-

servaron otras longitudes de onda cuyas intensidades fueron superiores al 50% de la absorbancia máxima, siendo estas las de 1751,00 nm, 2189,50 nm. Las restantes longitudes de onda tuvieron una contribución menor (<50% de la absorbancia máxima) en términos de intensidad en la huella espectral de las parafinas microcristalinas, destacando las longitudes de onda de 1810,50 nm, y 2041,00, que fueron superiores a 0,5 para las parafinas macrocristalinas. Como las intensidades y las relaciones de las señales son diferentes para cada tipo de ceras, dando así diferentes huellas dactilares, éstas pueden ser utilizadas para la discriminación de las diferentes parafinas en base a su composición molecular de forma rápida y sencilla.

4. CONCLUSIONES

La espectroscopia Vis-NIR combinada con herramientas de aprendizaje automático ha demostrado ser una metodología práctica adecuada para la caracterización y discriminación de parfinas en función de su composición molecular, ofreciendo una alternativa rápida, fiable y respetuosa con el medio ambiente a los métodos oficiales establecidos. Los resultados obtenidos mediante la aplicación de algoritmos de aprendizaje automático no supervisados, como el HCA, sugirieron una fuerte tendencia a agrupar las muestras según fueran ceras macrocristalinas o microcristalinas. Los modelos de aprendizaje automático supervisado desarrollados en base a ambos algoritmos (SVM y RF) han demostrado su eficacia y robustez para la discriminación de los dos tipos de parafinas más comunmente comercializadas, obteniendo un excelente rendimiento (100% de *accuracy* y 1 de *kappa* en el conjunto de pruebas). A su vez, el modelo RF permitió la extracción de las 7 longitudes de onda más relevantes en esta discriminación. Así, se construyó la huella espectral característica de cada tipo. En consecuencia, estas huellas espectrales pueden utilizarse como un método de rutina adecuado para la identificación rápida, precisa y directa de los tipos de parafinas. Además, si se desarrolla una base de datos, se podrían construir aplicaciones web con soporte para ordenadores y tablets a partir de los modelos desarrollados, simplificando así el análisis de datos dentro de la cadena de producción en la refinería e incluso extendiendose a su aplicación en el control de calidad interno que se realiza a este producto dentro de la propia industria agroalimentaria. Esto, junto con la portabilidad de la técnica y su facilidad de uso, permitiría un control de calidad aún más sencillo y automatizado de esta PDP.

Financiación: Esta investigación ha sido financiada por la Universidad de Cádiz y la Cátedra Fundación CEPSA.

Agradecimientos: Los autores agradecen a la Universidad de Cádiz y a la Cátedra Fundación CEPSA el contrato predoctoral (FPI UCA/TDI-4-19) concedido a Marta Barea-Sepúlveda. El agradecimiento se extiende a la Refinería CEPSA-San Roque y al Instituto de Investigación Vitivinícola y Agroalimentario (IVAGRO) por la prestación del material e infrastructuras necesarias para llevar a cabo esta investigación.

Conflictos de intereses: Los autores declaran no tener ningún conflicto de intereses.

FIGURAS Y TABLAS

Figura. 1. (A) Espectros Vis-NIR brutos para las muestras de parafinas macrocristalinas y microcristalinas estuidadas ($D_{2400x60}$); (B) Espectros de primera derivada mediante método de Savitzky-Golay ($D_{1491x60}$).

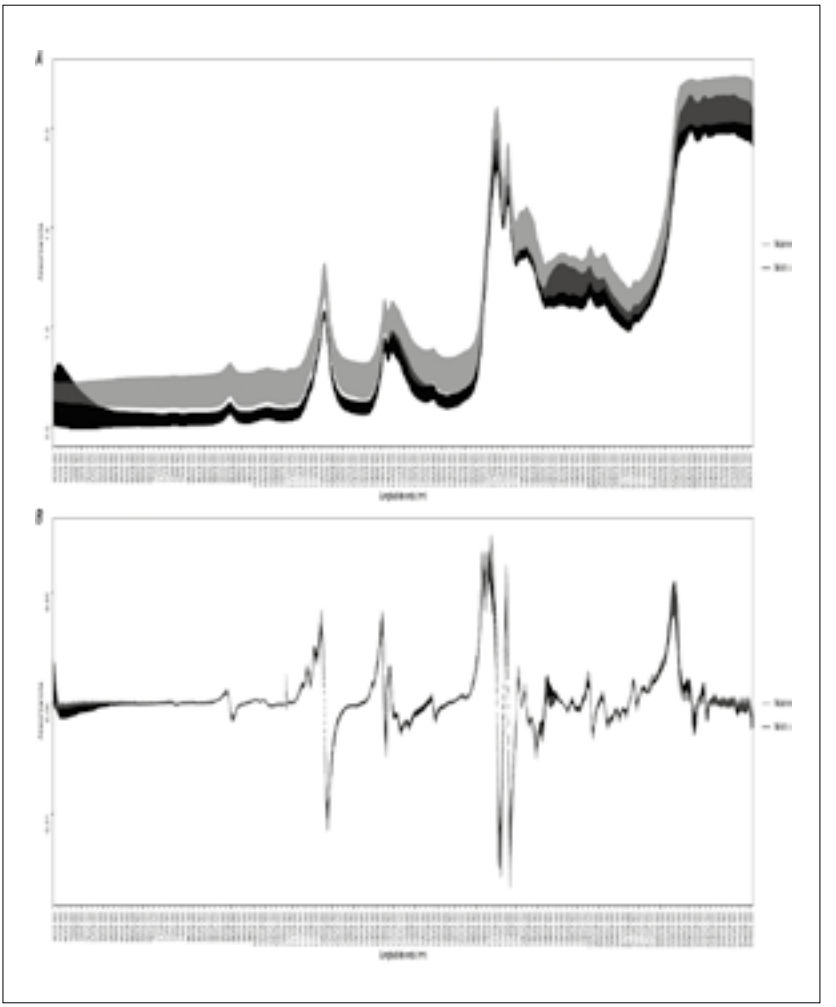


Figura. 2. Dendograma resultante del análisis jerarárquico de conglomerados (HCA) realizado sobre los datos espectrales pretratados ($D_{1491x60}$).

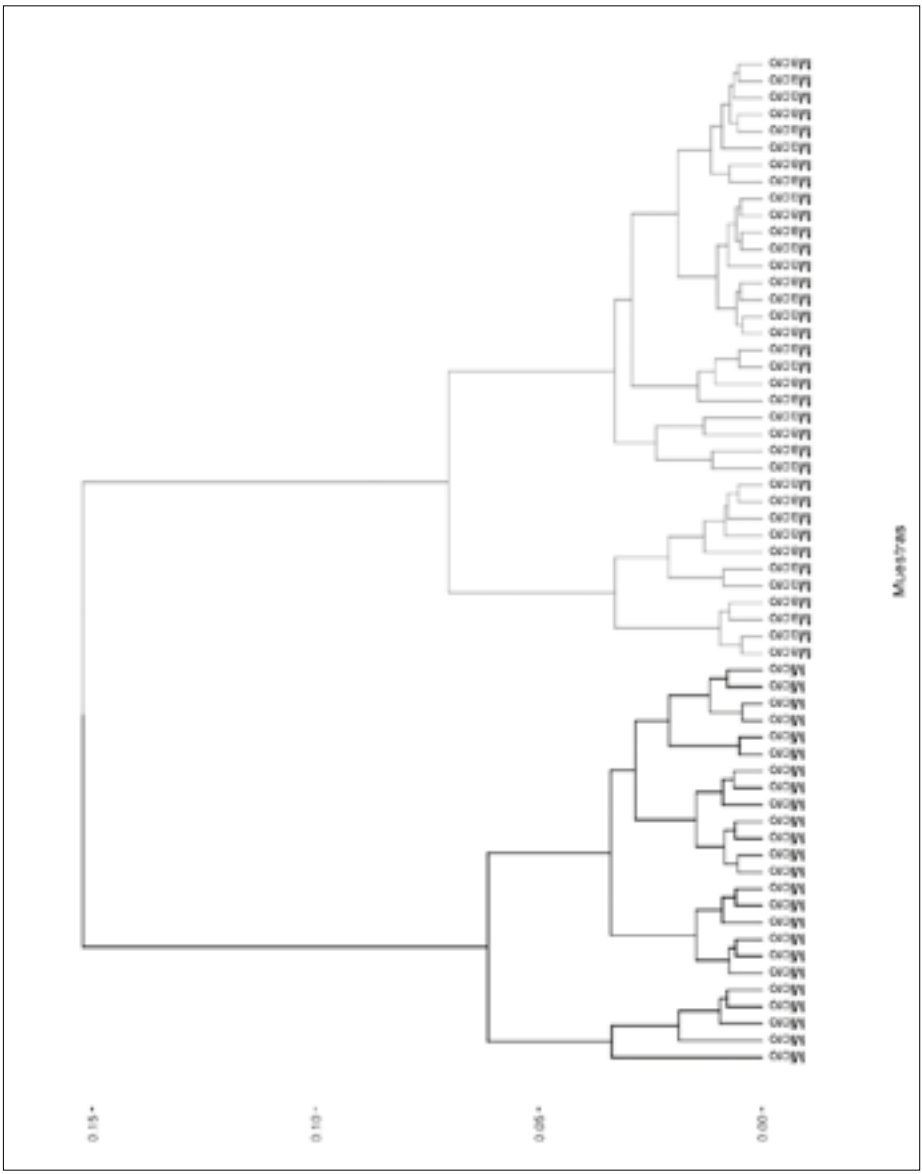


Figura. 3. Gráfico de contornos para la búsqueda del mejor valor de C y σ en términos de accuracy.

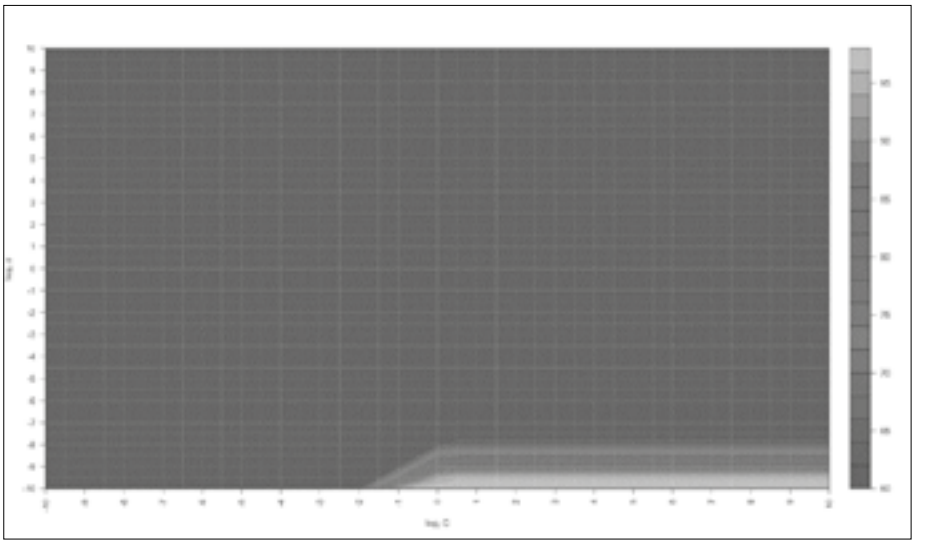
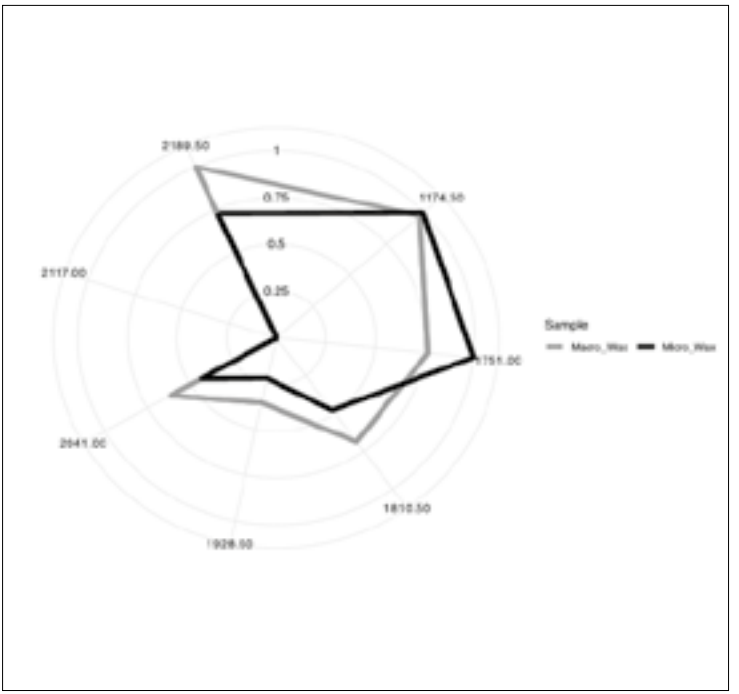


Figura. 4. Huella espectral obtenida mediante el modelo de RF para la discriminación de parafinas macrocristalinas y microcristalinas.



REFERENCIAS BIBLIOGRÁFICAS

1. A. Palou, J. Cruz, M. Blanco, R. Larraz, J. Frontela, C.M. Bengoechea, J.M. González, M. Alcalà, "Characterization of the composition of paraffin waxes on industrial applications", *Energy and Fuels*, 28 (2014) 956–963. https://doi.org/10.1021/ef4021813.
2. J.G. Speight, "Hydrocarbons from Petroleum", in: Handb. Ind. Hydrocarb. Process., Elsevier, Oxford, United Kingdom, 2011: pp. 122–125. https://doi.org/10.1016/b978-0-7506-8632-7.10003-9.
3. U.S. Food and Drug Administration (FDA). "Food additives permitted for direct addition to food for human consumption". Code of Federal Regulations; FDA: Silver Spring, MD, 2013; Part 172, Title 21, Vol. 3.
4. U.S. Food and Drug Administration (FDA). "Indirect food additives: Adjuvants, production aids, and sanitizers". Code of Federal Regulations; FDA: Silver Spring, MD, 2013; Part 178, Title 21, Vol. 3.
5. European Directorate for the Quality of Medicines and HealthCare (EDQM) Council of Europe. European Pharmacopoeia, 7th ed.; EDQM Council of Europe: Strasbourg, France, 2011.
6. W.P. Cottom, "Waxes," *Kirk-Othmer Encycl. Chem. Technol* (2000). https://doi.org/10.1002/0471238961.2301240503152020.A01.
7. U. Gupta, A.K. Mishra, "Study of microcrystalline and macrocrystalline structure based on Cambay basin crude oils", *Upstream Oil Gas Technol.*, 8 (2022) 100067. https://doi.org/10.1016/J.UPSTRE.2022.100067.
8. O. Saber, N. Hefny, A.A. Al Jaafari, "Improvement of physical characteristics of petroleum waxes by using nano-structured materials", *Fuel Process. Technol.*, 92 (2011) 946–951. https://doi.org/10.1016/J.FUPROC.2010.12.015.
9. J.G. Speight, "Pharmaceuticals", *Handb. Ind. Hydrocarb. Process.*, (2020) 553–595. https://doi.org/10.1016/B978-0-12-809923-0.00013-8.
10. R. Morello, C. De Capua, "Infrared thermographic investigation of the use of microcrystalline wax to preserve apples from thermal shocks", *Measurement*, 152 (2020) 107304. https://doi.org/10.1016/J.MEASUREMENT.2019.107304.

11. ASTM International. ASTM Standard D87: Standard Test Method for Melting Point of Petroleum Wax (Cooling Curve). https://www.astm.org/d0087-09r18.html.

12. ASTM International. ASTM Standard D938: Standard Test Method for Congealing Point of Petroleum Waxes, Including Petrolatum. https://www.astm.org/d0938-12r17.html.

13. ASTM International. ASTM Standard D1321: Standard Test Method for Needle Penetration of Petroleum Waxes. https://www.astm.org/d1321-16a.html.

14. ASTM International. ASTM Standard D721: Standard Test Method for Oil Content of Petroleum Waxes. https://www.astm.org/d0721-17.html.

15. ASTM International. ASTM Standard D1833: Standard Test Method for Odor of Petroleum Wax https://www.astm.org/d1833-87r17.html.

16. ASTM International. ASTM Standard D156: Standard Test Method for Saybolt Color of Petroleum Products (Saybolt Chromometer Method). https://www.astm.org/d0156-15.html.

17. ASTM-TAPPI (1.963). "Petroleum Waxes: Characterization, Performance, and Additives". The Proceedings of the Symposium on Petroleum Waxes. Special Technical Association Publication. STAP No 2.

18. ASTM International. ASTM Standard D5442: Standard Test Method for Analysis of Petroleum Waxes by Gas Chromatography. https://www.astm.org/d5442-17r21.html.

19. J.J. Espada, J.A.P. Coutinho, J.L. Peña, "Evaluation of Methods for the Extraction and Characterization of Waxes from Crude Oils", *Energy and Fuels*, 24 (2010) 1837–1843. https://doi.org/10.1021/EF901378U.

20. M. Chen, S. Khare, B. Huang, H. Zhang, E. Lau, E. Feng, "Recursive Wavelength-Selection Strategy to Update Near-Infrared Spectroscopy Model with an Industrial Application", *Ind. Eng. Chem. Res*., 52 (2013) 7886–7895. https://doi.org/10.1021/IE4008248.

21. R.E. Morris, M.H. Hammond, J.A. Cramer, K.J. Johnson, B.C. Giordano, K.E. Kramer, S.L. Rose-Pehrsson, "Rapid Fuel Quality Surveillance through Chemometric Modeling of Near-Infrared Spectra", *Energy and Fuels*, 23 (2009) 1610–1618. https://doi.org/10.1021/EF800869T.

22. K. El Bouchefry, R.S. de Souza, "Learning in Big Data: Introduction to Machine Learning", *Knowl. Discov. Big Data from Astron. Earth Obs. Astrogeoinformatics*, (2020) 225–249. https://doi.org/10.1016/B978-0-12-819154-5.00023-0.

23. A. Géron, "Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems", O'Reilly Media. (2019) 851. https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/.

24. P. Tong, Y. Du, K. Zheng, T. Wu, J. Wang, "Improvement of NIR model by fractional order Savitzky–Golay derivation (FOSGD) coupled with wavelength selection", *Chemom. Intell. Lab. Syst*,. 143 (2015) 40–48. https://doi.org/10.1016/J.CHEMOLAB.2015.02.017.

25. Å. Rinnan, F. van den Berg, S.B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra", *TrAC Trends Anal. Chem*., 28 (2009) 1201–1222. https://doi.org/10.1016/J.TRAC.2009.07.007.

26. A.C. Müller, S. Guido, "Introduction to Machine Learning with Python and Scikit-Learn", O'Reilly Media, Inc. (2015) 1. https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/ (accessed January 23, 2022).

27. D. Zamora, M. Blanco, M. Bautista, R. Mulero, M. Mir, "An analytical method for lubricant quality control by NIR spectroscopy", *Talanta*, 89 (2012) 478–483. https://doi.org/10.1016/J.TALANTA.2011.12.067.

28. R.M. Balabin, R.Z. Safieva, E.I. Lomakina, "Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques", *Anal. Chim. Acta.*, 671 (2010) 27–35. https://doi.org/10.1016/j.aca.2010.05.013.

29. M. Ferreiro-González, J. Ayuso, J.A. Álvarez, M. Palma, C.G. Barroso, "Gasoline analysis by headspace mass spectrometry and near infrared spectroscopy", *Fuel*, 153 (2015) 402–407. https://doi.org/10.1016/j.fuel.2015.03.019.

30. M. Barea-Sepúlveda, M. Ferreiro-González, J.L.P. Calle, G.F. Barbero, J. Ayuso, M. Palma, "Comparison of different processing approaches by SVM and RF on HS-MS eNose and NIR Spectrometry data for the discrimination of gasoline samples", *Microchem. J.*, 172 (2022) 106893. https://doi.org/10.1