X Congreso Ibérico de Agroingeniería
X Congresso Ibérico de Agroengenharia
Huesca, 3-6 septiembre 2019

Identificación optimizada de las longitudes de onda relevantes en espectros NIR de aceituna

Natalia Hernández-Sánchez[1] y María Gómez-del-Campo[2]

[1] Laboratorio de Propiedades Físicas-Tecnologías Avanzadas en Agroalimentación (LPF-TAGRALIA), ETSIAAB, Universidad Politécnica de Madrid, Av. Puerta de Hierro, 2 - 4, 28040 Madrid, España; n.hernandez@upm.es

[2] CEIGRAM, ETSIAAB, Universidad Politécnica de Madrid, Av. Puerta de Hierro, 2 - 4, 28040 Madrid, España; maria.gomezdelcampo@upm.es

Resumen: En el presente trabajo se analiza un novedoso procedimiento de selección de las longitudes de onda que contienen la información más relevante para la estimación de aceite y humedad en aceitunas intactas. Para ello se partió de espectros NIR completos con 700 variables, en los que se analizó el efecto de los pre-tratamientos clásicos. Los resultados pusieron de manifiesto que estos pre-tratamientos pueden tener un efecto negativo, pues llegan a eliminar información espectral relacionada con los contenidos de aceite y agua. Además, los datos crudos, no tratados, ofrecieron estimaciones adecuadas. Este resultado es de gran relevancia a la hora de la simplificación de los modelos de estimación, pues promueve el uso individual de longitudes de onda, en contraposición a los pre-tratamientos clásicos, que requieren normalmente el espectro completo para realizar la corrección. A continuación se llevó a cabo una selección jerárquica de longitudes de onda aplicando un criterio basado en la covarianza entre variables y en la ortogonalización de los espectros respecto de la variable que se va seleccionando, método CovSel. Esta ortogonalización optimiza la selección, ya que los espectros NIR se caracterizan por la altísima correlación entre las longitudes de onda contiguas. La identificación de un número reducido de longitudes de onda permite avanzar en el desarrollo de dispositivos multiespectrales, sencillos y portátiles, como cámaras, que podrían ser utilizados tanto en campo, como en ambientes industriales y de laboratorio.

Palabras clave: Espectroscopia; Inspección de alimentos; Pre-tratamientos espectrales; Selección de variables

1. Introducción

El conocimiento del contenido de aceite y agua de las aceitunas es fundamental para el manejo del cultivo, la decisión del momento de cosecha y el ajuste de las máquinas que intervienen en el proceso de extracción de aceite.

La espectroscopia en el infrarrojo cercano (NIR) ha demostrado su capacidad para determinar el contenido de aceite y agua de las aceitunas intactas [1, 2, 3]. Sin embargo, la simplificación de los procedimientos y de los equipos es de gran interés para el sector oleícola.

El análisis de los datos espectrales NIR generalmente involucra el manejo de cientos de variables que están altamente correlacionadas. Además, la necesidad de un preprocesamiento espectral previo al desarrollo de modelos de regresión aumenta los requisitos computacionales.

El presente trabajo recoge los principales resultados de un estudio sobre la viabilidad de estimar el contenido de aceite y agua mediante métodos computacionales más simples enfocados en algunas longitudes de onda en lugar de todo el espectro desarrollado por

doi: 10.26754/c_agroing.2019.com.3425

Hernández-Sánchez y Gómez-del-Campo [4]. La identificación de las longitudes de onda de interés se realizó mediante el método propuesto por Roger et al. [5] de selección de variables denominado CovSel (Covariance Selection).

2. Materiales y métodos

Se utilizaron aceitunas de la variedad Arbequina obtenidas en diferentes plantaciones, diferentes alturas dentro del árbol y diferentes orientaciones de las líneas de cultivo para incrementar la variabilidad del estado de madurez y, en consecuencia, del contenido de aceite y agua, y, con ello, la variabilidad en los datos espectrales.

El conjunto completo comprendió un total de 95 muestras, con alrededor de cien aceitunas en cada muestra. Las muestras se dividieron aleatoriamente en dos conjuntos: conjunto de calibración con 80 muestras para el desarrollo del modelo; y test set con 15 muestras para validación externa.

La humedad se determinó gravimétricamente y se expresó como porcentaje del peso fresco. El contenido de aceite se midió a partir de frutos de oliva secos utilizando un NMR Minispec NMS100 (Bruker Optik GmbH).

Los espectros NIR se obtuvieron con un equipo FOSS NIRSystems 5000 en el rango de 1100-2500 nm a una resolución de 2 nm (total de 700 longitudes de onda) en modo de reflectancia, luego se transformaron en valores de absorbancia como log (1/R).

Se evaluaron diferentes técnicas de preprocesamiento espectral, incluido el no tratamiento previo de los datos. Se aplicaron las técnicas más comunes, como MSC, algoritmo de suavizado y derivación SavGol con ventana de 21 longitudes de onda, polinomio de tercer grado y primera derivada (SavGol21_3_1); SavGol con ventana de 21 longitudes de onda, polinomio de tercer grado y segunda derivada (SavGol21_3_2), SNV y DT. Para combinar la reducción de los efectos multiplicativos y aditivos debidos a la dispersión, SavGol21_3_2 con SNV; y DT con SNV también se aplicaron.

La selección de la longitud de onda se llevó a cabo de acuerdo con la metodología explicada por Roger et al. [5]. Se definieron tres enfoques: a) selección de longitudes de onda independientes para cada modelo de estimación (aceite y humedad); b) selección de longitudes de onda para modelo que estime ambos parámetros a la vez; c) Selección de índices de longitudes de onda comunes para modelos independientes.

Para el primer enfoque, CovSel se ejecutó en los espectros NIR sin tratamiento previo (matriz X) y los vectores de contenido de aceite (base húmeda) y el vector de contenido de agua (base húmeda) independientemente, con un límite de 15 pasos. Esto produjo una selección de 15 longitudes de onda ordenadas. Las variables se introdujeron paso a paso en los modelos de mínimos cuadrados clásicos. Se construyeron hasta 15 modelos con un número creciente de variables en el orden previamente obtenido (desde una variable hasta 15). La validación cruzada produjo una curva de SEC y una curva de SECV que guió la elección del modelo final.

Para el segundo enfoque, CovSel se ejecutó en la matriz X y la matriz Y con el contenido de aceite y de agua, con un límite de 15 pasos. Esto produjo una selección de 15 longitudes de onda ordenadas también. CovSel se ejecutó una segunda vez para cada parámetro de forma independiente para producir tres clasificaciones de las 15 variables seleccionadas. Análogamente, se calcularon series de 15 modelos de regresión por mínimos cuadrados, una serie para cada respuesta. Los modelos óptimos fueron elegidos estudiando la evolución de la SECV.

Para el tercer enfoque, los índices se calcularon como combinaciones de alturas de pico relativas. CovSel se aplicó para identificar los índices con la mejor capacidad para estimar el contenido de aceite en base al peso fresco (% de materia fresca). Los índices seleccionados se utilizaron para generar modelos para estimar el contenido de agua (% de materia fresca). Al hacerlo, las longitudes de onda requeridas se limitarían a un número óptimo bajo.

Todos los modelos seleccionados se aplicaron al conjunto de validación. Los indicadores de desempeño fueron R^2, SEP y RPIQ.

3. Resultados y discusión

Se partió de espectros NIR completos con 700 variables, en los que se analizó el efecto de los pre-tratamientos clásicos. Los resultados pusieron de manifiesto que estos pre-tratamientos pueden tener un efecto negativo, pues llegan a eliminar información espectral relacionada con los contenidos de aceite y agua (Figura 1). Además, los datos crudos, no tratados, ofrecieron estimaciones adecuadas. Este resultado es de gran relevancia a la hora de la simplificación de los modelos de estimación, pues promueve el uso individual de longitudes de onda, en contraposición a los pre-tratamientos clásicos, que requieren normalmente el espectro completo para realizar la corrección.

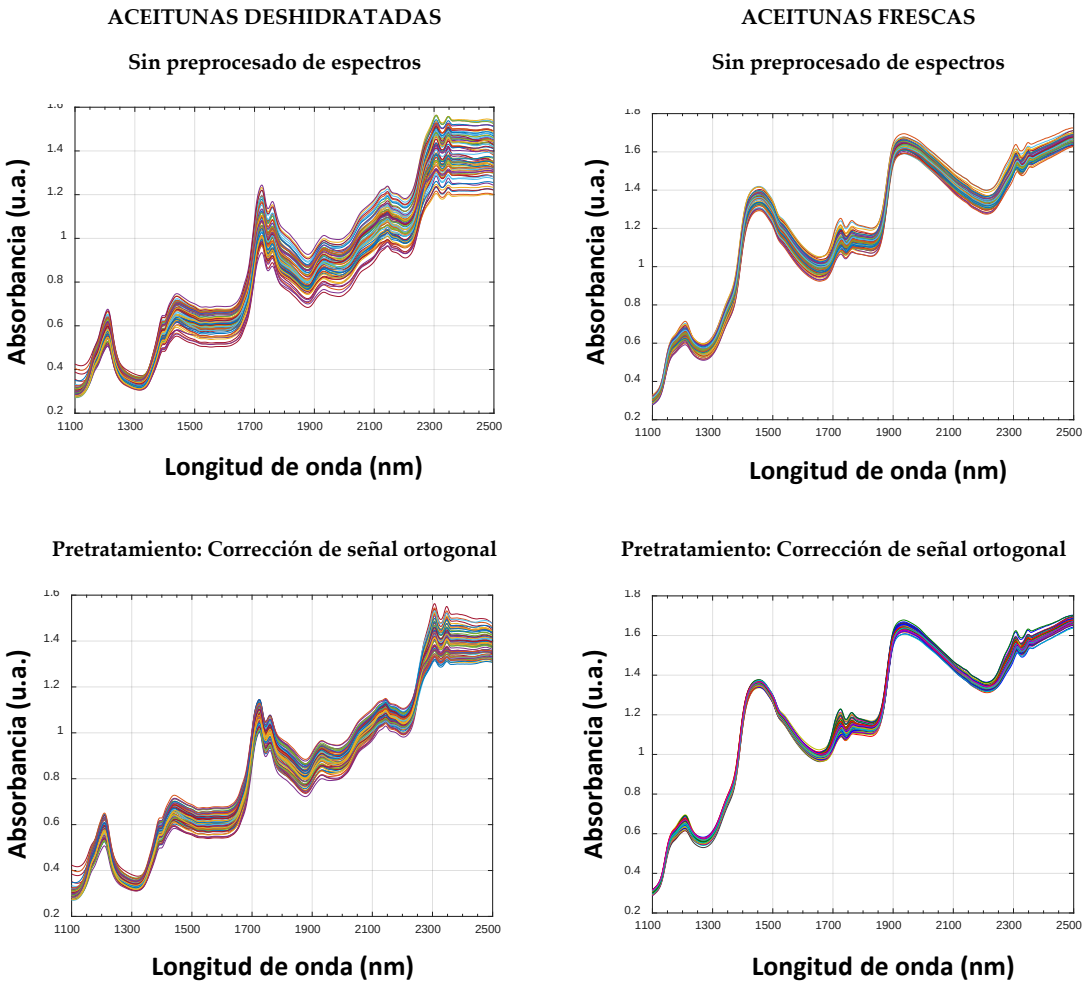

Figura 1. Espectros de absorbancia de aceitunas deshidratadas y aceitunas frescas. Arriba: espectros crudos. Abajo: espectros con un pretratamiento de ortogonalización para eliminar la variabilidad que no está relacionada con el parámetro a estimar, disminuyendo la dispersión de los espectros.

A continuación se llevó a cabo una selección jerárquica de longitudes de onda aplicando un criterio basado en la covarianza entre variables y en la ortogonalización de los espectros respecto de la variable que se va seleccionando. Esta ortogonalización optimiza la selección, ya

que los espectros NIR se caracterizan por la altísima correlación entre las longitudes de onda contiguas.

CovSel clasifica las k variables más útiles de X (matriz de datos) en el orden decreciente de su interés. La variable más útil se selecciona en cada paso. Covsel presenta la ventaja de maximizar la covarianza entre X e Y (matriz de propiedades a estimar) en lugar de la correlación. Para dos variables con la misma correlación con Y, se elegirá la que tenga la mayor covarianza. A continuación, los datos se proyectan ortogonalmente a esta variable seleccionada, lo que elimina la información que se correlaciona con ella. Como las variables vecinas en los espectros NIR están altamente correlacionadas, esta proyección disminuye drásticamente la varianza de las variables adyacentes a la seleccionada. Por lo tanto, la identificación de la siguiente variable en el siguiente paso no tendrá en cuenta dichas variables adyacentes. La consecuencia es que las variables que muestran altas variaciones desempeñan un papel importante en el modelo de regresión (Figura 2). CovSel también trata con Y que contiene múltiples respuestas y realiza la selección de variables en función de su covarianza global con todas las propiedades que se deben predecir.

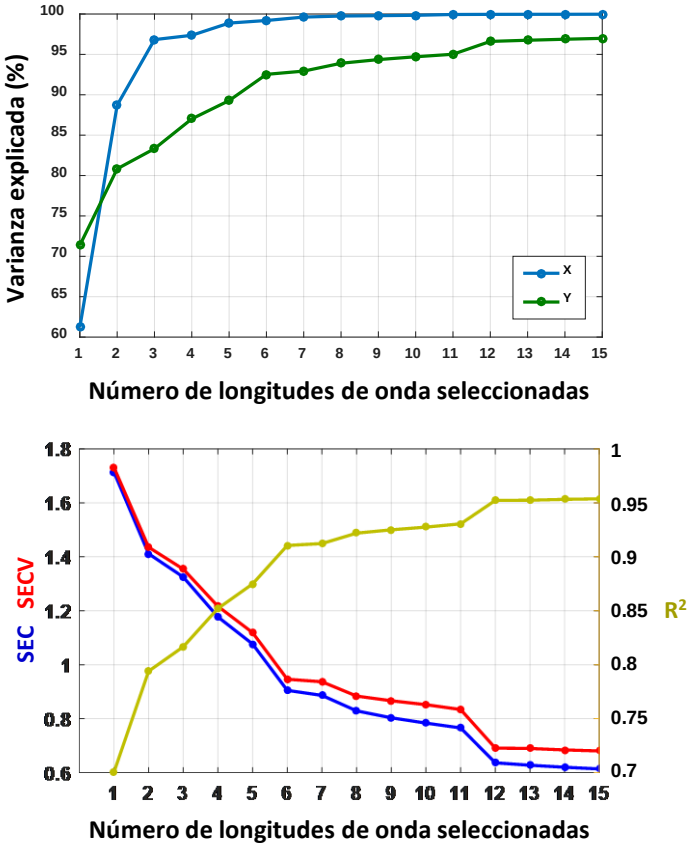

Figura 2. Evolución de la varianza explicada, errores de estimación y R^2 para modelos con número creciente de longitudes de onda seleccionadas mediante el procedimiento CovSel.

Los modelos de estimación desarrollados con los datos espectrales NIR al completo (700 longitudes de onda) sin pretratameintos obtuvieron un r^2 para el conjunto de validación externa de 0.9 para el contenido de aceite y 0.92 para el agua; RPIQt fue 4,9 y 4,3 respectivamente. La identificación de una relación entre la absorbancia a 1724 nm y a 1760 nm con el contenido de aceite, permitió restringir las longitudes de onda a tres. Para el contenido de aceite, el r^2 alcanzó 0,88 y RPIQt fue 4,4. Para el contenido de agua, el valor de r^2 fue de 0,84 y el RPIQt fue de 3,1

(Figura 3). La calidad de la estimación con solo tres longitudes de onda fue comparable al obtenido con PLSR en 700 variables.

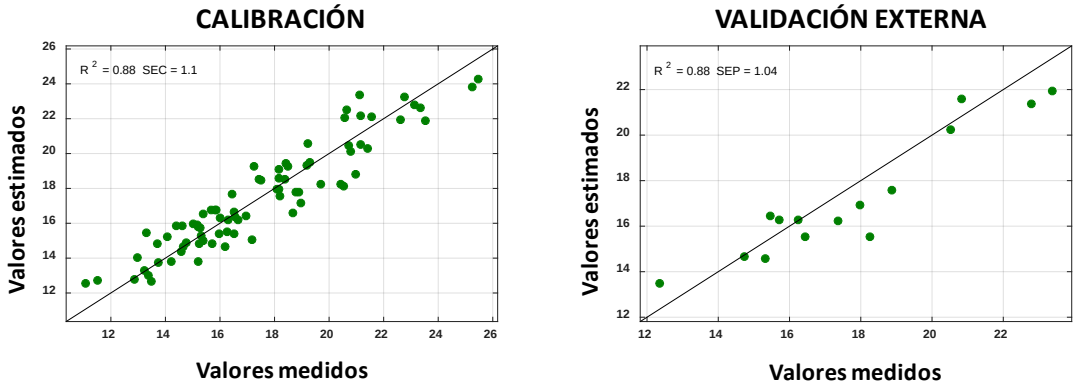

Figura 3. Valores estimados del contenido de aceite (% peso fresco) frente a los valores medidos con métodos de referencia. El modelo de estimación incluye tres longitudes de onda: 1206 nm, 1724 nm y 1760 nm.

4. Conclusiones

La identificación de un número reducido de longitudes de onda para la estimación de los contenidos de aceite y agua en aceituna mediante procedimientos basados en la covarianza permite avanzar en el desarrollo de dispositivos multiespectrales, sencillos y portátiles, como cámaras, que podrían ser utilizados tanto en campo, como en ambientes industriales y de laboratorio.

5. Agradecimientos

Los autores agradecen a Jacinto Cabetas de El Carpio de Tajo (Toledo), Antonio Capitán de Écija (Sevilla), Casas de Hualdo de Puebla de Montalbán (Toledo) y Todolivo de Pedro Abad (Córdoba) por el acceso a los olivares donde se llevó a cabo esta investigación.

Agradecemos a Beatriz Somoza-Rodríguez por su asistencia en la recopilación de datos de espectros NIR. Además, los autores agradecen a Jean Michel Roger de Irstea, UMR ITAP, (Francia) su valioso asesoramiento en quimiometría.

Referencias

1. Cayuela J A and Camino, M dCP. 2010. Prediction of quality of intact olives by near infrared spectroscopy. Eur. J. Lipid Sci. Technol. 112, 1209-1217.
2. Salguero-Chaparro L., Baeten V., Fernández-Pierna J.A., Peña-Rodríguez, F. Near infrared spectroscopy (NIRS) for on-line determination of quality parameters in intact olives, Food Chem. 2013, 139, 1121-1126.
3. Salguero-Chaparro L, and Peña-Rodríguez F. On-line versus off-line NIRS analysis of intact olives. LWT - Food Sci. Technol. 2014, 56, 363-369.
4. Hernandez-Sanchez N., Gomez-del-Campo M. From NIR spectra to singular wavelengths for the estimation of the oil and water contents in olive fruits. GRASAS Y ACEITES. 2018, 69 (4), 1-13.
   Roger J.M., Palagos B., Bertrand D., Fernandez-Ahumada E. CovSel: Variable selection for highly multivariate and multi-response calibration. Application to IR spectroscopy. Chemometr. Intell. Lab. Syst. 2011, 106, 216-223.