RIAA
Revista de Investigación Agraria y Ambiental

Área Agrícola
Recibido: 05-06-2019
Aceptado: 18-02-2020

DETERMINACIÓN DE NITRÓGENO FOLIAR EN PALMA DE ACEITE CON ESPECTROSCOPÍA EN EL INFRARROJO MEDIO (MIR) Y CERCANO (NIR) POR EL MÉTODO DE REGRESIÓN DE MÍNIMOS CUADRADOS PARCIALES DE COMPONENTES PRINCIPALES (PLS)

DETERMINATION OF FOLIAR NITROGEN IN PALM OF OIL WITH SPECTROSCOPY IN THE MIDDLE INFRARED (MIR) AND NEAR (NIR) BY THE REGRESSION METHOD OF MINIMAL PARTIAL SQUARES OF MAIN COMPONENTS (PLS).

[1] Jhoan Jose Crespo Gonzalez, [2]Orlando Simón Ruiz Villadiego, [3]Karen Stefanie Ospino Villalba

[1] Ingeniero Químico, MSc en Ciencias Agrarias.
[2]Quimico, Msc en Ciencia y Técnica del Carbón, Profesor asociado, universidad Nacional de Colombia Sede Medellín.
[3]ingeniera Agroindustrial, MSc. En Ingeniería.

[1]jjcrespo@unal.edu.co; [2]osruiz@unal.edu.co;
[3]ksospinov@unal.edu.co

RESUMEN

Contextualización: la determinación de nitrógeno foliar se utiliza como uno de los índices que mide la necesidad nutricional de la planta en los cultivos de palma de aceite. Para esta investigación, también fue de igual importancia enfocarse en la tendencia en general de los laboratorios llamada "química verde", la cual se centra en minimizar el uso de reactivos químicos en los diferentes análisis de laboratorio.

Vacío de conocimiento: mediante el uso de espectroscopía de infrarrojo medio y cercano (MIR y NIR) se pretendió minimizar en gran medida la generación de contaminantes producidos por el método de Kjeldahl, además de reducir los tiempos de análisis.

Propósito del estudio: determinar la cantidad de nitrógeno foliar mediante la construcción modelos predictivos a partir de los espectros de infrarrojo medio y cercano para la determinación de nitrógeno foliar usando como referencia el método Kjeldahl.

Metodología: en el desarrollo del experimento se analizaron 198 muestras foliares de

DOI: https://doi.org/10.22490/21456453.3206

palma y se tomaron sus respectivos espectros infrarrojos MIR y NIR. Cada uno de los espectros fue pretratado por diferentes métodos matemáticos para corregir efectos de dispersión de la radiación. En total se realizaron 8 pretratamientos a cada uno de los espectros, incluyendo los espectros crudos, que se tomaron a fin de elegir el mejor modelo de predicción tanto para los espectros NIR como para los espectros MIR.

Resultados y conclusiones: utilizando el pretratamiento de variable normal estándar (SNV) en el modelo se obtuvo un menor error de la predicción (RMSE) de 0,265 y un R2 de 0,51 para el infrarrojo cercano y para el infrarrojo medio, el modelo formado con la absorbancia de los espectros sin pretratar arrojó valores de RMSE de 0,245 y un R2 de 0,46. Aunque puede utilizarse de una forma general como modelo de predicción, se pueden observar puntos anómalos que amplían el error y disminuyen el R2, a partir de estos datos se puede evidenciar la necesidad de clasificar de una mejor forma los grupos de muestras foliares y si es necesario realizar modelos de predicción para cada uno de los grupos.

Palabras clave: espectroscopia infrarrojo cercano (NIR); espectroscopia infrarrojo medio (MIR); análisis de nitrógeno foliar, palma de aceite; regresión de mínimos cuadrados parciales (PLS); quimiometría.

ABSTRACT

Contextualization: The determination of foliar nitrogen is one of the criteria that measure the nutritional needs of the plant in oil palm crops. It is also of equal importance in this research to focus on a general trend in laboratories called "green chemistry", which focuses on minimizing the use of chemical reagents in different laboratory analyzes.

Knowledge gap: Using medium and near infrared spectroscopy (MIR and NIR), the intention was to greatly minimize the generation of contaminants produced by the Kjeldahl method, in addition to reducing analysis times.

Purpose: Determine the amount of foliar nitrogen by designing predictive models from the mid and near infrared spectra for the determination of foliar nitrogen using Kjeldahl as a reference method.

Methodology: In the development of the experiment, 198 palm leaf samples were analyzed and their respective MIR and NIR infrared spectra were taken. Each of the spectra was pretreated by different mathematical methods to correct for scattering effects of radiation. In total, 8 pretreatments were performed on each of the spectra, including the raw spectra. These were taken to choose the best prediction model for both NIR and MIR spectra.

Results and conclusions: Using the standard normal variable (SNV) pre-treatment in the model, an RMSE of 0.265 and an R^2 of 0.51 were obtained for the near-infrared and for the mid-infrared, the model formed with the absorbance of the untreated spectra yielded root mean square error (RMSE) values of 0.245 and an R^2 of 0.46. Although it can be used in a general way as a prediction model, anomalous points can be observed that increase the error and decrease the R2, from these data the need to classify the groups of foliar samples in a better way and if it is it is necessary to make prediction models for each of the groups.

Key words: near infrared spectrophotometry (NIR); medium infrared spectrophotometry (MIR); foliar nitrogen analysis; oil palm; partial least squares regression (PLS); chemometric.

RESUMEN GRÁFICO

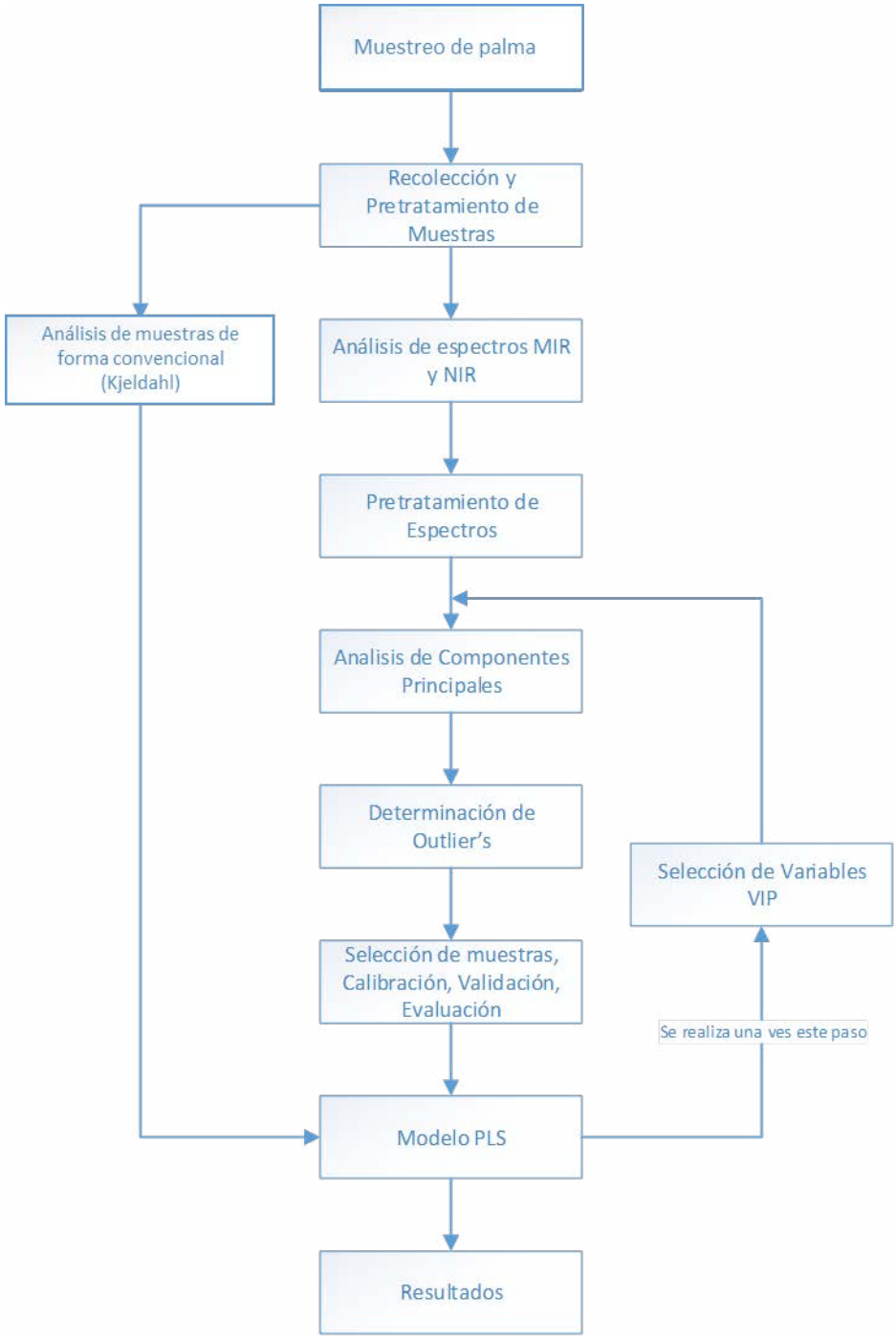

Figura 1. Diagrama del diseño del modelo de predicción. **Fuente:** autores.

1. INTRODUCCIÓN

Las hojas son estructuras fundamentales para el funcionamiento de las plantas, dado que realizan una función esencial para el mantenimiento de todas sus funciones fisiológicas, particularmente, la fotosíntesis; el elemento esencial de este proceso es el

nitrógeno, que forma parte de una de las moléculas más importantes de la tierra, la clorofila. Por lo tanto, este elemento se utiliza como un indicador fundamental del grado de nutrición y del estado general de la planta, junto con el fósforo (Lorén, 2013).

El análisis del tejido foliar es importante, ya que a través de este se define el estado nutricional de la planta y el nitrógeno es uno de los parámetros clave en los análisis foliares, puesto que está relacionado directamente con la cantidad de clorofila en las hojas, lo cual es fundamental en su crecimiento y nutrición (Lorén, 2013). El nitrógeno se puede encontrar en gran parte en las rocas ígneas, este no está disponible para las plantas en el mediano o corto plazo; también está presente en los océanos y en la atmósfera, puesto que se puede encontrar en gran proporción como N_2 que constituye el 78% del aire (Perdomo, 1999). Para que las plantas accedan a este son necesarios ciertos procesos de fijación mediante microorganismos, los cuales toman el nitrógeno atmosférico y lo transforman en formas de amonio de fácil absorción para la planta. Otra fuente de nitrógeno es la materia orgánica presente en el suelo que también lo contiene de forma disponible para las plantas (Perdomo, 1999).

La utilización de espectroscopía de infrarrojo medio (MIR) y cercano (NIR) ha tomado gran importancia en diferentes análisis como foliares y de suelos, estos métodos no destructivos no generan residuos y por tanto contribuyen al cuidado del medio ambiente, además reducen los costos del análisis asociado a los reactivos (PINEDA, 2007). Dichos métodos analíticos relacionan la absorbancia de la energía electromagnética con respecto a la concentración por medio de la ecuación de Lambert beer (Zafra, 2014; Mishra et al., 2017; Nicolaï et al., 2007).

La espectroscopia infrarroja cercana se enfoca en el rango del espectro electromagnético entre los 770 nm hasta los 2500 nm y se caracteriza por mostrar sobretonos y vibraciones moleculares, enfocándose en los enlaces de ciertas moléculas con hidrogeno, como C-H, N-H, O-H y S-H. En cuanto a la espectroscopia infrarroja media se basa en vibraciones fundamentales, además de las vibraciones C-H, N-H y O-H. También se muestran enlaces dobles y triples C=C, C≡C, C=O, C=N entre otros (Forouzangohar, 2009).

Para la comprensión y el análisis de los datos obtenidos es necesario el uso de la quimiometría, que permite la creación de modelos descriptivos y predictivos mediante métodos matemáticos y estadísticos (Olivieri, 2011).

En el análisis de espectros se producen diferentes efectos aditivos y multiplicativos de la señal generalmente ocasionados por el tamaño de partículas, humedad, temperatura o características del equipo. Estos efectos pueden ser disminuidos por pretratamientos matemáticos del grupo de muestras (Galea, 2015), métodos como la corrección del efecto multiplicativo de la dispersión (MSC) (Maleki, et al., 2007), variable normal estándar (SNV) (Cárdenas, 2012) o suavizado de Savitzky-Golay con 1° y 2° derivada (SG-1, SG-2), que son las más utilizadas y ayudan a disminuir el ruido de la señal producido por el equipo (Sila, 2016).

Los métodos descriptivos se definen generalmente como métodos no supervisados, es decir, no se tiene un conocimiento previo sobre las muestras, por tanto, el investigador deberá explicar los grupos formados, como por ejemplo el análisis de componentes principales (PCA) y análisis Clúster (Roggo, et al., 2007). Los métodos predictivos permiten modelar propiedades de un sistema para predecir el comportamiento de una o más variables

(Olivieri, 2011), estos métodos se consideran supervisados ya que se tiene un conocimiento previo del valor de referencia de la variable por predecir, para lo cual se utiliza un conjunto de muestras, llamadas de entrenamiento o calibración, que permiten la construcción del modelo predictivo, entre estos, el modelo de mínimos cuadrados parciales (PLS)(Roggo, et al., 2007; Mevik, 2016).

Es importante resaltar que se deben analizar todas las muestras deben para identificar aquellas que están muy diferenciadas del grupo mayoritario, con el fin de eliminarlas y obtener un modelo sea más preciso. Esta identificación se realiza con los valores de las muestras en los componentes principales, mediante la distancia entre muestras. Para este propósito, existen estadísticos como T^2 de Hotelling o el estadístico Q, que identifican muestras anómalas o *Outliers* (Penha 2001). Es importante además de la identificación de las muestras anómalas, la selección de grupos de calibración y evaluación del modelo, pues la utilización de las mismas muestras, para calibración y para validación del modelo, tiende a generar un sobreajuste de este, debido a que si se evalúa con una nueva muestra, el modelo es ineficiente en su predicción. En este sentido, es recomendable la implementación de métodos de selección de grupos como Kenard-Stone y Duplex, que ayudan a disminuir el sobreajuste y evitar un sesgo en la selección de muestras (Borovicka, 2012).

Para la construcción de un modelo robusto es necesario implementar los pasos anteriormente mencionados, además de incluir la selección de variables con un método de importancia de la variable de proyección (VIP), el cual se encarga se eliminar las variables que no están relacionadas con la variable por predecir. Todos estos procedimientos quimiométricos colaboran en el desarrollo de un modelo robusto y confiable capaz de analizar cada detalle evaluado que permita seleccionar el modelo que tenga la mejor correlación (r^2) y el menor error de la predicción (RMSE). Con este sistema se pretende crear un modelo robusto, donde todo el sistema de selección, clasificación, y agrupación de las muestras sea bajo criterios que garanticen el no sobreajuste del sistema, lo cual hace que el modelo sea mucho más ajustado a la realidad en el caso de tomar nuevas muestras.

2. MATERIALES Y MÉTODOS

Muestras y pretratamientos

Las muestras de palma de aceite fueron suministradas por el Laboratorio certificado "Dr. Calderon Labs." (http://www.drcalderonlabs.com/), en total se recolectaron 198 muestras seleccionadas de diferentes partes del país, debido a la política de protección de la información del laboratorio, no se puede obtener características específicas que pudieran ser útiles.

Según el procedimiento de preparación de muestras del laboratorio "Dr Calderon Labs", se tomaron 100 gr de muestra y se sometieron a un secado en estufa durante 24 a 48 horas a una temperatura entre 60°C y 80°C. Después, las muestras se molieron y tamizaron con un tamiz de 2 mm, seleccionando fracciones < 2 mm.

Análisis Químico

Los análisis de nitrógeno total de las muestras foliares de palma de aceite se realizaron en el laboratorio de suelos de la Universidad Nacional, aplicando el método de Kjeldahl referenciado en el Soil Survey Laboratory Methods Manual (Versión 03,1996): 3A1, 6B. Los equipos utilizados en el análisis fueron un digestor de muestras marca VELP, un destilador automático Kjeltec 8200 marca FOSS y por último una bureta de titulación digital marca BRAND.

Adquisición de Datos Espectrales

Para obtener los datos espectrales, se tomaron los espectros MIR y NIR. Para los espectros MIR, se requirió la formación de la pastilla de KBr. La relación de KBr con respecto a la muestra fue de 100:1 mg, respectivamente, que previamente se pesó y maceró en el laboratorio de suelos. Para la lectura de los espectros se peletizó la muestra con un equipo diseñado para tal fin, el cual proporciona una presión de 2 toneladas/cm2 por 30 segundos. Se utilizó un equipo FT-IR "Spectrum Two DTGS" de la empresa PerkinElmer, con un rango de medición de 4000 cm-1 a 500 cm-1, una resolución (Δx=1 cm-1) y se programó para que tomara 16 espectros de la misma muestra y reportara el espectro promedio.

La obtención de los espectros NIR se realizó en un equipo NIRMaster, de la marca Buchi, el cual es controlado mediante los programas NIRWare y NIRCal. Para cada muestra se tomaron 16 espectros, reportándose como resultado el espectro promedio. El rango de medición es de 10000 cm-1 a 4000 cm-1, con una resolución (Δx=4 cm-1). Las muestras se colocaron en un portamuestras de vidrio y se esparcieron con el fin de formar una capa de material que no permitiera el paso de luz a través de la muestra. Es decir, se utilizó el modo de reflectancia difusa.

Modelamiento

El modelamiento para la determinación de nitrógeno foliar se realizó mediante el seguimiento del diagrama que se muestra en la Figura 1. Se recolectaron las muestras y se analizaron por el método tradicional y se tomaron los espectros NIR y MIR para finalmente proceder al pretratamiento de los espectros.

Todo el proceso quimiométrico se realizó a través del programa estadístico "R". Para la corrección de efectos de dispersión de la señal, se utilizaron en total 8 pretratamientos de señales adicionales a la señal de absorbancia sin pretratamiento.

Los pretratamientos utilizados fueron: 1° y 2° derivada de Savitsky-Golay, en las cuales se utilizó un tamaño de ventana de 13 puntos y se siguió un polinomio de ajuste de segundo orden (Nicolaï, 2007; Jiang, 2017); Corrección del efecto multiplicativo de la dispersión (MSC); variable normal estándar (SNV) y las combinaciones de pretratamientos ("Savitzky-Golay 2 + MSC", "Savitzky-Golay 2 + SNV", "MSC + Savitzky-Golay 2", "SNV + Savitzky-Golay 2") (Rinnan, et al., 2009).

Con los datos recolectados del PCA, se realizaron las pruebas estadísticas para la determinación de muestras atípicas (*outliers*) que fueron el estadístico Q y el estadístico T2 de Hotelling, cada uno con funciones particulares. El estadístico Q se encarga de reconocer los valores atípicos fuera del modelo, mientras que el estadístico T2 se encarga de los valores atípicos dentro del modelo. Con esto se busca rechazar las muestras que estén por fuera del límite de confianza del 95% (α=0,05). Estas muestras se consideran *outliers* y se descartan de la formación del modelo. Este conjunto de muestras varía dependiendo del pretratamiento (Penha 2001).

La selección de las muestras de calibración, validación y evaluación se seleccionaron mediante una variante del método de Kennard-Stone. La selección de muestras en los grupos anteriormente mencionados es importante para la calidad del modelo de predicción y evitar en sobreajustes del modelo (Borovicka, 2012). El método de Kennard-Stone selecciona las muestras por medio de las distancias entre estas (Kennard, 1969), las cuales se calculan con los puntajes del análisis de componentes principales y adicionalmente con la ecuación de Mahalanobis la cual incluye el valor propio de cada componente

(Cao, 2013).Se utilizó el método Dúplex (Snee, 1977), que selecciona de forma intercalada las muestras de calibración y validación.

Para la validación cruzada del modelo se utilizó "Leave One Out" (LOO) (Botero, et al., 2009), del cual se toma el mejor modelo de predicción. Para obtener el número óptimo de componentes principales y se realizó a través de la Raíz cuadrada de los errores cuadráticos Medios (RMSE) del grupo de datos de validación (Riccioli, 2011).

Obtenido el modelo con el número óptimo de componentes se procedió a determinar valores predichos con el grupo de datos de evaluación. Como criterio de evaluación se tomó el RMSE y el coeficiente de determinación R2 (Riccioli, 2011), después se seleccionaron las variables independientes con mayor importancia mediante el método de "importancia de la variable de proyección" (VIP). De igual forma se determinó el RMSE y el R2 del grupo de evaluación.

Para cada uno de los pretratamientos utilizados, se determinó un modelo óptimo, siguiendo los pasos mencionados anteriormente de forma individual. Se recopilaron todos los datos de cada uno de los pretratamientos, con y sin la selección de variables importantes y posteriormente se eligió el mejor modelo, teniendo como criterio de selección los valores bajos de RMSE y Valores altos de R2.

El proceso de modelamiento se realizó siguiendo las bases teóricas mediante el programa estadístico "R", este programa posee diferentes paquetes estadísticos para el procesamiento de los datos (Hewson, 2009). Se realizó un código que sigue los pasos establecidos en la Figura 1. Debido al código realizado, en cada paso del algoritmo, los resultados se almacenan en carpetas según su proceso y según el pretratamiento espectral, obteniendo toda la información que pueden arrojar los procedimientos estadísticos.

3. RESULTADOS Y DISCUSIÓN

Sobre las muestras de palma de aceite no se tiene ningún tipo de información: manejo agronómico, posibles especies diferentes, diferencias en el muestreo, fases fenológicas, etc. Esto con la intención de que el modelo tuviera un rango más amplio sobre las muestras a predecir. La distribución de frecuencias del contenido de nitrógeno se presenta en la **Figura 2**. El rango de variación es el siguiente: 1,52 – 4,38 %, con un valor promedio de 2,642%, y una desviación estándar de 0,39.

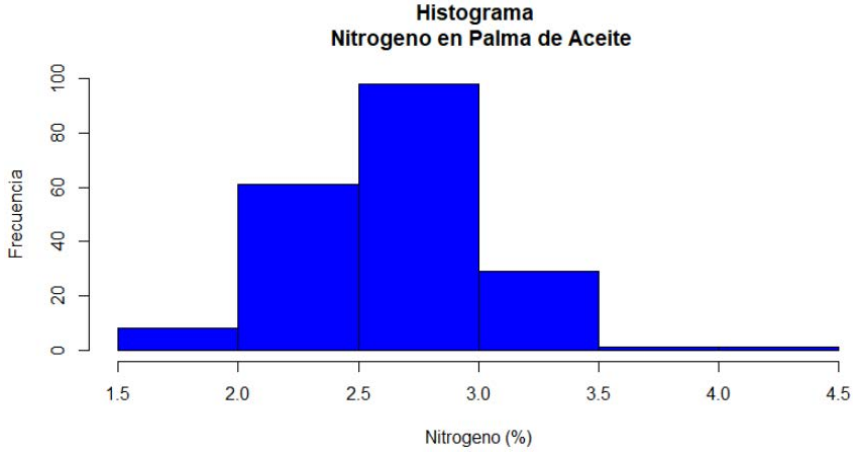

Figura 2. Histograma de frecuencias del contenido de nitrógeno en muestras foliares de palma de aceite.

Fuente: Autores

En la **Figura 2** se muestra la distribución de las concentraciones de nitrógeno en las muestras foliares, donde se observa que la distribución de las muestras presenta un comportamiento cercano a la distribución normal, lo que puede influir de manera relevante en el ajuste del modelo, y por tanto su resulte R^2.

En la **Figura 3**, Los espectros NIR presentan un cambio notable en la línea base debido a los efectos multiplicativos de la señal. En general las señales presentan un comportamiento homogéneo, es decir que los efectos aditivos son muy pequeños en comparación a los efectos multiplicativos. Esto indica que es necesario realizar pretratamientos de línea base para disminuir el error en la selección de muestras anómalas (Rinnan et al., 2009).

Los espectros NIR en general, la zona 9091 a 7194 cm−1, se relacionan con la presencia de carbohidratos, lípidos y proteínas; la zona de 7194 cm-1 a 6024 cm-1 muestra una banda de absorción muy amplia asociada con el primer sobretono de la vibración del enlace O-H y de los puentes de hidrógeno intermoleculares de moléculas de agua; en la región de 6024 y 5348 cm−1 se muestra otra banda de absorción que se asocia con el primer sobretono del estiramiento simétrico y asimétrico del enlace C-H en grupos CH2 y CH3, y se relacionan con la presencia de lípidos y proteínas. La siguiente banda de 5348 a 4963 cm−1, se relaciona principalmente con el contenido de carbohidratos, primer sobretono de los enlaces OH; en la región de 4963 a 4484cm−1 se presentan las combinaciones de las frecuencias de vibración de los enlaces C-H, N-H and C=O presentes en carbohidratos, lípidos y proteínas; igualmente, las bandas en las regiones de 4484 a 4237 cm−1 y 4237 a 4000 cm−1 (Westad, 2008).

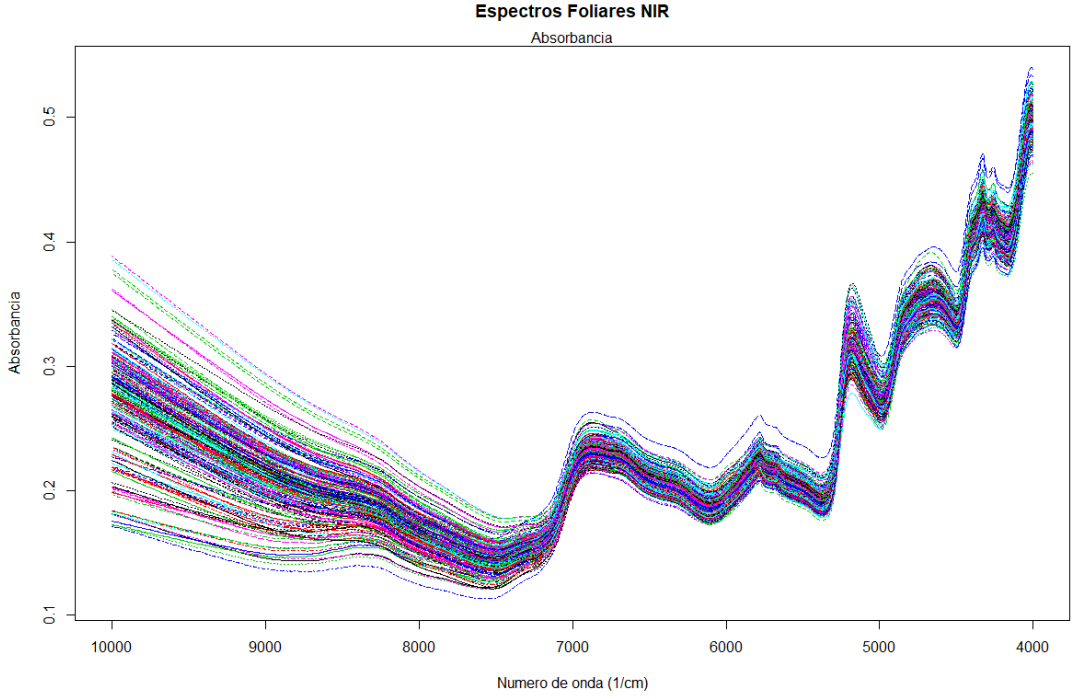

Figura 3. Espectros de Absorbancia NIR de las muestras de foliares de palma de aceite.

Fuente: Autores

El espectro de infrarrojo medio (MIR) de los tejidos foliares de las palmas de aceite (**Figura *4***) muestra las bandas de absorción características de los enlaces C-H (CH2, CH3, CH, aromático) en la región centrada a 2900 cm-1; seguido por una banda muy amplia e intensa que se asocia con los estiramientos del enlace OH, formando puentes de hidrógeno intra e intermoleculares en la región centrada en 3400 cm-1; en la región desde el 1800 cm-1 hasta los 1000 cm-1 se encuentran las señales de los estiramientos y flexiones de los enlaces: C=N,N-H y C-N, C-O, C=O (Westad, 2008).

La línea base de los espectros MIR posee un pequeño efecto multiplicativo en comparación a los efectos aditivos, dichos efectos se resaltan en 3 muestras del grupo, las cuales sufren un desplazamiento aditivo muy pronunciado.

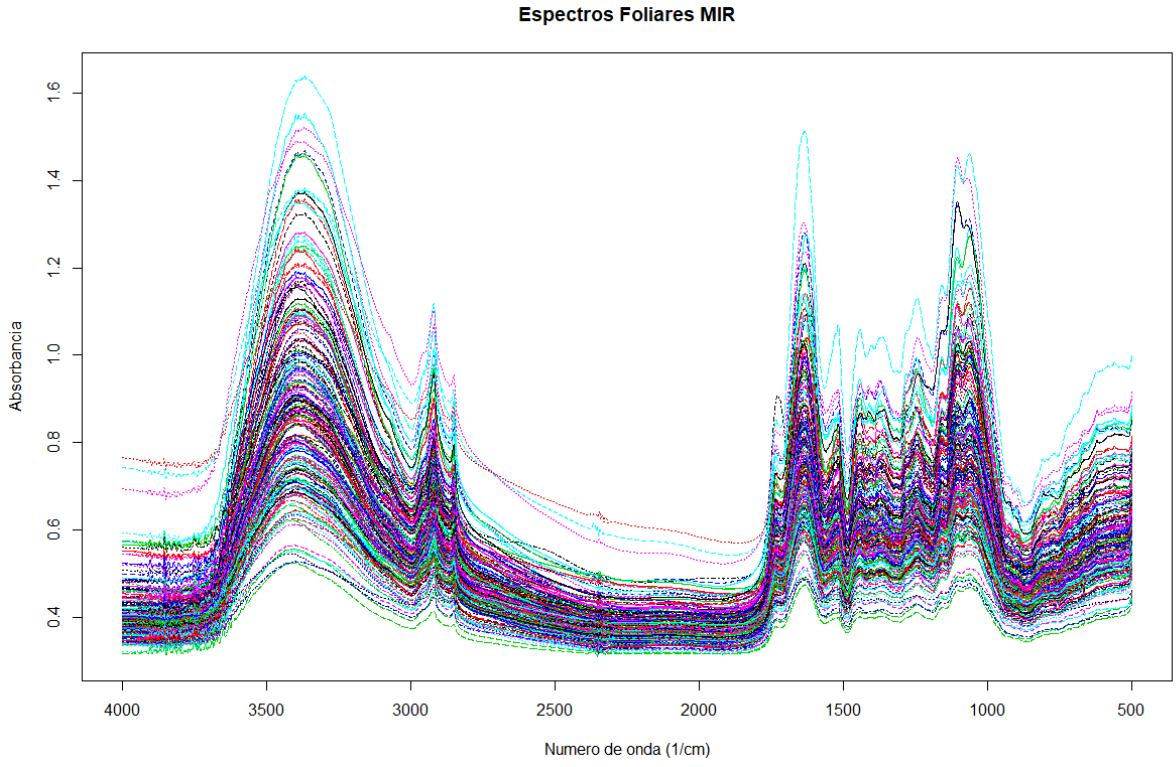

Figura 4. Espectros MIR de las muestras foliares de palma de aceite.

Fuente: Autores

En el análisis de componentes principales no se pudo realizar alguna comparación con respecto a alguna característica en común del grupo de muestras por no tener alguna información cualitativa de ellas, como lugar de procedencia, características del suelo, especie, entre otras. Lo que para el modelo puede tomar dos posturas, por un lado, al incluir dicha aleatoriedad en el sistema indica un rango más amplio de muestras a predecir con características diferentes lo que impactaría de forma positiva el modelo, por otro lado, dependiendo de las características propias de las muestras, si difieren mucho entre sí, le agregan mayor variabilidad al modelo disminuyendo el R^2 afectando de maneta negativa en modelo.

En la **Figura 5** los componentes principales 1 y 2 explican el 96,87% de la variabilidad de los datos y las muestras foliares presentan

un patrón de dispersión muy homogéneo. Sin embargo, en la parte central de la gráfica se pueden apreciar algunos grupos de muestras, pero su diferenciación en el gráfico no es muy clara. Adicionalmente, se pueden observar varias muestras en la periferia del grupo señalado con la esfera de acotación; las cuales, posiblemente, presentan características anómalas.

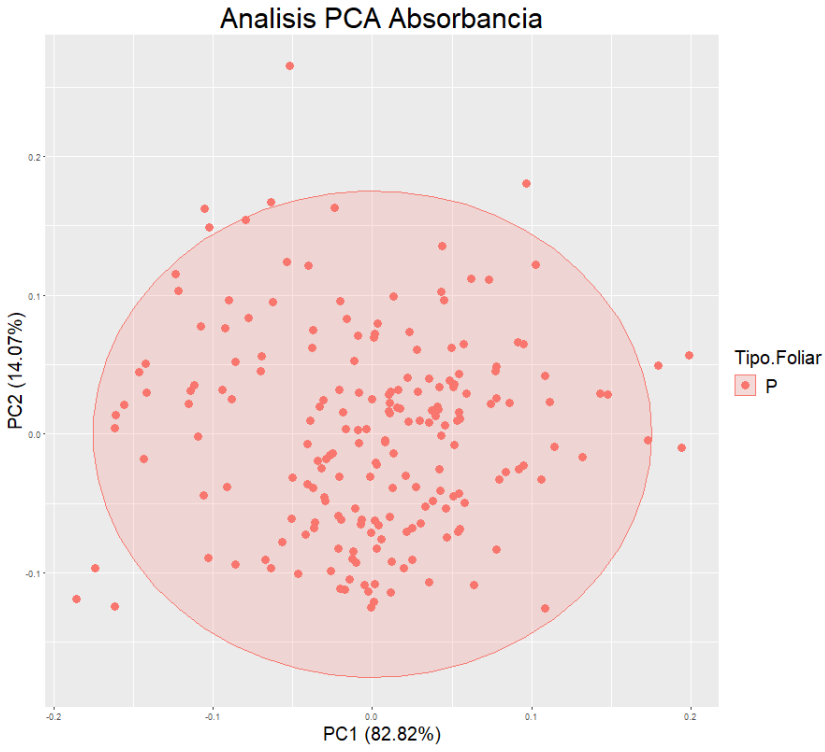

Figura 5. Gráfico de Scores del PCA utilizando los espectros NIR de muestras foliares de Palma de Aceite.

Fuente: Autores

Con el fin de analizar corroborar lo descrito en el análisis PCA se realizó el análisis clúster, para lo cual se utilizó el método de Ward como método de aglomeración. En esta etapa se evidencia tanto para el análisis clúster del NIR como del MIR un gran grupo de muestras que está formado por el 50% aproximadamente. En muchos casos, cuando se presentan este tipo de situaciones, estas muestras definen el comportamiento global del modelo. Si los otros grupos se diferencian mucho de este grupo principal, se recomienda construir modelos separados para cada grupo. Generalmente esto mejora el poder predictivo de los modelos para cada grupo de muestras.

En este trabajo se realizaron los diferentes pretratamientos con el fin de eliminar los efectos aditivos y multiplicativos de la señal, además de las diferentes combinaciones de pretratamientos como lo aconsejan Rinnan, et al., (2009) y Huang (2010), los cuales presentan diferentes alternativas de combinaciones. La combinación de los pretratamientos de corrección de efectos de dispersión (MSC, SNV), seguidos de la segunda derivada de Savitsky-Golay dieron mejores resultados al disminuir en gran medida la señal de ruido.

A cada uno de los 9 pretratamientos se les extrajo las muestras anómalas y se formaron los

grupos de calibración validación y evaluación. Las muestras de calibración y validación fueron seleccionadas mediante el método Dúplex, mientras que el grupo de evaluación, que corresponde al 20% de las muestras, se forma de las muestras sobrantes de la selección de los otros grupos. Esto puede ser peligroso para la evaluación del modelo ya que puede sesgar la información.

En el diseño del modelo PLS se utilizó el grupo de calibración para la creación del modelo, el grupo de validación para escoger el número de componentes principales que necesita el modelo y disminuir el error de la predicción; y finalmente, el grupo de evaluación para observar la correlación entre los valores originales de nitrógeno obtenidos por el método kjeldahl y los obtenidos mediante la predicción del modelo.

Tabla 1. Resultado de la predicción para cada uno de los pretratamientos utilizando la espectroscopia NIR, sin selección de variables y con selección de variables

Pretratamientos	RMSE	R^2	RMSE con VIP	R^2 con VIP
Absorbancia	0,276	0,2658	0,252	0,3747
Savitzky-Golay 1	0,344	-0,0877	0,261	0,3939
Savitzky-Golay 2	0,308	-0,263	0,31	0,0785
MSC	0,265	0,5123	0,304	0,5021
SNV	0,276	0,5109	0,301	0,4907
Savitzky-Golay 2 + MSC	0,283	0,1307	0316	-1,038
Savitzky-Golay 2 + SNV	0,292	-0,1049	0,28	0,0452
MSC+Savitzky-Golay 2	0,339	0,3626	0,322	0,3263
SNV+Savitzky-Golay 2	0,281	0,4375	0,384	0,204

Fuente: autores

Tabla 2. Resultado de la predicción para cada uno de los pretratamientos utilizando la espectroscopia MIR, sin selección de variables y con selección de variables

Pretratamientos	RMSE	R^2	RMSE con VIP	R^2 con VIP
Absorbancia	0,245	0,4581	0,279	0,293
Savitzky-Golay 1	0,254	0,2286	0,291	0,4464
Savitzky-Golay 2	0,343	0,1089	0,318	0,2891
MSC	0,306	-0,1053	0,227	-0,0803
SNV	0,287	-0,3219	0,318	0,017
Savitzky-Golay 2 + MSC	0,293	0,2871	0,372	-0,1476
Savitzky-Golay 2 + SNV	0,459	-0,1109	0,381	0,2156
MSC+Savitzky-Golay 2	0,342	0,3395	0,326	0,3913
SNV+Savitzky-Golay 2	0,28	0,3923	0,383	0,2127

Fuente: Autores

En las Tabla 1 y Tabla 2 se muestran los valores de la correlación entre los datos originales sobre el nitrógeno y los predichos por los modelos de cada uno de los pretratamientos realizados, tanto para los datos obtenidos con infrarrojo cercano (NIR) como para infrarrojo medio (MIR). En ambas tablas se puede notar que las correlaciones con VIP son generalmente menores que aquellas a las cuales no se le realizó la selección de variables. Por tanto, para este conjunto de datos se evidencia que la selección de variables no mejora la predicción del modelo, lo cual, es indiferente ya que el modelo toma el mejor acercamiento de los valores predichos, en todo caso, no se puede eliminar este paso del modelo ya que solo ha sido evaluado en un solo grupo de muestra, y no hay evidencia suficiente de que esta selección de muestra sea innecesaria.

El mejor modelo en el caso de la espectroscopia NIR fue con el pretratamiento MSC con un r^2 de 0,5123 y un RMSE de 0,265. Para el caso de espectroscopia MIR, el caso donde se creó el modelo con los espectros crudos obtuvo los mejores resultados, un r^2 de 0,4581 y un RMSE de 0,245. Tanto en los resultados de los modelos NIR como MIR se constata lo propuesto por Rinnan, et al., (2009). Es decir, para las combinaciones de pretratamientos es recomendable utilizar primero un pretratamiento de corrección de efectos de línea base seguido de las derivadas de Savitsky-Golay.

Los modelos obtenidos en general poseen un coeficiente de correlación pobre. Hay escasos documentos que relacionan los análisis de nitrógenos en muestras foliares, James y colaboradores (Jayaselan et al., 2017) muestra que en la elaboración de un modelo de determinación de nitrógeno foliar de palma de aceite a partir de métodos espectroscópicos es fundamental el muestreo, ya que dependiendo de la hoja donde se tome la muestra puede o no haber una buena correlación de la predicción. Estos realizan diferentes correlaciones en diferentes tipos de hojas los cual se muestra una excelente predicción en la hoja 17, es decir que se evidencia una variabilidad dependiendo de la hoja donde se tome el análisis foliar.

4. CONCLUSIONES

En general, la aplicación de la técnica descriptiva no supervisada PCA, en conjunto con el análisis clúster, evidencia que las muestras poseen poca relación. Cerca de la mitad de las muestras si tienen relación entre sí, lo que afecta notoriamente al modelo de predicción. Por otro lado, estos análisis cualitativos muestran de forma clara las diferencias entre las muestras, esto es de gran importancia ya que se puede ver con anticipación como la selección de muestras influye en un modelo adecuado.

Al tener un grupo con mucha variación, el procedimiento de muestras anómalas no es eficiente, ya que, dada la existencia de tantas muestras con alta variación, el sistema no alcanza a separarlas todas, quedando dentro del modelo parte de ellas. La selección de muestras anómalas de manera positiva selecciona de forma eficiente las muestras basándose en los criterios del Análisis de componentes principales.

Para la selección de los grupos de calibración, validación y evaluación el modelo garantiza una selección variada que garantiza que haya homogeneidad en los datos.

La determinación del nitrógeno foliar en palma de aceite, mediante la construcción de un modelo predictivo con las muestras analizadas, arrojó datos regulares sobre los análisis evaluados, probablemente debido a diferencias entre las muestras, o por factores como el muestreo, manejo agronómico, especie o ubicación del cultivo, entre otros.

A pesar de lo anterior, el modelo sí muestra robustez y una gran cantidad de información relevante en la caracterización y predicción de la modelo ajustada a la realidad y a la variabilidad de las muestras. Por consiguiente, es necesario obtener información específica acerca de las muestras y un mayor tamaño muestral, lo que garantiza un rango amplio en la concentración de nitrógeno, a fin de que el modelo tenga una mejor predicción, ya que al concentrarse una cantidad de muestras en un rango tan corto de nitrógeno hace que el R^2 del modelo sea muy sensible a pequeñas variaciones en los datos predichos.

En estos resultados también se detalla que el modelo PLS, al utilizar todas las variables predictoras, no disminuye de manera significativa el coeficiente de determinación, es decir que la variabilidad que aporta estas variables es menor; por tanto, queda a elección del investigador utilizar todas las variables o las VIP. Esto es de gran relevancia ya que muestra que con una cantidad menor de variables los valores de predicción se mantienen, dando paso a diseño de equipos más compactos y de menor costo para dicho análisis. Este proyecto abre de igual forma las puertas a la utilización de modelos más complejos de clasificación y predicción como SIMCA, PLS-DA, o Redes Neuronales.

CONTRIBUCIÓN DE LA AUTORÍA

Primer autor: metodología, investigación, análisis de datos, conceptualización, escritura – borrador original. **Segundo autor:** conceptualización, análisis de datos, supervisión, escritura – revisión y edición. **Tercer autor:** investigación, logística, revisión y edición.

AGRADECIMIENTOS

Este trabajo fue apoyado por el Centro de Desarrollo Tecnológico del Cesar (CDT).

LITERATURA CITADA

Borovicka, T. (2012). Training Set Construction Methods. … Sciences & Technologies: Bulletin of the ACM. Czech Technical University. Retrieved from https://bit.ly/2zIoGsH

Botero Herrera, J. M., Parra Sánchez, L. N., & Cabrera Torres, K. R. (2009). Determinación del nivel de nutrición foliar en banano por espectrometría de reflectancia. *Revista Facultad Nacional de Agronomía Medellín*, *62*(2), 5089–5098. Retrieved from http://revistas.unal.edu.co/index.php/refame/article/view/24919

Cao, N. (2013). Calibration optimization and efficiency in near infrared spectroscopy, 183. Retrieved from http://lib.dr.iastate.edu/etd/13199/

Cárdenas, V. (2012). Use of NIR spectroscopy and multivariate process spectra calibration methodology for pharmaceutical solid samples analysis (Tesis de Master). Universidad Autónoma de Barcelona, Catalunya, España.

Forouzangohar, M. (2009). Infrared Spectroscopy and Advanced Spectral Data Analyses to Better Describe Sorption of Pesticides in Soils (Tesis de Doctorado). University of Adelaide, Australia.

Galea, F. (2015). Desarrollo de un modelo predictivo usando tecnología NIRs para determinar las extracciones del triticale de doble aptitud (forraje y grano). Universidad de Extremadura.

Hewson, P. J. (2009). Multivariate Statistics with R. Multivariate Statistics with R. Recuperado de https://www.semanticscholar.org/paper/Multivariate-Statistics-with-R-Hewson/3cc07941fecb82ae6cb3ec97ca4375c1b82c0b34

Huang, J., Romero-Torres, S., y Moshgbar, M. (2010). Practical considerations in data pre-treatment for NIR and Raman spectroscopy. *American Pharmaceutical Review*. Recuperado de: https://www.americanpharmaceuticalreview.com/Featured-Articles/116330-Practical-Considerations-in-Data-Pre-treatment-for-NIR-and-Raman-Spectroscopy/

Jayaselan, H., Nawi, N., Ismail, W., Shariff, A., Rajah, V., y Arulandoo, X. (2017). Application of Spectroscopy for Nutrient Prediction of Oil Palm. *Journal of Experimental Agriculture International*, 15(3), 1–9. https://doi.org/10.9734/JEAI/2017/31502

Jiang, Q., Li, Q., Wang, X., Wu, Y., Yang, X., y Liu, F. (2017). Estimation of soil organic carbon and total nitrogen in different soil layers using VNIR spectroscopy : Effects of spiking on model applicability. *Geoderma*, 293, 54–63. https://doi.org/10.1016/j.geoderma.2017.01.030

Kennard, R. W., y Stone, L. A. (1969). Computer Aided Design of Experiments. *Technometrics*, 11(1), 137–148. https://doi.org/10.1080/00401706.1969.10490666

Lorén, F. J. (2013). Estudio de la fertirrigación nitrogenada con el inhibidor de la nitrificación 3,4 DIMETILPIRAZOLFOSFATO (DMPP) en melocotonero 'Miraflores' (Tesis de Doctorado). Universidad de Zaragoza, España.

Maleki, M. R., Mouazen, A. M., Ramon, H., y De Baerdemaeker, J. (2007). Multiplicative Scatter Correction during On-line Measurement with Near Infrared Spectroscopy. *Biosystems Engineering*, 96(3), 427–433. https://doi.org/10.1016/j.biosystemseng.2006.11.014

Mevik, B.-H., Wehrens, R., y Liland, K. H. (2016). Partial Least Squares and Principal Component Regression. Packages R CRAN. Retrieved from https://cran.r-project.org/web/packages/pls/pls.pdf

Mishra, P., Asaari, M. S. M., Herrero-Langreo, A., Lohumi, S., Diezma, B., y Scheunders, P. (2017). Close range hyperspectral imaging of plants: A review. *Biosystems Engineering*, 164, 49–67. https://doi.org/10.1016/j.biosystemseng.2017.09.009

Nicolaï, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., y Lammertyn, J. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biology and Technology*, 46(2), 99–118. https://doi.org/10.1016/j.postharvbio.2007.06.024

Olivieri, AC, Rivas, GA, (2011). Química Analítica en el siglo XXI: modelado de datos instrumentales y miniaturización de sistemas analíticos. *Ciencia Hoy*, 21, 51-56.

Penha, R., & Hines, J. (2001). Using principal component analysis modeling to monitor temperature sensors in a nuclear research reactor. Recuperado de http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.5158

Perdomo, C., y Barbazán, M. (1999). Nitrógeno, 70. Universidad de la República, Uruguay. Recuperado de http://www.fagro.edu.uy/~fertilidad/publica/Tomo N.pdf

Pineda, V. M. (2007). Determinacion Del Contenido De Materia Organica En Suelos Guatemaltecos Por Medio De La Tecnica De Reflectancia Con Espectroscopia De Infrarrojo Cercano (Tesis de Pregrado). Universidad de San Carlos de Guatemala, Guatemala.

Riccioli, C. (2011). Detección y cuantificación de la especie en harinas proteicas de origen animal mediante el uso de de sensores hiperspectrales (Tesis de Doctorado). Universidad de Córdoba, Córdoba, España.

Rinnan, Å., Berg, F. van den, y Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. TrAC - *Trends in Analytical Chemistry*, 28(10), 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007

Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., y Jent, N. (2007). A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *Journal of Pharmaceutical and Biomedical Analysis*, 44(3 SPEC. ISS.), 683–700. https://doi.org/10.1016/j.jpba.2007.03.023

Sila, A. (2016). Multivariate calibration techniques for infrared spectroscopy data (Tesis de Doctorado). University of Nairobi, Kenya.

Snee, R. D. (1977). Validation of Regression Models: Methods and Examples. *Technometrics*, 19(4), 415–428. https://doi.org/10.1080/00401706.1977.10489581

Westad, F., Schmidt, A., & Kermit, M. (2008). Incorporating chemical band-assignment in near infrared spectroscopy regression models. *Journal of Near Infrared Spectroscopy*, 16(3), 265–273. https://doi.org/10.1255/jnirs.786

Zafra, I. (2014). Potencial de datos espectrales NIRS "on-site" para la deteccion de desviaciones de producto en leche respecto al estandar de calidad y seguridad (Tesis de Maestría). Universidad de Oviedo, España.

Conflicto de Intereses
Los autores declaran no tener ningún conflicto de intereses