Volver al contenido

Estudio de firmas espectrales de palmas de aceite afectadas con la Marchitez letal, usando análisis estadísticos de datos funcionales*

Study of the Spectral Signatures of Oil Palms Affected with Lethal Wilt, Using Functional Data Statistic Analysis

AUTORES: Ramón Giraldo, Angie Molina Villarreal, Jorge Luís Torres León, María Claudia Acosta y Sergio Martínez.

CITACIÓN: Giraldo, R., Molina, A., Torres-Leon, J. L., Acosta, M., & Martínez, S. (2016). Estudio de firmas espectrales de palmas de aceite afectadas con la Marchitez letal, usando análisis estadísticos de datos funcionales. *Palmas 37*(Especial Tomo I), pp. 131-139.

PALABRAS CLAVE: análisis de datos funcionales, Marchitez letal, firmas espectrales, detección temprana.

KEYWORDS: Functional data analysis, Lethal wilt, spectral signatures, early detection.

*Artículo original recibido en español.

Ramón Giraldo
Profesor asociado, Departamento de Estadística, Universidad Nacional de Colombia. Associate Professor, Department of Statistics, National University of Colombia
rgiraldoh@unal.edu.co

Resumen

El análisis de datos funcionales (ADF) se ha convertido en la última década en un área de creciente impulso y desarrollo dentro de la estadística. En este trabajo se explora su uso en la modelación de firmas espectrales de palma de aceite. A manera de ilustración se muestran los resultados obtenidos con curvas de reflectancia de palmas localizadas en lotes con alta incidencia de la enfermedad Marchitez letal (ML) y lotes en donde históricamente no se han reportado casos, los cuales hacen parte de una plantación de palma de aceite localizada en la Zona Oriental palmera. Se estudian las diferencias en el comportamiento espectral de las curvas bajo tres condiciones de observación: palmas que presentan síntomas visibles de la enfermedad (ML), palmas aparentemente sanas dentro de foco (SDF) y palmas sanas fuera de foco (SFF).

Inicialmente se realizó un proceso de suavizado por B-splines para convertir los datos puntuales en curvas de reflectancia. Con ellas se llevó a cabo un análisis descriptivo y exploratorio funcional con el objetivo de identificar tendencias y variabilidad de las mismas. Se determinaron a través de técnicas de profundidad de datos funcionales las curvas atípicas, las cuales fueron posteriormente eliminadas de la muestra y de los análisis subsecuentes. Se usaron ANOVAS no paramétricas (pruebas de Kruskal-Wallis) en cada longitud de onda para determinar diferencias entre los valores medios de reflectancia de los tres grupos, en cada nivel foliar. Así mismo, se hicieron pruebas de comparación múltiple no paramétricas (pruebas de Nemenyi) para establecer entre cuáles grupos se presentaban las diferencias detectadas con las pruebas de Kruskal-Wallis. También se efectuaron pruebas de muestras pareadas (Wilcoxon) para establecer diferencias entre las observaciones de campo y laboratorio, en cada condición experimental (ML, SDF y SFF).

Las pruebas de Kruskal-Wallis indican, tanto con la información de campo como con la de laboratorio, que hay diferencias significativas en los valores medios de reflectancia. Con los datos de campo estas diferencias se dan en todo el espectro, mientras que con los de laboratorio dichas diferencias se presentan en las regiones del verde, rojo e infrarrojo cercano. En general, las firmas espectrales de palmas con ML tienen reflectancia media mayor que las SDF y SFF, y la reflectancia media de campo (en los tres grupos y las dos hojas) es mayor que la registrada en laboratorio.

Desde un punto de vista metodológico, los resultados indican que las técnicas de ADF consideradas brindan en el contexto de las firmas espectrales muchas posibilidades de modelación, facilitando la interpretación de datos que son complejos de analizar a través de sus contrapartes tradicionales.

Abstract

Functional Data Analysis (FDA) has become an area of growing momentum and development in statistics during the past decade. This paper explores its use in oil palm spectral signature modeling. By way of illustration, the results obtained with reflectance curves of palms located in a plot with a high incidence of lethal wilt disease are shown. The differences in the spectral behavior of curves are studied under three observation criteria: palms showing visible signs of the disease (LW), apparently healthy palms that are in the focus of the disease (HIF), and apparently healthy palms outside of its focus (HOF).

An initial exploratory functional analysis (EFA) of the reflectance curves was performed, with the purpose of identifying their trend and variability, taking into account each of the conditions (LW, HIF and HOF) and the foliar levels (9 and 17) considered. Multivariate functional techniques (main component analysis, FMCA, and functional classification, FC) were employed to find reflectance curve groups with similar characteristics. An analysis of functional variance (FANOVA) was applied to determine the differences between the three groups (LW, HIF and HOF) considering their mean spectral curves and functional regression (FR) curves in order to establish the association between curves and temperature and humidity.

Results show that the FDA techniques considered (EFA, FMCA, FC, FANOVA and FR) provide several modeling possibilities within the context of spectral signatures, easing the interpretation of data that are difficult to analyze through their traditional counterparts (regression, Anova, main components and classification). Engaging in statistical spectral signature research from a functional perspective, in addition to providing a more general view, will prevent considering classical statistical models based on assumptions that are difficult to comply with in practice.

Introducción

El diagnóstico temprano y no destructivo de enfermedades en palmas de aceite y, en general, de cualquier tipo de cultivo, es un tema de creciente importancia en la agricultura.

Debido a las relaciones entre las propiedades ópticas de las plantas, la concentración de pigmentos de las hojas y los cambios estructurales de las mismas ante la presencia de un patógeno, la espectroscopia de reflectancia se ha usado como una metodología para la detección de enfermedades vegetales. Varios estudios se han realizado en este campo, en los cuales se comparan diferentes herramientas estadísticas para el análisis de este tipo de información. Por ejemplo, Zhang *et al*. (2001) usan el análisis de componentes principales y un análisis de clasificación para la discriminación, a partir de información espectral, de tomates enfermos y sanos. Whang and Sousa (2005) aplican pruebas de análisis de varianza (ANOVA) y análisis lineal discriminante (LDA) sobre curvas de espectrometría con el fin de establecer diferencias entre especies de mangle; Vaiphasa (2005) hace un estudio similar pero usando la distancia de Jeffries-Matusita para cuantificar la diferencia entre las curvas de diversas especies. Lelong *et al*. (2010) usan regresiones por mínimos cuadrados parciales y luego un LDA para clasificar entre palmas sanas y varios niveles de afectación por *Ganoderma*. También, Shafri *et al*. (2011) estudian palmas afectadas por este hongo a partir de curvas de espectrometría, para lo cual efectúan pruebas ANOVA por pares de tratamientos y posteriormente un LDA con las longitudes de onda en donde se encontraron diferencias significativas.

En este trabajo se estudia el comportamiento de curvas espectrales de reflectancia tomadas sobre las hojas 9 y 17 de palmas de aceite localizadas en una plantación del Meta, Colombia. Se consideran tres tipos de condición: palmas que presentan síntomas visibles de la enfermedad (ML), palmas aparentemente sanas que se encuentran dentro del foco de la enfermedad (SDF) y palmas sanas fuera de foco (SFF). Para llevar a cabo el análisis, se hace un suavizado de las curvas usando una base de B-Splines, se detectan curvas anómalas o *outliers* usando una medida de profundidad de datos funcionales, se aplican pruebas de homogeneidad de varianza y normalidad para cada longitud de onda y a partir de sus resultados se hacen pruebas ANOVA o de Kruskal-Wallis (dependiendo de si se cumplen o no los supuestos). Por último, se usan pruebas de comparaciones múltiples paramétricas o no-paramétricas y se establece, primero en qué longitudes de onda se presentan diferencias entre los grupos y luego entre qué combinación de pares de grupos se encuentran las diferencias.

El presente documento está organizado de esta manera: en la siguiente sección se exponen los datos y la metodología de análisis. Luego, se muestran y discuten los resultados del suavizado, la detección de *outliers*, la verificación de supuestos y los valores P de las pruebas sobre la hipótesis de igualdad de las curvas medias de los grupos (ANOVA o Kruskal-Wallis) y de las pruebas de comparación múltiple. Posteriormente se presenta una sección de conclusiones, en donde se resumen los principales resultados obtenidos con la información analizada y finalmente se da un listado de las referencias bibliográficas consultadas.

Materiales y métodos

Este trabajo se realizó en una plantación comercial localizada en el departamento del Meta, Colombia. Los lotes seleccionados se caracterizan por tener el mismo material de palma y una edad de 6 años aproximadamente. Dichos lotes fueron elegidos teniendo en cuenta las tres condiciones de observación propuestas (palmas con ML, palmas SDF y palmas SFF). Para cada grupo se consideró una muestra de 40 palmas, sobre las cuales se recolectaron valores de reflectancia en las regiones del visible y el infrarrojo cercano (entre 400 y 1.075 nm). Las mediciones se hicieron en las hojas 9 y 17, efectuándose dos tomas por hoja, las cuales constaron de 15 repeticiones cada una, y fueron adquiridas en fase de campo y laboratorio.

Inicialmente se propone elaborar gráficos descriptivos con los datos originales, y luego mediante el uso de bases de B-splines (Ramsay and Silverman, 2005) se obtienen curvas suavizadas y la curva media correspondiente de cada grupo. A partir de los resultados obtenidos anteriormente, se efectúa

un análisis de profundidad de datos funcionales (López-Pintado and Romo, 2009) para cada grupo, con el propósito de detectar y eliminar las curvas *outliers* (extremas).

A continuación se realizan pruebas de normalidad y de homogeneidad de varianzas sobre los datos suavizados, para cada longitud de onda, a través del test de Lillifors (Dallal and Wilkinson, 1986) y del test de Levene (Levene, 1960), respectivamente. Con base en estos resultados se determinan si se pueden usar análisis de varianzas paramétricos (ANOVA), o si es necesario acudir a pruebas no paramétricas (Kruskal-Wallis).

Posteriormente, se obtienen gráficos de dispersión entre la longitud de onda (en el eje x) y los valores P (en el eje y) de las pruebas de comparación (ANOVA o Kruskal-Wallis), para identificar en cuáles longitudes de onda se dan diferencias significativas en los valores medios de espectrometría de los grupos.

Por último, se aplican, en cada longitud de onda, pruebas de comparación múltiple de Tukey (Hinkelmann and Kempthorne, 1994) si se satisfacen los supuestos de normalidad y homogeneidad de varianzas, o pruebas de comparación múltiple no paramétricas de Nemenyi (Lothar, 1997), en caso contrario.

Resultados y discusión

A continuación se presentan los resultados de los análisis de campo y de laboratorio en cada una de las hojas evaluadas (9 y 17). Inicialmente se hizo un análisis descriptivo de los datos por grupo, el cual consistió en graficar los valores de reflectancia para cada longitud de onda (líneas grises claras continuas), las curvas grises oscuras continuas o promedio, e identificar (en al menos uno de los grupos), a través del estudio de profundidad de datos funcionales, curvas extremas en función del comportamiento global de las curvas consideradas en cada grupo (líneas punteadas) (Figuras 1 y 2).

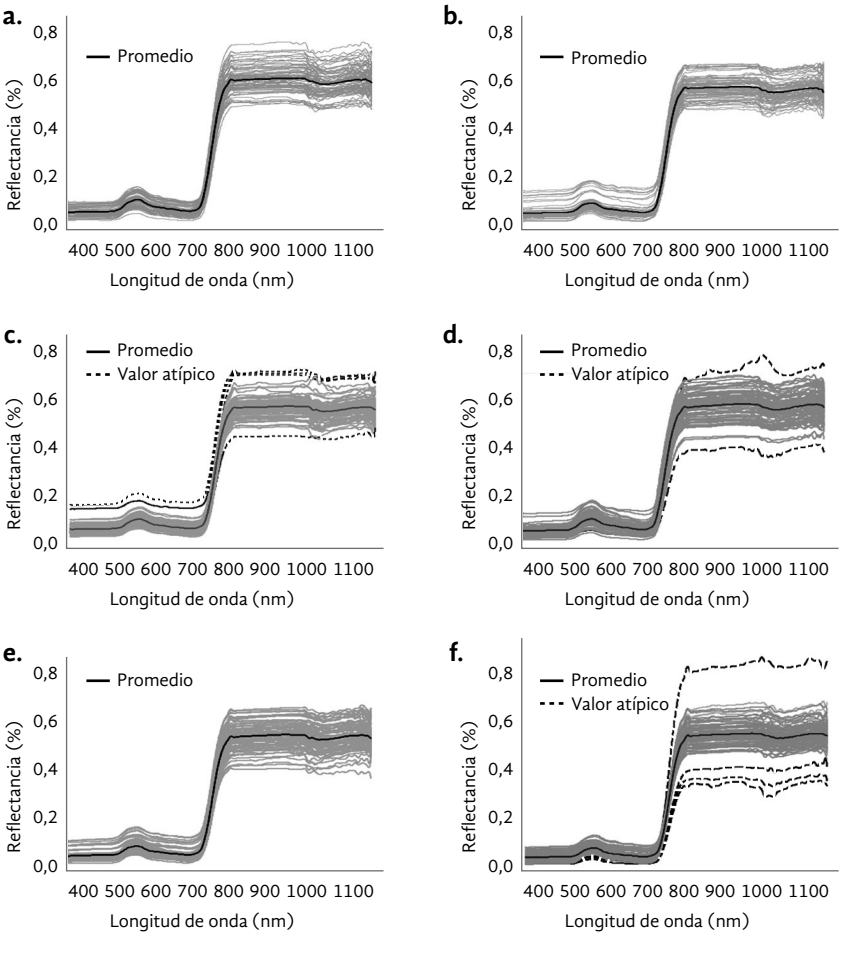

Figura 1. Representación gráfica de los valores de reflectancia medidos sobre las hojas 9 y 17 de las palmas consideradas en cada uno de los grupos definidos (ML, SDF, SFF) en fase de campo.

a. Curvas suavizadas ML Nv. 9. b. Curvas suavizadas SDF Nv. 9. c. Curvas suavizadas Campo SFF Nv. 9. d. Curvas suavizadas Campo ML Nv. 17. e. Curvas suavizadas SDF Nv. 17. f. Curvas suavizadas Campo SFF Nv. 17.

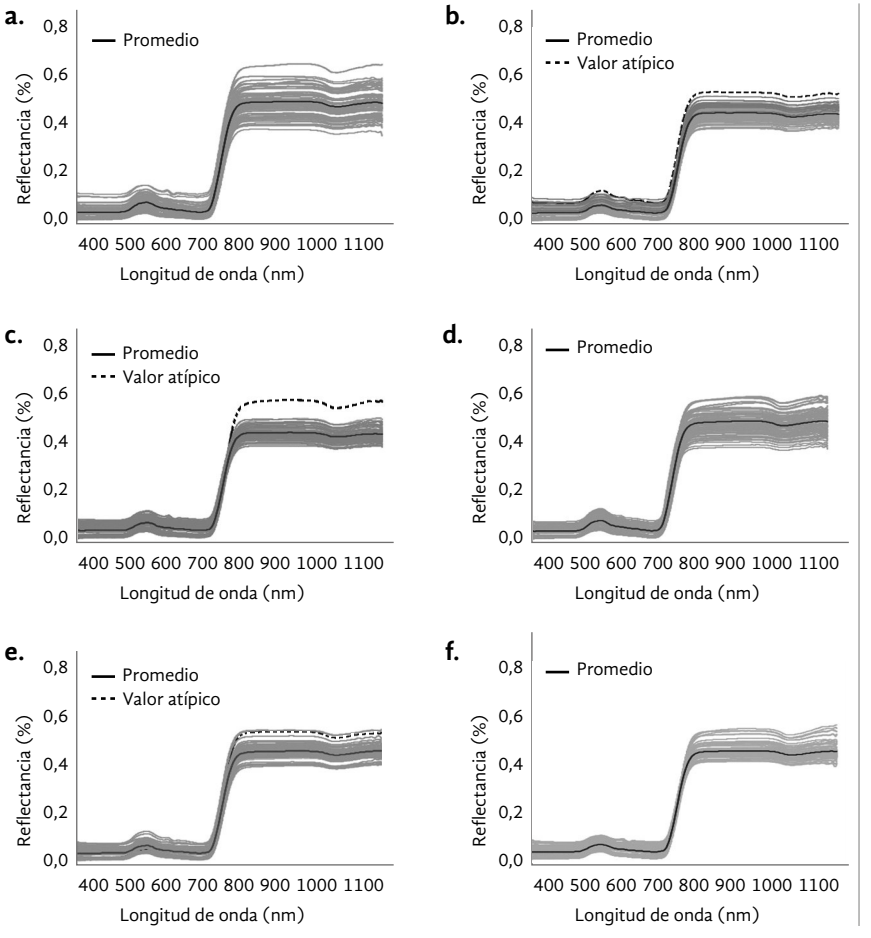

Figura 2. Representación gráfica de los valores de reflectancia medidos sobre las hojas 9 y 17 de las palmas consideradas en cada uno de los grupos definidos (ML, SDF, SFF) en fase de laboratorio.

a. Curvas suavizadas ML Nv. 9. b. Curvas suavizadas SDF Nv. 9. c. Curvas suavizadas Campo SFF Nv. 9. d. Curvas suavizadas Campo ML Nv. 17. e. Curvas suavizadas SDF Nv. 17. f. Curvas suavizadas Campo SFF Nv. 17.

Dado que las curvas atípicas pueden enmascarar las tendencias generales, estas fueron excluidas de los análisis posteriores.

Una comparación gráfica de las curvas medias (Figura 3) muestra que en ambas hojas evaluadas, tanto en campo como en laboratorio, la reflectancia media en palmas con ML es mayor que las de las palmas sanas. Esto es particularmente evidente en longitudes de onda superiores a 700 nm. A su vez, es posible apreciar que en campo las curvas espectrales medias son mayores que en laboratorio (más adelante se presentará una prueba formal de las diferencias entre los grupos).

Para establecer estadísticamente si existen o no diferencias significativas entre las medias de los grupos, usualmente se aplica una técnica de análisis estadístico llamada ANOVA, la cual se fundamenta en supues-

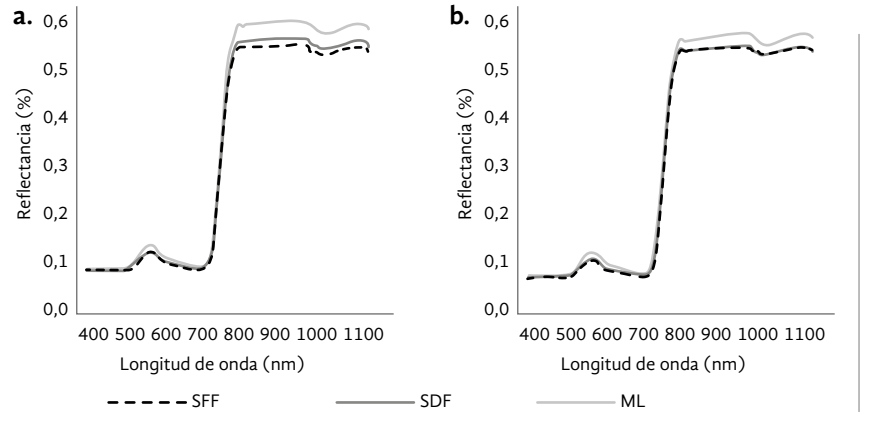

Figura 3. Gráfica de las curvas medias calculadas para los tres grupos (ML, SDF, SFF) en fase de campo y laboratorio.

tos de normalidad y homogeneidad de varianzas. En el caso considerado, esto supone que para cada longitud de onda los datos de reflectancia en el interior de los grupos se ajustan a un modelo de probabilidad normal y que las varianzas de los tres grupos, en cada longitud de onda, no son significativamente diferentes.

A continuación se presentan los valores P de las pruebas de hipótesis de normalidad (test de Lilliefors) y de homogeneidad de varianzas (test de Levene) efectuadas en cada longitud de onda con los datos de los tres grupos, en los dos niveles foliares (Figuras 4-7).

En todos los casos (datos de campo y de laboratorio en los dos niveles foliares), los valores P de las pruebas indican que estos supuestos requeridos para el uso de la ANOVA no se satisfacen. Se puede observar que en algunas longitudes de onda los valores P están por debajo del nivel de significancia de 5 % (línea punteada), lo que lleva al rechazo de las hipótesis correspondientes (normalidad u homogeneidad de varianzas).

Teniendo en cuenta lo anterior, se decidió emplear pruebas no paramétricas. En primera instancia se usaron pruebas de Kruskal-Wallis para determinar en qué longitudes de onda se dan diferencias entre los grupos, y posteriormente se emplearon pruebas de comparación múltiple de Nemeny para establecer entre cuáles de los grupos (ML, SDF y SFF) se daban específicamente dichas diferencias.

Los resultados de las pruebas de Kruskal-Wallis se presentan en las Figuras 8a y 8b. De acuerdo con estas pruebas, se puede establecer que para los datos tomados en campo correspondientes a la hoja 9 existen diferencias entre al menos un par de grupos a lo largo de todo el espectro evaluado (400 a 1.075 nm) (Figura 8a). Para el caso de la hoja 17, se observa que las diferencias se presentan en las longitudes de onda comprendidas entre 506 nm y 1.075 nm (Figura 8b).

Por otro lado, los datos de laboratorio muestran que para la hoja 9, en longitudes de onda superiores a 700 nm, se presentan diferencias entre los grupos (Figura 9a), y para el caso de la hoja 17, en las longitudes de onda comprendidas entre los rangos de 535 a 583 nm, y de 695 a 1.075 nm (Figura 9b).

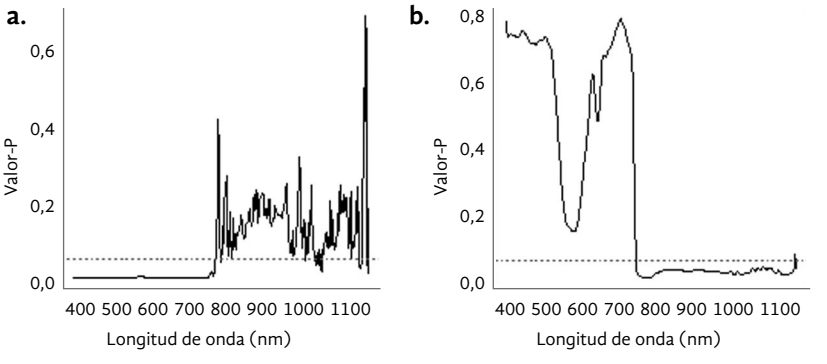

Figura 4. a. Valores P de las pruebas de hipótesis de normalidad en fase de campo para la hoja 9. La línea punteada corresponde al nivel de significancia de la prueba 5 %.

b. Valores P de las pruebas de homogeneidad de varianzas en fase de campo para la hoja 9.

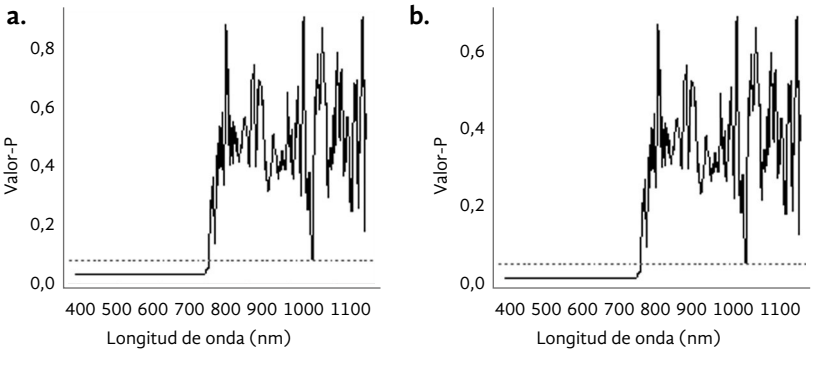

Figura 5. a. Valores P de las pruebas de hipótesis de normalidad en fase de campo para la hoja 17. La línea punteada corresponde al nivel de significancia de la prueba 5 %.

b. Valores P de las pruebas de homogeneidad de varianzas en fase de campo para la hoja 17.

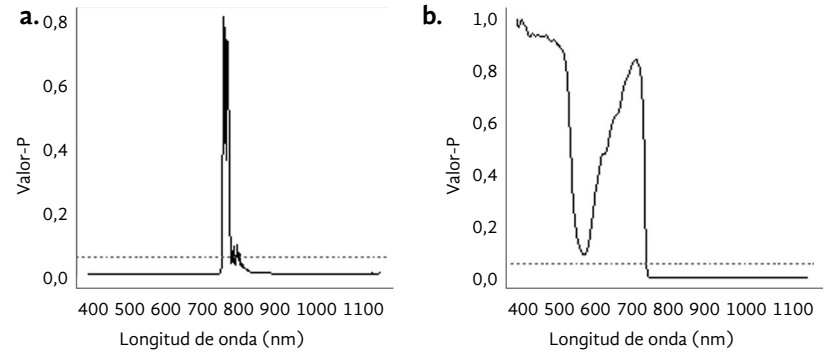

Figura 6. a. Valores P de las pruebas de hipótesis de normalidad en fase de laboratorio para la hoja 9. La línea punteada corresponde al nivel de significancia de la prueba 5 %.

b. Valores P de las pruebas de homogeneidad de varianzas en fase de laboratorio para la hoja 9.

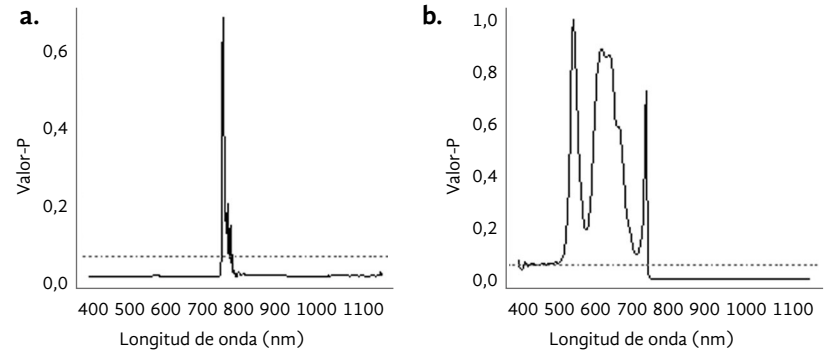

Figura 7. a. Valores P de las pruebas de hipótesis de normalidad en fase de laboratorio para la hoja 17. La línea punteada corresponde al nivel de significancia de la prueba 5 %.

b. Valores P de las pruebas de homogeneidad de varianzas en fase de laboratorio para la hoja 17.

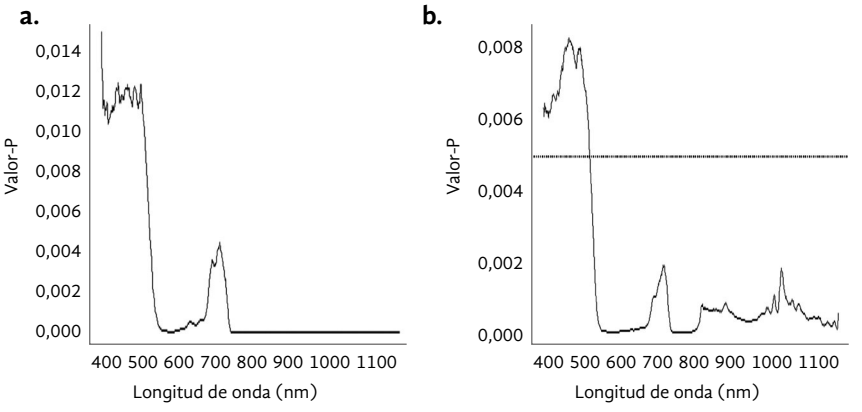

Figura 8. a. Valores P del test de Kruskal-Wallis para cada longitud de onda para datos de campo en la hoja 9.

b. Valores P del test de Kruskal-Wallis para cada longitud de onda para datos de campo en la hoja 17. La línea punteada corresponde al nivel de significancia de 5 %.

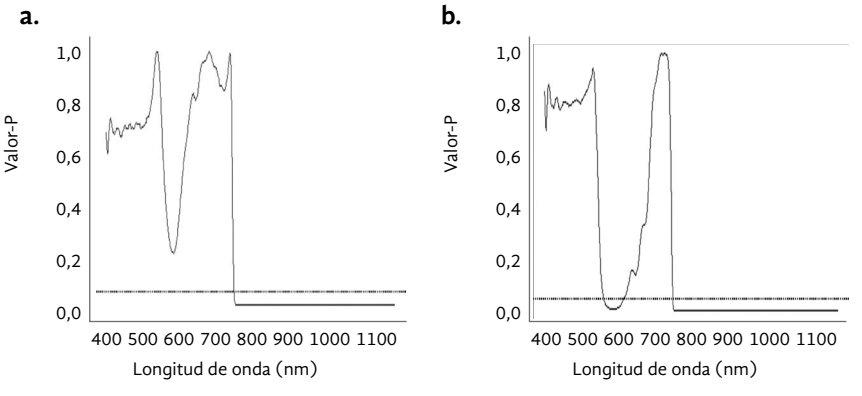

Figura 9. a. Valores P del test de Kruskal-Wallis para cada longitud de onda para datos de laboratorio en la hoja 9.

b. Valores P del test de Kruskal-Wallis para cada longitud de onda para datos de laboratorio en la hoja 17. La línea punteada corresponde al nivel de significancia de 5 %.

De manera general, se concluye que las diferencias entre los valores de reflectancia de los tres grupos objeto de estudio (ML, SDF y SFF) dependen, por un lado, de las condiciones bajo las cuales se hace la toma de los datos (campo o laboratorio), y por el otro, de qué tan avanzada se encuentra la expresión de síntomas en las hojas.

De manera complementaria, a continuación se presenta en la Tabla 1 el resumen de los resultados encontrados con las prueba de Nemenyi.

De acuerdo con los resultados descritos en la siguiente tabla, de manera general se puede interpretar que en las longitudes de onda en donde hay diferencias entre los grupos, la media de reflectancia de palmas afectadas por Marchitez letal es mayor que las medias de reflectancia obtenidas con las palmas que están dentro y fuera del foco de la enfermedad. Así mismo, salvo en los datos de nivel foliar 17 registrados en laboratorio, se observa que no hay, en ninguna longitud de onda, diferencias significativas entre los valores medios de reflectancia de palmas que están dentro y fuera del foco.

Por último, se llevaron a cabo pruebas de Wilcoxon de muestras pareadas (en cada palma hay dos mediciones, una en campo y otra en laboratorio) para determinar si hay diferencias significativas entre las medias de estas dos condiciones de medición. En todos los casos, la conclusión es que las medias obtenidas en campo son superiores a las medias registradas en laboratorio.

Conclusión y trabajo futuro

A partir de los análisis realizados puede concluirse que, tanto con la información de campo como con la de laboratorio, existen diferencias en los valores medios de reflectancia. Con los datos de campo estas diferencias se dan a lo largo de todas las longitudes de onda evaluadas, mientras que con los datos de laboratorio dichas diferencias se presentan en las longitudes de onda correspondientes a las regiones del verde, rojo e infrarrojo. En general, las firmas espectrales de palmas con ML tienen reflectancia media mayor que las SDF y que las SFF. Además, la reflectancia media obtenida con los

Tabla 1. Resumen de resultados encontrados en las pruebas de comparación múltiple de Nemenyi (nivel de significancia de 5 %).

Nivel de comparación	ML vs. SDF	ML vs. SFF	SDF vs. SFF
Datos de campo, nivel foliar 9	La reflectancia media de ML es mayor que la de SDF en todo el espectro.	La reflectancia media de ML es mayor que la de SDF en todo el espectro.	No hay diferencias significativas entre las medias de reflectancia de SDF y SFF en ninguna longitud de onda.
Datos de campo, nivel foliar 17	La reflectancia media de ML es mayor que la de SDF en longitudes de onda entre 506 a 1.075 nm.	La reflectancia media de ML es mayor que la de SFF en longitudes de onda entre 506 y 1.075 nm.	No hay diferencias significativas entre las medias de reflectancia de SDF y SFF en ninguna longitud de onda.
Datos de laboratorio, nivel foliar 9	La reflectancia media de ML es mayor que la de SDF en longitudes de onda superiores a 700 nm.	La reflectancia media de ML es mayor que la de SFF en longitudes de onda superiores a 700 nm.	No hay diferencias significativas entre las medias de reflectancia de SDF y SFF en ninguna longitud de onda.
Datos de laboratorio, nivel foliar 9	La reflectancia media de ML es mayor que la de SDF en longitudes de onda entre 535 a 583, y entre 693 y 1.075 nm.	La reflectancia media de ML es mayor que la de SFF en longitudes de onda entre 535 y 583 nm, y entre 693 y 1.075 nm.	La reflectancia media de SDF es mayor que la de SFF en longitudes de onda entre 727 y 1.075 nm.

datos de campo (en los tres grupos y los dos niveles foliares) es superior a la registrada en laboratorio.

Desde el punto de vista metodológico, se puede decir que los resultados obtenidos son bastante satisfactorios, dado que indican que las técnicas de análisis de datos funcionales brindan la posibilidad de tener una visión más general del patrón de comportamiento de las curvas de reflectancia. Aunque se aplicaron pruebas no paramétricas tradicionales para hacer las comparaciones en cada longitud de onda, la fase preliminar de análisis descriptivo y de identificación de observaciones atípicas desde la perspectiva funcional es fundamental para el posterior uso de las pruebas clásicas. Hay además un potencial de herramientas estadísticas comprendidas dentro del análisis de datos funcionales que aún no han sido aplicadas con este tipo de datos; por ejemplo, el análisis de regresión funcional para determinar la relación entre las curvas de reflectancia y covariables como temperatura o humedad, así como también el análisis discriminante funcional para identificar a partir de una curva de reflectancia si una palma está enferma, el uso de geoestadística funcional para hacer predicción espacial de la enfermedad a partir de las firmas espectrales.

Referencias bibliográficas

Dallal, G., and Wilkinson, L. (1986). An analytic approximation to the distribution of Lilliefors' test for normality. *The American Statistician*, 40, 294–296.

Hinkelmann, K., and Kempthorne, O. (1994). *Design and analysis of experiments. Volumen I: Introduction to Experimental Design*. John Wiley & Sons.

Kruskal, W., and Wallis, W. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47 (260): 583-621.

Lelong, C., Lanore, M., and Caliman, J. (2006). Evaluation of hyperspectral remote sensing relevance to estimate oil palm trees nutrition status. *Second Recent Advances in Quantitative Remote Sensing (RAQRS'II)*. A. Sobrino, José (ed.), Auditori de Torrent, Spain, 25-29 September 2006 (Valencia, Spain: Universitat de Valencia), pp. 147-152.

Levene, H. (1960). *Robust tests for equality of variances*. In Ingram O. and Hotelling, H. Stanford University Press. pp. 278-292.

Lothar, S. (1997). *Angewandte Statistik*. Berlin: Springer. pp. 395-397, 662-664.

López-Pintado, S., and Romo, J. (2009). On the Concept of Depth for Functional Data. *Journal of the American Statistical Association*, 104 (486), 718-734.

Ramsay, J., and Silverman, B. (2005). *Functional Data Analysis*. Springer.

Shafri, H., Anuar, M., Seman, I. and Noor, N. (2011). Spectral discrimination of healthy and *ganoderma*-infected oil palms from hyperspectral data. *International Journal of Remote Sensing*, 32(22), 7111-7129.

Vaiphasa, C., Skidmore, A., de Boer, W. and Vaiphasa, T. (2007). A hyperspectral band selector for plant species discrimination. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62, pp. 225-235.

Wang, L., and Sousa, W. (2009). Distinguishing mangrove species with laboratory measurements of hyperspectral leaf reflectance. *International Journal of Remote Sensing*, 30, pp.1267-1281.

Zhang, M., Qin, Q., Liu, X., and Ustin, S. (2003). Detection of stress in tomatoes induced by blight disease in California, USA, using hyperspectral remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 4, pp. 295-310.