Fenotipado masivo de nutrientes foliares y pigmentos fotosintéticos en dos especies de pinos coexistentes.

Máster Universitario en Restauración de Ecosistemas

Presentado por:

Sergio Núñez Castillo

Directores de TFM:

Ana Isabel García-Cervigón Morales

&

David Sánchez Pescador

Tutora académica

María Dolores Jiménez Escobar

En Madrid a 27 de enero de 2022

Contenido

Resumen 1

Palabras Clave 1

Abstract 2

Keywords 2

Introducción 3

Material y métodos 6

Descripción de la zona de estudio 6

Trabajo de campo 7

Trabajo de laboratorio 8

I. Caracterización espectral 8

II. Caracterización funcional de los individuos 9

III. Análisis de datos 10

Resultados 12

I. Espectros de absorción lumínica 12

II. Rasgos fenotípicos 13

III. Modelos predictivos 15

i. Pinus nigra 15

ii. Pinus sylvestris 17

Discusión 19

Conclusiones 22

Agradecimientos 23

Bibliografía 23

Anexos 27

Resumen

Los rasgos fenotípicos son claves a la hora de entender el funcionamiento, la estructura de los ecosistemas y las dinámicas existentes en sus poblaciones. Generalmente, estos rasgos se han estudiado y aplicado a nivel interespecífico, obviando la variación intraespecífica y usando sus valores promedios debido a la dificultad y costo de su medición. Por este motivo, se buscan métodos alternativos para realizar esta caracterización de forma sencilla, rápida y barata. Una alternativa consiste en utilizar modelos predictivos basados en la correlación del espectro de absorción lumínica del infrarrojo cercano y el visible (Vis-NIR) con los rasgos fenotípicos. En este trabajo se ha llevado a cabo un estudio en el Parque Natural del Alto Tajo, provincia de Guadalajara. Se midieron seis rasgos fenotípicos foliares (concentración de antocianinas, clorofila a, clorofila b, carbono, fósforo y nitrógeno orgánicos) y el espectro Vis-NIR en las hojas de 100 individuos de *Pinus nigra* y 102 de *Pinus sylvestris*. Con esta información se construyeron modelos predictivos usando tres algoritmos de machine learning (PLS, SVM y Random Forest), 12 pretratamientos y 7 rangos espectrales. Los resultados no fueron tan precisos como esperado, donde el mejor modelo apenas superó el 0,5 de precisión (R^2) para el fósforo orgánico en *Pinus nigra*. Los demás modelos rondaron el 0,3 – 0,1 de R^2. Sin embargo, se han podido plantear futuras líneas de trabajo para tratar de mejorar la precisión de los modelos, como ajustar mejor los rangos espectrales o aumentar el tamaño muestral. Estos modelos pueden suponer una revolución en la comunidad científica, permitiendo fenotipar masivamente grandes territorios obteniendo información aplicable a ámbitos como la ecología de comunidades o la restauración ecológica, proporcionando herramientas para conocer el estado fisiológico de los individuos de las comunidades y proporcionar claves que permitan recuperar los servicios ecosistémicos bloqueados en ambientes degradados.

Palabras Clave

Espectroscopía de infrarrojo cercano y visible, modelos predictivos, *Pinus nigra*, *Pinus sylvestris*, rasgos fenotípicos.

Abstract

Phenotypic traits play a key role in the understanding of ecosystem properties and community dynamics. However, these traits have been studied at the inter-specific level, considering average trait values per species and ignoring the intra-specific variation, due to the great amount of effort and cost needed to measure them. As a result, alternative approaches are being considered in order to achieve a fast, simple and affordable way of phenotypic characterization. One of these approaches is the use of predictive models based on the correlation between the visible and near infra-red spectrum (Vis-NIR) of biological samples and their phenotypic traits measured in the lab. In this study, we worked in the Alto Tajo Natural Park located in Guadalajara, Spain. Two species were considered: *Pinus nigra* and *Pinus sylvestris*. We measured six foliar phenotypic traits (anthocyanins, chlorophyll a, chlorophyll b, carbon, phosphorus and organic nitrogen concentrations) and collected the Vis-NIR spectra in 100 individuals of the former and 102 of the latter species. These data were used to build predictive models considering three different machine learning algorithms (PLS, SVM and Random Forest), 12 pre-treatments and 7 spectral ranges. The results obtained were not as precise as expected, since the best model –i.e., the one for the organic phosphorus in *Pinus nigra* using Random Forest–, roughly surpassed 0,5 of R^2. The other models reached 0,3 or 0,1 of R^2. Despite these results, the study allowed to establish future work areas related to the improvement of model precision. These models might represent a paradigm shift in the scientific community, as large communities could be phenotyped in shorter times providing valuable information that could be used in different areas such as community ecology or ecological restoration, providing tools to study in a easy and quick way the physiological status of the different individuals of communities and ecosystems.

Keywords

Phenotypic traits, *Pinus nigra*, *Pinus sylvestris*, predictive models, visible and near-infrared spectroscopy.

Introducción

Los rasgos funcionales son características morfo-fisiológicas, bioquímicas, estructurales fenológicas o de comportamiento que afectan directa o indirectamente al desempeño o "fitness" de los organismos (Violle et al., 2007) mediante su participación en el crecimiento, reproducción y supervivencia. Estos rasgos permiten caracterizar las plantas, y se pueden evaluar a dos niveles: intraespecífico, cuando se miden los rasgos de los distintos individuos de una misma especie, e interespecífico, cuando se consideran los rasgos funcionales entre las distintas especies que forman una comunidad. Estos rasgos funcionales son transversales ya que pueden ser utilizados en las diferentes ramas de la ecología (Herrera, 2017). De hecho, numerosos estudios han puesto de manifiesto la existencia de conexiones entre las diferentes subdisciplinas en la ecología; las características físico-químicas de los individuos que afectan su "fitness" afectan, a su vez, a las propiedades de los ecosistemas (Shipley et al., 2016). Debido a esta evidencia científica, se ha propuesto en numerosas ocasiones el establecimiento de otra subdisciplina que tenga en consideración todos aquellos trabajos que defienden que los individuos, a través de sus rasgos funcionales, tienen control sobre el ecosistema. Esta subdisciplina se conoce como "trait-based ecology" o ecología basada en rasgos funcionales o fenotípicos (Violle et al., 2007).

De acuerdo con Shipley et al. (2016), existen cuatro condiciones para esta subdisciplina que le otorgan una identidad y una razón de existencia. La primera es la importancia de los rasgos funcionales frente a la identidad específica de las plantas. A continuación, la comparación de valores de los rasgos entre varias especies para encontrar tendencias. El tercero es la comparación de estas tendencias con gradientes ambientales para observar cómo se afectan mutuamente. En último lugar, la existencia de un conjunto de rasgos pertenecientes a módulos de plantas o comunidades enteras que determinan la estructura y funcionalidad ecosistémica. En este sentido resulta curioso cómo se tuvieron que proponer esta serie de condiciones para consolidar esta rama de la ecología cuándo ya Darwin estableció esta idea en su obra *El origen de las especies*, donde ya hablaba de que una mayor variación intraespecífica se traducía en una mayor capacitación de la especie para sobrevivir en un mayor rango de condiciones ambientales. En esa obra ya se establecía la base para una ecología basada en rasgos, reconocía la importancia de ellos en la supervivencia de las especies y también su relación con la competencia y selección natural de estas especies (Darwin & Kebler, 1859; Sides et al., 2014).

Durante las últimas dos décadas diversos trabajos han reconocido el gran papel que ejercen los rasgos funcionales individuales a nivel de comunidad (Albert et al., 2011; Escudero et al., 2021; Violle et al., 2012). En este sentido, se quiere dejar atrás una visión que intentaba explicar la ecología de comunidades siguiendo un patrón general en el cual los valores usados de rasgos funcionales correspondían con las medias de estos a nivel de la especie, ignorando la variación

intraespecífica existente. Una de las razones que motivaban el uso de valores promedio en lugar de valores individuales es la gran dificultad a la que se enfrentan los investigadores a la hora de analizar los rasgos fenotípicos de los individuos presentes en la comunidad, ya que es un proceso largo y tedioso al que muchos científicos no se pueden enfrentar por falta de tiempo y/o de recursos y toman la solución más rápida, la toma de valores medios (Auger & Shipley, 2013; Escudero & Valladares, 2016). La consideración de la variación intraespecífica de los individuos permitiría además entender mejor la teoría de la coexistencia y al ensamblaje de comunidades, pudiendo así resolver la paradoja del principio de Gause la cual expone que dos especies no pueden compartir el mismo nicho (Violle et al., 2012). Existen diversos estudios que han tenido en cuenta esto y exponen interesantes teorías como la explicación de la coexistencia y que las diferencias de competitividad existentes entre individuos dejan de lado las posibles diferencias competitivas entre las especies y, junto a la variedad de nicho, disminuyen drásticamente la posibilidad de exclusión competitiva o, por lo menos, la ralentizan (Fridley et al., 2007; Hart et al., 2016; Hubbell, 2005).

Estos estudios, si bien nos proporcionan gran información referente a la teoría de coexistencia y de ensamblaje de comunidades, no explican la existencia de comunidades pobres en especies o monoespecíficas ya que hasta la fecha la mayoría han trabajado con valores promedio a nivel de especie, o midiendo sólo alguno de los individuos de cada especie, interpretando erróneamente que todos los individuos de una especie son iguales. De acuerdo con la teoría ecológica, especies y, por extensión, individuos iguales deberían tender a competir por los mismos recursos, ya que ocupan el mismo nicho. De manera que tiene que existir una variabilidad intraespecífica que permita una separación de nichos dentro de la misma especie (Auger & Shipley, 2013; Escudero & Valladares, 2016; Herrera, 2017). Es ahí donde radica la importancia de estudiar la variabilidad intraespecífica y los fenotipos individuales de cada especie dentro de una comunidad. Los escasos estudios sobre los efectos de estas variaciones intraespecíficas se centran, principalmente, en las consecuencias genéticas más que en la diversidad de rasgos existentes (Hart et al., 2016). Debido a esto, se hace cada vez más urgente el estudio de la variación de los fenotipos de los individuos coexistentes para observar cómo varían estos fenotipos en comunidades mono- o pluriespecíficas. Así, podremos confirmar los resultados de aquellos estudios que sí han trabajado con los fenotipos y han estudiado su variación, que afirman que la variación entre fenotipos es igual o mayor que la variación de rasgos total presente en la comunidad (Granda et al., 2012; Kraft et al., 2015).

A lo que se pretende contribuir con este estudio es a pasar de una ecología de comunidades basada en la medición de rasgos funcionales promedio o de unos pocos individuos a una ecología de comunidades basada en fenotipos, en la que se caracteriza funcionalmente cada individuo de la comunidad. Para ello, en este trabajo se propone el desarrollo de un modelo predictivo de rasgos funcionales a partir de un fenotipado masivo de individuos de una comunidad arbórea de modo

que se puedan predecir los valores individuales de distintos rasgos tediosos de medir a partir de una variable fácilmente medible.

Para desarrollar dicho modelo, el estudio se basó en un fenotipado masivo de poblaciones de *Pinus nigra* J.F Arnold *ssp. salzmannii* (Dunal) y *P. sylvestris* L. que cohabitan en el Parque Natural del Alto Tajo (Guadalajara), midiendo caracteres funcionales en el laboratorio y calibrando un modelo basado en la espectrometría del rango visible-infrarrojo cercano, en adelante denominado Vis-NIR. Se ha elegido esta técnica debido al bajo coste que supone y la facilidad de uso para calibrar diversos rasgos funcionales medidos en distintos órganos de la planta, ya que captura tanto características físicas como químicas. En este sentido las técnicas de espectrometría han demostrado en diversos estudios su aptitud para desarrollar modelos usados para crear bases de datos de rasgos funcionales a escalas antes imposibles (Costa et al., 2018). Estas técnicas nos permiten obtener el espectro de absorción lumínica de distintos materiales y a distintas longitudes de onda que pueden abarcar tanto el rango visible como el infrarrojo cercano, que después se pueden usar para relacionar con los rasgos funcionales medidos mediante modelos predictivos. Los resultados obtenidos suelen ser altamente satisfactorios y se ha demostrado su funcionamiento para predecir rasgos funcionales tanto en hojas frescas como prensadas e incluso en muestras de suelo, siendo los más precisos los obtenidos de hojas prensadas (Kothari, 2021) pudiéndose hasta predecir estos rasgos en las estaciones de crecimiento e identificar la especie a la que pertenece cada individuo, lo que puede ahorrar mucho más trabajo aún (Chen et al., 2021; Costa et al., 2018).

El éxito de este tipo de modelos predictivos puede suponer una gran revolución en el mundo de la ecología ya que permitiría realizar un fenotipado completo de comunidades muy grandes en una fracción de tiempo de lo que se tardaba antes. En todos los proyectos de restauración ecosistémica o evaluaciones ambientales primero se debe realizar un análisis de la zona de estudio, lo cual en muchas ocasiones no se realiza de forma acorde al protocolo establecido debido a la falta de tiempo y de presupuesto. Por lo tanto, el desarrollo de un modelo predictivo puede suponer una reducción de estos factores limitantes consiguiendo así datos fiables y unos proyectos más acertados con resultados muy satisfactorios. En definitiva, ser capaces de realizar un fenotipado masivo de una comunidad vegetal de una forma tan efectiva supondría un gran avance para las investigaciones científicas actuales y futuras ya que se reducirían de forma considerable los tiempos de recogida y procesado de muestras y el presupuesto destinado a ello. Por consiguiente, esto desembocaría en un incremento notable de la información disponible de los diversos ecosistemas del planeta, facilitando así la realización de proyectos de temática ambiental como son las restauraciones ecológicas que se dan actualmente de forma muy proactiva.

El objetivo general de este trabajo es construir un modelo para predecir rasgos fisiológicos relacionados con el contenido en pigmentos (antocianinas y clorofilas) y nutrientes (fósforo, nitrógeno y carbono orgánicos) en las hojas a partir de su espectro de absorción lumínica. Estos rasgos proporcionan información valiosa sobre las estrategias de adquisición de recursos, traduciéndose directamente en información sobre la capacidad fotosintética de los individuos, el potencial de crecimiento tanto a nivel aéreo como radical, el potencial reproductivo o la competitividad frente a otros individuos cercanos, entre otros (Costa et al., 2018; Escudero & Valladares, 2016; Kothari, 2021). Para lograr este objetivo se trabajará con dos especies dominantes de una misma comunidad arbórea y se testarán diversos algoritmos predictivos, rangos espectrales y pretratamientos de los espectros. Los objetivos específicos son: (1) evaluar cuál es la combinación de algoritmo, rango espectral y pretratamiento que permite crear unos modelos con mayor poder predictivo para cada rasgo, (2) valorar la existencia de diferencias entre especies en cuanto al ajuste de los distintos modelos y (3) valorar la existencia de diferencias en el poder predictivo de los distintos modelos a la hora de predecir los diferentes rasgos funcionales estudiados.

Material y métodos

Descripción de la zona de estudio

El estudio se llevó a cabo en un pinar dentro del Parque Natural del Alto Tajo. Este pinar se encuentra en las coordenadas 40°43'28.6"N 2°07'00.4"W, y ocupa una superficie de 177,433 hectáreas, localizándose entre las zonas sureste de la provincia de Guadalajara y noreste de la provincia de Cuenca, correspondiendo con el piso de vegetación supramediterráneo, ya que se encuentra a 1330 metros de altitud (Ferrero et al., 2006). Esta altitud y localización determinan en gran parte el clima existente en el lugar, correspondiente al típico presente a lo largo de toda la meseta central, un clima mediterráneo con influencia continental que presenta una amplitud térmica muy grande entre las distintas estaciones (inviernos largos y fríos y veranos cortos y suaves), determinando así las precipitaciones, las cuales se concentran en torno a los meses que constituyen la primavera y el otoño. La principal característica de la parcela de trabajo son las bajas temperaturas que se registran en los meses de invierno, coincidiendo con las temperaturas medias anuales más bajas de la Península Ibérica, siendo estas de 10,2 ° C mientras que las temperaturas máximas y mínimas en enero y agosto son de 8,2º y -3,5ºC, y 28,5º y 10,30ºC respectivamente en Molina de Aragón, localidad que se encuentra a 24 km de distancia del pinar (Granda et al., 2012). Respecto a las precipitaciones, rondan los 500 mm anuales, distribuyéndose entre los meses de primavera y los de otoño. Por otro lado los meses de verano apenas se superan los 30 mm de precipitación mensual, correspondiendo principalmente a tormentas veraniegas

(Bastias et al., 2013). Esta falta de precipitaciones se debe principalmente a la reducción que ocurre en el transporte de las precipitaciones desde el Oeste hacia el Este por la degradación paulatina de los flujos atlánticos (Ferrero et al., 2006).

Dentro del área de estudio se estableció una parcela de muestreo de 125 x 125 m que a su vez fue dividida en 25 subparcelas de 25 x 25 m. La parcela de muestreo está dominada principalmente por dos especies arbóreas, *Pinus nigra* J.F Arnold ssp. *salzmannii* (Dunal) Franco y *Pinus sylvestris* L. En menor proporción aparecen, en orden descendente: *Buxus sempervirens* L., *Juniperus thurifera* L. y *Quercus faginea* Lam. (datos propios). En el caso concreto de este trabajo nos centraremos exclusivamente en la población de *Pinus nigra* y *P. sylvestris*.

Trabajo de campo

Tras la elección de la zona de estudio se procedió a geolocalizar cada individuo de *P. nigra* y *P. sylvestris* presentes en la parcela mediante el uso del GPS Leica Viva GS 15, y a numerarlos mediante un marcado con chapa de metal, para poder reflejarlo en un mapa para facilitar la posterior localización de los individuos (Fig. 1). El estudio se centró en individuos adultos vivos, por lo que solo se incluyeron en el mapa todos aquellos árboles que superaban los 2,5 cm de DBH ("diameter at breast height" o diámetro a la altura del pecho), y se marcaron con chapa aquellos que superaban los 7,5 cm de DBH. Esto resultó en un total de 525 individuos adultos de *P. nigra* y 520 individuos adultos de *P. sylvestris* marcados, dando un total de 1045 adultos.

Puesto que los rasgos funcionales a analizar eran referentes a las acículas de los individuos, la toma de muestras se basó en la recogida de hojas. Debido a la gran altura que pueden alcanzar los individuos de estas de esta especies en nuestra zona, hasta 19,7 m para los *P. nigra* y hasta 24,0 m los *P. sylvestris*, la recogida de hojas se llevó a cabo mediante una tijera de podar unida a una pértiga de 12 metros de altura que facilitó el acceso a las ramas de los individuos más elevados, aunque algunos siguieron estando fuera del alcance y se descartaron por lo que, finalmente solo se pudieron coger muestras de 507 *P. nigra* (18 individuos menos) y de 456 *P. sylvestris* (64 individuos menos). De cada uno de estos individuos se separaron 3 subconjuntos de 10 acículas, marcadas como A, B y C, que se conservaron en sobres de papel y en nevera a 4ºC, lo que suman un total de 2889 submuestras [(507 ind. *P. nigra* + 456 ind. *P. sylvestris*) × 3 subconjuntos/ind.].

Con el objeto de caracterizar funcionalmente la población, se seleccionaron al azar 100 individuos de *P. nigra* y 102 de *P. sylvestris* que fueron utilizados para la medición de los rasgos funcionales de interés en el laboratorio para, a continuación, calibrar un modelo predictivo. De cada uno de estos 202 individuos, se recogió una muestra de hojas extra para la caracterización de los pigmentos fotosintéticos, la cual se mantuvo a – 80ºC y en oscuridad para evitar una posible

degradación de los pigmentos. Por lo que, finalmente se extrajeron un total de 3091 submuestras de los individuos de la zona de estudio.

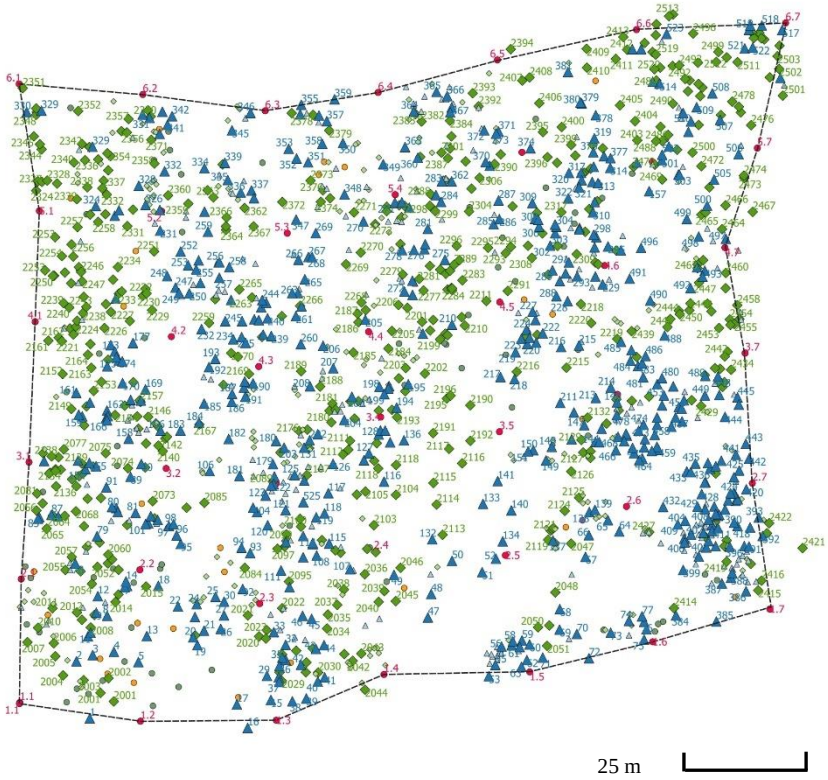

Figura 1. Mapa del área de estudio. Los puntos rojos marcan los límites de las distintas subparcelas. Los distintos símbolos corresponden con individuos de las diferentes especies: *P. nigra* adultos (triángulos azules grandes), *P. nigra* por debajo de 7,5 cm de DBH (triángulos azules pequeños), *P. sylvestris* adultos (rectángulos verdes grandes), *P. sylvestris* por debajo de 7,5 cm de DBH (rectángulos verdes pequeños), *Quercus faginea* (círculos amarillos) y *Juniperus thurifera* (círculos verdes).

Trabajo de laboratorio

I. Caracterización espectral

Diversos estudios han demostrado que los mejores resultados se obtuvieron con hojas previamente secadas (Costa et al., 2018; Kothari, 2021) por lo que las tres muestras A, B y C, se guardaron durante 48 horas en una estufa a 60º para conseguir una deshidratación completa. A continuación, se procedió a la obtención del dato espectral mediante utilización de un Analizador de laboratorio ASD LabSpec 4 Standard - Res i (Malvern Panalytical) (Fig. 2). Para ello se colocaron las 10 acículas de cada sobre en una pinza con una pequeña separación entre ellas y se realizaron dos mediciones con el espectrofotómetro, una por el haz y otra por el envés. De esta forma, se tomaron espectros de las 3 submuestras de cada individuo (6 mediciones por individuo), lo que equivale a 6182 espectros de absorción medidos en total.

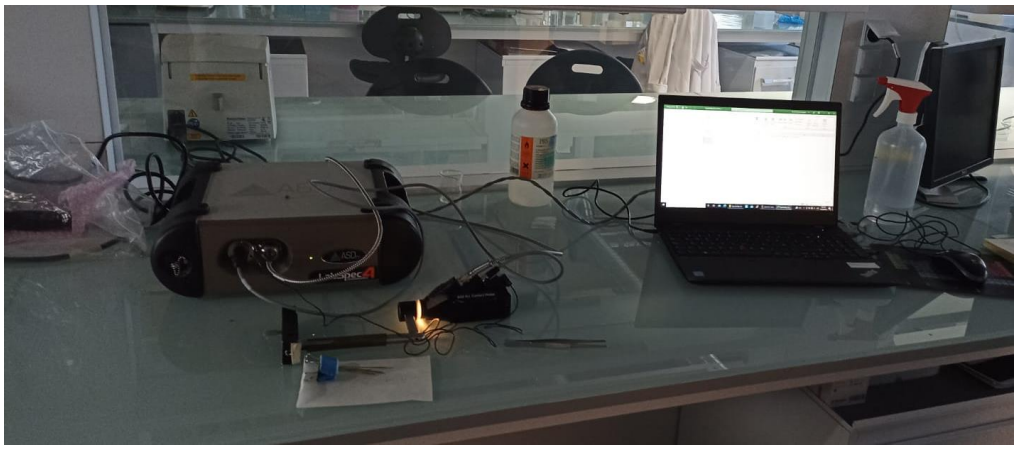

Figura 2. Espectrofotómetro utilizado para tomar los espectros de absorción lumínica de las muestras recogidas. Se trata de un modelo de Analizador de laboratorio ASD LabSpec 4 Standard - Res i de Malvern Panalytical.

II. Caracterización funcional de los individuos

Los rasgos funcionales seleccionados fueron elegidos en base a la facilidad de obtenerlos, que tengan la suficiente información bibliográfica como para comparar los resultados con otros estudios previos y debido a que son los que forman parte de los procedimientos analíticos habituales de NUTRILAB, el laboratorio de análisis de compuestos químicos y nutrientes de la URJC, donde se realizaron las mediciones (https://nutrilab-urjc.es/). Estos rasgos fueron: concentración de carbono, nitrógeno y fósforo orgánico y concentración de pigmentos fotosintéticos (antocianinas y clorofilas a y b).

- Concentración de carbono orgánico. Se obtuvo mediante el protocolo descrito en Anderson & Ingram (1990) mediante la digestión con ayuda de ácido sulfúrico, centrifugación y medición de su abundancia en espectrofotómetro a 600 nm.

- Concentración de nitrógeno y fosforo. Se procedió mediante el protocolo de Kjeldahl sobre las muestras de los sobres A. Se realizó con la ayuda de una digestión de ácido sulfúrico y adición de un catalizador en un tubo Kjeldahl y la posterior medición de su abundancia en un analizador de nutrientes (espectrofotómetro UV-VIS).

- Concentración de pigmentos (antocianinas, clorofilas a y b). Se procedió utilizando el método descrito por García-Plazaola & Becerril (2001) sobre las acículas del sobre adicional que fue previamente guardado en el congelador de -80ºC. Para evitar la pérdida de características de esos pigmentos, se trabajó en hielo y con acetona para conseguir extraer todo el material. Después se procedió a centrifugar las muestras para más tarde filtrarlas en un vial HPLC. Por último, se utilizó un UHPLC (Cromatógrafo de Líquidos

de Ultra Alta Resolución) de Shimadzu con dos fases para obtener las concentraciones buscadas de los pigmentos.

III. Análisis de datos

Una vez obtenidos todos los datos espectrales y fenotípicos, se procedió a la construcción de los modelos para relacionar los espectros de absorción medidos en las hojas de cada individuo con sus correspondientes rasgos fenotípicos.

Para poder preparar el modelo, primero se realizó un promediado de los espectros de absorción medidos para cada muestra. Tras esto se aplicaron hasta 12 pretratamientos a los espectros promediados (Tabla 1.). Estos pretratamientos modifican los espectros de acuerdo a diferentes transformaciones permitiendo resaltar diferentes aspectos (picos, áreas) y/o eliminar ruido de los mismos. Finalmente, para intentar maximizar la precisión de los modelos y discriminar áreas espectrales más informativas se trabajó con 7 rangos espectrales diferentes (Tabla 1).

Cada combinación de rango espectral y pretratamiento fue usada como variable independiente para construir los modelos predictivos de cada uno de los rangos funcionales considerados. En concreto se usaron tres algoritmos de "machine learning": PLS (Regresión por mínimos cuadrados), SVM (Máquinas de vector de soporte) y RF (Random forest). El algoritmo PLS es el método más usado en quimiometría ya que es especialmente adecuado cuando la matriz de predictores tiene más variables que observaciones, como es el caso, lo que permite flexibilizar tanto las hipótesis de partida como el tamaño muestral a emplear. Este algoritmo es ampliamente usado en diversas disciplinas y campos para relacionar la variabilidad en los datos fisiológicos medidos con los espectros infrarrojos (Caballero, 2006). Es un método relacionado con el análisis de componentes principales (PCA) en el que todas las variables predictivas (longitudes de ondas del espectro) son incluidas en el modelo reduciéndolas a un número de variables latentes o número de componentes (linealmente independientes) como combinaciones lineales de las variables predictivas (Alciaturi et al., 2003; Geladi & Kowalski, 1986). En nuestro caso, se establecieron un máximo de 15 componentes principales, tras lo cual se seleccionó el número óptimo de componentes principales según la menor tasa de error que se observó en cada caso.

El SVM es un algoritmo no-linear que permite una clasificación discriminativa que separa distintos grupos observacionales mediante un hiperplano para diferenciar las diferentes clases dentro de los datos. En el caso de su uso para ajuste a muestras cuantitativas (como es el caso en este trabajo), se usa un tipo de regresión épsilon (eps-svm) que permite predecir valores concretos de rasgos cuantitativos. Además, en nuestro caso se empleó un kernel lineal (SVM-L) para entrenar y validar el modelo que proyecta y transforma la información a un espacio de

características de mayor dimensión el cual aumenta la capacidad computacional. Estos métodos no lineales no han demostrado tener más efectividad que los algoritmos lineales como el PLS (Costa et al., 2018).

Por último, el Random Forest (RF) es un algoritmo propuesto por Breiman en 2001. Es un algoritmo de aprendizaje supervisado que divide los datos en grupos pequeños y ajusta árboles de decisión a cada uno de los grupos y los agrega, basando la estimación final en un promedio (regresión) y una mayoría (clasificación). Se ha convertido en un algoritmo muy popular debido a la gran capacidad de que ha demostrado en problemas variados de clasificación y regresión y, el hecho de que solo sea necesario ajustar unos pocos parámetros ha contribuido a su popularidad. Además, ha demostrado ser capaz de obtener buenos resultados con muestras pequeñas y espacios dimensionales altos. Encontramos dos componentes que tienen un papel crítico en RF: el "bagging" el cual es un procedimiento muy efectivo especialmente cuando la dimensionalidad de los datos es alta, y el "split criterion", método empleado para construir cada uno de los árboles individualmente. Uno de los mayores beneficios de RF es su capacidad para calcular información útil sobre los errores: estimación del error y de la importancia de las variables, medidas de proximidad y valoración de casos anómalos, información útil para evaluar la bondad del modelo y hacer cambios en los datos de entrenamiento si fuese necesario. La dificultad para analizar RF radica en su carácter de caja negra, por lo que es difícil justificar su desempeño, además de que se pierde la facilidad de interpretación de los árboles de decisión y que no se pueden extrapolar los valores (regresión) (Jin et al., 2020; Wright & Ziegler, 2017). Para el caso del RF los componentes principales se seleccionaron de acuerdo al menor error de predicción, para la regresión el error cuadrado medio, y la mejor combinación de los siguientes parámetros: número de árboles, el número de variables en que se puede dividir cada nodo (i.e. mtry) y el tamaño mínimo de nodo (i.e. min.node.size). Las combinaciones que se testaron fueron: 100, 500, 1000 y 1500 para el número de árboles; raíz cuadrada (redondeada a la baja) del número de variables, raíz cuadrada dividida entre 2 y raíz cuadrada multiplicada por 2 para el número de variables en que se puede dividir cada nodo; y 1, 5, 10 y 20 para el tamaño mínimo de nodo.

En cada caso se empleó el 70% de los datos obtenidos para calibrar el modelo y el 30% restante para validarlo de manera independiente de acuerdo con el ajuste por mínimos cuadrados (R^2) entre los valores observados y los valores predichos por el modelo. Hasta un total de 20 simulaciones distintas fueron empleadas con el objetivo de incluir en los modelos la mayoría de las combinaciones de muestras posibles. Esto supuso un total de 252 modelos estadísticos distintos para cada rasgo funcional (12 pretratamientos × 7 rangos espectrales × 3 algoritmos), un total de 3024 para todos los rasgos y especies, de los cuales se seleccionan los que mayor R^2 tuvieron. El mejor modelo para cada rasgo vino representado por aquel que tuvo un mayor R^2 en la validación independiente.

Los datos se analizaron en el entorno de R (R Core Team, 2021). Se utilizó el paquete asdreader (Roudier, 2017) para leer los espectros obtenidos con el espectrofotómetro y en formato asd. Para los pretratamientos se utilizaron las funciones *movav* (transformación Moving averages) del paquete prospectr (Stevens & Ramirez-Lopez, 2021), *savitzkyGolay* del paquete prospectr (Stevens & Ramirez-Lopez, 2021), *diff* del paquete base (R Core Team, 2021), *msc* del paquete pls (Mevik et al., 2011), la función *scale* del paquete base (R Core Team, 2021), *prep.snv* del paquete mdatools (Kucheryavskiy, 2020), *detrend* del paquete prospectr (Stevens & Ramirez-Lopez, 2021), *train* del paquete caret (Kuhn et al., 2021) para ajustar los PLS, *ksvm* perteneciente al paquete kernlab (Karatzoglou et al., 2004) para los SVM y *ranger* del paquete del mismo nombre (Wright et al., 2021) para los Random Forest.

Tabla 1. Algoritmos de machine learning utilizados para construir los modelos, pretratamientos aplicados sobre los espectros de absorción medidos y rangos espectrales seleccionados para ajustar los modelos.

Algoritmos machine learning	**Pretratamientos**	**Rangos espectrales**
PLS (Regresión por mínimos cuadrados)	Datos brutos (Raw)	400-2450 nm (eliminación de colas)
SVM (Máquinas de vector de soporte)	Primera derivada (D1)	400-700 nm (visible)
RF (Random forest)	Segunda derivada (D2)	400-1000 nm (detector VNIR)
	Detrend (Dt)	1001-1800 nm (detector SWIR 1)
	Primera derivada (D1)	1801-2450 nm (detector SWIR 2)
	Segunda derivada (D2)	701-2450 nm (infrarrojo)
	Mean centering scaling (MCS)	1001-2450 (detectores SWIR)
	Multiple scatter correction (MSC)	
	Moving averages (MVA)	
	Moving averages + primera derivada (MVA+D1)	
	Savitzky Golay (SG)	

Resultados

I. Espectros de absorción lumínica

Los espectros de absorción sin haber sido sometidos a ningún pretratamiento (Fig. 3a), apenas mostraron un perfil marcado y específico para cada especie. Los únicos picos de absorción que se pudieron destacar son aquellos cercanos a los 400 y 600 nm de longitud de onda. Los espectros de absorción sometidos al pretratamiento detrend (Fig. 3b) mostraron un perfil mucho más marcado en comparación con el previo ya que los picos de absorbancia eran mucho más claros. Estos picos correspondían con las longitudes de onda; 400, 600, 1450, 1700 y 1950 los cuales fueron aquellos más destacables. Este pretratamiento se muestra como ejemplo para visualizar los

efectos de las transformaciones aplicadas a los datos y pone de manifiesto la existencia de diferencias entre los espectros de ambas especies.

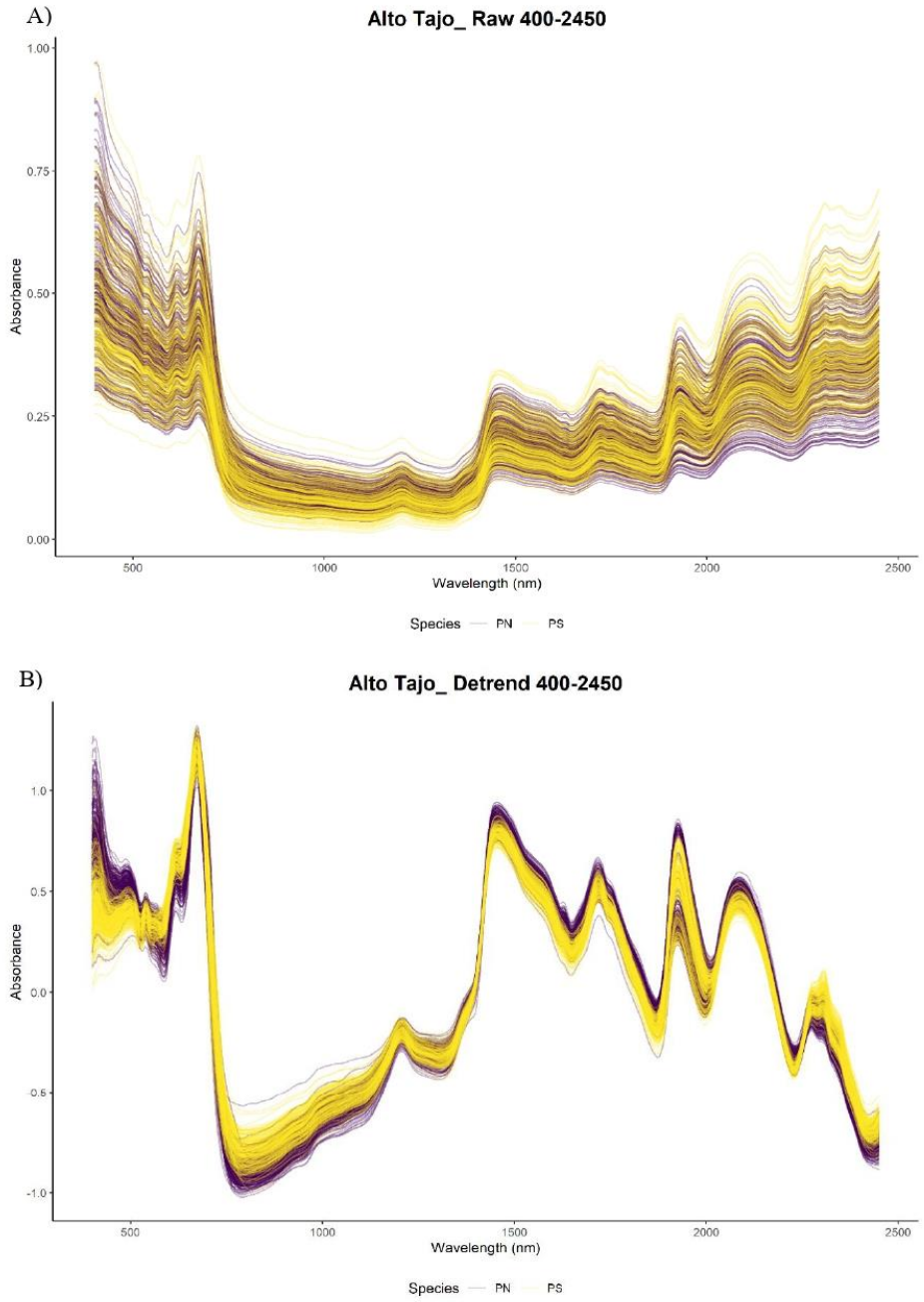

Figura 3: Espectros de absorción en el rango Vis – NIR (400 – 2500 nm) obtenidos mediante espectofotómetro para *P. nigra* y *P. sylvestris*. a) Espectros brutos y b) espectros tras ser sometidos al pretratamiento "detrend".

II. Rasgos fenotípicos

Los rasgos fenotípicos mostraron un comportamiento bastante homogéneo entre especies. El valor de los cuartiles (Fig. 4) no varió en gran medida aunque, por norma general, el valor del segundo cuartil, correspondiente con la media, fue mayor para *P. sylvestris*. Las desviaciones de los rasgos se asemejaron mucho entre especies, aunque destaca el contenido foliar en fósforo que presentó

una mayor desviación para *P. nigra*. Para confirmar las posible diferencias entre especies en los valores observados de los rasgos, se realizó un t-test en el entorno de R (R Core Team, 2021) donde se encontraron diferencias significativas para antocianinas (t = -3.011, df = 406.70, p-value = 0.003) , clorofila a (t = -12.171, df = 587.57, p-value < 0.001), clorofila b (t = -7.665, df = 585.23, p-value <- 0.001) y carbono orgánico (t = -3.790, df = 180.38, p-value < 0.001). Sin embargo, no se encontraron diferencias significativas para fósforo orgánico (t = -0.519, df = 169.87, p-value = 0.604) y nitrógeno orgánico (t = -1.485, df = 199.85, p-value = 0.139).

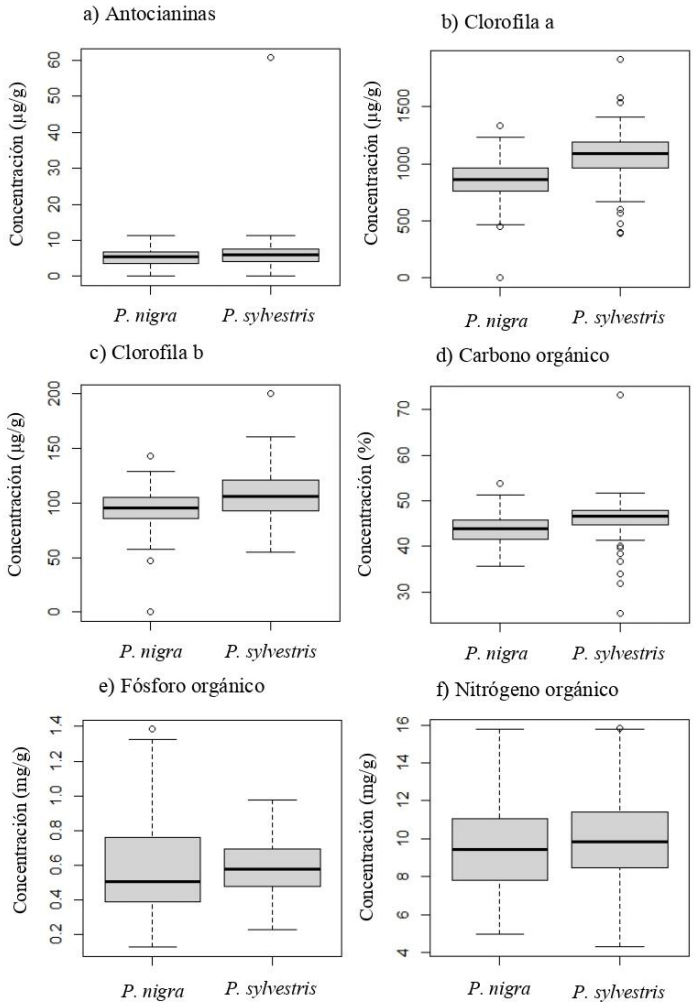

Figura 4. Box-plot de los valores medidos para los seis rasgos funcionales estudiados en *P. nigra* y *P. sylvestris*. Se presentan los resultados en forma de cuatro percentiles (cuartiles) junto a su desviación estándar (representado por las barras de error) y sus "outliers". El primer cuartil equivale al borde bajo de la caja principal (percentil 25), el segundo cuartil es el percentil 50, coincidiendo con la media. El borde superior de la caja equivale al percentil 75, y el bigote superior al percentil 100. Destaca la presencia de "outliers" en el caso de las antocianinas, clorofila a y carbono orgánico para *P. sylvestris*.

III. Modelos predictivos

Se obtuvieron un total de 252 modelos distintos por cada uno de los 6 rasgos de cada especie a los que se aplicaron los tres algoritmos de machine learning (PLS, SVM-L y RF) para los 7 rangos espectrales y los 12 pretratamientos.

i. Pinus nigra

Los modelos utilizados para predecir los rasgos funcionales de *Pinus nigra* alcanzaron valores de R^2 entre 0,0 y 0,5 en promedio (Fig. 5). El algoritmo que mejores ajustes produjo fue SVM-L, obteniendo valores de R^2 más altos para cuatro de los seis rasgos estudiados: antocianinas, clorofila a y b y carbono orgánico, mientras que el RF fue el más preciso para fósforo y nitrógeno inorgánico. Por otro lado, PLS no mostró un mayor R^2 en ninguno de los rasgos estudiados. En general, los rangos espectrales con una menor precisión alcanzada fueron los que corresponden al rango 400 – 1000 nm, 400 – 2450 nm, 701 – 2450 nm y 1001 – 2450 nm.

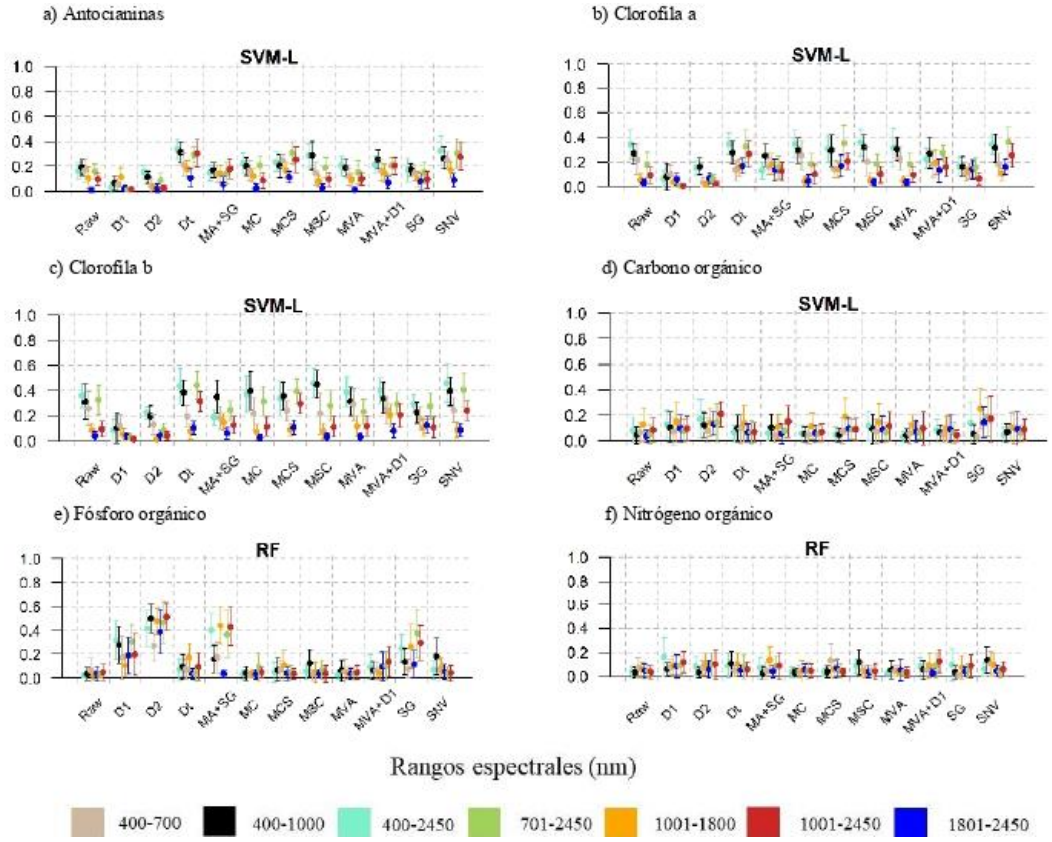

Figura 5: Precisión (R^2) de los modelos predictivos ajustados para *P. nigra* donde se representa la precisión promedio en el eje y (marcada con puntos) junto a su desviación estándar (representada por las barras de error) para a) antocianinas, b) clorofila a, c) clorofila b, d) carbono orgánico, e) fósforo orgánico y f) nitrógeno orgánico. En cada gráfica se presentan los 7 rangos espectrales seleccionados (representados con distintos colores, ver leyenda) y todos los pretratamientos aplicados: datos brutos (Raw), primera derivada (D1), segunda derivada (D2), detrend (Dt), datos brutos (Raw), primera derivada (D1), segunda derivada (D2), mean centering scaling (MCS), multiple scatter correction (MSC), moving averages (MVA), moving averages + primera derivada (MVA+D1), savitzky Golay (SG) y standard Normal Variate (SNV). Para facilitar la lectura e interpretación de los datos, se presentan los algoritmos que han mostrado una mayor precisión a la hora de predecir los rasgos bioquímicos en base al R^2 obtenido. Para visualizar los resultados obtenidos con los demás algoritmos para cada rasgo, dirigirse al Anexo.

Los modelos ajustados para las antocianinas (Fig. 5 a) mostraron un bajo poder de predicción bajo el algoritmo SVM - L, en el cual algunos modelos superaron el 0,35 de precisión media con una desviación que llegó a sobrepasar el umbral del 0,4. Estos modelos corresponden con el rango espectral 400 - 2450 nm y los pretratamientos Dt y SNV. Por otro lado, los peores modelos que se observaron son aquellos asociados al rango espectral 1001 – 1800 nm y 1801 – 2450 nm, los cuales apenas superaron el 0,1 de precisión y en ciertos pretratamientos no despegaron de la base del gráfico.

En cuanto a la clorofila a (Fig. 5 b) hubo varios modelos con los que se obtuvieron R^2 similares, siendo aquellos que corresponden con el rango espectral 400 – 2450 nm y pretratamiento DT y MSC, y para el espectro 701 – 1450 nm y pretratamiento MCS y SNV. Estos modelos mostraron un R^2 promedio cercano a 0,4. Además las desviaciones observadas fueron mayores para el segundo tramo del espectro de absorción. Por otro lado, los modelos con una menor R^2 fueron los correspondientes a los rangos 1001 – 2450 y 1801 – 2450 nm, en concreto para los pretratamientos D1 y D2 donde apenas se observó un poder de predicción mayor que cero.

Los modelos referentes a la clorofila b (Fig. 5 c) mostraron un poder de predicción que superó el 0,4 en varias ocasiones en los modelos presentes para el algoritmo SVM-L. Estos modelos más exactos corresponden con el rango espectral 400 - 2450 nm para los pretratamientos MSC Y SNV, donde el último presentó una desviación que superaba el 0,6 de precisión. Coincidiendo con los dos rasgos anteriores, los modelos con un poder de predicción más bajo fueron los asociados a los rangos espectrales 1001 – 1800 nm y 1801 – 2450 nm, cercanos al valor 0,1 e incluso por debajo de este.

Se puede destacar que los mejores modelos ajustados para las clorofilas coincidieron o incluían el espectro donde estos pigmentos muestran una mayor pico de absorbancia, el 400 - 700 nm y los peores modelos coincidían con aquellos en los que no estaba incluido este rango. Sin embargo, se observó un caso en el que no se cumplía, el rango espectral 701 – 2450 nm el cual mostró modelos con una precisión mayor del 0,3, superando incluso a aquellos modelos que incluían el rango 400 – 700 nm, que apenas superaron el 0,3.

Los modelos ajustados para el carbono orgánico (Fig. 5 d) presentaron un bajo poder de predicción, raramente superando el 0,2 en todos los pretratamientos y rangos espectrales. Sin embargo, destacó el modelo realizado con el rango espectral 1001-1800 nm y el pretratamiento SG, ya que presentó un poder de predicción cercano al 0,3 en promedio y poseía una desviación que lograba superar el umbral del 0,4.

Los modelos obtenidos para el fósforo orgánico (Fig. 5 e) son aquellos donde se encontró el mayor poder de predicción para la especie *P. nigra*. El algoritmo que se utilizó es RF y varios modelos superaron el 0,5 de predicción en promedio, estos se concentraron en dos pretratamientos

específicos, el D2 y MA + SG. Sin embargo, destacó entre todos aquel que utilizó el espectro 1001-2450 para el pretratamiento D2, ya que su predicción media se posicionaba por encima del 0,5 y su desviación superaba el 0,6. El pretratamiento SG también destacó entre los demás por tener un poder de predicción que sobresalía entre los demás, ya que muchos apenas superaron el 0,1 como los que fueron sometidos al pretratamiento MC, MVA y los que se quedaron en Raw.

En cuanto a los modelos ajustados para el nitrógeno orgánico, (Fig. 5 f) los más fiables fueron aquellos en los que se utilizó RF, aunque presentaba el menor poder de predicción de todos ya que no mostraba ningún dato por encima del 0,2, salvando las desviaciones de algunos modelos. El modelo más fiable para este rasgo correspondía con el asociado al espectro 400 – 2450 nm para el pretratamiento D1, aunque su poder de predicción era 0,19 y la desviación apenas superaba el 0,3. Los demás modelos se encontraban cercanos al 0,1 pero destacaron varios que apenas lograban alejarse de la base del gráfico.

ii. Pinus sylvestris

Respecto a los modelos ajustados para *P. sylvestris* (Fig. 6), su precisión no fue muy alta ya que se pudo ver que los valores promedio de R^2 nunca superaron el 0,4. De forma general se observó una predominancia del algoritmo SVM-L que fue el más preciso en los rasgos de clorofila a y b y en fósforo orgánico, seguido de RF para antocianinas y carbono orgánico y finalmente PLS para nitrógeno orgánico.

Respecto a aquellos modelos ajustados con datos de antocianinas (Fig. 6 a), los más precisos fueron aquellos realizados con Random forest, uno con el pretratamiento MVA+D1 y en el rango espectral de 1801-2450 nm y el otro con el pretratamiento SG y en el rango espectral de 400-700 nm. Sin embargo, la precisión de estos modelos fue relativamente baja ya que apenas lograron superar el 0,25 de media y su desviación estándar varió desde el 0,0 hasta casi rozar el 0,5 de precisión.

Los modelos ajustados para ambas clorofilas compartieron una preferencia por un mismo algoritmo, el SVM-L pero con resultados diferentes. Por un lado, entre aquellos ajustados para la clorofila a (Fig. 6 b) apenas sobresalieron dos modelos que lograron superar el 0,2 de precisión en promedio, uno fue aquel con el pretratamiento MCS en el rango espectar de los 400-2450 nm y otro con el pretratamiento MSC en el rango de los 400-1000. Por otro lado, en los modelos desarrollados para la clorofila b (Fig. 6 c) hubo varios cuyas medias rozaban el 0,4 de precisión y los bigotes de la desviación el 0,45. Estos modelos correspondían a un mismo rango espectral, 400-1000 nm, y a los pretratamientos Dt y MSC.

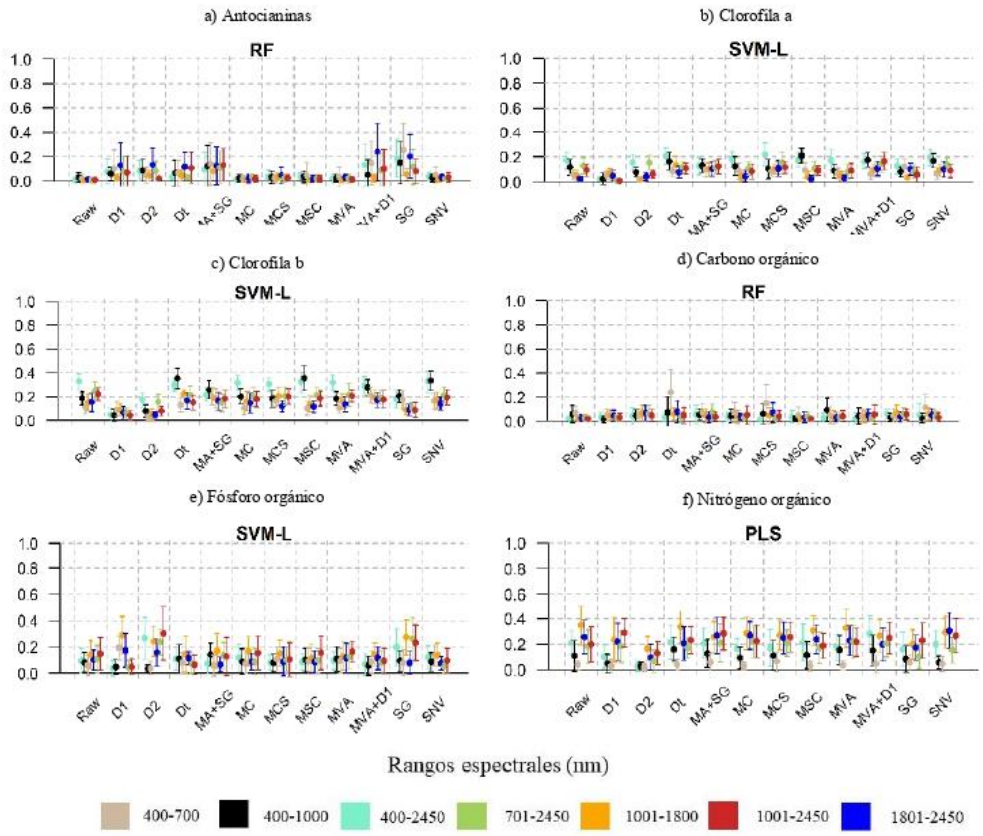

Figura 6: Precisión (R^2) de los modelos predictivos ajustados para *P. sylvestris* donde se representa la precisión promedio en el eje y (marcada con puntos) junto a su desviación estándar (representada por las barras de error) para a) antocianinas, b) clorofila a, c) clorofila b, d) carbono orgánico, e) fósforo orgánico y f) nitrógeno orgánico. En cada gráfica se presentan los 7 rangos espectrales seleccionados (representados con distintos colores, ver leyenda) y todos los pretratamientos aplicados: datos brutos (Raw), primera derivada (D1), segunda derivada (D2), detrend (Dt), datos brutos (Raw), primera derivada (D1), segunda derivada (D2), mean centering scaling (MCS), multiple scatter correction (MSC), moving averages (MVA), moving averages + primera derivada (MVA+D1), savitzky Golay (SG) y standard Normal Variate (SNV). Para facilitar la lectura e interpretación de los datos, se presentan los algoritmos que han mostrado una mayor precisión a la hora de predecir los rasgos bioquímicos en base al R^2 obtenido. Para visualizar los resultados obtenidos con los demás algoritmos para cada rasgo, dirigirse al Anexo.

En el caso del carbono orgánico (Fig. 6 d), el algoritmo que presentó mayor precisión fue el Random forest. Aunque la mayoría de los modelos apenas llegó a superar el 0,1 de precisión, destacó el asociado al pretratamiento Dt y al rango espectral 400-700nm, con una precisión media cercana al 0,25 y una desviación que superaba el 0,4.

Los modelos para el fósforo orgánico (Fig. 6 e) mostraron unos resultados algo mejores que los anteriores. Los mejores resultados se obtuvieron con el algoritmo SVM-L donde se alcanzó una precisión del 0,3 y una desviación que llegaba a superar el 0,5 asociadas al rango espectral 1001-2450 y el pretratamiento D2. Finalmente, aquellos modelos ajustados para el nitrógeno orgánico (Fig. 6 f) mostraron su mayor precisión con el algoritmo PLS. La precisión media fue muy cercana al 0,4 y las desviaciones llegaron al 0,5 en el rango espectral 1001 - 1800 para los pretratamientos Dt, MVA e incluso RAW.

Discusión

A lo largo del transcurso del presente trabajo, se ha estado analizando el potencial que tiene la espectrometría Vis-NIR a la hora de predecir diversos rasgos fenotípicos de las hojas de dos especies arbóreas coexistentes, *Pinus nigra* y *P. sylvestris*.

Tras medir los espectros de absorción foliares, determinar los rasgos en el laboratorio y aplicar tres diferentes algoritmos de machine learning para tratar de predecirlos, se observó que el algoritmo que demostró una mayor precisión fue el SVM, ya que resultó ser el más fiable en cerca del 60% de los casos, es decir, considerando los modelos para las dos especies. Sin embargo, los modelos predictivos ajustados en general fueron poco precisos, ya que ningún modelo en el mejor de los casos superó el R^2 de 0,6. Estos umbrales suelen ser superados en estudios similares con algunos de los rasgos estudiados, por ejemplo Chen et al. (2021) obtuvo valores de $R^2 \geq 0.86$ para clorofilas y nitrógeno orgánico, siendo bastante superiores. Costa et al. (2018) proponen en su trabajo una clasificación de los modelos en función de su poder de predicción, considerando que R^2>0,7 son modelos buenos, R^2 = 0.5–0.7 moderados y R^2<0,5 malos. Si aplicásemos esta clasificación en nuestro estudio, solo tendríamos modelos moderados para un rasgo, el del fósforo orgánico en *P. nigra*, mientras que para el resto de los rasgos en ambas especies los modelos serían malos. Aun así, y de forma general, se pudo observar un mayor poder de predicción en aquellos modelos referentes a la especie *P. nigra* donde se encontraron modelos que superaban el 0,4 de media de precisión incluso llegando al 0,5. Por otro lado, para la especie *P. sylvestris* ningún poder de predicción superó el 0,4 de media, lo cual supuso que ninguno de los modelos construidos llegó al umbral de precisión deseado. Esto nos hace plantear una serie de cuestiones como si es posible que los rasgos fenotípicos de ambas especies sean difícilmente predecibles con la técnica de la espectroscopia de Vis-NIR o si es posible que hayan existido errores humanos a la hora de tomar las muestras y procesarlas.

Estas cuestiones pueden parecer muy independientes entre sí, pero pueden responderse de manera conjunta. Recordemos que los individuos pertenecientes a la especie *P. sylvestris* tenían una mayor altura frente a los de *P. nigra*, lo que propició el descarte de ciertos árboles a la hora de realizar el muestreo debido a su gran tamaño: mientras que de *P. nigra* se descartaron 18 individuos, de *P. sylvestris* se descartaron 64, más del triple. La toma de muestras de *P. sylvestris* en los individuos de mayor altura se realizó en las ramas más bajas, posiblemente en un peor estado fisiológico frente a aquellas muestras tomadas en las ramas altas de los individuos más bajos, que eran, por tanto, más accesibles. Esto conllevó la existencia de una mayor variabilidad natural en las muestras a la hora de determinar los rasgos fenotípicos en el laboratorio, lo que pudo afectar al ajuste de los modelos. En cambio, los individuos de *P. nigra* fueron más accesibles al no tener tanta altura, permitiendo coger de manera más homogénea hojas más frescas y que

habían recibido un mayor aporte de luz. Esta hipótesis relacionada con la existencia de diferencias en cuanto a la absorción de luz entre las hojas muestreadas para ambas especies se apoya en los resultados de los t-test, que fueron significativos para los valores de los pigmentos fotosintéticos (antocianinas, clorofila a y b) y carbono orgánico, cuya concentración está directamente relacionada con la actividad fotosintética (Bishop, 1971).

La precisión de los modelos ajustados para las clorofilas rondó el 0,3 aunque en algunos casos lograron rozar el 0,5. Esto coincide con otros estudios en los que los resultados de las clorofilas corresponden con predicciones de 0,3 (Costa et al., 2018). Los autores de este estudio comentan la posibilidad de que estos malos resultados fueran resultado de una elección incorrecta de protocolos a la hora de analizar estos rasgos o una posible limitación en el equipo que se utilizó. En el caso del presente trabajo, debemos hacer una labor de introspección para lograr entender qué ha podido suceder para obtener unos valores tan bajos. El proceso de recogida de muestras y el posterior análisis de pigmentos fotosintéticos es bastante delicado debido a que son sustancias fotosensibles y se pueden degradar muy fácilmente. La recogida y análisis de muestras se realizó de la forma más precavida posible para evitar su posible degradación; se almacenaron en frío y oscuridad nada más empaquetar, se conservaron en congelador a – 80ºC para impedir una pérdida de sus cualidades y se analizaron siempre sobre hielo y en la máxima oscuridad posible. A pesar de todos estos esfuerzos, hay varios factores a tener en cuenta que pudieron afectar a la calidad de las muestras. Uno es la gran distancia existente entre la zona de estudio y el laboratorio donde se analizaron las muestras, la cual es de aproximadamente de 200 km por carretera. Esta gran distancia se traduce en un largo periodo de tiempo desde que se recogieron las muestras hasta que fueron depositadas en el congelador, conllevando un posible decremento en la concentración de las clorofilas o un deterioro de sus características. Otro factor que puede haber influenciado en estos malos resultados son las condiciones de trabajo en el laboratorio. Aunque se actuó y adaptó el área de trabajo para tener una condiciones con una temperatura y oscuridad lo más constantes posible, estas condiciones no se pudieron mantener siempre. Esto es debido a que al trabajar de día y en un laboratorio compartido con otros investigadores, no se pueden tener unas condiciones de oscuridad ideal que impidan la estimulación de los pigmentos fotosintéticos. Esto se puede observar en los "outliers" que se observaron en la Figura 4, ya que la mayoría de ellos corresponden con los pigmentos fotosintéticos. El caso de los "outliers" presentes para el carbono orgánico, podrían deberse a alguna posible contaminación en la zona de trabajo y en el procesado de las muestras. Por último, cabe la posibilidad de que los protocolos utilizados o el equipo usado no tuvieran la capacidad de medir con exactitud estos rasgos, ya que es posible que la sensibilidad de la técnica no permita caracterizar la variabilidad intraespecífica de las muestras. El rango de la concentración de cada rasgo en las especies medidas puede ser muy bajo y es posible que la técnica empleada no permita diferenciar esto (Costa et al., 2018). Otra posibilidad de fallo puede

radicar en el número de muestras caracterizadas y usadas para construir los modelos, ya que puede que sea demasiado bajo como para proporcionar suficientes datos que permitan construir modelos con una alta exactitud.

Volviendo a los algoritmos, y como se ha mencionado anteriormente, SVM - L es un algoritmo no linear, lo que nos puede indicar que existe información no linear que se puede estar perdiendo al utilizar otros algoritmos que utilizan métodos lineales como el PLS. Esto es algo que no ha sido demostrado ya que diversos autores establecieron que la tecnología Vis-NIRs es una tecnología linear, donde no existe posibilidad que ofrezcan ventajas frente a algoritmos que utilizan métodos lineales (Costa et al., 2018; Nicolaï et al., 2007). Esta información puede servir como base para una investigación futura la cual intente explicar si, en efecto, los algoritmos que utilizan métodos no lineares (como es el SVM – L) son más aptos a la hora de analizar los datos espectrales de materiales biológicos frente a aquellos algoritmos que utilizan métodos lineales como el PLS y tienen pérdidas de datos no lineares.

Resulta interesante comentar la importancia de usar pretratamientos a la hora de utilizar espectros de absorción ya que se utilizan para detectar información que de otra manera no se conseguiría y eliminar ruido. Esto se confirma en el presente trabajo ya que los espectros fueron bastante similares para ambas especies cuando no habían sido sometidos a ningún pretratamiento (RAW), pero la aplicación de distintos pretratamientos permitió detectar algunas diferencias entre especies. Además, el hecho de que las especies muestren espectros con diferencias marcadas e identificativas (Fig. 3) nos puede hacer pensar si es posible diferenciar entre especies simplemente por su espectro de absorbancia, como sí se ha podido hacer en otros estudios (Sohn et al., 2021). En nuestro ejemplo de la figura 3b, el primer pico de absorbancia a los 400 nm es un rasgo identificativo de la especie *P. nigra*, ya que la especie *P. sylvestris* apenas se ve estimulada en este rango. Esto puede suponer una herramienta para futuros estudios con otras especies o subespecies que presenten dificultad a la hora de ser determinadas, como por ejemplo aquellas en que las diferencias son muy sutiles o deben hacerse por análisis genético (Li et al., 2019). Diversos estudios han concluido que esta es una útil y potente herramienta debido a la facilidad y rapidez de este método además de la alta fiabilidad que proporcionan ya que, en ciertas especies, se ha demostrado un porcentaje de precisión de identificación del 99,7 % (± 0,006), incluso en etapas de crecimiento utilizando el algoritmo SVM y el pretratamiento Derivative Savitzky- Golay (Sohn et al., 2021). Aunque para aquellos modos construidos con el espectro sin ser modificado (Raw) los autores de este estudio también obtuvieron un buen porcentaje de predicción (R^2 = 98.0 ± 0.008).

A través de esta discusión se han puesto de manifiesto varias líneas de investigación futuras. Una de ellas es la eliminación de los “outliers” presentes en nuestros datos para tratar de aumentar el

poder de precisión de los modelos. Por otra parte, debido a que en los resultados se vio cómo en ciertos rangos espectrales utilizados los modelos eran muy pobres, sería recomendable variar estos rangos probando con otros más restringidos. En aquellos casos en que se obtuvieron modelos más precisos, la utilización de bandas espectrales más específicas podría ayudar a mejorar la precisión. Por otro lado, en el planteamiento del proyecto se podría considerar seleccionar un mayor número de muestras a caracterizar ya que, como se ha dicho anteriormente, un mayor número de muestras ayudaría a ampliar la información de partida para ajustar los modelos, muy posiblemente mejorando su precisión. Por último, también se sugiere considerar en conjunto los modelos predictivos de las dos especies ya que un objetivo futuro sería obtener modelos aplicables para predecir rasgos funcionales en distintas especies de forma rápida y precisa. Si esto se consiguiese, permitiría fenotipar de forma masiva comunidades completas en muy poco tiempo permitiendo a los investigadores abarcar comunidades mucho más grandes en mucho menos tiempo y con un presupuesto mucho más bajo. Por ende, la comunidad científica gozaría de un mayor aporte de información que podría ser aplicada a la vida ciudadana. Por ejemplo, este conocimiento podría ser aplicable en el ámbito de la restauración ecológica de ecosistemas degradados que han sufrido un deterioro de sus procesos y funciones y que, por tanto, proporcionan una serie de servicios más limitados, ya que permitiría conocer de forma rápida y sencilla el estado funcional de los individuos que los componen.

Conclusiones

- Los modelos predictivos basados en el espectro de absorción del infrarrojo cercano tienen un gran potencial a la hora de predecir datos fenotípicos aun cuando los resultados que se han obtenido en el estudio no son satisfactorios. Se deberían tomar otros caminos y analizar los posibles fallos que hayan podido ocurrir para evitarlos y tratar de corregirlos en futuros estudios.

- El algoritmo que demostró un mayor poder de predicción fue el SVM-L, dando pie a una posible investigación futura para valorar la existencia de información no linear a la hora de realizar estos modelos, o de su mayor aptitud para analizar datos referentes a rasgos fisicoquímicos.

- Se ha comprobado cómo, aun con dos especies muy parecidas del mismo género, existen diferencias entre los valores observados para los rasgos de los pigmentos fotosintéticos (antocianinas, clorofilas a y b) y el carbono orgánico, lo que podría estar afectando al ajuste de los modelos. Esto puede arrojar evidencia sobre la posible diferencia en la capacidad fotosintética de estos o, de las diferentes estrategias de coexistencia de cada especie, pero sería necesario encontrar modelos que fueran capaces de predecir de forma adecuada estas diferencias a nivel individual.

Agradecimientos

En primer lugar me gustaría agradecer la constante dedicación y apoyo de mis dos tutores Ana Isabel García-Cervigón y David Sánchez Pescador, quienes me han guiado.

Este trabajo ha sido financiado por la Agencia Estatal de Investigación a través del proyecto Phenotypes (PGC2018-099115-B-I00), por lo que me gustaría agradecer a Adrián Escudero Alcántara y a la agencia haberme permitido participar en este gran proyecto de investigación.

Por último, agradecer a todos los compañeros del proyecto por su incansable labor investigadora, en especial a Manuel Rojo Valencia, Carlos Díaz Palomo y Javier Pajares Pérez.

Bibliografía

Albert, C. H., Grassein, F., Schurr, F. M., Vieilledent, G., & Violle, C. (2011). When and how should intraspecific variability be considered in trait-based plant ecology? *Perspectives in Plant Ecology, Evolution and Systematics*, *13*(3), 217–225. https://doi.org/10.1016/j.ppees.2011.04.003

Alciaturi, C. E., Escobar, M. E., De La Cruz, C., & Rincón, C. (2003). Partial least squares (PLS) regression and its application to coal analysis. *Revista Tecnica de La Facultad de Ingenieria Universidad Del Zulia*, *26*(3), 197–204.

Anderson, J. M., & Ingram, J. S. I. (1990). Tropical Soil Biology and Fertility: A Handbook of Methods. *The Journal of Ecology*, *78*(2), 547. https://doi.org/10.2307/2261129

Auger, S., & Shipley, B. (2013). Inter-specific and intra-specific trait variation along short environmental gradients in an old-growth temperate forest. *Journal of Vegetation Science*, *24*(3), 419–428. https://doi.org/10.1111/j.1654-1103.2012.01473.x

Bastias, C. C., Benavides, R., Sansevero, J. B. B., GODOY, A. ., GARCÍA- RABASA, S., & Valladares, F. (2013). *Efecto de la diversidad del dosel en la regeneración de diferentes especies arbóreas*. 1–9.

Bishop, N. I. (1971). *Photosynthesis: The electron transport system of green plants*.

Caballero, J. (2006). SEM vs . PLS : Un enfoque basado en la práctica. *IV Congreso de Metodología de Encuestas*, 57–66. http://www.unavarra.es/congreso/encuestas/index.htm

Chen, L., Zhang, Y., Nunes, M. H., Stoddart, J., Khoury, S., Chan, A. H. Y., & Coomes, D. A. (2021). Predicting leaf traits of temperate broadleaf deciduous trees from hyperspectral reflectance: can a general model be applied across a growing season? *Remote Sensing of*

Environment, *October*, 112767. https://doi.org/10.1016/j.rse.2021.112767

Costa, F. R. C., Lang, C., Almeida, D. R. A., Castilho, C. V., & Poorter, L. (2018). Near-infrared spectrometry allows fast and extensive predictions of functional traits from dry leaves and branches. *Ecological Applications*, *28*(5), 1157–1167. https://doi.org/10.1002/eap.1728

Darwin, C., & Kebler, L. (1859). *On the Origin of Species by Means of Natural Selection.* (Retrieved). J. Murray. http://hdl.loc.gov/loc.rbc/General.17473.1

Escudero, A., Matesanz, S., Pescador, D. S., De, M., Fernando, C., & Cavieres, L. A. (2021). Every bit helps : The functional role of individuals in assembling any plant community , from the richest to monospecific ones. *Journal of Vegetation Science*, *July 2020*, 1–9. https://doi.org/10.1111/jvs.13059

Escudero, A., & Valladares, F. (2016). Trait-based plant ecology: moving towards a unifying species coexistence theory: Features of the Special Section. *Oecologia*, *180*(4), 919–922. https://doi.org/10.1007/s00442-016-3578-5

Ferrero, L. M., Montouto, O., & Herranz, J. M. (2006). *Flora Amenazada y de interés del parque natural del alto tajo*.

Fridley, J. D., Grime, J. P., & Bilton, M. (2007). Genetic identity of interspecific neighbours mediates plant responses to competition and environmental variation in a species-rich grassland. *Journal of Ecology*, *95*(5), 908–915. https://doi.org/10.1111/j.1365-2745.2007.01256.x

García-Plazaola, J. I., & Becerril, J. M. (2001). Seasonal changes in photosynthetic pigments and antioxidants in beech (Fagus sylvatica) in a Mediterranean climate: Implications for tree decline diagnosis. *Australian Journal of Plant Physiology*, *28*(3), 225–232. https://doi.org/10.1071/pp00119

Geladi, P., & Kowalski, B. R. (1986). An example of 2-block predictive partial least-squares regression with simulated data. *Analytica Chimica Acta*, *185*(C), 19–32. https://doi.org/10.1016/0003-2670(86)80029-0

Granda, E., Escudero, A., de la Cruz, M., & Valladares, F. (2012). Juvenile-adult tree associations in a continental Mediterranean ecosystem: No evidence for sustained and general facilitation at increased aridity. *Journal of Vegetation Science*, *23*(1), 164–175. https://doi.org/10.1111/j.1654-1103.2011.01343.x

Hart, S. P., Schreiber, S. J., & Levine, J. M. (2016). How variation between individuals affects species coexistence. *Ecology Letters*, *19*(8), 825–838. https://doi.org/10.1111/ele.12618

Herrera, C. M. (2017). The ecology of subindividual variability in plants: Patterns, processes, and prospects. *Web Ecology*, *17*(2), 51–64. https://doi.org/10.5194/we-17-51-2017

Hubbell, S. P. (2005). Neutral theory in community ecology and the hypothesis of functional equivalence. *Functional Ecology*, *19*(1), 166–172. https://doi.org/10.1111/j.0269-8463.2005.00965.x

Jin, Z., Shang, J., Zhu, Q., Ling, C., Xie, W., & Qiang, B. (2020). Random Forests. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12343 LNCS*, 503–515. https://doi.org/10.1007/978-3-030-62008-0_35

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). *kernlab -- An {S4} Package for Kernel Methods in {R}* (pp. 1--20). Journal of Statistical Software. http://www.jstatsoft.org/v11/i09/

Kothari, S. A. (2021). *Reflectance spectroscopy allows rapid , accurate , and non-destructive estimates of functional traits from pressed leaves*.

Kraft, N. J. B., Godoy, O., & Levine, J. M. (2015). Plant functional traits and the multidimensional nature of species coexistence. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(3), 797–802. https://doi.org/10.1073/pnas.1413650112

Kucheryavskiy, S. (2020). mdatools --- R package for chemometrics. In *Chemometrics and Intelligent Laboratory Systems*. https://doi.org/10.1016/j.chemolab.2020.103937

Kuhn, M., Wing, J., & Weston, S. (2021). *Classification and Regression Training* (6.0-90). https://github.com/topepo/caret/

Li, Y., Via, B. K., & Young, T. (2019). *Visible-Near Infrared Spectroscopy and Chemometric Methods for Wood Density Prediction and Origin / Species Identification*. 1–19.

Mevik, B.-H., Wehrens, R., & Liland, K. H. (2011). *Partial Least Squares and Principal Component regression* (2.3-0). http://mevik.net/work/software/pls.html

Nicolaï, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & 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

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Roudier, P. (2017). *asdreader: Reading ASD Binary Files in R*. https://cran.r-

project.org/package=asdreader

Shipley, B., De Bello, F., Cornelissen, J. H. C., Laliberté, E., Laughlin, D. C., & Reich, P. B. (2016). Reinforcing loose foundation stones in trait-based plant ecology. *Oecologia*, *180*(4), 923–931. https://doi.org/10.1007/s00442-016-3549-x

Sides, C. B., Enquist, B. J., Ebersole, J. J., Smith, M. N., Henderson, A. N., & Sloat, L. L. (2014). Revisiting darwins hypothesis: Does greater intraspecific variability increase species ecological breadth? *American Journal of Botany*, *101*(1), 56–62. https://doi.org/10.3732/ajb.1300284

Sohn, S., Oh, Y., Pandian, S., Lee, Y., & Zaukuu, J. Z. (2021). *Identification of Amaranthus Species Using Visible-Near-Infrared (Vis-NIR) Identification of Amaranthus Species Using Visible-Near-Infrared (Vis-NIR) Spectroscopy and Machine Learning Methods*. *October*. https://doi.org/10.3390/rs13204149

Stevens, A., & Ramirez-Lopez, L. (2021). *An introduction to the prospectr package* (R package version 0.2.2). R package Vignette.

Violle, C., Enquist, B. J., McGill, B. J., Jiang, L., Albert, C. H., Hulshof, C., Jung, V., & Messier, J. (2012). The return of the variance: Intraspecific variability in community ecology. *Trends in Ecology and Evolution*, *27*(4), 244–252. https://doi.org/10.1016/j.tree.2011.11.014

Violle, C., Navas, M.-L., Vile, D., Kazakou, E., Fortunel, C., Hummel, I., & Garnier, E. (2007). Let the concept of trait be functional! *Oikos*, *116*(5), 882–892. https://doi.org/10.1111/j.2007.0030-1299.15559.x

Wright, M. N., Wager, S., & Probst, P. (2021). *A Fast Implementation of Random Forests*. https://github.com/imbs-hl/ranger

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1). https://doi.org/10.18637/jss.v077.i01

Anexos

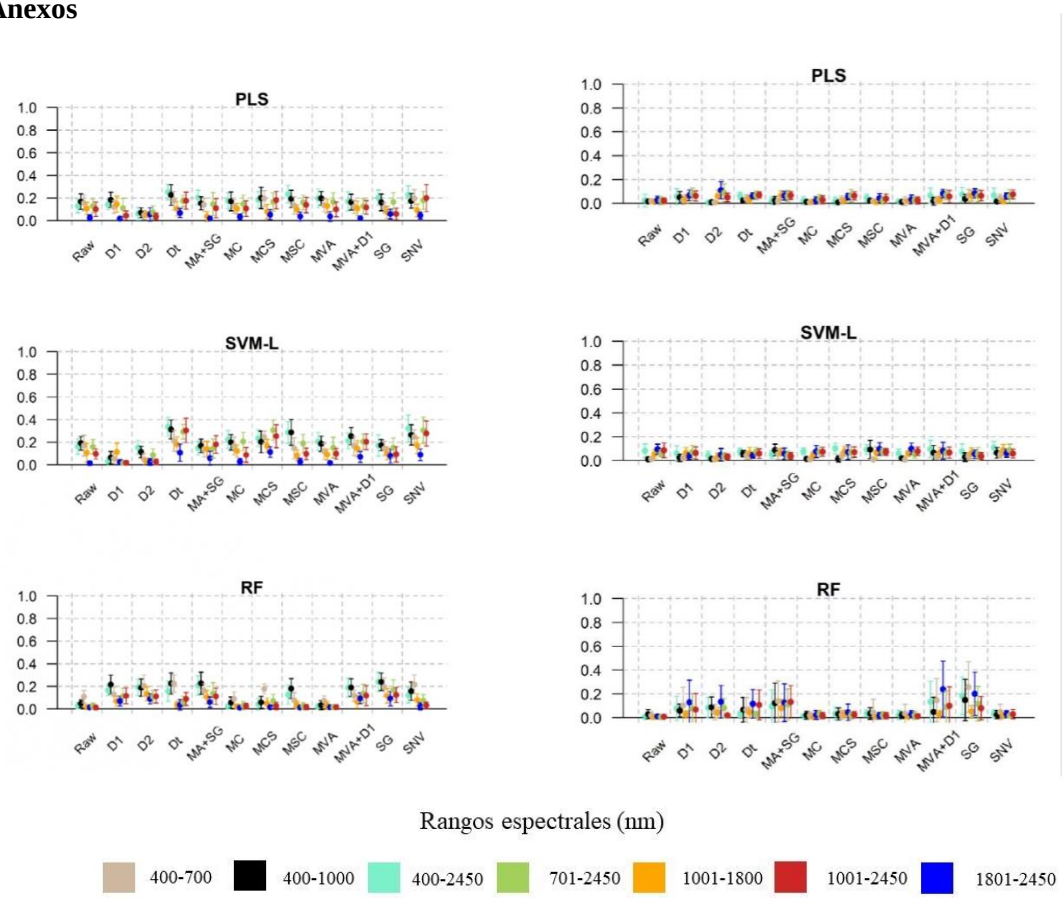

Figura A1. Modelos predictivos seleccionados para el rasgo de antocianinas en las dos especies; *P. nigra* (a) y *P. syilvstris (b)*, donde se representa la precisión promedia (punto) junto a su desviación estándar representada por las líneas. En cada caso se tienen en cuenta los 7 rangos espectrales seleccionados (representados con distintos colores, ver leyenda) y todos los pretratamientos aplicados: datos brutos (Raw), primera derivada (D1), segunda derivada (D2), detrend (Dt), datos brutos (Raw), primera derivada (D1), segunda derivada (D2), mean centering scaling (MCS), multiple scatter correction (MSC), moving averages (MVA), moving averages + primera derivada (MVA+D1), savitzky Golay (SG) y standard Normal Variate (SNV).

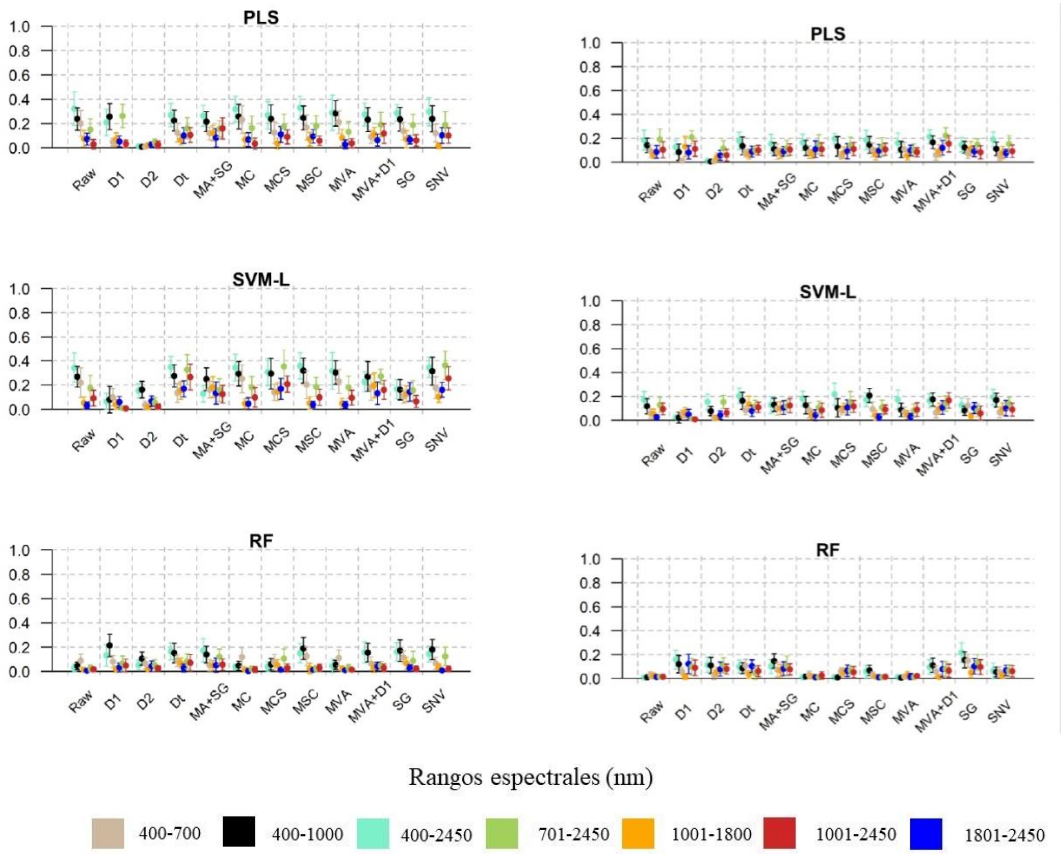

Figura A2. Modelos predictivos seleccionados para el rasgo de clorofila a en las dos especies; *P. nigra* (a) y *P. syilvstris* (b), donde se representa la precisión promedia (punto) junto a su desviación estándar representada por las líneas. En cada caso se tienen en cuenta los 7 rangos espectrales seleccionados (representados con distintos colores, ver leyenda) y todos los pretratamientos aplicados: datos brutos (Raw), primera derivada (D1), segunda derivada (D2), detrend (Dt), datos brutos (Raw), primera derivada (D1), segunda derivada (D2), mean centering scaling (MCS), multiple scatter correction (MSC), moving averages (MVA), moving averages + primera derivada (MVA+D1), savitzky Golay (SG) y standard Normal Variate (SNV).

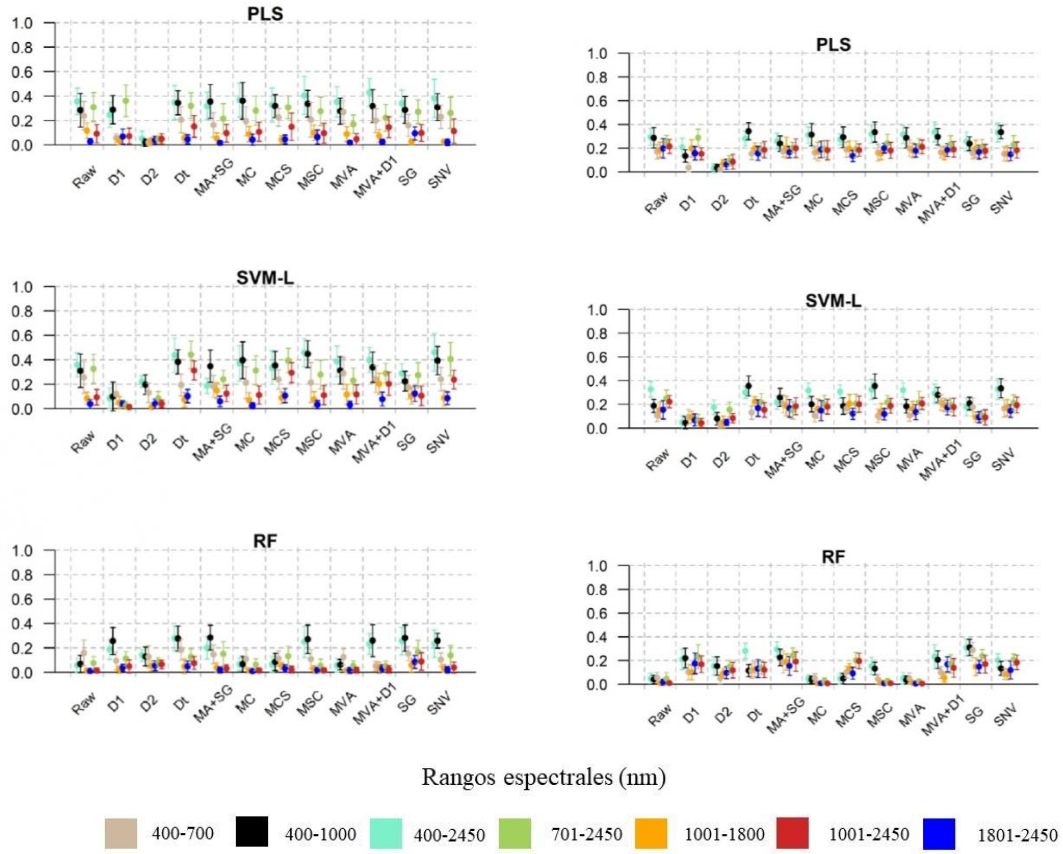

Figura A3. Modelos predictivos seleccionados para el rasgo de clorofila b en las dos especies; *P. nigra* (izquierda) y *P. syilvstris* (derecha), donde se representa la precisión promedia (punto) junto a su desviación estándar representada por las líneas. En cada caso se tienen en cuenta los 7 rangos espectrales seleccionados (representados con distintos colores, ver leyenda) y todos los pretratamientos aplicados: datos brutos (Raw), primera derivada (D1), segunda derivada (D2), detrend (Dt), datos brutos (Raw), primera derivada (D1), segunda derivada (D2), mean centering scaling (MCS), multiple scatter correction (MSC), moving averages (MVA), moving averages + primera derivada (MVA+D1), savitzky Golay (SG) y standard Normal Variate (SNV).

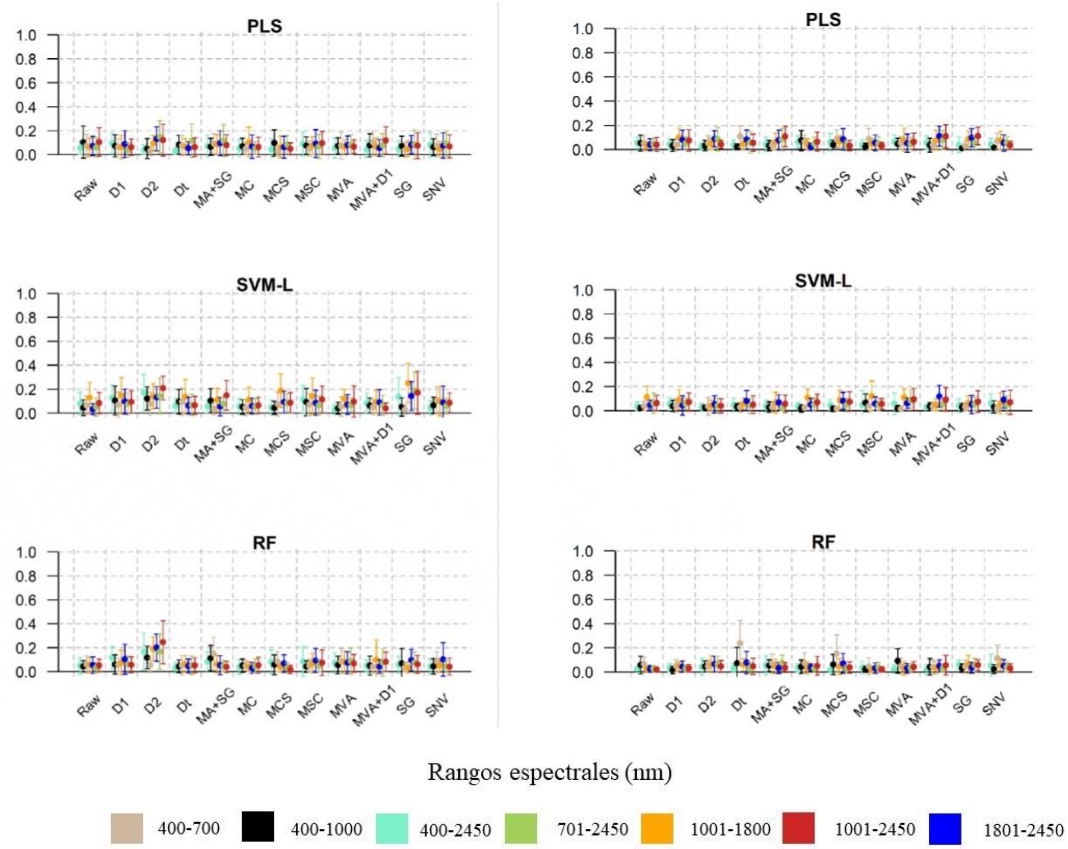

Figura A4. Modelos predictivos seleccionados para el rasgo de carbono orgánico en las dos especies; *P. nigra* (izquierda) y *P. syilvstris* (derecha), donde se representa la precisión promedia (punto) junto a su desviación estándar representada por las líneas. En cada caso se tienen en cuenta los 7 rangos espectrales seleccionados (representados con distintos colores, ver leyenda) y todos los pretratamientos aplicados: datos brutos (Raw), primera derivada (D1), segunda derivada (D2), detrend (Dt), datos brutos (Raw), primera derivada (D1), segunda derivada (D2), mean centering scaling (MCS), multiple scatter correction (MSC), moving averages (MVA), moving averages + primera derivada (MVA+D1), savitzky Golay (SG) y standard Normal Variate (SNV).

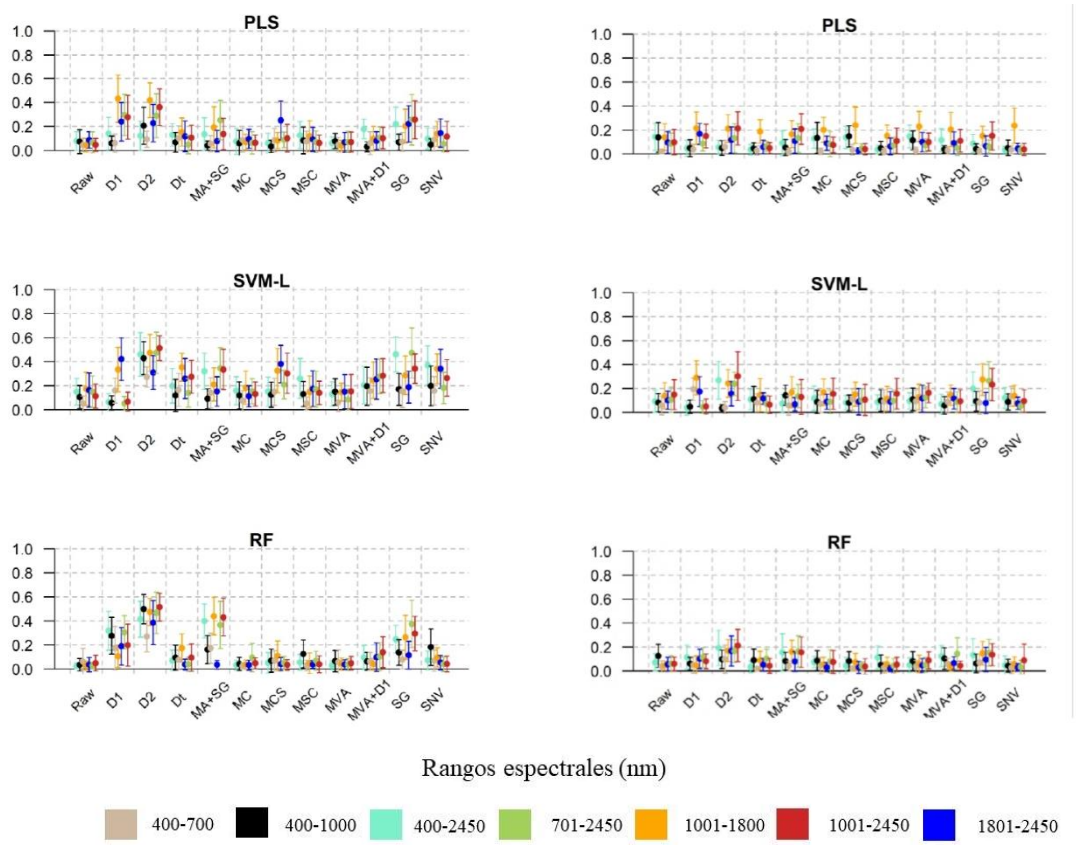

Figura A5. Modelos predictivos seleccionados para el rasgo de fósforo orgánico en las dos especies; *P. nigra* (izquierda) y *P. syilvstris* (derecha), donde se representa la precisión promedia (punto) junto a su desviación estándar representada por las líneas. En cada caso se tienen en cuenta los 7 rangos espectrales seleccionados (representados con distintos colores, ver leyenda) y todos los pretratamientos aplicados: datos brutos (Raw), primera derivada (D1), segunda derivada (D2), detrend (Dt), datos brutos (Raw), primera derivada (D1), segunda derivada (D2), mean centering scaling (MCS), multiple scatter correction (MSC), moving averages (MVA), moving averages + primera derivada (MVA+D1), savitzky Golay (SG) y standard Normal Variate (SNV).

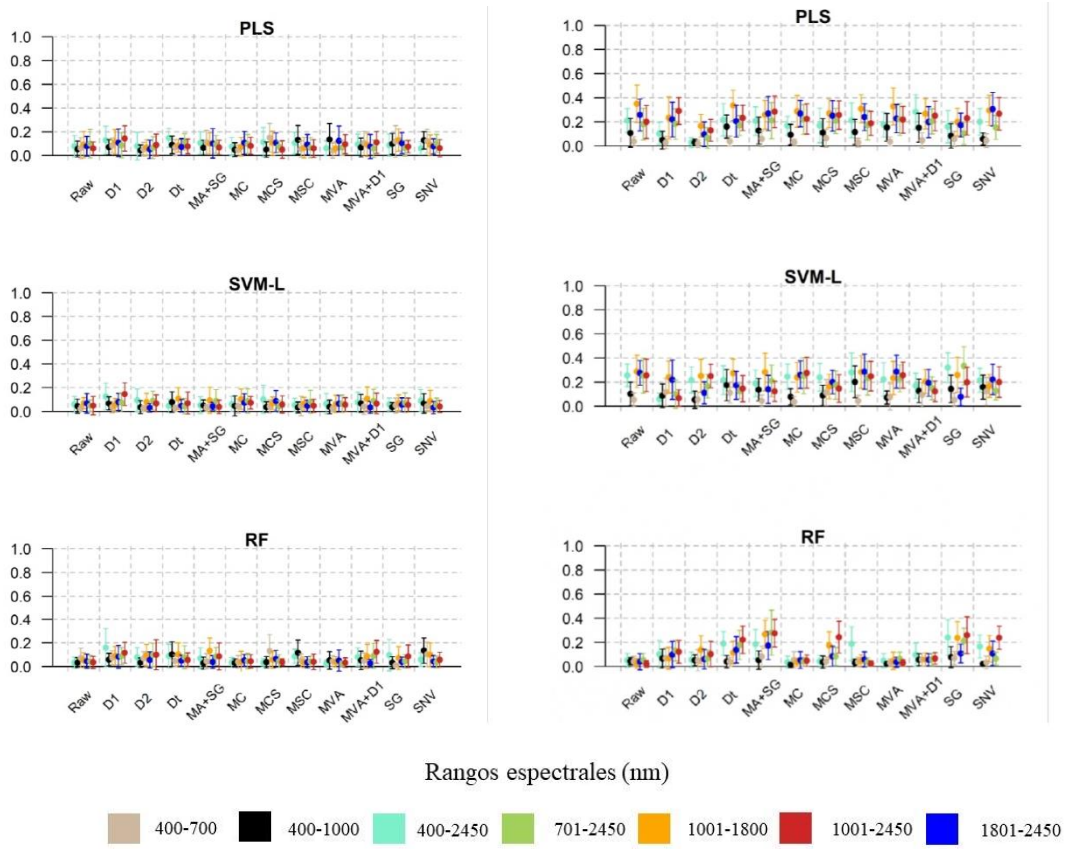

Figura A6. Modelos predictivos seleccionados para el rasgo de nitrógeno orgánico en las dos especies; *P. nigra* (izquierda) y *P. syilvstris* (derecha), donde se representa la precisión promedia (punto) junto a su desviación estándar representada por las líneas. En cada caso se tienen en cuenta los 7 rangos espectrales seleccionados (representados con distintos colores, ver leyenda) y todos los pretratamientos aplicados: datos brutos (Raw), primera derivada (D1), segunda derivada (D2), detrend (Dt), datos brutos (Raw), primera derivada (D1), segunda derivada (D2), mean centering scaling (MCS), multiple scatter correction (MSC), moving averages (MVA), moving averages + primera derivada (MVA+D1), savitzky Golay (SG) y standard Normal Variate (SNV).