Universidad de Costa Rica

Facultad de Ingenierías

Escuela de Ingeniería en Biosistemas

Investigación Dirigida para optar por el título de Licenciatura en Ingeniería Agrícola y de Biosistemas

DESARROLLO DE MODELOS PREDICTIVOS PARA LA CUANTIFICACIÓN DE NUTRIENTES EN EL SUELO MEDIANTE ESPECTROSCOPIA DE INFRARROJO MEDIO EN EL CANTÓN DE NICOYA, COMO UNA ALTERNATIVA AL ANÁLISIS DE LABORATORIO TRADICIONAL

Jose Andrés Parajeles Herrera

Ciudad Universitaria Rodrigo Facio

San José, Costa Rica

2025

Acta de la Presentación Oral del Trabajo Final de Graduación ante el Tribunal Examinador para Optar por el Grado de Licenciatura en Ingeniería Agrícola y de Biosistemas

Modalidad: Investigación Dirigida

Lugar: Sala de Audiovisuales 1, Facultad de Ingeniería	Fecha: 30/07/2025	Hora: 01:30 pm	Consecutivo: 14-2025

Artículo 1: Presentación de los miembros del Tribunal del Trabajo Final y del estudiante. Se explica el procedimiento que consiste en la presentación oral de 45 min, el espacio para preguntas, la deliberación del Tribunal y la comunicación del acuerdo, según los artículos 26 y 27 del Reglamento de Trabajos Finales de Graduación.

Tribunal Examinador

MIEMBROS DEL TRIBUNAL	ROL
Dra. Alejandra Rojas González	Directora TFG
Dr. Manuel Camacho Umaña	Miembro del Comité Asesor
Lic. José Carlos Lacayo Vega	Miembro del Comité Asesor
Lic. Juan Carlos Hernández Lewis	Lector Invitado
Dra. Marianela Alfaro Santamaría	Presidenta

Artículo 2: Exposición oral del estudiante

Nombre del TFG: Desarrollo de modelos predictivos para la cuantificación de nutrientes en el suelo mediante espectroscopía de infrarrojo medio en el cantón de Nicoya, como una alternativa al análisis de laboratorio tradicional.

NOMBRE DEL POSTULANTE	CARNÉ
José Andrés Parajeles Herrera	B65276

Artículo 3: Sesión de preguntas sobre aspectos propios del trabajo presentado

Artículo 4: Deliberación del Tribunal. Sale de la sala de sesiones el sustentante para la deliberación del Tribunal. Se le avisará al sustentante vía llamada telefónica o el medio acordado con el estudiante.

Observaciones:

1. Análisis de resultados: graficos de dispersión incorporar línea de tendencia.
2. Análisis de resultados: complementar y contrastar los resultados contra literatura científica.
3. Metodología: eliminar mucha literatura para reducir
4. Revisar citación en el documento (Faltan citas)

Artículo 5: Producto de la deliberación se realiza la votación pública:

Votación pública obteniendo: 5 votos a favor y 0 votos en contra.

Voto disidente: no

Calificación del Trabajo Final de Graduación:	Aprobado	X	No aprobado	
Aprobación con distinción el Trabajo Final de Graduación:	Sí		No	X

Artículo 6: La persona que preside el Tribunal comunica a la persona postulante el resultado de la deliberación y se le declara acreedora al grado de **Licenciatura en Ingeniería Agrícola y de Biosistemas.** Se le indica la obligación de presentarse al acto público de graduación , para ser juramentado y recibir el diploma correspondiente.

Se da lectura al acta que firma la persona que preside el Tribunal Examinador a las 16:10 horas.

Marianela Alfaro S. Dra. Marianela Alfaro Santamaría, persona que preside

Nota: De conformidad con los artículos 28, 29 y 36, el estudiante deberá entregar las copias con la versión final, incluyendo las modificaciones formuladas en ocasión de la presentación pública, un resumen de 200-500 palabras y la autorización de publicación del trabajo en el repositorio Kérwá. El director del trabajo final de graduación será responsable de que se realicen las correcciones propuestas en la presentación pública y de que la versión final del documento quede debidamente confeccionada.

Dedicatoria

A mis papás y a mis hermanas, por su amor y apoyo incondicional.

Y a mi versión de niño, mira lo lejos que hemos llegado.

Agradecimientos

A la Universidad de Costa Rica, por abrirme sus puertas y ser un lugar seguro donde cada persona puede ser realmente quien quiere ser.

Al Departamento de Estudios Básicos de Tierras (DEBT) del INTA, por facilitarme los datos de laboratorio necesarios para realizar este trabajo.

Al CICIMA y a todo su personal, por permitirme usar sus equipos para la obtención de los datos espectrales.

A Alejandro Bermúdez, del INTA, por su apoyo constante y seguimiento durante todo este proceso.

A José Lacayo, del INTA, por formar parte del comité de este trabajo y por todas sus ideas y apoyo a lo largo del camino.

Al Dr. Manuel Camacho, del CIA, por ser parte del comité y brindar todo su respaldo.

A la Dra. Alejandra Rojas, por su dirección y apoyo durante este proceso.

A mis amigos que me acompañaron durante todo este camino, Glori y José Adrián.

A ”mi rey”, que me ha acompañado desde el primer año, y fuimos una roca el uno para el otro.

Y, por supuesto, a Mariel, esto es de los dos.

Índice

Índice de Abreviaturas

- **DEBT**: Departamento de Estudios Básiscos de Tierras
- **ATR**: Reflectancia Total Atenuada
- **CIA**: Centro de Investigaciones Agronómicas
- **CICIMA**: Centro de Investigación en Ciencia e Ingeniería de Materiales
- **FAO**: Organización de las Naciones Unidas para la Alimentación y la Agricultura
- **INTA**: Instituto Nacional de Innovación y Transferencia en Tecnología Agropecuaria
- **MIR**: Espectroscopía de Infrarrojo Medio
- **ODS**: Objetivos de Desarrollo Sostenible
- **PCA**: Análisis de Componentes Principales
- **PLSR**: Regresión de Mínimos Cuadrados Parciales
- **R^2**: Coeficiente de Determinación
- **RPD**: Ratio de Desviación de Rendimiento
- **RMSE**: Raíz del Error Cuadrático Medio
- **SNV**: Variación Estándar Normalizada (Standard Normal Variate)

Índice de figuras

Índice de tablas

Resumen

Este trabajo se enfocó en desarrollar modelos predictivos basados en espectroscopía MIR para cuantificar nutrientes en suelos del cantón de Nicoya, Guanacaste, como alternativa a los análisis de laboratorio convencionales. Se realizó un análisis estadístico descriptivo de los nutrientes, la exploración de sus firmas espectrales, la evaluación de métodos de pretratamiento y la creación de modelos de regresión por mínimos cuadrados parciales (PLSR).

Se emplearon 1000 muestras de suelo de Nicoya, proporcionadas por el Departamento de Estudios Básicos de Tierras (DEBT) del INTA, que incluyen diversos tipos de suelo (Alfisoles, Inceptisoles, Entisoles, Vertisoles, Ultisoles) y diferentes usos, generando variabilidad en la concentración de nutrientes. Los promedios obtenidos fueron: Ca 20.74 cmol/L, K 0.30 cmol/L, P 5.66 mg/L, Mg 6.16 cmol/L y materia orgánica (MO) 5.35 %, con diferencias notables entre distritos y usos. Las distribuciones de Ca, Mg y MO fueron simétricas, lo que favoreció el modelado, mientras que K y P presentaron sesgos muy marcados hacia valores bajos y colas largas con valores atípicos, dificultando su predicción.

Las firmas espectrales MIR se obtuvieron en el Centro de Investigación en Ciencia e Ingeniería de Materiales (CICIMA) usando un espectrómetro FTIR PerkinElmer, con un promedio cada 20 mediciones para reducir ruido y mejorar la relación señal-ruido. Un análisis PCA reveló que el primer componente explicaba el 93 % de la variabilidad, lo que subraya la necesidad de una señal limpia para captar diferencias sutiles. Se probaron pretratamientos como Detrend, Normalización Estándar de Variables (SNV) y filtro Savitzky-Golay (SG) en primera y segunda derivada, solos y combinados. La combinación SNV + SG fue especialmente efectiva para normalizar y mejorar la resolución espectral, beneficiando la calibración PLSR. Se verificó la homogeneidad entre conjuntos de entrenamiento y validación mediante PCA.

Los modelos PLSR mostraron distintos niveles de desempeño. Para Calcio, el mejor modelo (SNV + SG 2ª derivada) alcanzó un RMSE de 6.03, r^2 de 0.66 y RPD de 1.71, considerado aceptable, apoyado por su distribución simétrica. Para Magnesio, el mejor resultado fue RMSE 2.25, r^2 0.56 y RPD 1.5, también aceptable. La Materia Orgánica mostró potencial pero con menor precisión (RMSE 1.89, r^2 0.32 y RPD < 1.4). En contraste, Potasio y Fósforo tuvieron un

desempeño pobre (r^2 cercanos a 0.05 y -0.03, RPD < 1.4), principalmente por sus distribuciones sesgadas y la influencia de valores extremos, dificultando que el PLSR extraiga correlaciones fiables. Esto evidencia que la calidad y características de los datos de referencia de laboratorio son clave para el éxito del modelado.

En conclusión, aunque el Calcio presentó mejores métricas, la espectroscopía MIR combinada con PLSR todavía tiene limitaciones, ofreciendo una precisión moderada que restringe su uso en aplicaciones que requieren alta exactitud. Para mejorar los resultados, se recomienda explorar técnicas avanzadas de aprendizaje automático que capturen relaciones no lineales y complejas mejor que PLSR, ampliar la cobertura y cantidad de muestras, y fomentar colaboraciones interdisciplinarias. Además, integrar MIR con tecnologías como teledetección y SIG podría potenciar su aplicación en agricultura de precisión.

Abstract

This study focused on developing predictive models based on mid-infrared (MIR) spectroscopy to quantify soil nutrients in the Nicoya canton, Guanacaste, as an alternative to conventional laboratory analysis. A descriptive statistical analysis of nutrients, exploration of their spectral signatures, evaluation of preprocessing methods, and creation of partial least squares regression (PLSR) models were carried out.

A total of 1000 soil samples from Nicoya, provided by the Basic Soil Studies Department (DEBT) of INTA, were used. These samples represent various soil types (Alfisols, Inceptisols, Entisols, Vertisols, Ultisols) and land uses, which generated variability in nutrient concentrations. The average values obtained were: Ca 20.74 cmol/L, K 0.30 cmol/L, P 5.66 mg/L, Mg 6.16 cmol/L, and organic matter (OM) 5.35 %, with notable differences between districts and land uses. The distributions of Ca, Mg, and OM were symmetric, favoring modeling, while K and P showed strong skewness towards low values and long tails with outliers, making prediction more difficult.

MIR spectral signatures were obtained at the Center for Research in Science and Engineering of Materials (CICIMA) using a PerkinElmer FTIR spectrometer, averaging every 20 measurements to reduce noise and improve signal-to-noise ratio. Principal component analysis (PCA) revealed that the first component explained 93 % of variability, highlighting the need for clean signals to capture subtle differences. Preprocessing methods such as Detrend, Standard Normal Variate (SNV), and Savitzky-Golay (SG) filtering on first and second derivatives, alone and combined, were tested. The combination of SNV + SG was particularly effective in normalizing and improving spectral resolution, benefiting PLSR calibration. Homogeneity between training and validation sets was verified by PCA.

PLSR models showed varying performance levels. For Calcium, the best model (SNV + SG second derivative) achieved an RMSE of 6.03, r^2 of 0.66, and RPD of 1.71, considered acceptable due to its symmetric distribution. For Magnesium, the best result was RMSE 2.25, r^2 0.56, and RPD 1.5, also acceptable. Organic Matter showed potential but with lower accuracy (RMSE 1.89, r^2 0.32, and RPD < 1.4). In contrast, Potassium and Phosphorus performed poorly (r^2 near

0.05 and -0.03, RPD < 1.4), mainly due to skewed distributions and outlier influence, which limited PLSR's ability to find reliable correlations. This highlights the importance of the quality and characteristics of laboratory reference data for successful modeling.

In conclusion, although Calcium showed better metrics, MIR spectroscopy combined with PLSR still has limitations, offering moderate accuracy that restricts its use in applications requiring high precision. To improve results, it is recommended to explore advanced machine learning techniques that can capture nonlinear and complex relationships better than PLSR, increase sample size and coverage, and encourage interdisciplinary collaborations. Additionally, integrating MIR with technologies like remote sensing and GIS could enhance its application in precision agriculture.

Capítulo 1

1. Introducción

La fertilidad del suelo es uno de los factores más importantes que afectan la productividad agrícola. Conocer los niveles de nutrientes como potasio, fósforo, calcio, magnesio y materia orgánica es fundamental para tomar decisiones de manejo de cultivos y optimizar la producción agrícola (Martínez y Gutierrez, 2021).

El análisis de laboratorio tradicional para identificar estos nutrientes, aunque preciso, es costoso y requiere una considerable inversión de tiempo y recursos. Por ello, se plantea la necesidad de implementar alternativas que no solo sean más accesibles en términos económicos y logísticos, sino que también mantengan altos estándares de exactitud, precisión y reproducibilidad en los resultados. (Bonilla Segovia, Dávila Rojas, y Villa Quishpe, 2021).

La espectroscopia de infrarrojo medio es una técnica analítica que ha demostrado ser eficaz para evaluar el contenido de nutrientes del suelo. Esta tecnología permite analizar múltiples parámetros de manera rápida y a menor costo. No obstante, su aplicabilidad práctica depende de la calidad de los modelos predictivos desarrollados y de su capacidad para ofrecer resultados confiables, comparables con los métodos tradicionales (Nath, Laik, Meena, Pramanick, y Singh, 2021).

Para lograr modelos verdaderamente precisos, es indispensable contar con bases de datos locales robustas y bien caracterizadas que sirvan como referencia para la calibración. La calidad, representatividad y tamaño de estos datos de referencia impactan directamente en la exactitud y reproducibilidad de los modelos generados, además de la selección del modelo adecuado. (Nath et al., 2021).

En este sentido, se justifica el desarrollo de modelos predictivos específicos para las condiciones edafoclimáticas del cantón de Nicoya. Guanacaste fue seleccionado como sitio estratégico debido a que la sinergia entre el EIB de la UCR y el INTA/DEBT permitió aprovechar datos de referencia ya existentes en esta región. Cabe aclarar que la elección del cantón no obedece a una necesidad agronómica superior respecto a otras zonas del país, sino a consideraciones prácticas

relacionadas con el acceso a información confiable y validada.

Tal como se indica en el presente trabajo, los métodos tradicionales de análisis de suelos son costosos, requieren mucho tiempo y recursos, lo que refuerza la necesidad de desarrollar alternativas más económicas y rápidas. Por ello, estos modelos estarán orientados a la detección precisa, exacta y reproducible de nutrientes clave mediante espectroscopía de infrarrojo medio, garantizando así una herramienta útil y confiable para los agricultores de la región. De esta manera, se provee una alternativa que no solo es eficiente y rentable, sino también científicamente válida para mejorar la gestión de la fertilidad del suelo.

Este trabajo de investigación está alineado con los Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas, especialmente con el Objetivo 2 (Hambre cero) y el Objetivo 15 (Vida de ecosistemas terrestres) (ONU, 2021).

El Objetivo 2 busca acabar con el hambre, lograr la seguridad alimentaria y mejorar la nutrición. Alcanzar este propósito requiere información precisa sobre la calidad del suelo y la disponibilidad de nutrientes, lo cual permite a los productores optimizar sus prácticas agrícolas y aumentar los rendimientos (ONU, 2021).

Por su parte, el Objetivo 15 promueve el uso sostenible de los ecosistemas terrestres, la restauración de suelos degradados y la conservación de los recursos naturales. Una evaluación precisa de los nutrientes del suelo es fundamental para implementar prácticas sostenibles y prevenir el agotamiento de estos recursos (ONU, 2021).

Este estudio tiene como objetivo desarrollar y validar modelos predictivos altamente precisos y reproducibles para la cuantificación de nutrientes clave del suelo en la región del Cantón de Nicoya, Guanacaste mediante espectroscopia de infrarrojo medio. Se espera que estos modelos brinden una alternativa confiable, eficiente y accesible para el manejo de la fertilidad del suelo, y que además sirvan como base para investigaciones futuras a nivel nacional.

1.1. Objetivos

1.1.1. General

Desarrollar modelos predictivos utilizando espectroscopía de infrarrojo medio para la detección y cuantificación de nutrientes en el suelo, con el propósito de ofrecer una alternativa al análisis de laboratorio convencional.

1.1.2. Específicos

- Realizar un análisis detallado de las muestras de suelo procesadas de laboratorio (INTA), con el fin de caracterizar las diferentes variables de interés (Calcio, Magnesio, Potasio, Fósforo y Materia orgánica).
- Efectuar un análisis exploratorio de las firmas espectrales de las muestras de suelo en el rango espectral MIR.
- Evaluar 3 técnicas de pretratamiento de datos y sus combinaciones para mejorar la calidad de los modelos con el fin de reducir la interferencia de señales no relacionadas a los nutrientes.
- Formular modelos predictivos mediante regresión por mínimos cuadrados parciales (PLSR), para predecir la presencia y concentración de nutrientes en el suelo.

Capítulo 2

2. Marco teórico

2.1. Nutrientes en el suelo

El suelo es un importante recurso natural para la producción de alimentos y la sostenibilidad ambiental. Los nutrientes del suelo son esenciales para el crecimiento y desarrollo de las plantas y, por lo tanto, para la productividad agrícola. En general, las plantas necesitan 16 nutrientes esenciales para crecer, los cuales se dividen en macronutrientes y micronutrientes (Martínez y Gutierrez, 2021).

Los macronutrientes son nutrientes que las plantas necesitan en grandes cantidades: nitrógeno (N), fósforo (P), potasio (K), calcio (Ca) y azufre (S). Estos nutrientes son importantes para la producción de clorofila, la formación de proteínas y el desarrollo de raíces (Martínez y Gutierrez, 2021). Por otro lado, los oligoelementos son nutrientes que las plantas necesitan en cantidades muy pequeñas, pero no menos importantes para su crecimiento. Los principales oligoelementos son: hierro (Fe), manganeso (Mn), zinc (Zn), cobre (Cu), boro (B), cloro (Cl), molibdeno (Mo) y níquel (Ni) (Martínez y Gutierrez, 2021).

Es importante que los nutrientes en el suelo se encuentren en cantidades adecuadas para contribuir a un crecimiento saludable de las plantas. Tanto el exceso como la deficiencia de nutrientes pueden afectar negativamente su desarrollo y rendimiento. No obstante, es importante aclarar que, si bien este estudio se enfoca en variables de fertilidad química disponible, la salud del cultivo también depende de otros factores del suelo, como sus propiedades microbiológicas y físicas. Por ello, la medición y el seguimiento de los nutrientes del suelo representan una parte fundamental dentro del manejo integral del cultivo, pero no deben interpretarse como el único componente determinante del estado de salud vegetal.

Se utilizan varias técnicas analíticas para medir los nutrientes del suelo, como la extracción de nutrientes por solución química, el análisis de la composición y más. Los resultados obtenidos por estos métodos permiten determinar la concentración y disponibilidad de nutrientes en el suelo y brindan recomendaciones para una adecuada fertilización (Bonilla Segovia et al., 2021).

2.2. Estudios sobre metodologías de análisis de suelos en Costa Rica

En Costa Rica, el análisis de la fertilidad del suelo ha sido un tema de interés constante debido a la diversidad edafológica del país y a la importancia de la agricultura como actividad productiva. A lo largo de los años, diversas investigaciones han abordado la necesidad de contar con métodos analíticos confiables, accesibles y adecuados a las condiciones locales.

Uno de los trabajos pioneros es el de (Bertsch, 2012), quien publicó un manual técnico para interpretar la fertilidad de los suelos costarricenses, el cual se convirtió en una referencia fundamental para los laboratorios nacionales. Este manual proporciona criterios para evaluar la disponibilidad de nutrientes esenciales (N, P, K, Ca, Mg, micronutrientes) y define rangos críticos según el tipo de suelo y cultivo, basados en estudios de correlación entre datos analíticos y respuesta agronómica.

A nivel metodológico, uno de los debates más relevantes en el país ha sido la elección de soluciones extractoras que permitan una medición adecuada de la disponibilidad de nutrientes. Tradicionalmente, se ha utilizado el método Olsen modificado con bicarbonato para fósforo, y el KCl 1M para la extracción de cationes intercambiables como Ca y Mg. Sin embargo, estudios posteriores como el de (Bertsch, Bejarano, y Corrales, 2005) compararon estos métodos con el Mehlich 3, una solución extractora multielemental. La investigación encontró altas correlaciones estadísticas entre ambos métodos, lo que sugiere que el Mehlich 3 podría ser una alternativa viable, siempre que se generen nuevas curvas de calibración adaptadas al contexto edafológico del país.

Por su parte, (Molina y Cabalceta, 1990) evaluaron diversas soluciones extractoras en Vertisoles y Ultisoles, dos de los órdenes de suelo más comunes en el país. Su estudio mostró que las diferentes soluciones (incluidas Bray I, Mehlich 1 y Olsen) presentan variaciones significativas en la extracción de fósforo, lo que resalta la importancia de seleccionar métodos en función del tipo de suelo y de establecer niveles críticos diferenciados. Esta línea fue ampliada por (Cabalceta y Cordero, 1994), quienes determinaron niveles críticos específicos de fósforo disponible para varios órdenes de suelo, estableciendo umbrales que permiten una mejor interpretación de resultados.

En cuanto a micronutrientes, (Molina y Bornemisza, 2006) realizaron un estudio enfocado en el nivel crítico de zinc (Zn) en diferentes suelos del país, utilizando DTPA como solución extractora. Su trabajo estableció que concentraciones inferiores a 0.8 mg/kg pueden limitar el crecimiento de cultivos sensibles como el maíz. Este tipo de estudios es esencial para interpretar adecuadamente las necesidades de fertilización con micronutrientes.

A nivel institucional, (Corrales, Bertsch, y Bejarano, 2005) elaboraron un informe técnico sobre los laboratorios de análisis de suelos, foliares y aguas en Costa Rica, en el que identificaron tanto fortalezas como desafíos. Entre los principales retos mencionan la falta de estandarización de métodos, diferencias en los niveles críticos empleados, y la necesidad de incorporar programas sistemáticos de control de calidad.

Más recientemente, (Guerrero y Bertsch, 2020) reportaron los resultados del primer ejercicio de intercomparación de la Red Latinoamericana de Laboratorios de Suelos (LATSOLAN), en el cual participaron laboratorios costarricenses. Este estudio consistió en el análisis de muestras ciegas y la comparación de resultados entre laboratorios, con el fin de evaluar su exactitud y consistencia. El informe reveló la necesidad de seguir fortaleciendo la formación técnica del personal y la trazabilidad de los métodos empleados, así como de incorporar tecnologías analíticas emergentes.

2.3. Metodologías de análisis de suelo

2.3.1. Extracción de potasio con Olsen Modificado

El método de Olsen fue desarrollado por Olsen, Cole, Watanabe y Dean en 1954 con el objetivo de estimar el fósforo disponible en suelos, especialmente aquellos con pH neutro a ligeramente alcalino. El fundamento químico del método radica en la capacidad de una solución de bicarbonato de sodio ($NaHCO_3$) a 0.5 mol/L y pH 8.5 para liberar el fósforo adsorbido en las superficies de partículas del suelo, particularmente en formas de fosfato de calcio y fosfato adsorbido sobre óxidos de hierro y aluminio. El bicarbonato actúa como un tampón alcalino que reduce la actividad del ion calcio (Ca^{2+}), promoviendo así la solubilización del fósforo que de otro modo permanecería retenido.

Aunque el propósito original del método era la determinación de fósforo disponible, su mecanismo extractante basado en la modificación del equilibrio químico entre los nutrientes adsorbidos y la solución del suelo también permite movilizar otros elementos, como potasio, calcio y magnesio. En el caso del potasio, éste se encuentra principalmente en forma intercambiable, es decir, débilmente retenido en los sitios de intercambio catiónico de las arcillas y materia orgánica del suelo, o incluso en fracciones no intercambiables atrapadas entre capas de minerales tipo 2:1 como la illita o la vermiculita.

Para mejorar la eficiencia del método en la extracción de nutrientes distintos al fósforo, especialmente el potasio, se han propuesto modificaciones que dan lugar al denominado método de Olsen Modificado. Una de las referencias más citadas sobre esta modificación es la de Hunter (1975), quien exploró nuevas técnicas para análisis rutinario de suelos y plantas en zonas tropicales. Hunter propuso ajustes en la concentración del reactivo, el tiempo de contacto entre la muestra y el extractante, la proporción suelo:solución y las condiciones físicas del procedimiento, como la agitación constante (Hunter, 1975).

Estas modificaciones buscan aumentar la capacidad del reactante para desorber el potasio intercambiable y parte del potasio no intercambiable, lo cual es particularmente importante en suelos donde este nutriente está firmemente retenido o parcialmente fijado. El extractante alcalino sigue siendo el bicarbonato de sodio, pero bajo condiciones que favorecen una extracción más completa del potasio disponible sin recurrir a reactivos ácidos o altamente agresivos que podrían extraer formas no disponibles (Hunter, 1975).

En síntesis, el método de Olsen y su versión modificada se basan en principios de equilibrio químico entre la fase sólida y la fase líquida del suelo. Su aplicación para potasio permite estimar con buena precisión la fracción que puede ser absorbida por las plantas a corto plazo, siempre que se consideren adecuadamente las condiciones fisicoquímicas del suelo y se apliquen las modificaciones adecuadas al método original según el tipo de análisis requerido (Hunter, 1975).

2.3.2. Extracción con KCl 1M de Ca y Mg

La extracción de calcio (Ca^{2}) y magnesio (Mg^{2}) con cloruro de potasio 1 molar (KCl 1M) es una técnica estándar en análisis de suelos para estimar los cationes intercambiables. Este método no

mide los cationes disueltos en la fracción acuosa del suelo (solución del suelo), sino aquellos que están adsorbidos en las superficies de las partículas del suelo, principalmente arcillas y óxidos de hierro y aluminio. Estos cationes están débilmente retenidos en el complejo de intercambio catiónico y pueden ser desplazados por el ión potasio (K), presente en alta concentración en la solución extractora.

La solución de KCl actúa entonces como un desplazante que reemplaza a los cationes Ca^{2} y Mg^{2} del complejo de cambio, permitiendo que pasen a la solución y puedan ser cuantificados posteriormente mediante técnicas como la espectroscopía de absorción atómica (AAS) o espectrometría de emisión óptica con plasma acoplado inductivamente (ICP-OES). Este procedimiento es ampliamente reconocido por su simplicidad, repetibilidad y por ofrecer una buena estimación del reservorio de nutrientes disponibles para las plantas a corto plazo (Thomas, 1982).

2.3.3. Cuantificación Elemental por Espectroscopía de Absorción Atómica

Una vez extraídos los cationes del suelo, como el potasio (K), el calcio (Ca) y el magnesio (Mg), su cuantificación se realiza mediante espectroscopía de absorción atómica (AAS, por sus siglas en inglés). Esta técnica se basa en la capacidad que tienen los átomos libres en estado gaseoso de absorber luz a una longitud de onda específica, correspondiente a la transición electrónica característica de cada elemento.

El procedimiento comienza con la atomización de la muestra líquida en una llama o en un horno de grafito. Una lámpara específica para el elemento a analizar emite luz monocromática, que atraviesa el vapor atómico generado. La intensidad de la luz absorbida es proporcional a la concentración del elemento presente en la muestra, de acuerdo con la ley de Beer-Lambert.

La espectroscopía de absorción atómica es una técnica altamente sensible y precisa, especialmente útil para la determinación de elementos traza en soluciones extractoras obtenidas mediante métodos como Olsen modificado (para K), o KCl 1M (para Ca y Mg). Su aplicabilidad en análisis rutinarios de laboratorio se ha consolidado por su reproducibilidad, bajo límite de detección y compatibilidad con múltiples elementos a través del uso de diferentes lámparas catódicas huecas (Sadzawka et al., 2006).

Esta técnica es preferida frente a métodos colorimétricos cuando se requiere mayor sensibilidad o menor interferencia química. Asimismo, permite establecer curvas de calibración con altos niveles de confiabilidad, fundamentales para el diagnóstico nutricional del suelo y la formulación de recomendaciones agronómicas.

2.3.4. Determinación de Materia Orgánica Oxidable del Suelo (Walkley y Black, 1934)

El método de Walkley y Black es una técnica clásica desarrollada para estimar el contenido de materia orgánica oxidable del suelo. Se basa en la oxidación del carbono orgánico presente en el suelo mediante dicromato de potasio en medio ácido (H_2SO_4), reacción que genera calor suficiente para oxidar parcialmente la materia orgánica. La cantidad de dicromato no reducido se valora posteriormente por titulación con sulfato ferroso, permitiendo calcular el carbono oxidado, y a partir de éste, estimar la materia orgánica total utilizando un factor de corrección (Walkley y Black, 1934).

Este método fue ampliamente utilizado por su relativa simplicidad y bajo requerimiento instrumental. Sin embargo, presenta limitaciones importantes: no todo el carbono orgánico es oxidado durante la reacción, y el factor de corrección (1.33) asumido originalmente puede no ser aplicable a todos los tipos de suelos, especialmente aquellos con altos contenidos de óxidos de hierro o materia orgánica resistente a la oxidación.

Dado su carácter parcial y la variabilidad inherente, este método ha sido progresivamente reemplazado en laboratorios acreditados por métodos más precisos y reproducibles, como la combustión seca según Dumas. Este último, utilizado en equipos como el *Elementar Vario Macro Cube*, implica la combustión completa de la muestra en una atmósfera rica en oxígeno, y la posterior cuantificación del CO_2 generado. Este enfoque permite medir el carbono total con alta exactitud, sin requerir factores de corrección empíricos (Staff, 2022).

A pesar de estar considerado obsoleto en metodologías modernas como el manual del Kellogg Soil Survey Laboratory (KSSL), el método de Walkley y Black sigue siendo referenciado en estudios históricos o comparativos, y proporciona un punto de partida útil para entender la evolución de las técnicas de análisis de materia orgánica en suelos.

2.4. Espectroscopía de infrarrojo medio (MIR)

El uso de la espectroscopía de infrarrojo medio (MIR) es una técnica no destructiva que permite la identificación y cuantificación de los constituyentes químicos del suelo. Este método se basa en el hecho de que los enlaces químicos de los componentes orgánicos e inorgánicos del suelo tienen una frecuencia de oscilación característica en el rango infrarrojo medio (Ji et al., 2016).

El proceso de caracterización espectral del suelo mediante espectroscopía de infrarrojo medio (MIR) consiste en colocar una muestra de suelo en un espectrómetro de infrarrojo medio que emite radiación infrarroja hacia la muestra. Los enlaces químicos de los compuestos presentes absorben parte de esta radiación en longitudes de onda específicas, generando un espectro de absorción que es detectado y registrado por el equipo (Ji et al., 2016).

La absorbancia (A) es una medida clave en la espectroscopía de infrarrojo medio, que se define como la cantidad de radiación absorbida por la muestra en función de la longitud de onda. Se expresa mediante la ley de Beer-Lambert:

$$A = \log\left(\frac{I_0}{I}\right) \tag{1}$$

donde I_0 es la intensidad de la radiación incidente y I la intensidad de la radiación transmitida. Un mayor valor de absorbancia indica que la muestra ha absorbido más radiación en esa longitud de onda específica, lo que puede correlacionarse con la presencia y concentración de ciertos compuestos (Nath et al., 2021).

La respuesta espectral resultante es una curva que representa la absorbancia en función de la longitud de onda. Esta curva permite identificar los grupos funcionales presentes en la muestra, ya que cada enlace químico tiene una absorción característica en el espectro infrarrojo medio. Por ejemplo, los grupos carbonilos ($C = O$) presentan picos de absorción entre 1700 y 1750 cm^{-1}, mientras que los enlaces O-H de los grupos hidroxilo absorben en la región de 3200-3600 cm^{-1}. La calidad de la respuesta espectral depende de la correcta preparación de la muestra, la calidad del espectrómetro utilizado y las condiciones de medida (Nath et al., 2021).

La espectroscopía de infrarrojo medio (MIR) es ampliamente utilizada en la caracterización de

suelos porque permite la identificación y cuantificación de los elementos y compuestos presentes en la muestra. Además, los espectros obtenidos se pueden analizar mediante métodos estadísticos multivariados como la regresión de mínimos cuadrados parciales (PLSR), que establece la correlación entre el espectro de absorción y los valores analíticos de los componentes químicos del suelo (Nath et al., 2021).

de Palaminy, Daher, y Moulherat (2022) resaltan que para obtener una firma espectral precisa en el espectro medio (MIR), la muestra debe ser colocada y presionada con un objeto contundente como un yunque, asegurando así un contacto óptimo con el detector. Además, el lente del equipo debe ser limpiado con etanol después de cada muestra para evitar la contaminación cruzada y garantizar mediciones precisas. También es fundamental que la muestra esté bien homogeneizada y deshidratada, ya que la presencia de humedad puede afectar la absorción en ciertas regiones del espectro y generar interferencias en la interpretación de los resultados.

2.5. Pretratamientos de las firmas espectrales

El preprocesamiento de características espectrales es un paso importante en el procesamiento de datos espectrales. El objetivo principal de este paso es reducir el ruido y mejorar la calidad de las firmas espectrales para representar mejor la información química del suelo.

Tsagkaris et al. (2023) investigó la aplicación de pretratamientos para los datos espectrales, con el fin de poder obtener información sobre el origen botánico de la miel. El estudio se realizó con un equipo de espectroscopía infrarroja por Transformada de Fourier (FTIR) con el equipo de Nicolet iS50 FT-IR de la marca Thermo Fisher Scientific. El autor utilizó dos métodos de pretratamiento: corrección de escala y derivaciones de datos espectrales (corrección de ruido). Para la corrección de escala utilizó el Standard normal variate (SNV) y para la corrección de ruido Savitzky-Golay (SG).

Tsagkaris et al. (2023) sugiere el pretratamiento SG como una herramienta apta para el tratamiento de los datos espectrales MIR. El modelo tuvo una capacidad de predicción del 81 %.

Silalahi, Midi, Arasan, Mustafa, y Caliman (2018) utiliza el Standart Normal Value (SNV) y el Detrent para la generación de un algoritmo robusto de correción de escala, donde meciona que el

SNV puede ser utilizado independientemente sin embargo, el Detrend es comúnmente utilizado en conjunto.

El estudio de Vestergaard et al. (2021) explora la aplicación de la espectroscopía visible y del infrarrojo cercano (VIS-NIR) para la predicción de las propiedades del suelo en Ontario, Canadá. El estudio compara diferentes algoritmos de preprocesamiento (1ª derivada, 2ª derivada, Savitzky-Golay, Gap, SNV y Detrend) y modelado (PLSR, Cubist, RF y ELM) para optimizar la predicción de la materia orgánica del suelo (SOM) y otras propiedades como el pH, la conductividad eléctrica y la textura del suelo. Los resultados indican que la combinación de la 1ª derivada + Gap y el algoritmo Random Forest (RF) proporciona las mejores predicciones para muchas propiedades del suelo. El estudio destaca la importancia de seleccionar algoritmos de preprocesamiento y modelado adecuados para mejorar la precisión de las predicciones de las propiedades del suelo mediante espectroscopía VIS-NIR.

En el contexto del preprocesamiento de los datos espectrales, este juega un papel crucial en la optimización del rendimiento del modelo. El estudio evaluó varios métodos de pretratamiento en combinación con el modelo de regresión de mínimos cuadrados parciales (PLSR). Los resultados indican que el pretratamiento de Savitzky-Golay demostró un buen rendimiento, con valores de R^2 de calibración y validación de 0.77 y 0.78, respectivamente. Sin embargo, la adición de la corrección de la normalización de la variable estándar (SNV) a Savitzky-Golay resultó en una disminución del rendimiento de la validación ($R^2 = 0{,}64$), mientras que la inclusión de la eliminación de la tendencia (*detrending*) redujo aún más el R^2 de validación a 0.52. El pretratamiento con SNV solo mostró un R^2 de calibración de 0.77, pero un R^2 de validación más bajo de 0.59. De manera similar, la combinación de SNV y *detrending* produjo resultados de validación deficientes. En general, Savitzky-Golay mostró la mejor capacidad predictiva con el modelo PLSR (Vestergaard et al., 2021).

2.6. Análisis de componentes principales (PCA)

El análisis de componentes principales (PCA) es una técnica de reducción de dimensionalidad utilizada para estudiar y visualizar grandes conjuntos de datos espectrales de suelos. Este método convierte un conjunto de variables (en este caso los valores de intensidad de los diferentes

picos espectrales) en un nuevo conjunto de variables más pequeñas y manejables llamadas componentes principales (Ditta et al., 2019).

El objetivo principal de PCA es encontrar combinaciones lineales de las variables originales que expliquen el valor máximo posible de la varianza total del conjunto de datos, de modo que las nuevas variables o componentes sean ortogonales entre sí. De esta forma, se pueden identificar patrones y tendencias en los datos que no son evidentes en su forma original (Ditta et al., 2019).

El PCA se utiliza en el análisis de datos espectroscópicos del suelo para reducir la cantidad de variables a considerar al modelar y analizar la química del suelo. Con PCA, es posible identificar variables espectrales o picos que contribuyen significativamente a los cambios químicos del suelo y eliminar aquellos que no brindan información relevante (Shan, Zhao, Wang, Ying, y Peng, 2020).

Además, PCA también permite representar datos gráficamente en un espacio más pequeño (generalmente 2D o 3D) trazando los datos como una función de los primeros dos o tres componentes principales. De esta forma, se pueden identificar patrones y grupos en los datos que pueden ser útiles para explicar la química del suelo (Shan et al., 2020).

2.7. La regresión de mínimos cuadrados parciales (PLSR)

La regresión de mínimos cuadrados parciales (PLSR) es una técnica estadística utilizada para modelar la relación entre dos conjuntos de variables, conocida como modelado de regresión. Cuando se aplica a la espectroscopia de infrarrojo medio (MIR), PLSR tiene como objetivo establecer la relación entre la señal espectral recibida y la concentración de nutrientes en el suelo (Metz, Abdelghafour, Roger, y Lesnoff, 2021).

El proceso de modelado PLSR implica seleccionar un conjunto de muestras de suelo para las cuales se dispone de datos espectrales y de concentración de nutrientes. Estos datos se dividen en dos conjuntos: uno para construir el modelo (conjunto de entrenamiento) y el otro para verificar la precisión del modelo (conjunto de validación) (Cao et al., 2023).

El propósito de construir un modelo PLSR es establecer una relación matemática entre los datos espectrales y los datos de concentración de nutrientes (un modelo por nutriente). Para ello, se

utiliza un algoritmo que tiene como objetivo encontrar una combinación lineal de variables espectrales con el fin de maximizar la covarianza con los datos de concentración de nutrientes. El modelo resultante es una ecuación matemática que relaciona los valores espectrales de la muestra con la concentración de nutrientes (Metz et al., 2021).

Una vez que se construye el modelo PLSR, se utiliza el conjunto de validación para evaluar su precisión en la predicción de las concentraciones de nutrientes. El modelo se ajusta para minimizar la diferencia entre las concentraciones de nutrientes medidas y los valores predichos por el modelo. La precisión del modelo se puede evaluar utilizando estadísticas como el coeficiente de determinación (R2), el error cuadrático medio (MSE) y el desvío residual de predicción (RPD) (Metz et al., 2021).

Perret et al. (2020) realizó un modelo predictivo usando el análisis PLSR con la herramienta MATLAB, la investigación se llevó a cabo con 1375 muestras de suelo de la zona de Guanacaste y Limón.

Este análisis lo realizó tomando en cuenta 29 propiedades, tales como: pH, acidez intercambiables, K, Ca, Mg, P, Fe, Cu, Zn, Mn, Na, Si, B, S, C, N, MO, textura (porcentaje de limo, arcilla, arena), densidad aparente, capacidad de intercambio catiónico (CICe), saturación de bases y demás.

Se tomó en cuenta que esta técnica es la asociación de una reducción de mínimos cuadrados parciales con una regresión lineal multivariada, donde explica la correlación que existe entre los espectros y la propiedad del suelo. Para la generación de este modelo, Perret et al. (2020) se basó en una ecuación que comprende variables como: la estimación de la propiedad, el coeficiente de regresión para cada longitud de onda, reflectancia y la inserción.

Se obtuvieron modelos con un coeficiente de determinación en el orden de 0.8 y un RMSE del 10 %, para las variables: Ca, Mg, Fe, C, N y CICe.

2.8. Desarrollo de modelos de calibración y validación cruzada.

El desarrollo de modelos de calibración y validación cruzada es un proceso importante para construir modelos de regresión para predecir las concentraciones de nutrientes en el suelo a

partir de datos espectrales obtenidos por espectroscopia de infrarrojo medio (MIR).

La calibración se refiere al proceso de construcción de un modelo estadístico que relaciona la firma espectral del suelo con las concentraciones reales de nutrientes. Este proceso utiliza un conjunto de datos de muestra de suelo que incluye medidas espectroscópicas y concentraciones conocidas de nutrientes. El modelo se adapta a estos datos para encontrar una relación matemática entre las características espectrales y las concentraciones de nutrientes. El modelo de medición suele ser una regresión multivariable, como la regresión de mínimos cuadrados parciales (PLSR) (Seema et al., 2022).

Una vez que se ha creado un modelo de calibración, es importante probar su capacidad para predecir las concentraciones de nutrientes en suelos extraños o diferentes. La validación cruzada es una técnica utilizada para evaluar la capacidad predictiva de un modelo. En la validación cruzada, el conjunto de datos de muestra se divide en dos grupos: el grupo de entrenamiento y el grupo de prueba. El modelo se empareja con el grupo de entrenamiento y se utiliza para predecir las concentraciones de nutrientes en el grupo de prueba. El poder predictivo del modelo se evalua comparando las predicciones del modelo con las concentraciones reales de nutrientes en el grupo experimental (Seema et al., 2022).

La validación cruzada se puede realizar de varias maneras, siendo la validación cruzada “k-fold” una de las más utilizadas. En la validación cruzada “k-fold”, un conjunto de datos de muestra se divide en k grupos o ”pliegues”del mismo tamaño. Luego, el modelo se ajusta a datos de veces k-1 y se usa para predecir las concentraciones de nutrientes en las veces restantes. Este proceso se repite k veces para que cada curva se use una vez como grupo de prueba. Los resultados de cada iteración se promedian para dar una medida general de la previsibilidad del modelo (Seema et al., 2022).

2.9. Importancia del desarrollo sostenible en la agricultura y la necesidad de adoptar tecnologías eficientes y sostenibles

La agricultura es una actividad económica importante a nivel mundial ya que proporciona los alimentos y las materias primas necesarias para la vida humana. Sin embargo, también es una

actividad que puede tener un impacto negativo significativo en el medio ambiente si se lleva a cabo de manera no sostenible. Por lo tanto, es importante que la agricultura se desarrolle de manera sostenible para garantizar la seguridad alimentaria y proteger el medio ambiente.

La sostenibilidad en la agricultura significa la introducción de tecnologías y prácticas agrícolas eficientes y sostenibles desde un punto de vista ecológico, económico y social.

Esto significa que las prácticas agrícolas deben poder satisfacer las necesidades del presente sin comprometer la capacidad de las generaciones futuras para satisfacer sus propias necesidades.

Para lograr este objetivo, se necesitan tecnologías y prácticas agrícolas eficientes y sostenibles. Esto puede incluir el uso de prácticas agrícolas de conservación, como la labranza cero y la labranza cero, que reducen la erosión del suelo y retienen la humedad. Esto también puede incluir el uso de métodos de riego eficientes, como el riego por goteo, que reducen la cantidad de agua que necesitan las plantas.

La introducción de tecnologías eficientes y sostenibles también puede implicar el uso de fertilizantes y plaguicidas orgánicos en lugar de productos químicos sintéticos. Esto no solo reduce la contaminación ambiental, sino que también puede mejorar la calidad de los cultivos y la salud del consumidor.

La agricultura sostenible también incluye proteger y conservar la biodiversidad, al tiempo que promueve prácticas agrícolas que mantienen la salud del suelo y reducen la necesidad de productos químicos sintéticos. Además, es importante implementar políticas y programas que alienten la adopción de prácticas agrícolas sostenibles y apoyen a los agricultores para que las adopten.

Capítulo 3

3. Materiales y métodos

3.1. Sitio de estudio

Las 1000 muestras que serán analizadas corresponden al cantón de Nicoya, número 2 de la provincia de Guanacaste, con una extensión territorial de 1333.68 km^2, dividido en 7 distritos y con una población total de 57125 habitantes.

La distribución de las muestras se realizó considerando una representación equitativa dentro del cantón; sin embargo, por razones de privacidad, la información detallada sobre su distribución espacial no está disponible. El distrito más grande es San Antonio, con 339.50 km^2, mientras que el más pequeño es Mansión, con 215.0 km^2, según se puede observar en la figura y en la Tabla 1.

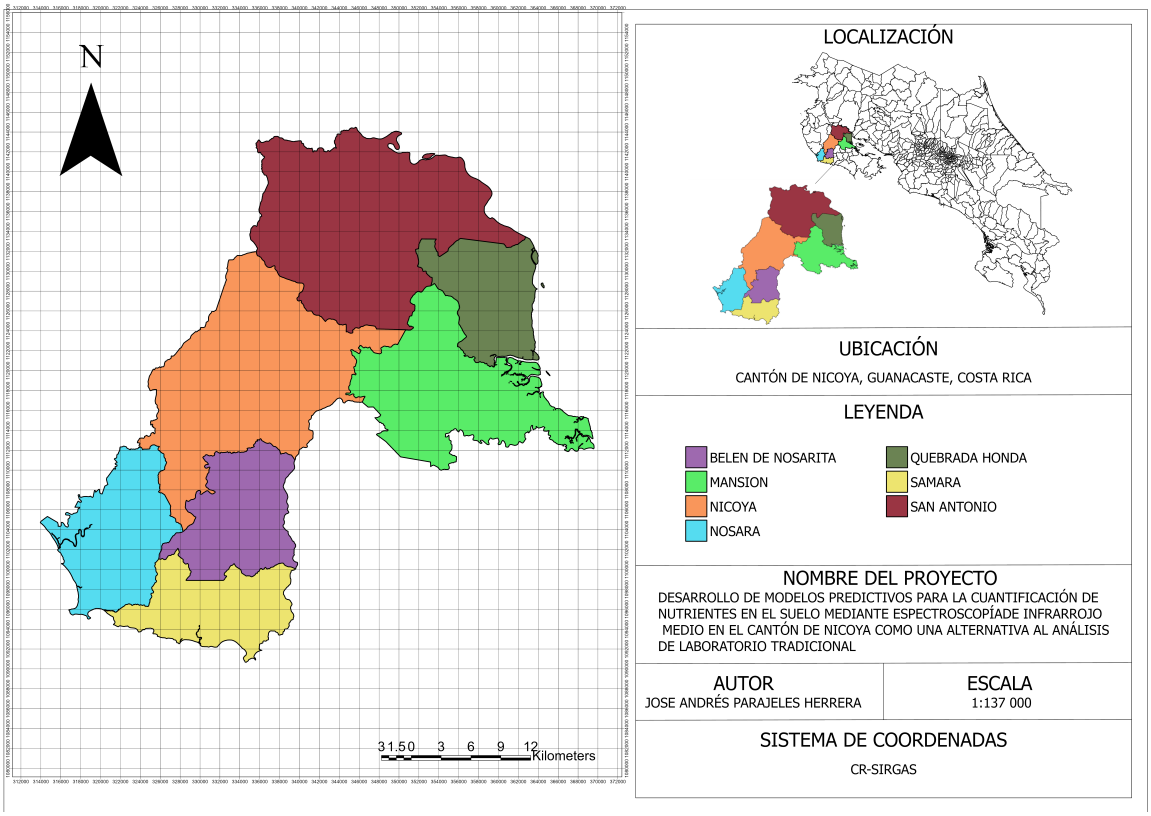


Figura 1: Ubicación de los distritos en estudio. Elaboración propia

Tabla 1: Extensión por distrito.

Distritos	Extensión (km^2)
Mansión	215.00
San Antonio	339.50
Quebrada Honda	109.17
Nicoya	310.66
Nosara	133.65
Belén de Nosarita	123.07
Sámara	109.36

3.2. Análisis estadístico

Se utilizaron 1000 muestras de suelo recopiladas del Departamento de Estudios Básicos de Tierras (DEBT) del Instituto Nacional de Innovación y Transferencia en Tecnología Agropecuaria (INTA), las cuales representan una amplia variedad de condiciones edáficas en el cantón de Nicoya. Estas muestras fueron obtenidas de diferentes localidades dentro de los distritos de Sámara, Nosara, Belén de Nosarita, Quebrada Honda, Mansión, San Antonio y Nicoya, abarcando una extensa área geográfica con diversos tipos de suelo y usos de la tierra en Belén de Nosarita, Quebrada Honda, Mansión, San Antonio y Nicoya.

El primer objetivo del estudio se centró en realizar un análisis estadístico detallado de las muestras de suelo disponibles en el conjunto de datos. Estas muestras estaban compuestas por diversos parámetros, incluyendo un identificador único (ID) y los valores de calcio, potasio, fósforo, magnesio y materia orgánica.

Para establecer una comprensión inicial del estado de los nutrientes del suelo, se llevó a cabo un análisis exhaustivo utilizando herramientas como Microsoft Excel. Este análisis implicó la computación de varios parámetros estadísticos clave, incluyendo promedios y desviaciones estándar, para cada uno de los elementos de interés, como calcio, fósforo, potasio, magnesio y materia orgánica.

Este proceso permitió obtener una visión general de las tendencias y variabilidades presentes en los datos de las muestras de suelo. Los promedios proporcionaron una medida central representativa de los niveles de nutrientes en el conjunto de datos, mientras que las desviaciones estándar ofrecieron información sobre la dispersión de los valores alrededor de esos promedios.

Análisis Descriptivo de las Distribuciones de Nutrientes (Histogramas):

Para cada uno de los cinco nutrientes analizados (CA, K, P, MG, MO), se generaron histogramas de frecuencia. El propósito de este análisis fue visualizar la distribución univariada de cada variable de respuesta, permitiendo la identificación de características clave como la forma de la distribución (simetría o asimetría), la presencia de sesgos (a la derecha o a la izquierda), la identificación de la moda (valor o rango de valores más frecuentes), la extensión del rango de valores y la presencia de posibles valores atípicos (outliers). Esta evaluación inicial fue fundamental para comprender la naturaleza de la variabilidad de cada nutriente y anticipar posibles desafíos en el modelado predictivo posterior. Los histogramas permitieron una inspección visual de la concentración de datos en ciertos rangos de valores, así como la dispersión general de las mediciones, proporcionando una base para entender el comportamiento individual de cada nutriente en el conjunto de muestras

Análisis de Componentes Principales (PCA):

Se realizó un Análisis de Componentes Principales (PCA por sus siglas en inglés) sobre el conjunto de datos multivariado que incluye las concentraciones de los cinco nutrientes. El PCA es una técnica de reducción de dimensionalidad que transforma un conjunto de variables correlacionadas en un nuevo conjunto de variables no correlacionadas, denominadas componentes principales (PC); estas componentes son combinaciones lineales de las variables originales y se construyen de tal manera que la primera componente principal (PC1) explica la mayor cantidad posible de varianza total de los datos, la segunda componente principal (PC2) explica la mayor varianza restante, y así sucesivamente.

El objetivo del PCA en este estudio fue, en primer lugar, visualizar la estructura de variabilidad global, donde el gráfico de dispersión de las dos primeras componentes principales (PC1 y PC2) permitió una representación bidimensional de la disposición de las muestras en el espacio

multivariado, revelando agrupaciones o patrones de dispersión. En segundo lugar, se buscó identificar las principales fuentes de variabilidad; para ello, los vectores de carga asociados a cada nutriente, proyectados en el mismo gráfico de las componentes principales, indicaron la dirección y la magnitud de la contribución de cada variable original a las componentes principales, siendo la longitud del vector proporcional a la varianza de la variable explicada por el plano de las componentes principales y su dirección indicativa de la correlación de la variable con cada componente, un análisis crucial para entender qué nutrientes impulsan las mayores diferencias entre las muestras.

Finalmente, el PCA también tuvo como objetivo explorar relaciones entre nutrientes y con las componentes, donde la proximidad y la dirección de los vectores de los nutrientes permitieron inferir posibles correlaciones entre ellos y con las principales dimensiones de variabilidad del sistema

3.3. Obtención de la firma espectral

Desde la Edafoteca del DEBT del INTA, se cuenta con un sistema de identificación de muestras, que incluye letras y números, las letras son referentes al lugar de donde proviene. Esta identificación sirvió como una guía para poder relacionar la muestra física de suelo con la información del análisis de laboratorio realizado por la institución.

Las muestras utilizadas estaban acondicionadas mediante secado al horno a 105 grados celsius y tamizadas siguiendo los procedimientos del LSF del INTA. Las muestras se tamizaron a un tamaño de partícula de 75 µm, siguiendo las recomendaciones establecidas en la literatura científica. El tamizado garantizó que las partículas de suelo fueran lo más uniformes posible, lo cual es crucial para evitar sesgos en las mediciones espectrales y asegurar la calidad de los datos obtenidos (FAO, 2022).

La obtención de los datos espectrales se llevó a cabo en el Centro de Investigación en Ciencia e Ingeniería de Materiales (CICIMA), ubicado en la Universidad de Costa Rica (UCR). CICIMA es un centro de referencia en nanotecnología y se especializa en el estudio microscópico de las propiedades físicas y químicas de los materiales. La infraestructura avanzada del CICIMA y su equipo altamente capacitado proporcionaron un entorno propicio para la recolección de los datos. El uso de estas instalaciones garantizó que el análisis se realizara con los más altos estándares de precisión y fiabilidad.

Para la obtención de la firma espectral, se utilizó un espectrómetro de infrarrojo por transformada de Fourier (FTIR) PerkinElmer modelo Frontier, el cual opera en la región del infrarrojo medio (MIR) y cercano (NIR). Este equipo destaca por su versatilidad, precisión y alta sensibilidad, convirtiéndolo en una herramienta esencial en el análisis químico de materiales. Este está diseñado para ofrecer resultados confiables y reproducibles, permitiendo la identificación precisa de compuestos orgánicos e inorgánicos, siendo ideal para aplicaciones en ciencia de materiales, farmacéutica, polímeros y odontología, entre otros.

Una de sus principales fortalezas es su amplio rango espectral, que permite adaptarse a distintos tipos de muestras y requerimientos analíticos. Su resolución de hasta $0{,}1\,\mathrm{cm}^{-1}$ asegura una detección precisa de bandas de absorción, lo que posibilita la caracterización detallada de es-

tructuras moleculares. Su tecnología avanzada de interferometría y algoritmos de transformada de Fourier optimizan la calidad del espectro, proporcionando datos de alta resolución y mínimo ruido.

La técnica utilizada fue Reflectancia Total Atenuada (ATR), la cual se presenta como ventajosa porque permite medir directamente la superficie de la muestra sin necesidad de una preparación adicional, lo cual simplifica el proceso y minimiza posibles fuentes de error (FAO, 2022).

La configuración del equipo se estableció para realizar 8 scans a una resolución de $1\,\mathrm{cm}^{-1}$. Esto significa que el espectroscopio, operando en el rango de $4000\,\mathrm{cm}^{-1}$ a $400\,\mathrm{cm}^{-1}$, efectuó 8 mediciones por cada incremento de $1\,\mathrm{cm}^{-1}$ en el espectro. Cada una de estas mediciones, conocidas como scans, se acumulan y se promedian para mejorar la relación señal-ruido y aumentar la precisión del espectro obtenido. En términos prácticos, por cada punto de datos a lo largo del rango espectral, el equipo realizó ocho capturas individuales de información espectral y luego las combinó en un único valor promedio. Este método de escaneo permite obtener un espectro final que es más preciso y fiable, al minimizar el impacto de posibles ruidos y variaciones aleatorias que puedan afectar a mediciones individuales (FAO, 2022).

El proceso completo de obtención de la firma espectral de las 1000 muestras de suelo de la Edafoteca, se extendió durante un período de 3 meses. Este tiempo incluyó no solo la preparación de las muestras, sino también múltiples mediciones repetidas para asegurar la consistencia.

3.4. Procesamiento de la firma espectral

Los archivos con la información espectral fueron exportados en formato TXT. Cada archivo representaba una muestra individual y constaba de dos columnas delimitadas por comas. La primera columna registraba la resolución espectral en cm^{-1}, mientras que la segunda columna contenía los valores de absorbancia medidos para cada punto de resolución.

Para organizar y gestionar eficientemente los datos espectrales, se desarrolló un código en Python que permite crear una matriz de espectros. En esta matriz, se asignaba la primera columna para los identificadores únicos de las muestras (ID), seguida de las resoluciones espectrales en la primera fila y los valores de absorbancia en las filas subsiguientes. Esta estructura de datos

unificada facilitaba la manipulación y el análisis posterior de los espectros.

Durante el proceso de creación de la matriz de espectros, se llevó a cabo una exhaustiva revisión para identificar y depurar posibles muestras que presentaran errores o anomalías en los datos. Se implementaron criterios de calidad estrictos para garantizar la integridad de los datos utilizados en el análisis posterior. Aquellas muestras que no contaran con la cantidad de datos correctos o presentaran anomalías en su estructura, fueron excluidas del conjunto de datos.

Para evaluar la calidad y la integridad de los espectros obtenidos, se realizó un análisis visual detallado mediante la visualización gráfica de los datos espectrales. Se observaron patrones consistentes de ruido en ciertos rangos espectrales, especialmente en las regiones de 700 a 800 cm^{-1} y de 1400 a 1600 cm^{-1}. Estas observaciones proporcionaron información valiosa sobre las áreas problemáticas que requerían atención adicional.

Con el fin de reducir la carga computacional y facilitar el procesamiento de los datos espectrales, se aplicó un promediado de cada 20 mediciones del espectro de absorbancia. Esta estrategia permitió suavizar las curvas y mejorar la homogeneidad de los datos, conservando las características espectrales relevantes. El promediado adecuado de datos espectrales mejora la relación señal-ruido al reducir las fluctuaciones aleatorias que pueden interferir con la señal real. Al promediar varias mediciones, se atenúan los picos de ruido y se obtiene una representación más precisa de la señal, lo que permite un análisis más fiable y estable (Anderson et al., 2023).

3.5. Análisis exploratorio de las firmas espectrales

Para la realización del análisis de componentes principales (PCA) en las firmas espectrales, se utilizaron las siguientes herramientas y paquetes en Python:

1. ***Pandas***: Para la manipulación y el análisis de datos.
2. ***Scikit-learn***: Para aplicar el modelo de PCA, específicamente el subpaquete `sklearn.decomposition` que contiene la clase `PCA`.
3. ***Matplotlib***: Para la visualización gráfica de los resultados.
4. ***mpl_toolkits.mplot3d***: Para la visualización 3D de los resultados del PCA.

Se cargaron los datos espectrales desde un archivo Excel previamente convertido en matriz. Se utilizó la función `read_excel` de Pandas para importar los datos en un *DataFrame*.

Se extrajeron los identificadores de muestra (IDs) y los datos espectrales del *DataFrame*. Los IDs se almacenaron en un array y los datos espectrales en una matriz.

Se configuró el modelo PCA para reducir los datos a diez componentes principales. Sin embargo, al analizar la variabilidad explicada, se observó que los primeros tres componentes principales capturaban la mayor parte de la variabilidad en los datos. Por esta razón, se decidió reducir la dimensionalidad a tres componentes principales para facilitar la visualización y mantener la mayor parte de la información relevante (Davies y Fearn, 2004).

Se generó una visualización en 3D de los resultados del PCA para los tres primeros componentes principales, incorporando colores y etiquetas para identificar las diferentes muestras en el espacio tridimensional, lo que permitió observar posibles outliers entre las muestras de suelo.

3.6. Aplicación de los pretratamientos a las firmas espectrales

Aplicación del Pretratamiento Detrend a las Firmas Espectrales

La aplicación del pretratamiento *Detrend* a las firmas espectrales se realizó como parte del proceso de procesamiento y mejora de los datos espectrales obtenidos inicialmente. Este pretratamiento tiene como objetivo principal eliminar tendencias lineales de los espectros, reduciendo así las variaciones sistemáticas que podrían no estar relacionadas con los fenómenos físicos o químicos de interés (FAO, 2022).

Inicialmente, los datos espectrales se almacenaron en un archivo Excel. Estos datos se organizaron en un *DataFrame* utilizando la biblioteca Pandas de *Python*, donde cada fila representaba una muestra y las columnas correspondían a los valores de absorbancia para diferentes longitudes de onda.

Para aplicar el pretratamiento *Detrend*, se utilizó el código desarrollado en *Python*, aprovechando las capacidades de las bibliotecas Pandas, NumPy y SciPy. El proceso comenzó extrayendo los IDs de las muestras y los espectros espectrales del *DataFrame* original. A continuación, se aplicó la función `detrend` de SciPy para eliminar cualquier tendencia lineal presente en cada espectro, realizando esta operación a lo largo del eje de las columnas (es decir, a lo largo de las longitudes de onda).

Una vez aplicado el pretratamiento *Detrend*, los espectros resultantes se organizaron en un nuevo *DataFrame* utilizando Pandas. Este *DataFrame* incluyó nuevamente los IDs de las muestras junto con los espectros procesados, listos para su análisis posterior.

Para verificar la efectividad del pretratamiento *Detrend*, se graficaron estos con los espectros originales. Esta visualización permitió comparar las formas de onda antes y después del pretratamiento, destacando cualquier mejora en la uniformidad y la eliminación de tendencias no deseadas.

Finalmente, el *DataFrame* que contenía los espectros con el pretratamiento *Detrend* se guardó en un nuevo archivo Excel, asegurando la preservación de los datos pretratados para análisis posteriores.

Aplicación del Pretratamiento SNV a las Firmas Espectrales

La aplicación del pretratamiento *Standard Normal Variate* (SNV) a las firmas espectrales se llevó a cabo como parte del proceso de mejora y normalización de los datos espectrales adquiridos inicialmente. El SNV es una técnica comúnmente utilizada para eliminar efectos sistemáticos no deseados de los espectros, como variaciones en la intensidad debidas a diferencias en la cantidad de muestra o en la geometría de medición (Dotto, Dalmolin, ten Caten, y Grunwald, 2018).

Los datos espectrales originales se obtuvieron de una matriz de Excel previamente conformada con todos los espectros. Utilizando la biblioteca Pandas de *Python*, se leyeron los datos y se organizaron en un *DataFrame* donde cada fila representaba una muestra y las columnas contenían los valores de absorbancia para diferentes longitudes de onda.

Para aplicar el pretratamiento SNV, se desarrolló una función en *Python* que restaba la media de cada espectro y luego dividía por la desviación estándar de cada espectro. Esto se realizó utilizando las funciones de las bibliotecas NumPy y Pandas, asegurando que el proceso fuera eficiente y aplicable a grandes conjuntos de datos.

Una vez aplicado el pretratamiento SNV a todos los espectros, se creó un nuevo DataFrame utilizando Pandas. Este *DataFrame* incluía nuevamente los IDs de las muestras junto con los espectros normalizados mediante SNV, listos para análisis posteriores.

Para verificar la efectividad del pretratamiento SNV, se graficaron los espectros originales y los espectros normalizados SNV para una selección de muestras. Esta visualización permitió comparar las formas de onda antes y después del pretratamiento, evaluando así la reducción de efectos sistemáticos no deseados y la mejora en la consistencia de los datos espectrales.

Finalmente, el *DataFrame* que contenía los espectros normalizados SNV se guardó en un nuevo archivo Excel. Este paso aseguró que los datos pretratados estuvieran disponibles para análisis posteriores y para la generación de informes científicos.

Aplicación del Filtro Savitzky-Golay a las Firmas Espectrales

La aplicación del filtro *Savitzky-Golay* para la primera derivada y segunda derivada a las firmas espectrales se realizó como parte del proceso de pretratamiento para mejorar la resolución de picos espectrales y reducir el ruido inherente a los datos. Este filtro es especialmente útil para suavizar datos y encontrar puntos de inflexión en las curvas de absorbancia (Zimmermann y Kohler, 2013).

Utilizando la biblioteca Pandas de *Python*, se leyeron los datos de la matriz de espectros y se organizaron en un *DataFrame* donde cada fila representaba una muestra y las columnas contenían los valores de absorbancia para diferentes longitudes de onda.

Se utilizó la función `savgol_filter` de la biblioteca SciPy para aplicar el filtro *Savitzky-Golay* a los espectros. Este filtro permite suavizar datos mediante la adaptación local de polinomios de bajo orden (en este caso, un polinomio de segundo orden) a una ventana de datos especificada (11 puntos en este caso).

Al aplicar el filtro *Savitzky-Golay* con el parámetro `deriv=1` o `deriv=2`, se calculó la primera o segunda derivada de los espectros. Esto permite identificar cambios rápidos en los espectros, que son indicativos de puntos de inflexión y características importantes en las muestras espectrales.

Se creó un nuevo *DataFrame* utilizando Pandas que incluía los IDs de las muestras junto con los espectros suavizados y derivados. Este *DataFrame* se preparó para análisis posteriores y para la generación de gráficos que visualicen las características mejoradas de los espectros.

Finalmente, los *DataFrame* que contenía los espectros suavizados y derivados se guardó en un nuevo archivo Excel.

Combinación de Pretratamientos de Datos Espectrales

Se realizaron combinaciones de los pretratamientos con el fin de determinar cuál de ellos proporciona los mejores resultados. No se consideró combinaciones que incluyeran el pretratamiento *detrend*, ya que, según la literatura, este método en combinación con otros pretratamientos da lugar a los peores valores de rendimiento. En total, se realizaron cuatro combinaciones diferentes de pretratamientos.

Primera derivada Savitzky-Golay y SNV

La combinación de la primera derivada Savitzky-Golay (SG) con Normalización Estándar de los Vectores (SNV) se utilizó para suavizar los espectros y corregir tendencias no deseadas, preservando la forma general de las curvas espectrales. SG para la primera derivada es efectivo para eliminar el ruido de baja frecuencia y resaltar características importantes, mientras que SNV ayuda a mejorar la comparabilidad entre espectros al reducir las variaciones debidas a la magnitud de las señales (FAO, 2022).

SNV y Savitzky-Golay primera derivada

La aplicación de SNV antes de SG para la primera derivada, maximiza la comparabilidad entre muestras al suavizar las señales, facilitando así la identificación de patrones espectrales comunes y reduciendo la influencia de factores externos como la intensidad de la señal (Huang et al., 2024).

Savitzky-Golay 2da derivada con SNV

Esta combinación permite mejorar la resolución de picos espectrales y detectar cambios más sutiles en los espectros. SG para la segunda derivada es útil para identificar puntos de inflexión y características espectrales más específicas, mientras que SNV normaliza los espectros, reduciendo la variabilidad no deseada (Huang et al., 2024).

SNV y Savitzky-Golay 2da derivada

Aplicar SNV después de SG para la segunda derivada corrige la amplitud de las señales y mejoró la comparabilidad entre espectros, especialmente útil con muestras que tienen diferentes intensidades de señal inicialmente.

3.7. Generación de modelos PLSR

Carga de datos

Los datos espectrales fueron cargados desde los archivos de Excel con los datos pretratados creando un DataFrame denominado `spectral_df`. Del mismo modo, se cargaron los datos objetivos (niveles de nutrientes) desde otro archivo Excel en un DataFrame llamado `target_df`.

División de datos

Los datos se dividieron en conjuntos de entrenamiento y prueba utilizando la función `train_test_split` de scikit-learn. Se asignó el 80 % de los datos para entrenamiento y el 20 % restante para prueba. Esta división se realizó asegurando la consistencia de las muestras entre los conjuntos utilizando el parámetro `random_state=42`. Utilizar esta función asegura la consistencia y reproducibilidad en la partición de los conjuntos de entrenamiento y prueba durante el desarrollo del modelo. Este parámetro permite controlar la aleatorización aplicada en la división de los datos, asegurando que cada ejecución del código produzca la misma partición de muestras. Esta consistencia es fundamental para la reproducibilidad de resultados en experimentos posteriores, facilitando la comparación justa de diferentes configuraciones de modelos y la evaluación precisa del rendimiento del modelo sin introducir sesgos aleatorios (Cheng y Sun, 2017).

Entrenamiento del modelo PLSR

Para cada nutriente de interés (Ca, K, P, Mg, MO), se implementó el modelo PLSR separadamente:

1. Se inicializa un objeto de Regresión por Mínimos Cuadrados Parciales (PLSR) configurado con un número de componentes (`n_components`) que determina cuántas relaciones subyacentes entre las variables explicativas (espectrales) y la variable objetivo (nutriente) serán capturadas por el modelo. Donde se eligen la cantidad ideal de componentes que minimice el RMSE.

2. El modelo de regresión de mínimos cuadrados parciales (PLSR) se ajusta a los datos de entrenamiento utilizando el método `fit`, con el objetivo de establecer una relación matemática entre los valores objetivo del nutriente y las variables latentes o componentes

del modelo.

3. Se toma un porcentaje de los datos disponibles para la fase de calibración, en la cual el modelo aprende a identificar patrones en los espectros que están asociados a la concentración del nutriente en estudio. El conjunto de datos restante se utiliza para la validación, permitiendo evaluar la capacidad del modelo para predecir valores en muestras desconocidas. La calibración del modelo implica la selección del número óptimo de componentes latentes, buscando minimizar el error de predicción y evitar el sobreajuste. Este procedimiento asegura que el modelo pueda generalizar correctamente a nuevas muestras y realizar predicciones precisas sobre el contenido del nutriente a partir de los espectros adquiridos.

Este proceso se repitió para cada nutriente, almacenando cada modelo en un diccionario (`models`).

Evaluación del modelo

Para cada modelo entrenado, se procedió con la evaluación en el conjunto de prueba:

1. **Selección de Datos de Prueba:** Se seleccionaron los datos relevantes para la evaluación del modelo PLSR. Estos datos (`X_test_nutrient` y `y_test_nutrient`) excluyen el ID de la muestra y la variable objetivo (nutriente). Este paso asegura que el modelo sea evaluado con datos independientes no utilizados durante el entrenamiento.

2. **Predicción con Modelos PLSR:** Utilizando los modelos PLSR previamente entrenados para cada nutriente, se realizaron predicciones sobre los datos de prueba (`X_test_nutrient`). El método `predict` de scikit-learn se empleó para calcular las estimaciones de los valores de los nutrientes basadas en las características espectrales de las muestras.

3. **Cálculo de Métricas de Evaluación:** Se calcularon métricas estándar para evaluar la calidad de las predicciones del modelo. Entre estas métricas se incluyen el Error Cuadrático Medio de la Raíz (RMSE) y el coeficiente de determinación (r^2). El RMSE proporciona una medida de la discrepancia entre los valores observados y los predichos, mientras que r^2 indica la proporción de la varianza en los datos explicada por el modelo. Adicionalmente, se calculó el Ratio de Desviación de Rendimiento (RPD). El RPD es un indicador práctico para comparar modelos predictivos de diferentes variables y estudios. Se calcula

como la relación entre la desviación estándar (DE) de los valores de referencia medidos en laboratorio (la variable que se intenta predecir) y el RMSE del modelo de predicción (Bellon-Maurel, Fernandez-Ahumada, Palagos, Roger, y McBratney, 2010).

4. **Almacenamiento de Resultados:** Todos los resultados obtenidos, incluyendo las predicciones de los nutrientes y las métricas de evaluación (RMSE y R^2), se organizaron y almacenaron en un diccionario (`results`). Esta estructura de datos facilita la revisión y análisis posterior de los resultados para cada nutriente evaluado.

Análisis de resultados

Se analizaron los resultados obtenidos para cada nutriente, evaluando la precisión de las predicciones mediante las métricas calculadas. Se utilizó visualización gráfica para comparar los valores reales y los predichos utilizando gráficos de dispersión con la recta de identidad.

Capítulo 4

4. Resultados y discusión

4.1. Mapa de ordenes de suelo

La presencia de distintos órdenes de suelo, tal como se ilustra en la Figura 2 de (Mata, Rosales, Sandoval, Vindas, y Alemán, 2020), evidencia la diversidad de las condiciones edáficas en los diferentes distritos. Esta diversidad se traduce en variaciones marcadas en las propiedades químicas y en el contenido de materia orgánica.

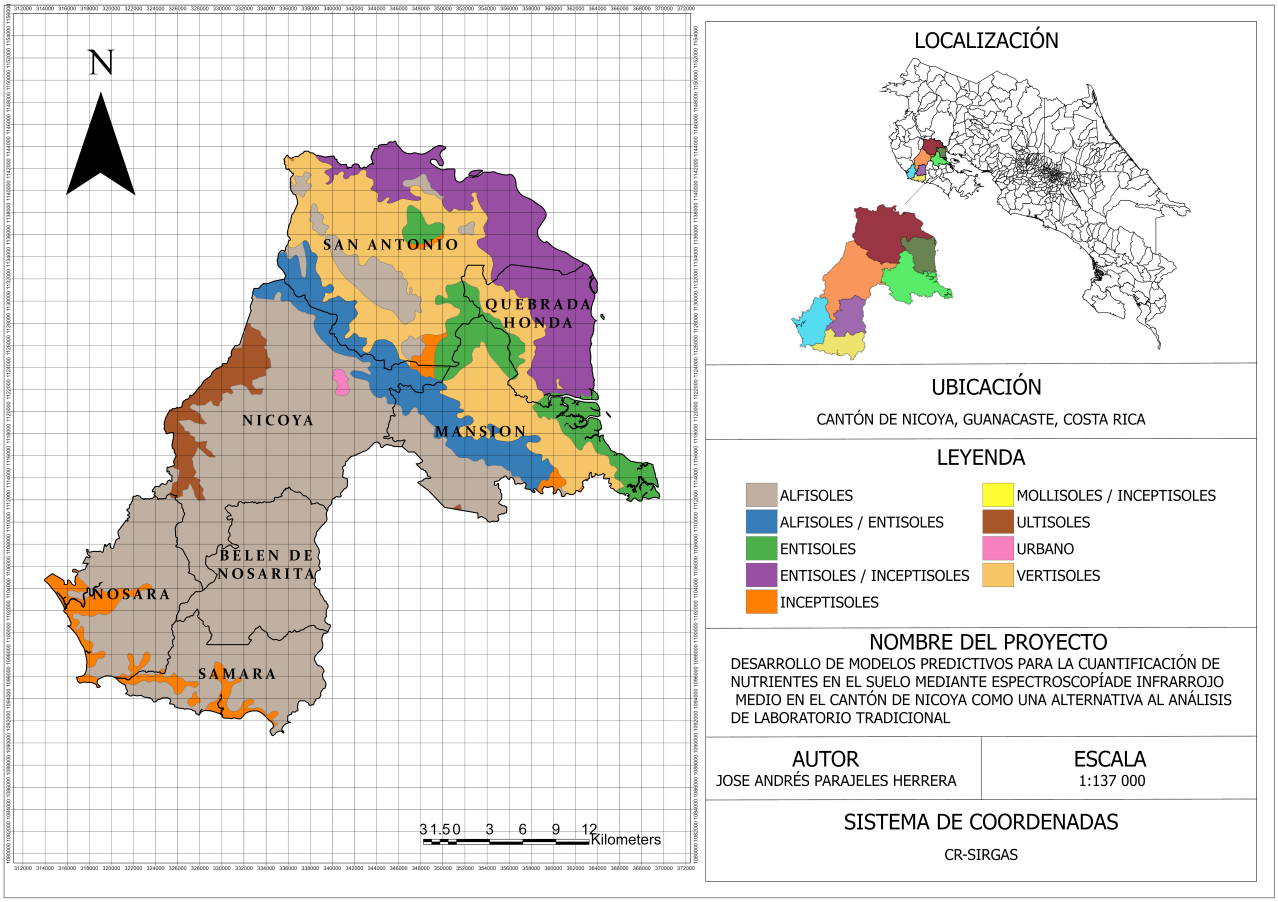


Figura 2: Órdenes de suelo cantón de Nicoya. Fuente: (Mata et al., 2020)

En el cantón de Nicoya, se presenta una notable diversidad de órdenes de suelo, entre los que predominan los Alfisoles, aunque también se encuentran Inceptisoles, Entisoles, Vertisoles y Ultisoles. Esta diversidad edáfica está estrechamente relacionada con la variabilidad en el conte-

nido de nutrientes como calcio (Ca), potasio (K), fósforo (P), magnesio (Mg) y materia orgánica (MO), observada en los diferentes distritos analizados. En particular, distritos del norte como San Antonio, Quebrada Honda y Mansión muestran una marcada diversidad en cuanto a la distribución de tipos de suelo, lo cual sugiere una influencia significativa de factores locales como las prácticas de manejo agrícola y las condiciones ambientales particulares.

Por el contrario, distritos como Sámara, Nicoya, Belén de Nosarita y Nosara presentan una menor diversidad en la clasificación edáfica, aunque, de forma interesante, mantienen una variabilidad similar en los datos de nutrientes. Esta situación indica que, más allá de las características intrínsecas del suelo, factores externos como el manejo, el tipo de cultivo y la historia de uso del suelo desempeñan un papel fundamental en la dinámica de los nutrientes y de la materia orgánica.

Los Alfisoles, predominantes en varias zonas de Guanacaste, son suelos con una fertilidad moderada a alta, especialmente adecuados para cultivos bajo riego y con prácticas sostenibles. En cambio, los Inceptisoles y Entisoles, suelos jóvenes con menor grado de desarrollo, pueden presentar retos en términos de manejo, no tanto por su baja capacidad de retención de nutrientes, sino por su variabilidad en propiedades físicas y químicas. De hecho, algunos Inceptisoles pueden ser ricos en nutrientes debido al material parental del que derivan. Los Vertisoles, con su elevada proporción de arcillas expansivas, presentan desafíos particulares en términos de labranza y drenaje, mientras que los Ultisoles, altamente meteorizados y ácidos, suelen necesitar enmiendas y fertilización para mantener su productividad (Chinchilla, Mata, y Alvarado, 2011).

Estos hallazgos resaltan la complejidad de los sistemas edáficos presentes en la región y la necesidad de incorporar un enfoque integral que considere tanto las propiedades físico-químicas del suelo como los factores de manejo. Solo así es posible comprender a fondo la dinámica de los nutrientes y promover un uso racional y sostenible de los recursos edáficos en los distintos ambientes productivos de Guanacaste.

4.2. Análisis estadístico descriptivo de las muestras de suelo procesadas de laboratorio

En este apartado se presenta el análisis de los datos que resumen la composición química y el contenido de materia orgánica de las muestras de suelo estudiadas. Los resultados se expresan mediante promedios y desviaciones estándar, lo que permite apreciar tanto el valor medio de cada elemento como la dispersión de los datos, indicador de la variabilidad propia de los diferentes entornos.Información que es crucial para comprender las diferencias regionales y el impacto de las prácticas de manejo sobre la fertilidad del suelo.

En primer lugar, la tabla 2 muestra el promedio general de cada elemento, proporcionando un marco de referencia global para la composición de los suelos en el cantón de Nicoya. Se observa que, en promedio, los suelos presentan aproximadamente 20.74 cmol/L de calcio, 0.30 cmol/L de potasio, 5.66 mg/L de fósforo, 6.16 cmol/L de magnesio y un contenido de materia orgánica de 5.35 %.

Tabla 2: Promedio general y desviación estándar de los elementos

Elemento	**Promedio**	**Desviación estándar**	**Unidad**
Calcio (Ca)	20.74	$\pm$ 10.09	cmol/L
Potasio (K)	0.30	$\pm$ 0.30	cmol/L
Fósforo (P)	5.66	$\pm$ 7.04	mg/L
Magnesio (Mg)	6.16	$\pm$ 3.26	cmol/L
Materia orgánica (MO)	5.35	$\pm$ 2.25	%

La tabla 3 segmenta los datos por distrito, revelando diferencias en los promedios y en la dispersión de los valores para cada elemento. Las desviaciones estándar indican la variabilidad de las mediciones dentro de cada región, la cual puede asociarse a diferencias en las condiciones geológicas, climáticas y en las prácticas de manejo local.

Tabla 3: Promedio con la Desviación Estándar por Distrito

Distrito	**Ca (cmol/L)**	**K (cmol/L)**	**P (mg/L)**	**Mg (cmol/L)**	**MO (%)**
Nosara	24.0 ± 4.5	0.30 ± 0.15	3.5 ± 1.8	7.5 ± 2.5	5.5 ± 1.8
Nicoya	20.0 ± 3.8	0.40 ± 0.12	4.5 ± 2.0	8.5 ± 2.2	5.0 ± 1.5
Belén de nosarita	18.0 ± 3.0	0.50 ± 0.20	5.0 ± 2.5	7.0 ± 2.0	6.0 ± 2.0
Sámara	30.0 ± 5.0	0.25 ± 0.10	3.0 ± 1.2	8.0 ± 2.8	4.0 ± 1.5
Mansión	22.0 ± 3.5	0.20 ± 0.08	4.0 ± 1.8	6.5 ± 2.0	5.5 ± 2.0
Quebrada honda	30.0 ± 4.0	0.30 ± 0.10	4.0 ± 1.5	7.5 ± 2.0	4.5 ± 1.3
San Antonio	25.0 ± 4.0	0.40 ± 0.15	5.0 ± 2.0	9.0 ± 2.5	5.0 ± 1.8

La tabla 4 presenta el resumen de los datos agrupados según el uso de suelo. Los resultados muestran que el tipo de manejo influye en la disponibilidad de nutrientes y en el contenido de materia orgánica. Por ejemplo, los Pastos limpios presentan un contenido de calcio y magnesio ligeramente superior en comparación con los suelos clasificados como Plantación forestal, mientras que los suelos en cobertura de Charral muestran una mayor variabilidad en fósforo y materia orgánica. Estos patrones sugieren que las prácticas agronómicas y el tipo de cobertura vegetal afectan significativamente la composición del suelo.

Tabla 4: Promedio $\pm$ Desviación Estándar por Uso de Suelo

Uso de Suelo	**Ca (cmol/L)**	**K (cmol/L)**	**P (mg/L)**	**Mg (cmol/L)**	**MO (%)**
Pastos limpios	23.0 ± 5.0	0.35 ± 0.20	4.0 ± 2.0	8.0 ± 3.0	5.5 ± 2.0
Pastos encharralados	24.0 ± 6.0	0.30 ± 0.15	3.5 ± 2.5	7.5 ± 3.5	5.0 ± 2.5
Bosque secundario	22.0 ± 4.0	0.25 ± 0.10	4.0 ± 2.0	8.0 ± 2.5	6.0 ± 2.0
Plantación forestal	20.0 ± 3.0	0.20 ± 0.10	3.5 ± 1.5	7.0 ± 2.0	5.0 ± 1.5
Charral	19.0 ± 4.0	0.30 ± 0.20	5.0 ± 3.0	7.5 ± 2.5	5.5 ± 2.5

Los datos analizados muestran diferencias notables en la composición del suelo tanto a nivel regional (por distrito) como en función del uso de suelo. Las variaciones en las medias y en las desviaciones estándar resaltan la importancia de los factores ambientales y de manejo, evidenciando que la fertilidad del suelo es modulada por condiciones locales específicas. Esta informa-

ción es esencial para orientar estrategias de manejo y conservación adaptadas a las condiciones particulares de cada área.

La representación gráfica mediante histogramas de las variables edáficas medidas en las muestras de suelo (Calcio, Potasio, Fósforo, Magnesio y Materia Orgánica) permite observar con claridad la distribución de frecuencia de cada nutriente en el conjunto de datos.

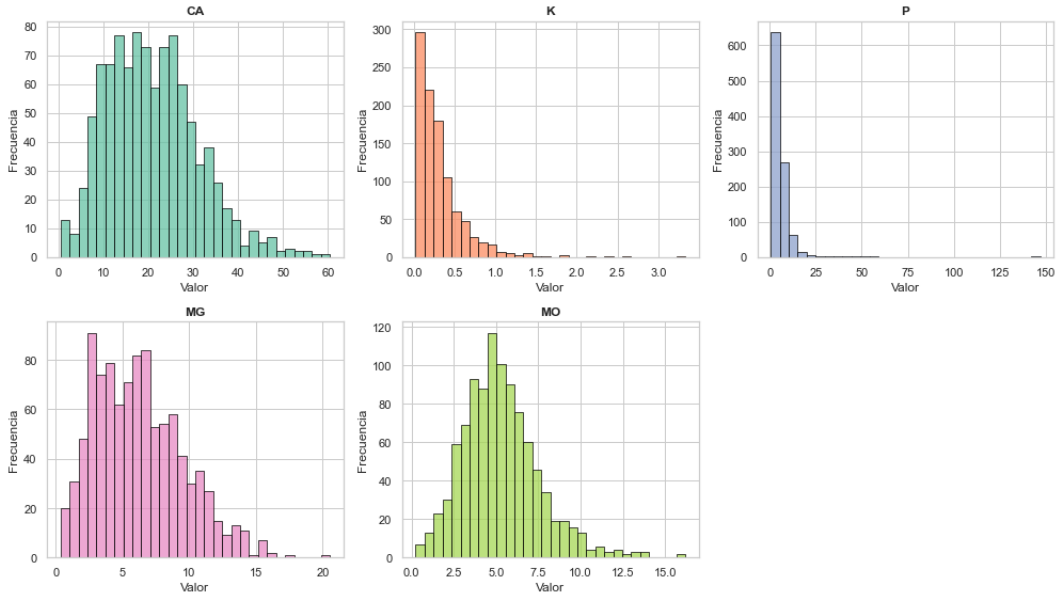


Figura 3: Histograma con la distribución de los datos por nutriente

Al examinar los histogramas, se observa una marcada diferencia en las distribuciones de los nutrientes. La distribución del Calcio es notablemente más simétrica y se asemeja a una campana, aunque con una ligera cola hacia la derecha. Su pico principal (moda) se sitúa aproximadamente entre 15 y 20 unidades, y la dispersión abarca un rango considerable de aproximadamente 0 a 60 unidades. La frecuencia de las observaciones disminuye gradualmente a medida que nos alejamos del centro, lo que sugiere una buena representatividad de valores medios y una variabilidad que el modelo de regresión puede explotar de manera efectiva. Esta distribución relativamente bien comportada es ideal para la mayoría de los métodos de modelado lineal.

La distribución del Potasio es fuertemente sesgada a la izquierda. Una abrumadora mayoría de las observaciones se concentra en valores muy bajos, prácticamente en el origen (0-0.5 unidades). La frecuencia cae drásticamente a medida que los valores aumentan, formando una çola"larga y muy dispersa que se extiende hasta aproximadamente 3.0 unidades. Este patrón

es indicativo de un fenómeno donde la mayor parte de las muestras poseen concentraciones basales o mínimas de K, y solo un número reducido de ellas presenta niveles elevados. Desde una perspectiva de modelado, esta asimetría extrema representa un desafío considerable: el ruido.º la falta de variabilidad significativa en el rango bajo puede enmascarar las verdaderas relaciones con los datos espectrales.

La distribución del Fósforo presenta el sesgo a la izquierda más extremo de todos los nutrientes. Más de 600 observaciones (la mayoría) se agrupan en el primer bin, indicando valores muy cercanos a cero. La cola es incluso más larga y extendida que la de K, alcanzando valores de hasta 150 unidades. Esta distribución hiperconcentrada en el extremo inferior subraya que la mayor parte de los datos contiene muy poca información diferenciadora en términos de concentración, mientras que los valores muy altos son raros y potencialmente actúan como valores atípicos. Esto implica que cualquier modelo que intente capturar la variabilidad de P tendrá que lidiar con la escasez de datos en los rangos medios y altos, y la dominancia estadística de los valores bajos.

La distribución del Magnesio también muestra un sesgo a la derecha, similar al Ca pero un poco más pronunciado. La moda se sitúa alrededor de 2-4 unidades, y la distribución se extiende hasta aproximadamente 20 unidades. Aunque hay una concentración de datos en los valores bajos, la cola no es tan extrema ni tan esparcida como la de K o P. Esto sugiere que, si bien puede requerir algún tratamiento, es una distribución más manejable que las de K y P.

Finalmente, la distribución de la materia orgánica es la más cercana a una distribución normal o simétrica entre todos los nutrientes. Su forma es claramente de campana, con la moda alrededor de 5 unidades y una dispersión que va de 0 a 15 unidades. Esta distribución equilibrada y bien definida, con una buena representación de valores en todo su rango, es la más favorable para el modelado predictivo, ya que el algoritmo tiene una variabilidad consistente y definida para aprender las relaciones.

El gráfico de Análisis de Componentes Principales (PCA) complementa esta visión, proyectando las muestras y la influencia de cada nutriente en un espacio bidimensional (PC1 y PC2). PC1 explica el 34.61 % de la varianza total, y PC2 el 26.46 %, sumando casi el 61 % de la varianza explicada por estas dos primeras dimensiones

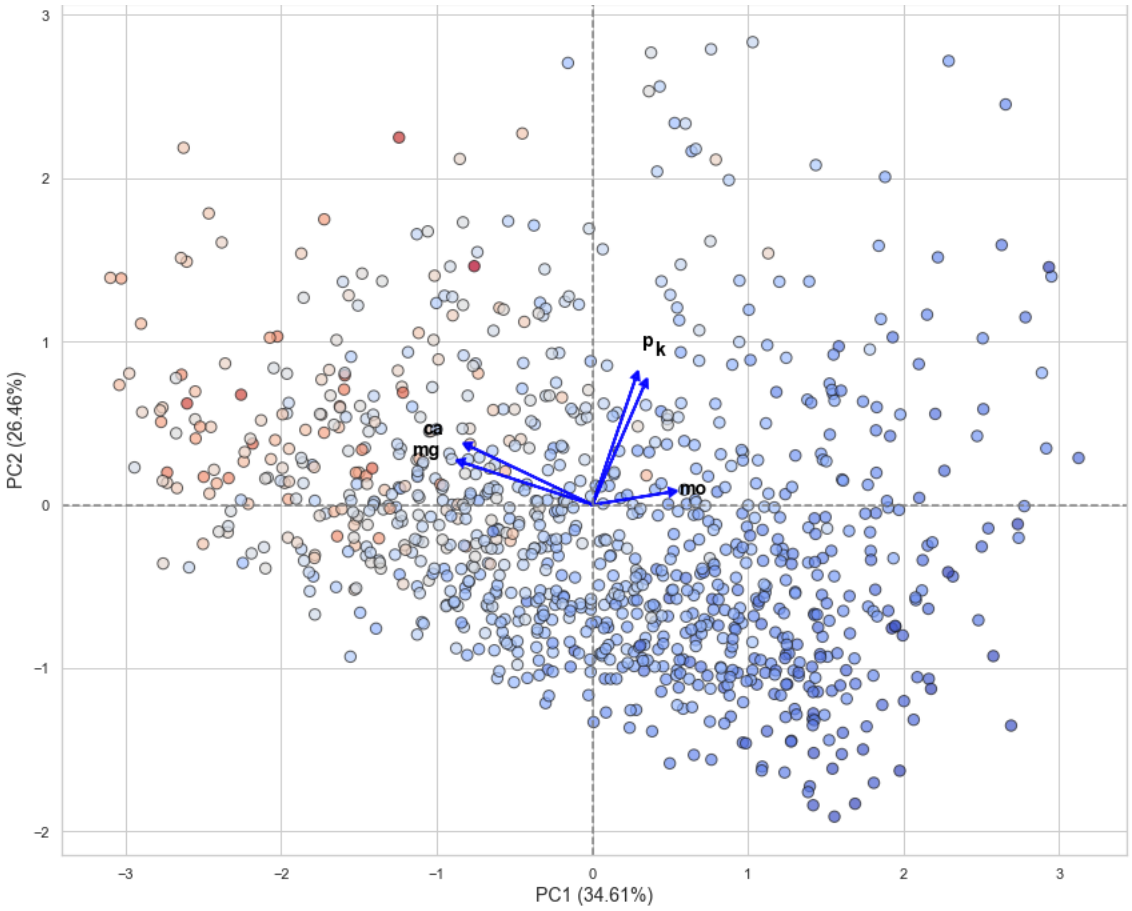


Figura 4: Análisis de componentes principales de los datos de laboratorio

Los puntos de las muestras se distribuyen a lo largo del plano definido por PC1 y PC2. El gradiente de color (de rojo a azul) a través de los puntos es una característica visual clave, observándose que las muestras en la parte izquierda del gráfico (valores bajos de PC1) son predominantemente rojas, mientras que las de la derecha (valores altos de PC1) son azules. Esto sugiere que PC1 captura una variación fundamental en el conjunto de datos que se correlaciona con la variable representada por el color.

En cuanto a la contribución de los nutrientes, los vectores de P y K son los más largos y apuntan con una fuerte inclinación hacia la parte superior derecha (PC1 positiva y PC2 positiva). La

longitud de estos vectores indica que P y K son los nutrientes que más contribuyen a la variabilidad total en el conjunto de datos, es decir, tienen la mayor varianza o dispersión aparente. Las muestras con valores altos de P y K tenderán a ubicarse en esta región del espacio PCA (superior derecha, predominantemente azul).

La alta varianza que se refleja en los vectores largos del PCA para P y K no proviene de una variabilidad homogénea a lo largo de su rango, sino de la gran diferencia entre la concentración masiva de datos en valores bajos y la existencia de unos pocos valores extremadamente altos (la cola del histograma). Estos valores atípicos o extremos inflan la varianza total y, por lo tanto, la longitud de los vectores en el PCA

El vector de MO está fuertemente alineado con el eje PC1 positivo y es relativamente largo. Esto implica que MO es un contribuyente importante a la variabilidad capturada por PC1, y que las muestras con alto contenido de MO se ubicarán hacia la derecha del gráfico. Por último, los vectores de Ca y Mg apuntan hacia la parte izquierda del gráfico (PC1 negativa), sugiriendo que las muestras con altas concentraciones de Ca y Mg se agrupan en la parte izquierda del espacio PCA (predominantemente roja), o que su variabilidad es inversamente correlacionada con la variabilidad principal capturada por PC1.

El gradiente de color de rojo a azul en el PCA, alineado con la dirección de los vectores de P, K y MO, sugiere que PC1 y PC2 capturan una tendencia general donde las muestras con altos niveles de P, K y MO tienden a agruparse en la región azul. Las muestras con valores bajos de estos nutrientes, y posiblemente con valores más altos de Ca y Mg (cuyos vectores apuntan en la dirección opuesta en PC1), se ubican en la región roja. Esta visualización es útil para entender cómo los nutrientes contribuyen a la diferenciación de las muestras, pero no debe confundirse directamente con la facilidad de modelado predictivo si las distribuciones subyacentes son problemáticas para el algoritmo.

La diversidad edáfica del cantón de Nicoya, reflejada en la presencia de Alfisoles, Inceptisoles, Entisoles, Vertisoles y Ultisoles, influye de manera fundamental en la distribución y disponibilidad de los nutrientes. Los boxplots que ilustran la concentración de cada nutriente por orden de suelo (Figuras 5, 6, 7, 8) complementan los análisis descriptivos previos y los histogramas, ofreciendo una visión más detallada de cómo las propiedades inherentes de cada tipo de suelo modulan la química.

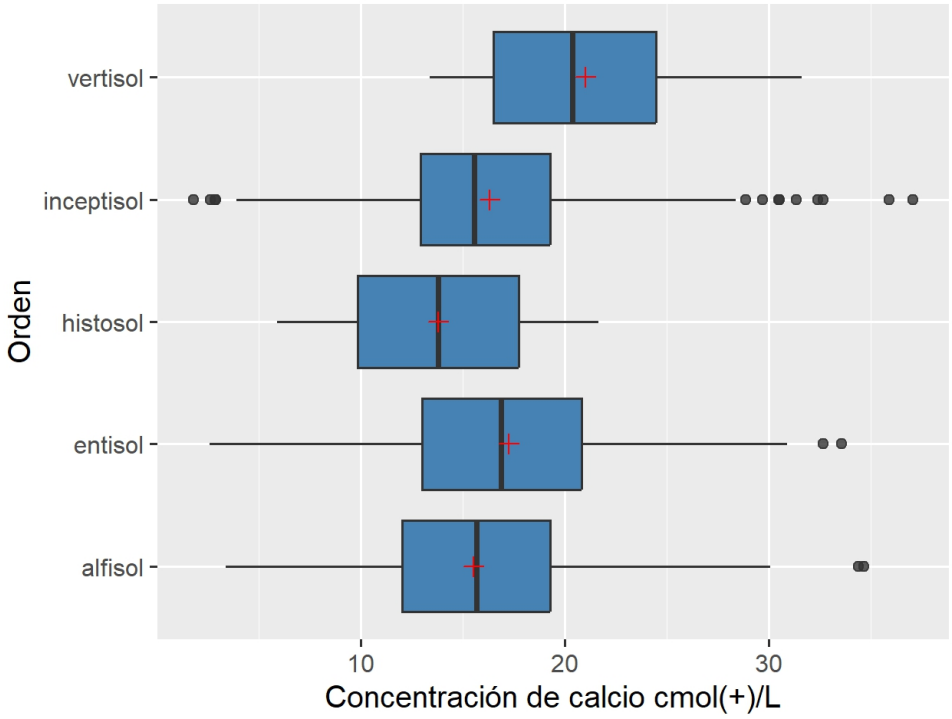


Figura 5: Concentración de Calcio por Orden de Suelo

Como se había observado en el histograma, el Calcio presenta una distribución relativamente simétrica y bien dispersa. Los boxplots confirman esta tendencia, mostrando que los Vertisoles exhiben las concentraciones más elevadas y una menor dispersión intercuartílica, lo que sugiere una alta disponibilidad de Ca y una mayor homogeneidad en este nutriente dentro de este orden de suelo. Esto es consistente con las características de los Vertisoles, que a menudo se desarrollan a partir de materiales parentales ricos en calcio o en ambientes con acumulación de bases. Los Alfisoles también presentan concentraciones considerables de Ca, aunque con una mayor variabilidad (cajas más grandes), lo cual se alinea con su naturaleza de suelos con moderada a

alta saturación de bases. En contraste, los Ultisoles y Entisoles muestran las concentraciones de Calcio más bajas, lo cual es esperable dado que los Ultisoles son suelos fuertemente lixiviados y ácidos, y los Entisoles, al ser suelos jóvenes, a menudo reflejan directamente el bajo contenido de Ca de su material parental. Los Inceptisoles se sitúan en un rango intermedio, con variabilidad considerable. Esta relación entre orden de suelo y concentración de Ca es vital para entender la fertilidad intrínseca de cada tipo edáfico en la región (Chinchilla et al., 2011).

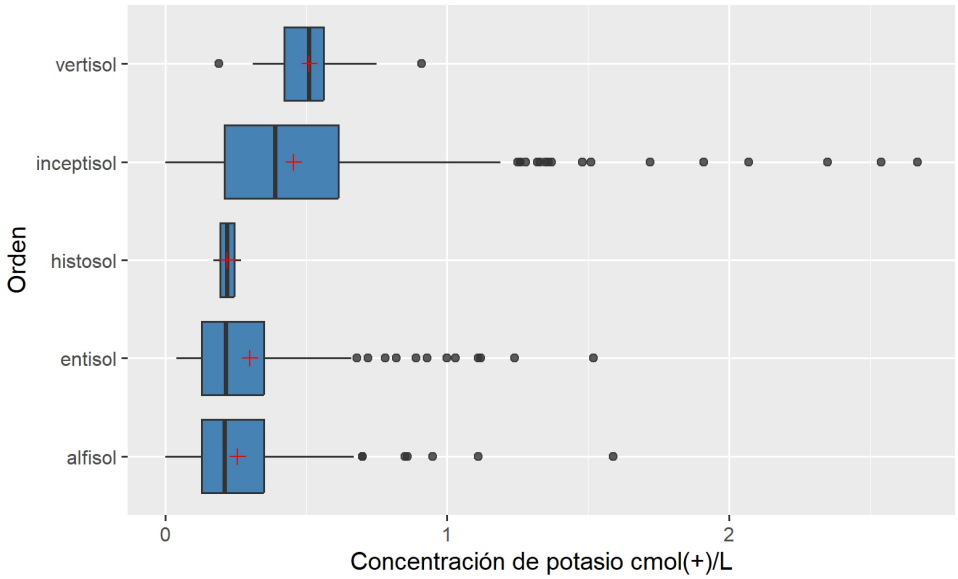


Figura 6: Concentración de Potasio por Orden de Suelo

La marcada asimetría hacia valores bajos, previamente evidenciada en el histograma de Potasio, se reitera en los boxplots. La mayoría de los órdenes de suelo, incluidos Alfisoles, Inceptisoles, Entisoles, Vertisoles y Ultisoles, muestran medianas de K muy cercanas a cero o en el rango bajo de concentración. Esto confirma la escasez general de potasio disponible en los suelos de Nicoya. Sin embargo, es notable la presencia de valores atípicos extremos (puntos individuales fuera de los bigotes) en varios órdenes, especialmente en los Alfisoles y en menor medida en los Vertisoles. Estos valores atípicos son los responsables de inflar la desviación estándar y el vector de carga en el PCA, creando una falsa percepción de variabilidad homogénea. La predominancia de cajas muy comprimidas en la parte inferior de la escala, incluso en órdenes teóricamente más fértiles, subraya el desafío en la modelización de este nutriente.

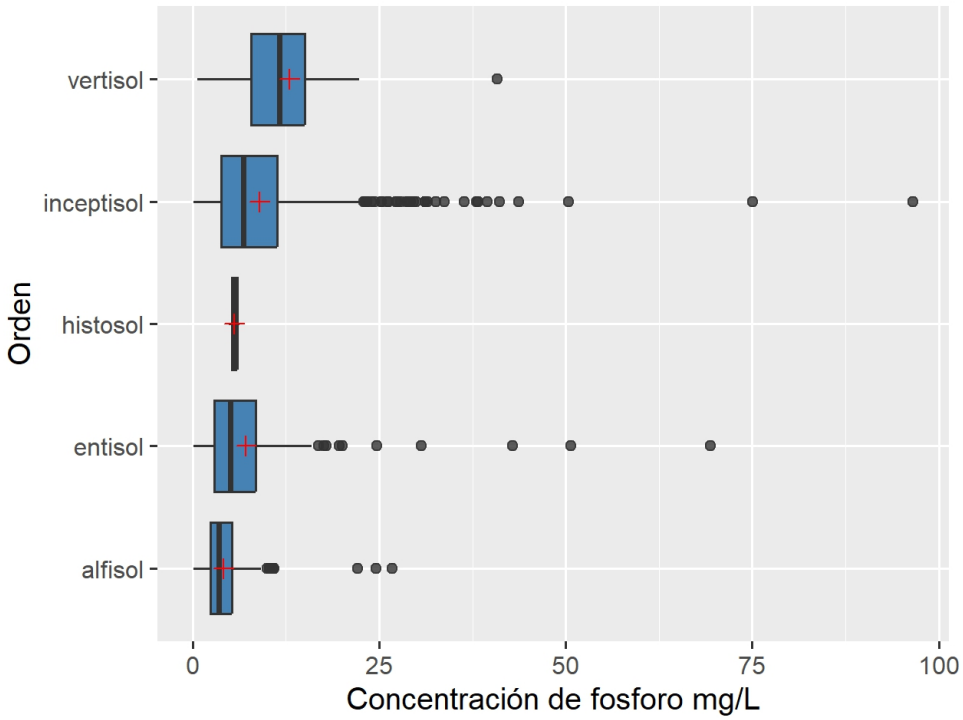


Figura 7: Concentración de Fósforo por Orden de Suelo

El Fósforo presenta la distribución más sesgada y concentrada en valores bajos, una característica ya destacada por su histograma. Los boxplots visualizan claramente este fenómeno: para todos los órdenes de suelo, las medianas de P son extremadamente bajas, y las cajas intercuartílicas están fuertemente comprimidas cerca del eje X. Esta situación es particularmente pronunciada en Alfisoles, Inceptisoles y Entisoles. Similar al potasio, la presencia de numerosos valores atípicos muy altos se observa en varios órdenes, especialmente en los Alfisoles y Vertisoles, lo que distorsiona las métricas de varianza y la percepción de su importancia en el PCA. La baja disponibilidad general de P y la extremada asimetría en su distribución por orden de suelo son factores críticos que explican el bajo rendimiento de los modelos predictivos para este nutriente, ya que la variabilidad útil es casi inexistente para la mayoría de las muestras.

La distribución del Magnesio, que se había clasificado como intermedia en términos de sesgo, encuentra confirmación en los boxplots. Los Vertisoles se destacan por tener las concentraciones más altas de Mg, con una dispersión moderada, lo cual es coherente con su génesis a partir de rocas básicas. Los Alfisoles también muestran niveles elevados de Magnesio y una variabi-

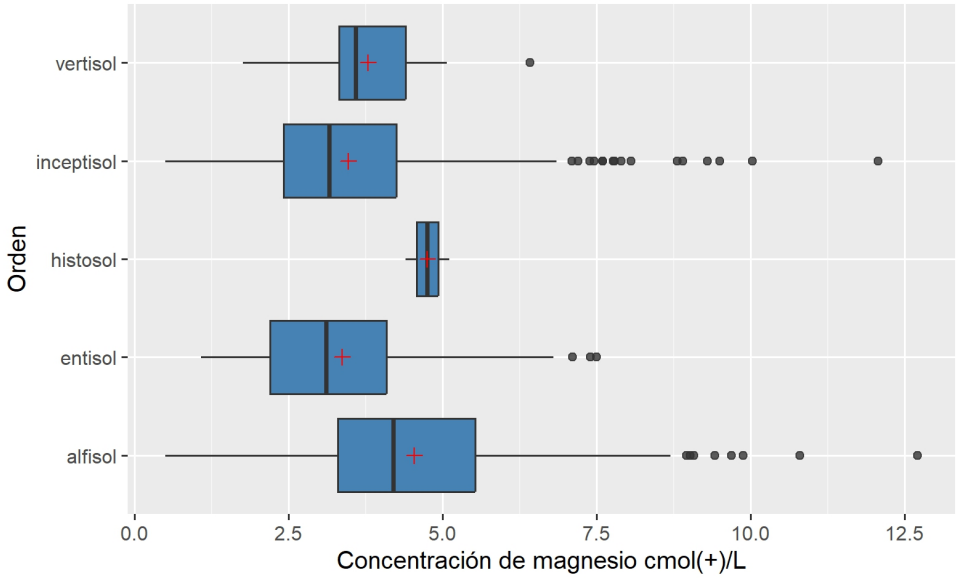


Figura 8: Concentración de Magnesio por Orden de Suelo

lidad considerable, indicando una buena disponibilidad. Los Inceptisoles y Entisoles presentan valores intermedios y una variabilidad comparable. Al igual que con el Calcio, los Ultisoles exhiben las concentraciones de Magnesio más bajas, reflejando su naturaleza de suelos lixiviados (Chinchilla et al., 2011).

La presencia de valores atípicos es menos pronunciada que en K o P, lo que contribuye a que el Magnesio sea un nutriente más manejable para el modelado, aunque no tan ideal como el Calcio o la Materia Orgánica.

4.3. Análisis exploratorio de las firmas espectrales de las muestras

En la figura 9 se pueden observar los datos espectrales graficados con la longitud de onda en el eje x y la absorbancia en el eje y. Algunos de los espectros originales mostraban variabilidad significativa en los rangos de menos de 500 y de 700 a 800 cm^{-1} y de 1400 a 1600 cm^{-1}, indicando la presencia de ruido para algunas muestras.

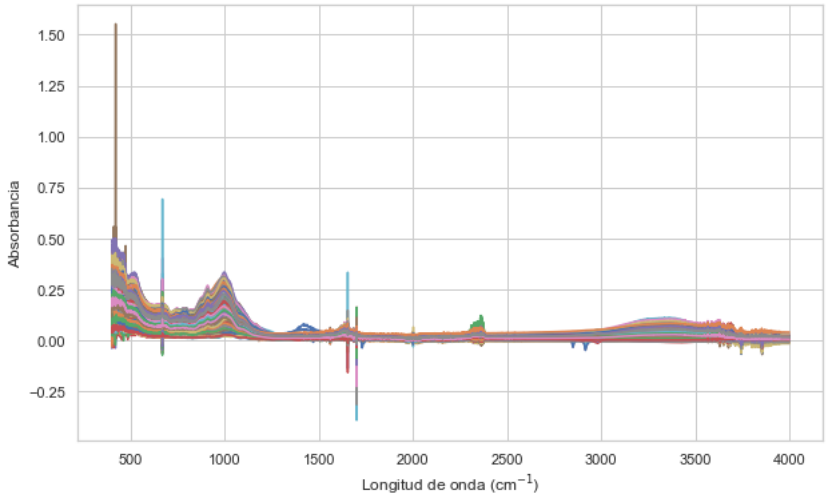


Figura 9: Firmas espectrales sin aplicar ningún procedimiento

Esta variabilidad en los espectros puede ser atribuida a varias razones. En primer lugar, la presencia de interferencias ambientales durante la adquisición de los datos espectrales puede introducir fluctuaciones aleatorias. Factores como cambios en la temperatura, la humedad, y la presencia de partículas en el aire pueden afectar la medición de la absorbancia en ciertos rangos espectrales, especialmente aquellos más sensibles como los mencionados (Shi et al., 2023).

En los rangos de 700 a 800 cm^{-1}, la variabilidad puede estar relacionada con el ruido instrumental. Además, la dispersión de luz, los efectos de difracción, y las imperfecciones en los componentes ópticos del espectrómetro pueden contribuir a las inconsistencias en la absorbancia medida. Es importante señalar que dicho ruido también podría estar asociado a la forma en la que se tomaron las muestras. Una preparación inadecuada o una mala homogeneización del material puede generar espectros poco representativos, por lo que se recomienda repetir o eliminar aquellas muestras que presenten comportamientos anómalos en esta región espectral.

Asimismo, es posible que la resolución espectral empleada haya sido excesivamente alta en

relación con el comportamiento del suelo, el cual responde de forma más generalizada y menos específica que compuestos químicos puros. En este sentido, una resolución muy fina puede introducir ruido innecesario, capturando variaciones que no son relevantes desde el punto de vista analítico y que, en lugar de mejorar la sensibilidad, afectan negativamente la calidad del modelo espectral. Por lo tanto, una resolución intermedia o baja podría ser más adecuada para analizar matrices complejas como el suelo, ya que permite suavizar señales espurias sin perder información significativa.

Por otro lado, en los rangos de 1400 a 1600 cm^{-1}, la variabilidad podría deberse a la superposición de bandas de absorción de diferentes compuestos presentes en las muestras. Este fenómeno puede causar picos y valles en el espectro que no corresponden necesariamente a la concentración real de las variables de interés, sino a la interferencia de otras sustancias presentes. Esta superposición puede ser particularmente problemática en muestras complejas donde múltiples componentes tienen absorbancias en rangos similares (Shi et al., 2023).

La preparación y manipulación de las muestras también pueden introducir variabilidad. Diferencias en la homogeneidad de las muestras, el grosor de la capa de la muestra en el camino óptico, y posibles contaminaciones pueden resultar en espectros inconsistentes. Además, cualquier variación en el alineamiento de la muestra en el espectrómetro entre mediciones puede llevar a diferencias en los datos espectrales obtenidos (Jakkan, Ghare, y Sakode, 2023).

Después de aplicar el procedimiento de promediado usando un total de 20 datos, se observó una reducción notable en el ruido, como se muestra en la figura 10. Los espectros promediados presentan una línea más suave y uniforme, con picos menos pronunciados en las áreas anteriormente afectadas por el ruido. Este suavizado mejora la claridad de los espectros y facilita la identificación de características espectrales significativas, lo que es esencial para un análisis más preciso y fiable.

El ruido en los espectros puede dificultar la interpretación correcta de los datos, especialmente en regiones donde la señal de interés es débil o está enmascarada por variaciones no deseadas. Para suelos, la resolución espectral extremadamente alta no siempre es necesaria porque la información relevante suele estar distribuida en características más amplias y menos definidas (Mammadov, Denk, Mamedov, y Glaesser, 2024).

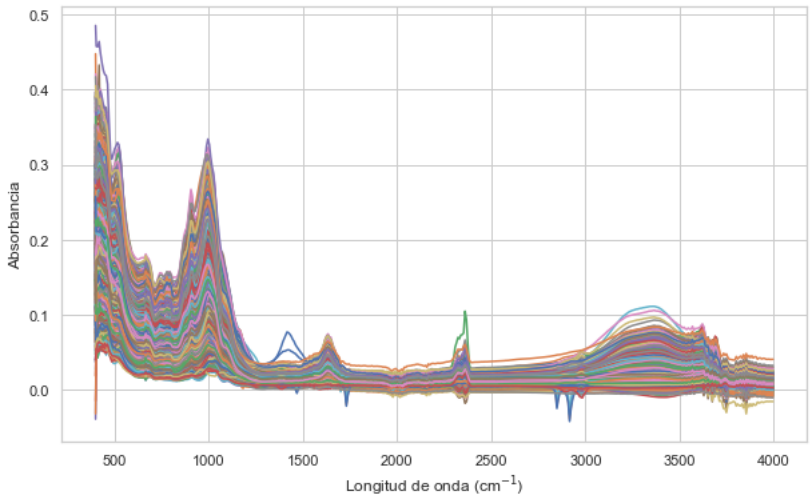


Figura 10: Firmas espectrales promediadas cada 20 mediciones

A diferencia de muestras biológicas o químicas complejas, los suelos presentan una variabilidad espectral que no justifica la resolución de 1 cm^{-1} con múltiples mediciones, ya que esto introduce un ruido considerable. La implementación del promediado de los espectros cada 20 muestras ha demostrado ser una estrategia efectiva para minimizar el impacto del ruido. Este método aprovecha el principio de que el ruido aleatorio tiende a cancelarse cuando se promedian múltiples mediciones, mientras que la señal verdadera se refuerza. Así, el resultado es una representación más precisa de las características espectrales inherentes a las muestras (Mammadov et al., 2024).

La reducción del ruido mediante el promediado no solo mejora la calidad visual de los espectros, sino que también aumenta la aceptabilidad de los análisis cuantitativos y cualitativos posteriores. En el contexto de la Regresión por Mínimos Cuadrados Parciales (PLSR), la calidad de los datos de entrada es fundamental para obtener modelos precisos y fiables. Los datos menos ruidosos permiten que los algoritmos de PLSR identifiquen y capturen más efectivamente las relaciones subyacentes entre las variables espectrales y las variables objetivo, mejorando así el rendimiento predictivo y la interpretabilidad de los modelos (Mammadov et al., 2024).

Al suavizar las fluctuaciones aleatorias y reducir las interferencias no deseadas, los espectros promediados permiten una identificación más clara de los picos de absorción correspondientes a los componentes químicos presentes en las muestras de suelo. Este aspecto es particularmente importante para la construcción de modelos predictivos mediante PLSR, ya que una mayor

claridad en los datos espectrales contribuye a una mejor correlación entre los espectros y las concentraciones de nutrientes en el suelo (Stenberg y Rossel, 2010).

Además, los espectros promediados presentan una mejor base para la construcción de modelos PLSR. En técnicas de análisis multivariado como PLSR, la reducción del ruido es crucial para maximizar la precisión del modelo. La disminución del ruido y la mejora en la claridad de los espectros facilitan un análisis más preciso y fiable, permitiendo que los modelos PLSR identifiquen patrones más claros y consistentes en los datos.

La implementación del promediado también tiene implicaciones prácticas en términos de eficiencia y reproducibilidad. Al reducir la necesidad de realizar múltiples mediciones y minimizar la variabilidad entre mediciones, se mejora la consistencia de los datos espectrales. Esto es especialmente beneficioso en estudios a largo plazo o en análisis comparativos, donde la reproducibilidad de los resultados es crucial (Stenberg y Rossel, 2010).

Debido a que los resultados arrojaron que los primeros tres componentes principales explicaban el 98 % de la variabilidad en los datos, se optó por utilizar solo estos tres componentes principales para el análisis posterior. Esta decisión facilita la visualización y asegura que la mayor parte de la información relevante esté contenida en estos componentes.

Los resultados obtenidos del PCA fueron los siguientes:

- Componente Principal 1: 93 % de la variabilidad explicada.
- Componente Principal 2: 3 % de la variabilidad explicada.
- Componente Principal 3: 2 % de la variabilidad explicada.

Estos resultados se pueden observar representados en la Figura 11, donde se pueden visualizar los tres componentes principales.

El PCA reveló que el primer componente principal (PC1) explica una abrumadora mayoría de la variabilidad en los datos (93 %), mientras que el segundo y tercer componentes principales (PC2 y PC3) explican un 3 % y un 2 % de la variabilidad espectral, respectivamente. Este resultado indica que la mayor parte de la información contenida en los datos espectrales puede ser capturada mediante el primer componente principal.

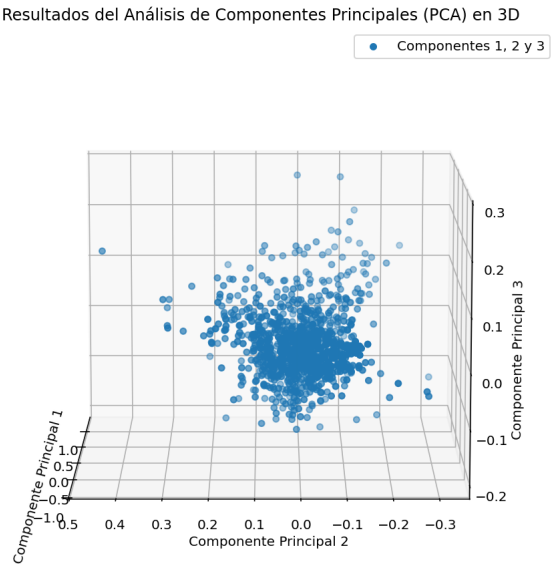


Figura 11: Análisis de componentes principales

Esta fuerte dominancia del primer componente principal (PC1) en el análisis de componentes principales (PCA) refleja una estructura subyacente en los datos espectrales donde una sola dirección en el espacio de características captura la mayoría de la variación significativa observada. En el contexto de la espectroscopía de suelos con infrarrojo medio (MIR), esta característica puede interpretarse como la presencia de una o pocas bandas espectrales predominantes que exhiben una absorción específica y consistente en la mayoría de las muestras de suelo analizadas (Abdel-Fattah et al., 2021).

La alta proporción de variabilidad explicada por PC1 (93 %) sugiere que estas bandas espectrales dominantes pueden estar asociadas con componentes químicos o minerales característicos del suelo, como óxidos metálicos, materia orgánica, arcillas, o minerales secundarios. Estos componentes suelen manifestarse con bandas de absorción bien definidas en ciertos rangos espectrales, lo cual les confiere una prominencia significativa en los datos de espectroscopía de suelos (Abdel-Fattah et al., 2021).

Además, la consistencia en la absorción de estas bandas sugiere una presencia uniforme o generalizada de estos componentes en las muestras de suelo estudiadas, lo que contribuye a la aceptabilidad y la reproducibilidad de las mediciones espectrales. Esta interpretación está res-

paldada por estudios previos en espectroscopía de suelos que han identificado características espectrales dominantes relacionadas con la composición mineralógica y química del suelo.

El hecho de que PC1 capture la mayoría de la variación también indica que estas características espectrales dominantes pueden servir como marcadores útiles para la identificación y cuantificación de componentes específicos del suelo en aplicaciones de modelado predictivo como PLSR. Por lo tanto, la focalización en PC1 no solo simplifica la interpretación de los datos, sino que también potencia la capacidad del modelo para correlacionar eficazmente los espectros con las concentraciones de nutrientes y otros parámetros de interés en el suelo (Zhao y Wan, 2023).

La alta variabilidad explicada por el primer componente principal (PC1) en el análisis de espectroscopía de suelos con infrarrojo medio (MIR) presenta también algunas desventajas significativas. PC1 tiende a capturar la variación más prominente y dominante en los espectros, simplificando la interpretación al identificar características espectrales clave, pero al mismo tiempo puede enmascarar variaciones más sutiles y menos evidentes que también son relevantes. Estas variaciones sutiles podrían incluir diferencias en la composición química a niveles más bajos de concentración o entre muestras con perfiles espectrales similares pero no idénticos. Por lo tanto, PC1 puede no diferenciar de manera precisa entre muestras químicamente diferentes pero con perfiles espectrales superpuestos(Abdel-Fattah et al., 2021).

Además, debido a su dominancia, PC1 puede tener dificultades para detectar y cuantificar compuestos presentes en concentraciones bajas en las muestras de suelo, lo cual es crucial en estudios de contaminación ambiental o evaluaciones de calidad del suelo. Esta limitación puede afectar la capacidad del modelo para identificar trazas de elementos o compuestos que son de interés específico. Asimismo, PC1 puede agrupar incorrectamente muestras que son químicamente similares pero tienen diferencias sutiles en sus perfiles espectrales, lo que compromete la precisión del modelado predictivo y la interpretación de los resultados (Abdel-Fattah et al., 2021).

En términos de su aplicación a la Regresión por Mínimos Cuadrados Parciales (PLSR), los resultados del PCA pueden ser tanto ventajosos como desventajosos. Por un lado, el hecho de que PC1 explique tanta variabilidad sugiere que PLSR podría beneficiarse de una reducción de dimensionalidad similar, centrándose en los componentes principales más importantes y, por lo

tanto, simplificando el modelo. Esta simplificación puede reducir el ruido y mejorar la estabilidad y la interpretabilidad del modelo.

Por otro lado, la concentración de tanta variabilidad en un solo componente puede significar que el modelo PLSR podría perder información importante si no se consideran adecuadamente los componentes adicionales. En otras palabras, mientras que PC1 captura la mayoría de la variabilidad, PC2 y PC3 aún contienen información relevante que podría contribuir a la precisión y la aceptabilidad del modelo predictivo (Abdel-Fattah et al., 2021).

La fuerte concentración de variabilidad en PC1 podría ser un reflejo de la naturaleza de las muestras de suelo y de las características espectrales dominantes en la región MIR. Estudios previos han mostrado que los espectros de suelos a menudo están dominados por bandas de absorción de minerales y materia orgánica que tienen características espectrales muy definidas. Estos estudios sugieren que una resolución muy alta en los datos espectrales puede no ser siempre necesaria, ya que la mayoría de la información relevante puede ser capturada por unas pocas características espectrales clave. En este contexto, promediar los espectros puede ser especialmente útil, ya que reduce el ruido sin perder información crítica, lo que es consistente con nuestros resultados del PCA (Zhao y Wan, 2023).

La distribución de un 98 % de la variabilidad indica que la mayor parte de la información relevante está capturada en el PC1, lo cual reduce el potencial impacto de outliers en la representación general de la estructura de los datos, por lo que para asegurar la integridad y representatividad del conjunto de datos analizados, ninguna de las muestras se consideró como un outlier.

4.4. Evaluación de 3 técnicas de pretratamiento de datos y sus combinaciones

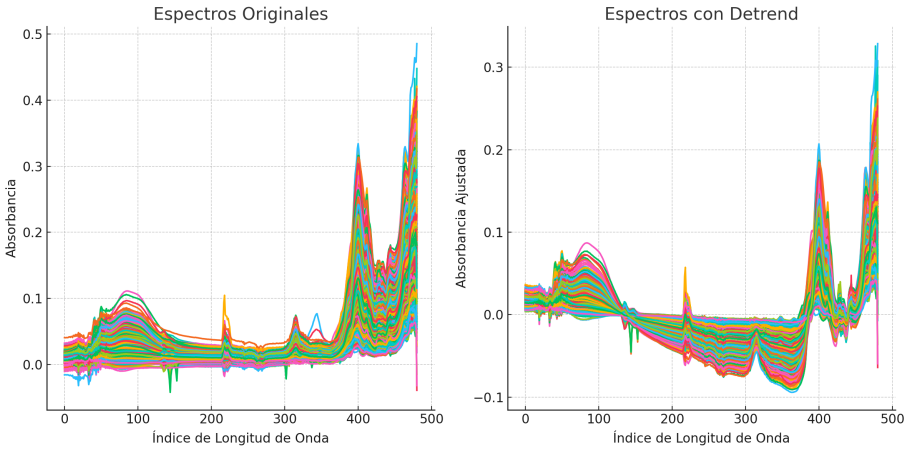


Figura 12: Aplicación del pretratamiento Detrent

Los espectros originales muestran una variabilidad considerable en la absorbancia a lo largo de las diferentes longitudes de onda (400-4000 cm^{-1}). Esta variabilidad se manifiesta en forma de picos y valles significativos, lo cual es típico en datos de espectroscopía MIR. Estos picos y valles son indicativos de las diferentes vibraciones moleculares presentes en los compuestos del suelo. Sin embargo, también se observa una tendencia general en los espectros originales, que podría estar influenciada por variaciones sistemáticas no relacionadas directamente con las propiedades químicas de las muestras de suelo. Estas variaciones pueden deberse a factores como la dispersión de la luz, el espesor de la muestra y otras inconsistencias experimentales que afectan la medición de la absorbancia (Li et al., 2022).

Después de aplicar el pretratamiento Detrend, se observa una disminución notable en las tendencias lineales presentes en los datos originales. La absorbancia ajustada muestra una mayor uniformidad, lo que sugiere que las variaciones sistemáticas han sido eliminadas eficazmente. Esta uniformidad es crucial porque permite que las diferencias en los espectros sean más representativas de las variaciones químicas reales en las muestras de suelo. La eliminación de estas tendencias lineales ayuda a resaltar las características inherentes de los espectros, facilitando un análisis más preciso de las propiedades químicas de las muestras (Li et al., 2022).

El pretratamiento Detrend ha demostrado ser efectivo en la eliminación de tendencias lineales

no deseadas en los datos espectrales. Al comparar las dos gráficas, se pueden destacar varios puntos críticos para la predicción de nutrientes en suelos. En primer lugar, la reducción de la variabilidad que no está relacionada con las propiedades químicas de las muestras ha sido significativa, permitiendo un análisis más claro y preciso de los espectros (Li et al., 2022). Además, los espectros detrendeados muestran una mayor uniformidad en comparación con los espectros originales. Esta uniformidad en los datos espectrales mejora la aceptabilidad de los modelos PLSR, facilitando la comparación y el análisis de datos entre diferentes muestras y condiciones experimentales. La uniformidad también reduce la influencia de factores externos que pueden distorsionar las mediciones espectrales, proporcionando una base más sólida para el análisis y la interpretación de los resultados.

A pesar de la eliminación de tendencias, las características importantes de los espectros, como los picos y valles, se han preservado. Esto es vital para el análisis cuantitativo y cualitativo de los nutrientes, ya que los picos específicos en los espectros MIR están directamente relacionados con las vibraciones moleculares de los compuestos presentes en el suelo. La preservación de estas características permite que los modelos PLSR identifiquen y cuantifiquen con precisión los nutrientes presentes en las muestras de suelo, mejorando la precisión y la fiabilidad de las predicciones (Li et al., 2022).

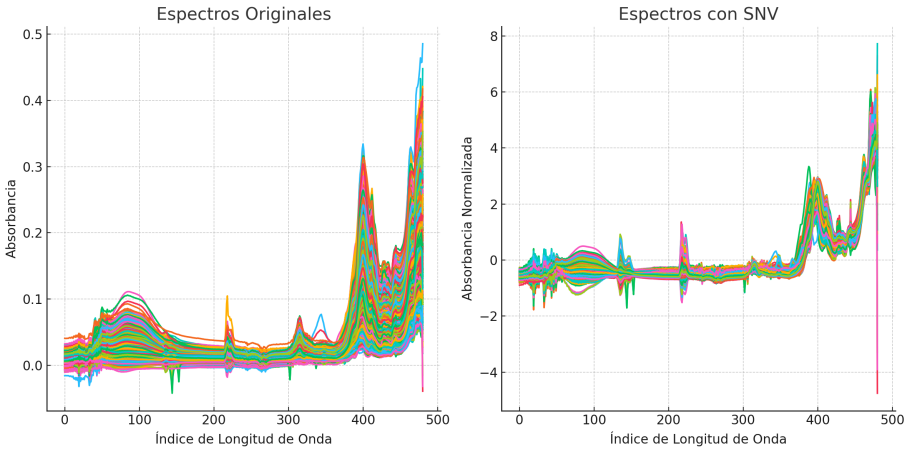


Figura 13: Aplicación del pretratamiento variación estándar normalizada (SNV)

La aplicación del pretratamiento SNV se enfoca en eliminar estas variaciones sistemáticas no deseadas, resultando en una absorbancia normalizada a lo largo de todas las longitudes de onda. Al observar los espectros procesados, se nota una reducción significativa de la dispersión en

la magnitud de las señales, lo que sugiere que el SNV ha normalizado efectivamente los datos (Shin, Kim, Cho, Yang, y Cho, 2024). Esta normalización facilita la identificación de las verdaderas diferencias químicas entre las muestras, mejorando la capacidad de distinguir entre las distintas concentraciones de nutrientes (Kandpal, Munnaf, Cruz, y Mouazen, 2022).

El SNV ha eliminado eficazmente las variaciones en la magnitud de las señales espectrales que no están relacionadas con las propiedades químicas del suelo. Esto es crítico para el análisis, ya que permite una interpretación más precisa de los espectros. La uniformidad lograda en los datos es especialmente beneficiosa para la construcción de los modelos PLSR, ya que estos modelos requieren datos consistentes y comparables para generar predicciones de los nutrientes: calcio (Ca), fósforo (P), potasio (K), magnesio (Mg) y materia orgánica (MO) (Shin et al., 2024).

Al eliminar las variaciones de magnitud, el SNV permite que los picos y valles característicos de los espectros se destaquen de manera más clara. Los picos mejor definidos son fundamentales para la cuantificación de los nutrientes, ya que cada pico corresponde a vibraciones moleculares específicas de los compuestos presentes en el suelo (Kandpal et al., 2022).

La normalización de los espectros que realiza el SNV facilita la comparación directa entre diferentes muestras de suelo. Esto es crucial para identificar los patrones y tendencias en la composición química del suelo. Al eliminar las variaciones sistemáticas, se obtiene un conjunto de datos más homogéneo, lo que permite comparar muestras en diversas condiciones y con el objetivo de mejorar la aceptabilidad de las conclusiones obtenidas (Shin et al., 2024).

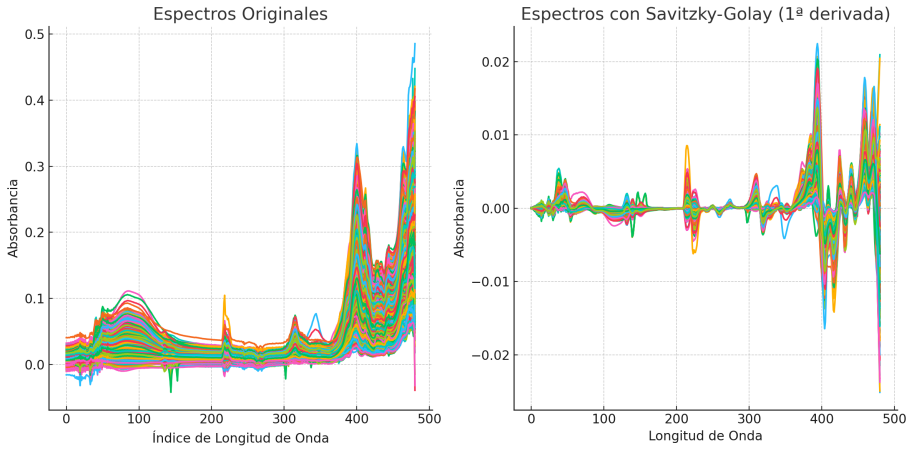


Figura 14: Aplicación del pretratamiento de la primera derivada (Savitzky-Golay

El filtro Savitzky-Golay ha mejorado la resolución de los picos espectrales, destacando transiciones que pueden estar asociadas con características específicas de los nutrientes en el suelo. Esto facilita la identificación y cuantificación de compuestos presentes en las muestras, como el calcio (Ca), fósforo (P), potasio (K), magnesio (Mg) y materia orgánica (MO). La capacidad de distinguir claramente estos picos es crucial para desarrollar modelos PLSR precisos (Zhang y Mouazen, 2023).

Al aplicar el filtro, se ha reducido el ruido de alta frecuencia en los espectros, lo que contribuye a obtener datos más limpios y confiables. Esta reducción de ruido es esencial para mejorar la aceptabilidad de los modelos, ya que el ruido puede interferir con la detección de señales químicas relevantes y disminuir la precisión de las predicciones.

La transformación de los espectros a través del filtro Savitzky-Golay facilita la comparación directa entre diferentes muestras de suelo. Al resaltar las diferencias en los espectros derivados, es posible identificar patrones y tendencias que no serían evidentes en los datos originales. Esto es especialmente útil en estudios de suelos donde se busca correlacionar las propiedades espectrales con parámetros agronómicos específicos (Zhang y Mouazen, 2023).

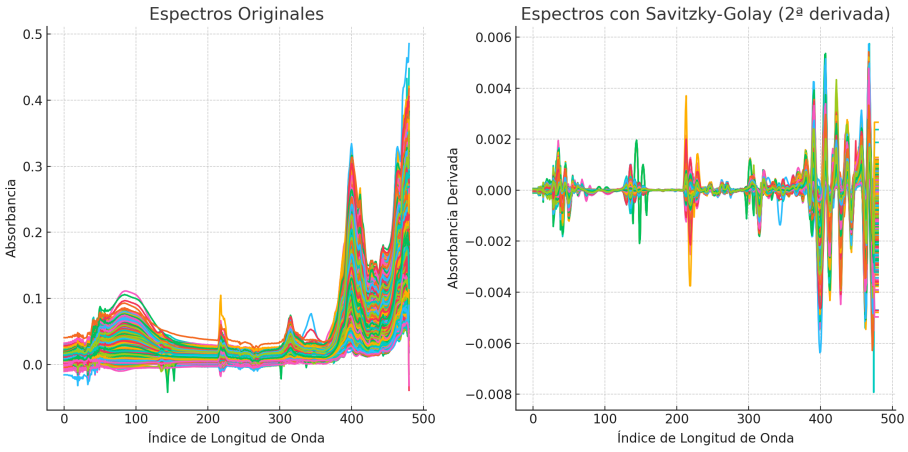


Figura 15: Aplicación del pretratamiento de la segunda derivada (Savitzky-Golay

Los espectros procesados con el filtro de la segunda derivada Savitzky-Golay, se nota un cambio significativo en la apariencia de los datos. Los picos y valles se han convertido en transiciones más pronunciadas, lo que permite una identificación más clara de los puntos de inflexión en los espectros (Kandpal et al., 2022).

El filtro ha mejorado la resolución de los picos espectrales, destacando transiciones que pueden estar asociadas con características específicas de los nutrientes en el suelo. Además, se ha reducido el ruido de alta frecuencia en los espectros, lo que contribuye a obtener datos más limpios y confiables.

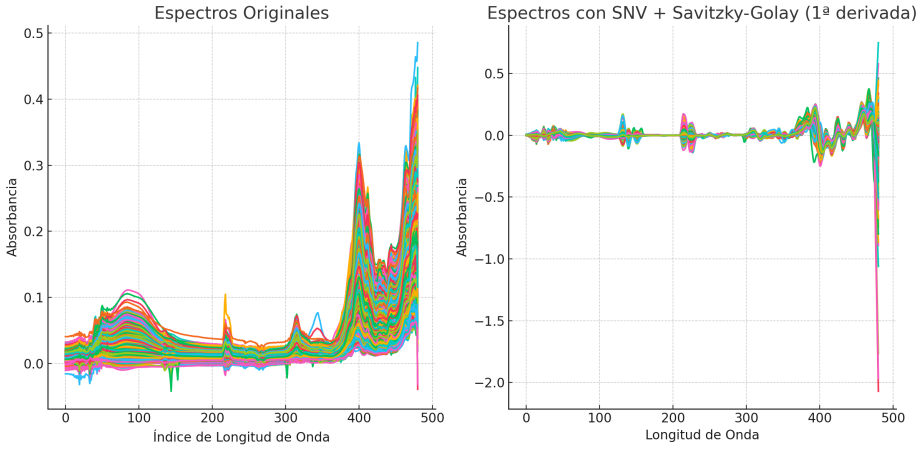


Figura 16: Aplicación del pretratamiento de SNV en combinación con la primera derivada (Savitzky-Golay

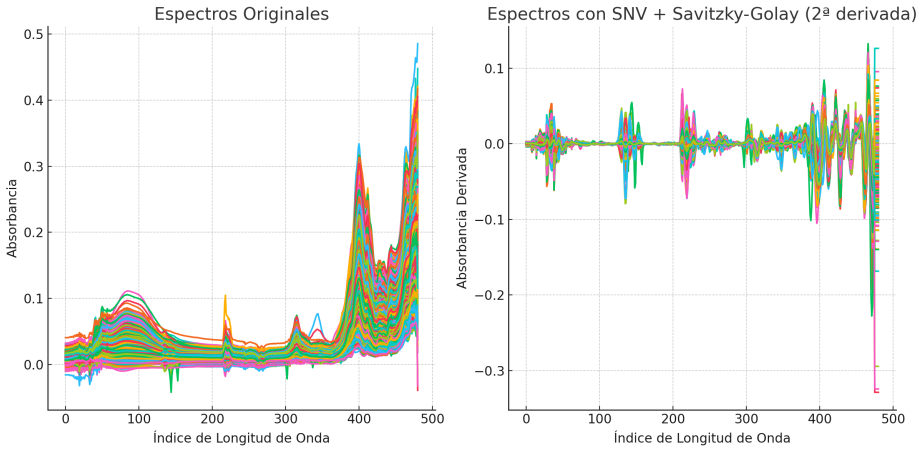


Figura 17: Aplicación del pretratamiento de SNV en combinación con la segunda derivada (Savitzky-Golay

Al comparar los pretratamientos SNV + Savitzky-Golay (1ª Derivada) y SNV + Savitzky-Golay (2ª Derivada), se observan varias diferencias y ventajas específicas.

En cuanto a la resolución de picos, la primera derivada del filtro Savitzky-Golay proporciona una mejora notable, adecuada para identificar transiciones importantes y reducir el ruido en los

espectros. Sin embargo, la segunda derivada va un paso más allá, mejorando aún más la resolución de los picos y haciendo las transiciones más pronunciadas. Esto facilita la identificación de detalles finos en los espectros, lo que puede ser crucial en ciertos análisis químicos (Wu, Huang, Zhao, Jin, y Ruan, 2024).

Ambos pretratamientos redujeron significativamente el ruido de alta frecuencia. No obstante, la segunda derivada podría ser más efectiva en este aspecto, ya que tiene una mayor capacidad para resaltar cambios menores y eliminar variaciones no deseadas en los datos espectrales.

La normalización, lograda mediante el SNV en ambas combinaciones, elimina eficazmente las variaciones sistemáticas debidas a la magnitud de las señales. Esto es fundamental para facilitar la comparación directa entre diferentes muestras de suelo, mejorando la consistencia y la interpretabilidad de los datos espectrales.

En cuanto a la aplicación en modelos PLSR, la primera derivada es ideal cuando se busca un equilibrio entre la reducción de ruido y la mejora de la resolución de picos. Por otro lado, la segunda derivada es más adecuada para estudios donde es crucial identificar transiciones finas y detalles específicos en los espectros. Esto puede llevar a una mejora en la precisión de los modelos PLSR, especialmente para ciertos nutrientes donde los detalles finos en los espectros son esenciales para una predicción precisa (Wu et al., 2024).

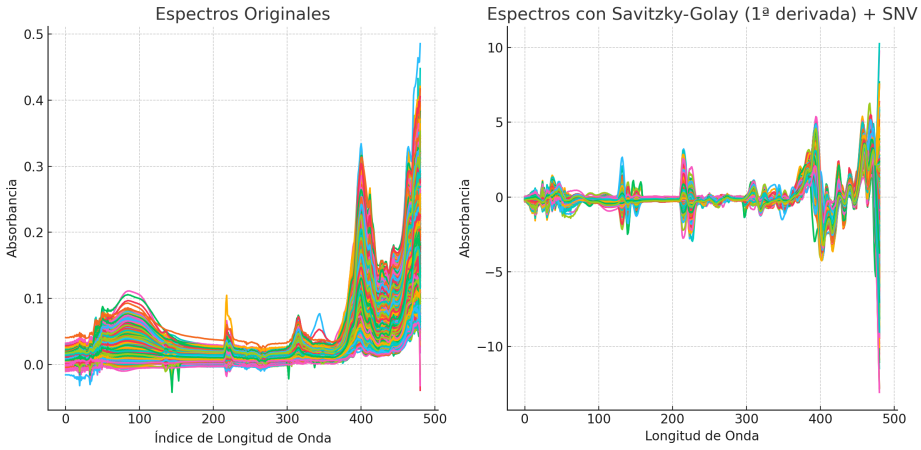


Figura 18: Aplicación del pretratamiento la primera derivada (Savitzky-Golay) en combinación con SNV

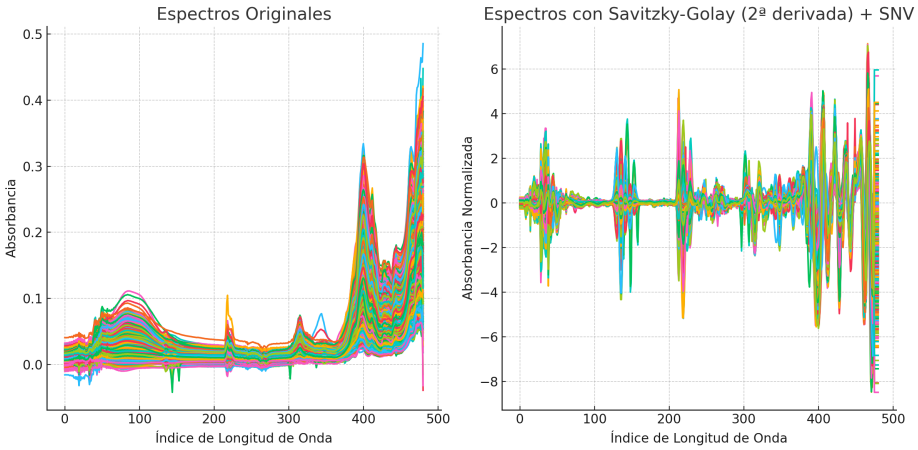


Figura 19: Aplicación del pretratamiento la segunda derivada (Savitzky-Golay) en combinación con SNV

Al comparar los espectros procesados con los métodos Savitzky-Golay de 1ª y 2ª derivada combinados con SNV, se observan diferencias notables en la resolución de picos, reducción de ruido, normalización y la aplicación en modelos PLSR.

En términos de resolución de picos, el método Savitzky-Golay (1ª derivada) + SNV mejora significativamente la resolución de los picos espectrales, facilitando la identificación de transiciones importantes en los espectros. Esto es particularmente útil para destacar características clave de los nutrientes en el suelo. Por otro lado, Savitzky-Golay (2ª derivada) + SNV proporciona una resolución aún mayor, haciendo que las transiciones sean más pronunciadas y los detalles más evidentes. Este enfoque es especialmente útil en estudios detallados donde se requiere identificar con precisión pequeñas variaciones en los espectros (Zhang y Mouazen, 2023).

En cuanto a la reducción de ruido, ambos métodos son efectivos, pero Savitzky-Golay (1ª derivada) + SNV reduce el ruido de alta frecuencia, lo cual es crucial para obtener datos más limpios y confiables. Esta reducción de ruido facilita el entrenamiento de modelos PLSR con datos consistentes. Savitzky-Golay (2ª derivada) + SNV es generalmente más efectivo en la reducción de ruido, ya que resalta cambios pequeños y elimina variaciones menores que podrían interferir en el análisis de los espectros (Zhang y Mouazen, 2023).

Ambos métodos, al combinarse con SNV, normalizan eficazmente las variaciones sistemáticas debidas a la magnitud de las señales, lo que facilita la comparación directa entre diferentes muestras de suelo. Esto asegura que las diferencias observadas reflejen variaciones químicas reales.

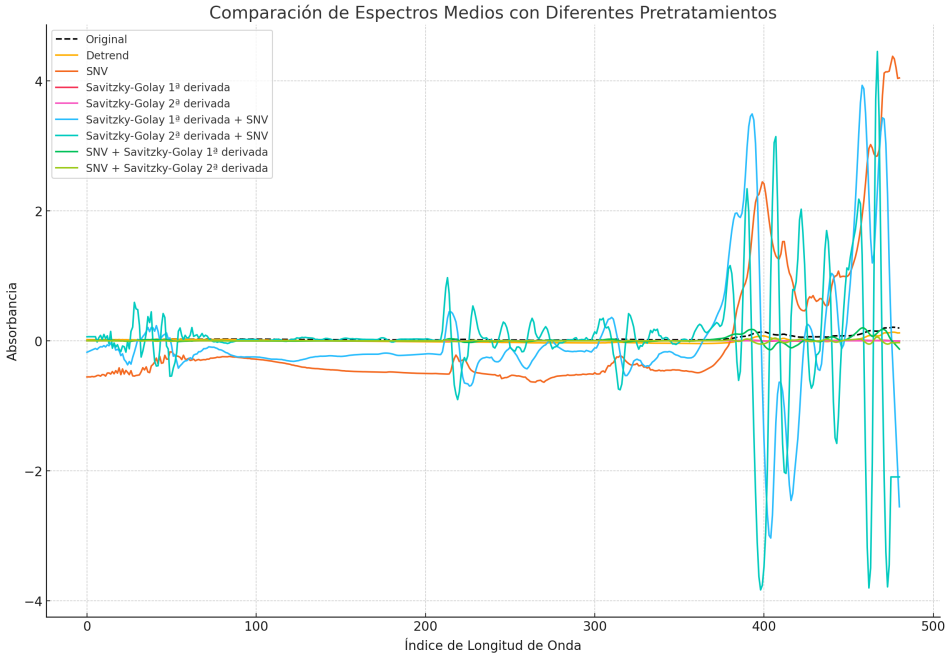


Figura 20: Comparación de la aplicación de los pretratamientos y combinaciones

Los espectros originales (línea negra discontinua) presentan una variabilidad considerable en la absorbancia, especialmente entre los índices de longitud de onda de 400 a 500. Estos espectros contienen tanto información relevante sobre las propiedades químicas de las muestras como ruido y posibles tendencias sistemáticas.

La línea amarilla muestra los espectros tras aplicar el pretratamiento Detrend. Este método elimina las tendencias lineales, resultando en espectros más planos y uniformes, pero no aborda el ruido. La eliminación de tendencias se observa claramente, ya que los valores de absorbancia se aproximan más a cero a lo largo de todo el espectro.

La línea verde muestra los espectros tras aplicar SNV. Este pretratamiento normaliza las variaciones sistemáticas en la absorbancia, lo que resulta en espectros más comparables entre sí. Se observa una reducción de la variabilidad general y una mejor alineación de los espectros.

La línea azul clara representa los espectros tras aplicar el filtro Savitzky-Golay para la primera derivada. Este pretratamiento mejora la resolución de picos y reduce el ruido de alta frecuencia. Los picos y valles se vuelven más pronunciados, facilitando la identificación de características espectrales importantes.

La línea fucsia muestra los espectros tras aplicar el filtro Savitzky-Golay para la segunda derivada. Este método proporciona una mayor resolución de los picos y una mayor complejidad espectral. Los picos son más definidos y se observan más detalles en el espectro.

Las combinaciones de Savitzky-Golay con SNV (líneas rosa, celeste, verde claro y verde oscuro) muestran una mejora significativa en la resolución de picos y una normalización efectiva de la absorbancia. Estos pretratamientos combinados reducen el ruido y mejoran la claridad y detalle de las señales químicas.

Los grandes picos observados en el rango de 400 a 500 en la gráfica se deben a la aplicación de derivadas (primera y segunda) con el filtro Savitzky-Golay. Estos picos representan transiciones rápidas en la absorbancia, que son características típicas de derivadas.

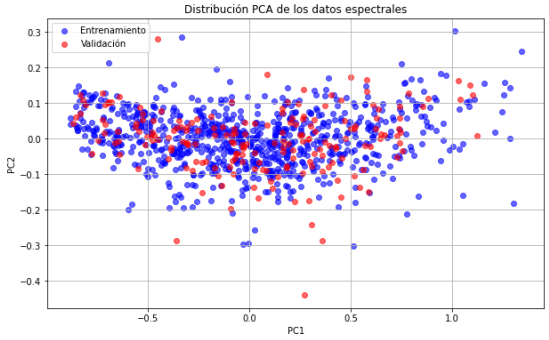
Estos picos también pueden resaltar la presencia de ruido de alta frecuencia que no ha sido completamente eliminado, especialmente en los pretratamientos que incluyen derivadas.

En comparación con los espectros originales, los pretratamientos con derivadas muestran una mayor complejidad y resolución de picos. Esto es útil para identificar características espectrales específicas, pero también puede introducir artefactos que deben interpretarse con cuidado.

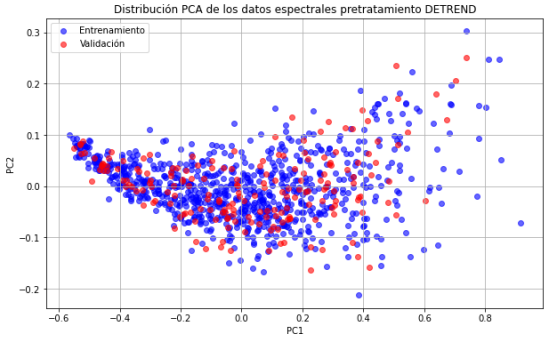
4.5. Formulación de modelos predictivos mediante regresión por mínimos cuadrados parciales (PLSR)

4.5.1. PCA conjuntos de entrenamiento y validación

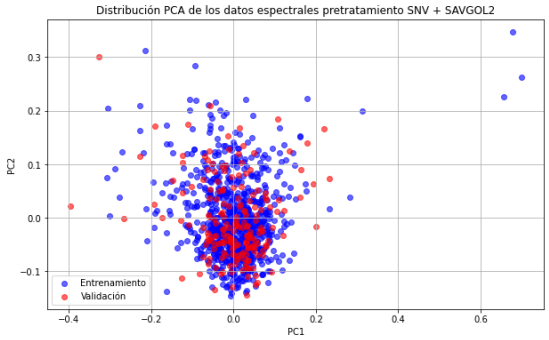
Las representaciones en el espacio PCA como en la figura 21 permiten visualizar cómo se distribuyen los conjuntos de entrenamiento y validación tras aplicar distintos pretratamientos a los datos espectrales. Ver una superposición o cercanía entre ambas nubes de puntos (azules para entrenamiento y rojos para validación) es esencial para comprobar que los datos de validación no difieren de forma significativa de los de entrenamiento, lo cual favorece una mayor generalización del modelo.



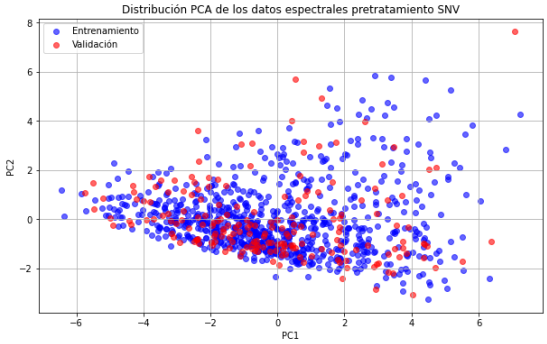
(a) PCA de los conjuntos de validación y entrenamiento del modelo para los datos sin pretratamiento



(b) PCA de los conjuntos de validación y entrenamiento del modelo para los datos con pretratamiento Detrend



(c) PCA de los conjuntos de validación y entrenamiento del modelo para los datos con pretratamiento SNV + Savitzky-Golay 2ª Derivada



(d) PCA de los conjuntos de validación y entrenamiento del modelo para los datos con pretratamiento SNV

Figura 21: PCA de los conjuntos de validación y entrenamiento para diferentes pretratamientos

Si el conjunto de validación apareciera muy separado en el espacio PCA, ello indicaría que las muestras de validación no están bien representadas por el modelo entrenado, o que existen diferencias sistemáticas en la adquisición de los datos

Es posible observar cómo, en los gráficos de PCA, se aprecia una correcta homogeneización de los puntos correspondientes a los conjuntos de entrenamiento y validación. Esta superposición indica que la variabilidad capturada en el entrenamiento también está presente en la validación, lo cual garantiza que el modelo pueda generalizar de manera confiable a nuevas muestras.

Cuando ambas nubes de puntos comparten una distribución similar, se refuerza la idea de que provienen de la misma población y que no existen discrepancias sustanciales en la adquisición o preprocesamiento de los datos. En consecuencia, la consistencia en la distribución de entrenamiento y validación refuerza la aceptabilidad de los modelos PLSR y aumenta la confianza en las predicciones que se deriven de ellos.

4.5.2. Variables latentes del modelo

Las variables latentes son fundamentales en los modelos PLSR porque permiten resumir y simplificar la información contenida en un gran número de variables predictoras, cuyo número ideal se determina cuando se obtiene una minimización del RMSE. Esto no solo facilita la interpretación del modelo, sino que también ayuda a evitar problemas cuando las variables originales están muy correlacionadas entre sí. Estas variables latentes permiten que el modelo sea más sencillo y preciso a la hora de realizar predicciones.

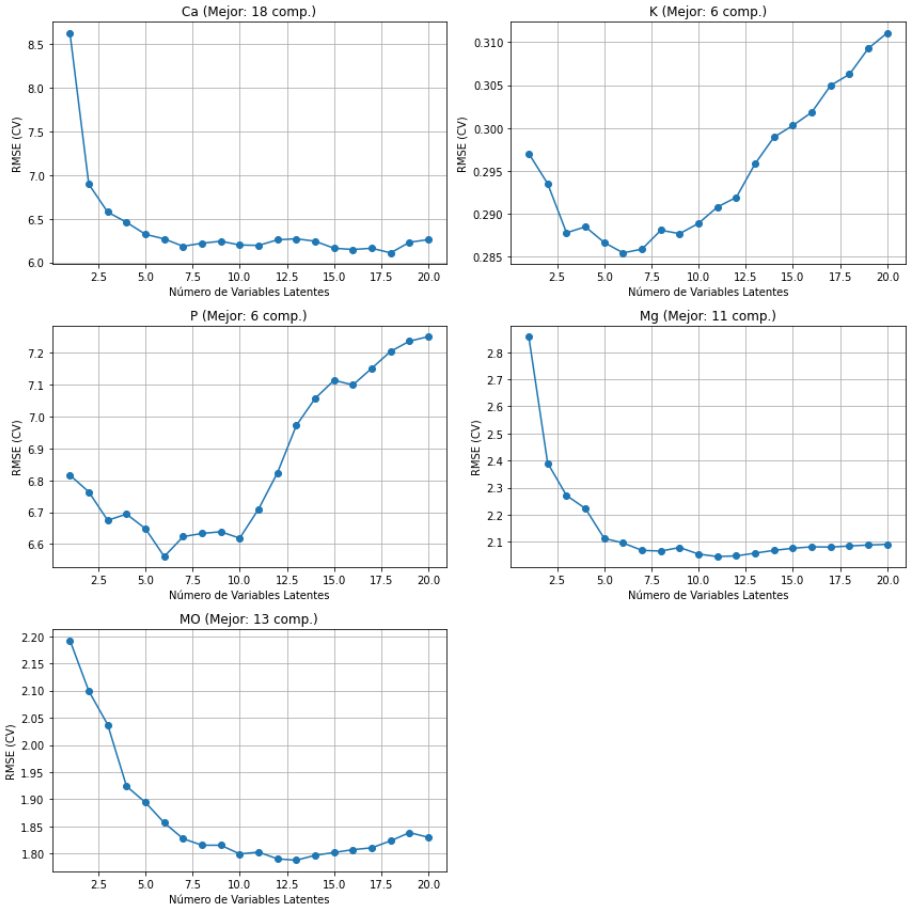


Figura 22: Cantidad de variables latentes por nutriente para la generación del modelo con el pretratamiento Savitzky-Golay 1ª Derivada + SNV

La figura 22 presenta el ejercicio realizado para la determinación del número ideal de variables latentes para el modelo con los datos pretratados con Savitzky-Golay 1ª Derivada + SNV , donde a cada nutriente se le probó un rango de valores de variables latentes y se estimó el error de predicción (RMSE) mediante validación cruzada. Las gráficas muestran cómo el error varía a

medida que se incrementa o se disminuye el número de componentes, lo cual permite identificar el punto en que el modelo logra un mejor equilibrio entre complejidad y precisión para cada nutriente.

Al aplicar el proceso de forma independiente a cada uno de los pretratamientos, evaluándose distintos números de componentes latentes en los modelos PLSR para predecir cada nutriente. Una vez identificada la cantidad óptima de componentes para cada combinación, los resultados se agruparon y se presentaron de manera comparativa en la Figura 23. En esta figura, puede observarse claramente cómo varía el número de componentes recomendados según el tipo de pretratamiento empleado y el nutriente en cuestión, lo que facilita la toma de decisiones sobre cuál combinación podría brindar el mejor desempeño predictivo.

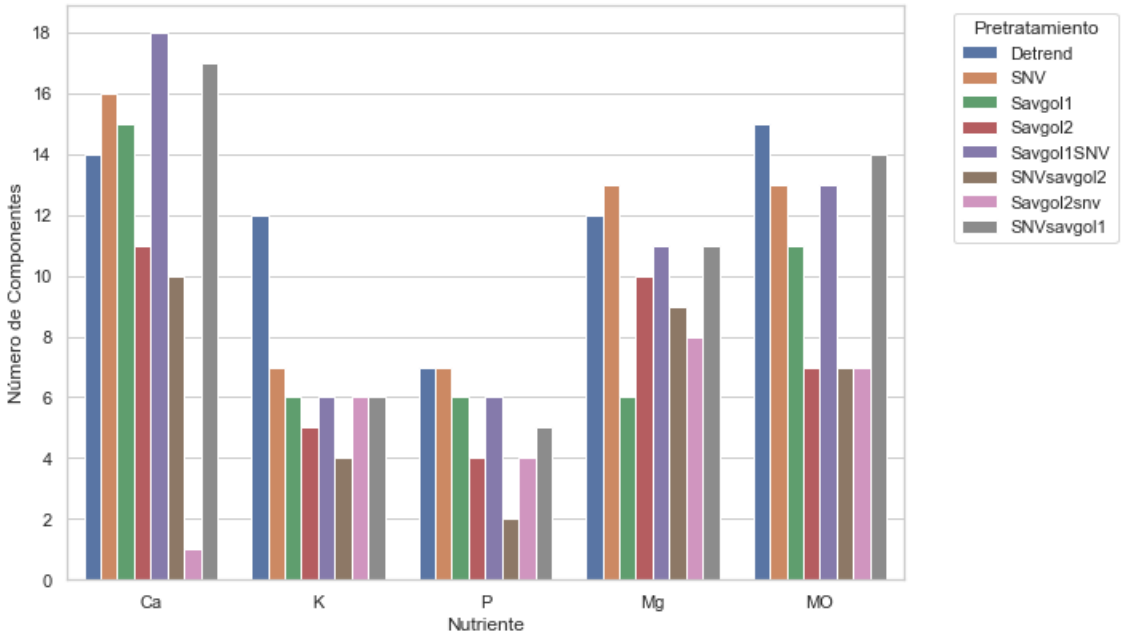


Figura 23: Cantidad de variables latentes por nutriente para generar el modelo para cada uno de los pretratamientos

Observando los resulados de la figura 23, se observa una marcada variabilidad entre nutrientes. En particular, tanto el calcio (Ca) como la materia orgánica (MO) suelen requerir un número de variables latentes relativamente elevado en la mayoría de los pretratamientos, lo cual sugiere que la relación entre la señal espectral y estos nutrientes está distribuida en varias regiones del espectro o se encuentra mezclada con otras señales, haciendo que el modelo necesite capturar

más dimensiones para describirla adecuadamente.

Por el contrario, el potasio (K) y el fósforo (P) presentan una tendencia a utilizar menos componentes, lo que indica que la información relevante para predecirlos podría concentrarse en pocas variables latentes. El magnesio (Mg) se ubica en un punto intermedio, pues muestra un rango de valores que oscila entre lo moderado y lo alto, dependiendo del pretratamiento específico.

En cuanto al efecto del pretratamiento, no se identifica un método único que garantice el menor o el mayor número de variables para todos los nutrientes. Cada combinación "nutriente-pretratamiento" ofrece resultados diferentes, lo que evidencia la complejidad propia de la señal y la necesidad de probar diversas estrategias de corrección y filtrado. Así, se ve que en algunos casos, métodos más avanzados como la combinación de SNV con Savitzky-Golay de segunda derivada (snvsavgol2) logran reducir de forma significativa el número de componentes en ciertos nutrientes (por ejemplo, P con tan solo 2 componentes), lo que podría interpretarse como un mejor realce de las características espectrales relevantes. En cambio, otros métodos, como la combinación de Savitzky-Golay de primera derivada con SNV (savgol1snv), pueden requerir un mayor número de componentes, lo que implica que, pese a reducir el ruido o corregir la línea base, todavía se necesitan múltiples dimensiones para capturar toda la variabilidad significativa de la señal.

Desde el punto de vista de la complejidad y especificidad de los modelos, el hecho de que algunos nutrientes necesiten sistemáticamente más variables indica que su información espectral está más dispersa o se superpone con otras señales, dificultando la separación y la extracción de las características relevantes en un espacio de baja dimensión, como lo es en este caso. De esta manera, se refuerza la idea de que la elección tanto del pretratamiento como del número de variables latentes debe ajustarse a cada caso particular, considerando la naturaleza de la señal y los objetivos del análisis.

4.5.3. Análisis de los modelos predictivos

Los resultados de los modelos predictivos están fuertemente condicionados por las técnicas específicas de extracción y cuantificación utilizadas para los nutrientes analizados (Ca, P, K, Mg y MO). Métodos como Olsen Modificado para fósforo y potasio, la extracción con KCl para calcio y magnesio, y el método de Walkley y Black para materia orgánica, influyen directamente en la eficiencia de extracción y, por tanto, en la composición de las muestras y los datos generados. Estas técnicas, al ajustar variables como pH, concentración del reactivo, volumen de muestra y tiempo de reacción, determinan la cantidad de nutrientes efectivamente liberados desde la matriz del suelo hacia la solución extractante.

Además, la cuantificación posterior mediante espectroscopía de absorción atómica (para Ca, Mg y K) depende críticamente de la etapa de extracción, ya que cualquier variabilidad en la concentración del extracto afecta directamente la precisión y sensibilidad del análisis. En el caso del carbono orgánico, la eficiencia de oxidación en el método de Walkley y Black también puede verse influida por el tipo de suelo y el estado de la materia orgánica, lo cual repercute en la calidad del dato y, en consecuencia, en el desempeño del modelo predictivo.

Por ello, la precisión, estabilidad y aplicabilidad de los modelos están estrechamente ligadas a la consistencia de los métodos empleados en laboratorio. Esta dependencia metodológica implica que los modelos generados son específicos a las condiciones bajo las cuales fueron calibrados, limitando su capacidad de generalización a otros contextos sin las debidas adaptaciones, normalización de datos o una nueva calibración con base en las técnicas analíticas correspondientes.

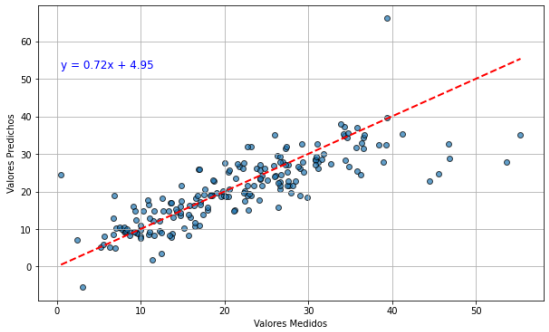
Mejores modelos para la predicción de Calcio

El histograma de frecuencia de los datos de laboratorio (ver figura 3) mostró que la distribución del calcio (Ca) es mucho más simétrica y amplia que las distribuciones fuertemente sesgadas a la derecha observadas en otros elementos como el potasio (K) y el fósforo (P). Esta amplitud refleja que las concentraciones de Ca abarcan un rango extenso de valores sin acumulaciones marcadas en los extremos, lo que lo distingue claramente de los nutrientes con distribuciones más restringidas o dominadas por valores atípicos.

Esta configuración estadística es especialmente favorable para los algoritmos de modelado predictivo, ya que ofrece una varianza más útil y menos concentrada. En consecuencia, las relaciones entre las concentraciones de Ca y las características espectrales tienden a expresarse de manera más lineal y consistente a lo largo de todo el rango de valores, reduciendo el riesgo de ajustes sesgados.

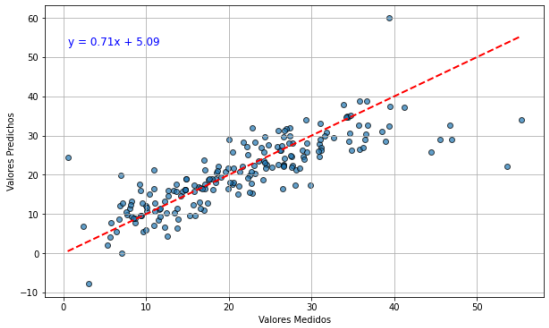
La relevancia de esta distribución uniforme se refleja directamente en la calibración de los modelos. Como señalan, la variabilidad en la concentración de carbonatos en los suelos es un factor determinante en la precisión de las predicciones. Una base de datos que representa de forma equilibrada todo el rango de concentraciones de Ca facilita la construcción de modelos generalizables, mientras que distribuciones sesgadas o con poca variabilidad limitan la capacidad del modelo para captar adecuadamente las señales espectrales asociadas (Oberholzer, Summerauer, Steffens, y Ifejika Speranza, 2023).

Por otro lado, el buen desempeño del Ca en espectroscopía MIR no puede atribuirse únicamente a la forma de su distribución estadística. Sus claras bandas de absorción vinculadas a carbonatos y arcillas aportan un respaldo adicional al proceso de calibración. Es la combinación entre estas propiedades espectrales intrínsecas y una representación equilibrada de las concentraciones en la base de calibración lo que potencia la capacidad predictiva del Ca frente a otros nutrientes más problemáticos (Shepherd y Walsh, 2002).



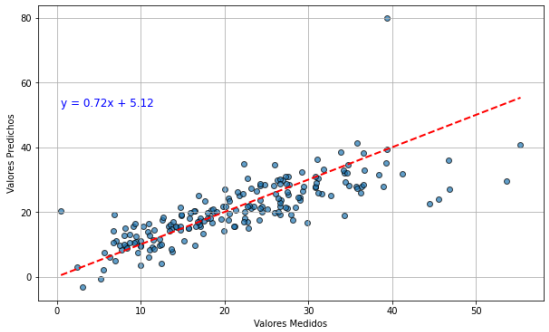
(a) Valores reales vs. predichos para Calcio con Savitzky-Golay 1ª Derivada + SNV

- **RMSE**: 6.03
- **R**^{2}: 0.66



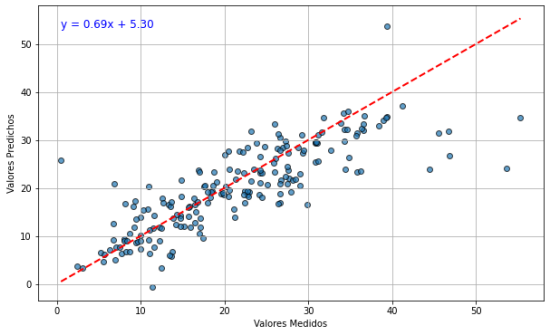
(b) Valores reales vs. predichos para Calcio con el pretratamiento SNV

- **RMSE**: 6.3
- **R**^{2}: 0.63



(c) Valores reales vs. predichos para Calcio con el pretratamiento Detrend

- **RMSE**: 6.23
- **R**^{2}: 0.64

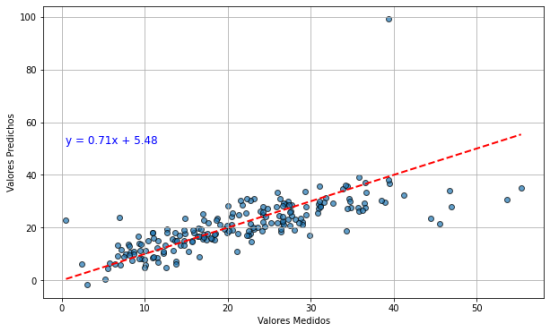


(d) Valores reales vs. predichos para Calcio con Savitzky-Golay 2ª Derivada + SNV

- **RMSE**: 6.22
- **R**^{2}: 0.64

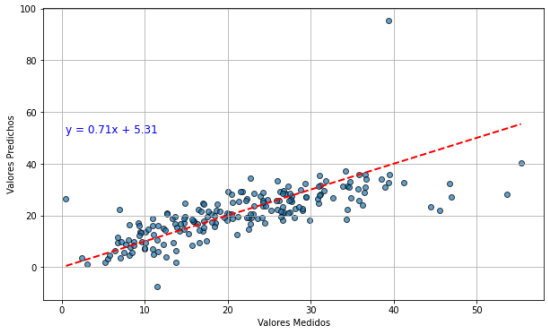
Figura 24: Mejores modelos para la predicción del calcio.

Modelos menos efectivos para la predicción de calcio



(a) Valores reales vs. predichos para Calcio con Savitzky-Golay 1ª Derivada

- **RMSE**: 7.14
- **R**^{2}: 0.53



(b) Valores reales vs. predichos para Calcio con Savitzky-Golay 2ª Derivada

- **RMSE**: 7.47
- **R**^{2}: 0.48

Figura 25: Peores modelos para la predicción del calcio.

Los modelos con pretratamientos SNV y combinaciones de SNV con Savitzky-Golay (1ª derivada y 2ª derivada) mostraron un mejor rendimiento, con valores de RMSE alrededor de 6.22-6.3 y R^2 en el rango de 0.62-0.66.

En la figura 24 se puede observar que los pretratamientos como SNV son conocidos por su capacidad para eliminar variaciones sistemáticas y ruidos no relacionados con la señal química del interés. SNV corrige las variaciones en la dispersión de la luz y las diferencias en el tamaño de las partículas, lo que mejora la uniformidad de los espectros y permite una mejor captura de las características espectrales relevantes para la predicción de nutrientes como el Ca (Karray et al., 2023).

La segunda y primera derivada de Savitzky-Golay, como se puede notar en la figura 25 si no se combina con otros pretratamientos como SNV, tiende a intensificar tanto el ruido como las características no relacionadas con el nutriente. Esto se debe a que la segunda derivada amplifica las fluctuaciones en los datos, lo que puede resultar en un aumento del ruido y una disminución en la precisión del modelo de predicción. Este efecto negativo se refleja en los valores más altos de RMSE (7.47) y los valores más bajos de R^2 (0.48) observados en los resultados sin el pretratamiento SNV (Karray et al., 2023).

En la literatura científica, la predicción del Ca mediante espectroscopía MIR ha sido ampliamente documentada con resultados altamente satisfactorios. (Lelago y Bibiso, 2022) reportaron coeficientes de determinación superiores a 0.85 y valores de RPD mayores a 2.5 en la predicción de Ca en diferentes suelos africanos y europeos, destacando la fuerte asociación de este elemento con minerales carbonatados y arcillosos que generan bandas espectrales claras en la región MIR. Asimismo, (Shepherd y Walsh, 2002) encontraron que el Ca es uno de los cationes de cambio con mejor desempeño predictivo a partir de espectros de reflectancia, atribuyendo este resultado a que la señal del Ca está indirectamente relacionada con propiedades edáficas de alta sensibilidad espectral, como la capacidad de intercambio catiónico y la textura del suelo.

Modelos para la predicción de Potasio

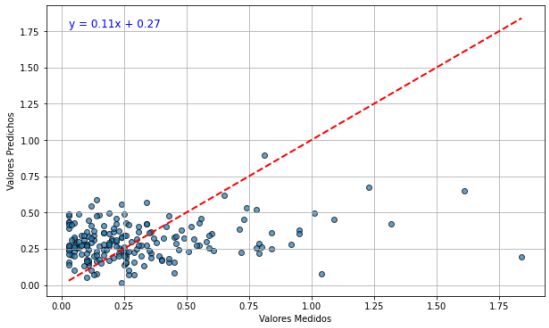
El potasio (K) representa un desafío considerable para la modelización quimiométrica debido a las características inherentes de su distribución de concentración y su comportamiento en el espacio multivariado. Aunque en el análisis de componentes principales aparece como una variable de relevancia aparente, su utilidad práctica en modelos como el PLSR se ve seriamente comprometida. En suelos agrícolas etíopes, la predicción de K disponible mostró un rendimiento deficiente, con valores de R^2 entre 0.25 y 0.31 y RPD entre 1.2 y 1.4. Este bajo desempeño ha sido atribuido a la alta movilidad de los iones de K en la solución del suelo, lo que provoca fluctuaciones rápidas en su contenido y dificulta la detección de patrones espectrales consistentes (Lelago y Bibiso, 2022).

El histograma de frecuencia del potasio (figura 3) revela una distribución altamente asimétrica, con un sesgo marcado hacia valores bajos. La mayoría de las muestras presentan concentraciones reducidas de K, con una acumulación significativa en el extremo inferior del rango, mientras que solo unas pocas muestras alcanzan valores elevados que generan una cola larga y dispersa. Esta configuración limita la variabilidad útil para la calibración, ya que la mayor parte del conjunto de datos presenta diferencias mínimas entre sí.

Este patrón también se refleja en la figura 26, donde se observa un fuerte agrupamiento de los datos. En el espacio de componentes principales (figura 4), el vector de carga correspondiente al potasio aparece entre los más largos, lo que sugiere una contribución significativa a la varianza total. Sin embargo, esta aparente importancia es engañosa: no deriva de una variabilidad ho-

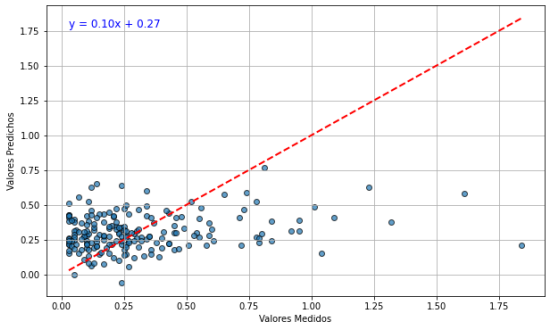
mogénea en todo el rango de concentración, sino de la marcada diferencia entre la gran cantidad de muestras con bajo contenido de K y unos pocos valores atípicos extremadamente altos, los cuales inflan artificialmente la varianza.

Como consecuencia, el modelo PLSR enfrenta serias limitaciones para calibrar esta variable. La densidad de muestras en concentraciones bajas reduce la posibilidad de identificar relaciones lineales estables entre las características espectrales y el contenido de K. En estos niveles, las señales espectrales asociadas al potasio son débiles y fácilmente enmascaradas por el ruido o por la influencia de otros componentes del suelo, lo que explica el bajo poder predictivo observado.



(a) Valores reales vs. predichos para Potasio con SNV

- **RMSE**: 0.29
- **R**^{2}: 0.05



(b) Valores reales vs. predichos para Potasio con Savitzky-Golay 1ª Derivada + SNV

- **RMSE**: 0.29
- **R**^{2}: 0.05

Figura 26: Modelos para la predicción de Potasio.

Modelos para la predicción de Fósforo

El fósforo (P) representa uno de los mayores retos en la modelización quimiométrica, mostrando obstáculos incluso más complejos que los observados en el potasio (K). Aunque en el análisis multivariado el P aparece con una relevancia destacada, su utilidad práctica en modelos de regresión como el PLSR es limitada. En suelos agrícolas etíopes, la predicción espectral del P disponible mostró valores de R^2 muy bajos (0.25–0.31) y RPD 1.2, lo que evidencia un desempeño deficiente. Este bajo rendimiento se asocia a la fuerte retención del P en la matriz del suelo (particularmente ligado a óxidos de Fe y Al) y a la complejidad de sus formas químicas en estado natural (Lelago y Bibiso, 2022).

El histograma de frecuencia del fósforo (figura 3) evidencia una distribución altamente asimétrica, caracterizada por una acumulación masiva de muestras en rangos bajos de concentración y una cola larga generada por unos pocos valores excepcionalmente altos. Esta configuración restringe la variabilidad útil para calibración, ya que la mayoría de las observaciones se concentran en un rango estrecho en el cual las diferencias espectrales son mínimas y difíciles de capturar por el modelo.

Este patrón se observa también en la figura 27, donde se aprecia un agrupamiento claro de datos en los niveles bajos. En el análisis de componentes principales (figura 4), el fósforo se proyecta con un vector de carga extenso, lo que sugiere una influencia elevada en la varianza explicada. Sin embargo, este efecto es engañoso, pues no refleja una variabilidad homogénea de la variable, sino la distorsión producida por unos pocos valores extremos que inflan artificialmente su importancia estadística.

Como consecuencia, los modelos PLSR enfrentan severas limitaciones al calibrar el fósforo. La combinación de un rango de concentración dominado por valores bajos y la presencia de atípicos extremos obliga al modelo a intentar ajustarse simultáneamente a dos patrones muy distintos, generando un desequilibrio en la calibración. Esto se traduce en una baja capacidad predictiva para la mayoría de las muestras, posicionando al P como una de las variables más problemáticas en espectroscopía MIR. En este sentido, se requieren enfoques alternativos de tratamiento de datos o estrategias de modelado avanzadas que mejoren su interpretabilidad y utilidad práctica.

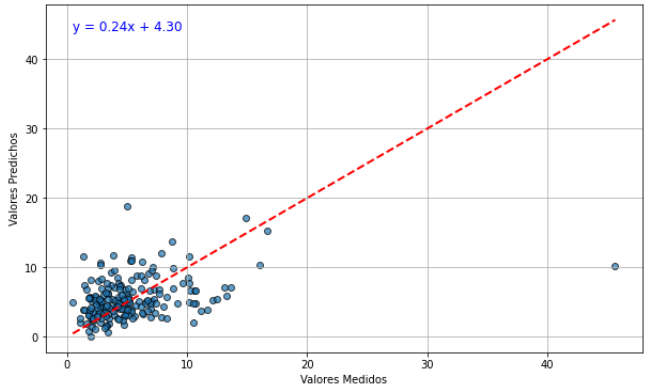


Figura 27: Valores reales vs. predichos para Fósforo con el pretratamiento Detrend

- **RMSE**: 4.28
- **R**^{2}: -0.03

Mejores modelos para la predicción de Magnesio

El magnesio (Mg) constituye un caso intermedio en la modelización quimiométrica mediante PLSR. Comparte ciertas características favorables con el calcio (Ca), como la posibilidad de correlacionarse con minerales del suelo que presentan bandas espectrales claras, pero presenta particularidades que lo diferencian de este y de elementos más problemáticos como el potasio (K) y el fósforo (P). En suelos agrícolas etíopes, la predicción espectral del Mg alcanzó valores de R^2 moderadamente altos (0.84) y RPD 2.6, lo que lo sitúa como un nutriente de desempeño aceptable dentro del contexto de análisis multivariado (Lelago y Bibiso, 2022).

El histograma del Mg (figura 3) evidencia un sesgo hacia valores bajos, aunque menos pronunciado que en K y P, con una cola extendida y regular hacia concentraciones más altas y sin valores atípicos extremos. Esta distribución moderadamente equilibrada ofrece una variabilidad suficiente para la calibración del modelo, permitiendo que el PLSR capture relaciones espectrales consistentes a lo largo del rango de concentración. A diferencia de K y P, donde la mayoría de las observaciones se concentra en rangos muy bajos y unos pocos valores extremos inflan artificialmente la varianza, el Mg proporciona información más representativa y homogénea, lo que fortalece la capacidad del modelo de generalizar a nuevas muestras.

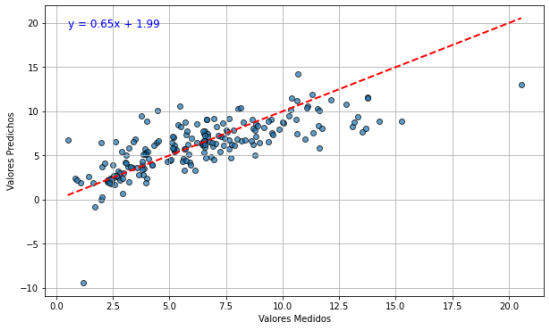
En el espacio de componentes principales (figura 4), el Mg se proyecta claramente hacia el eje

PC1, reflejando una contribución estructurada a la varianza total. Esta proyección indica que, aunque no domine la varianza global como el Ca, su influencia es significativa y proviene de la variabilidad real de los datos, no de valores atípicos extremos. Esto refuerza su posición como un nutriente con comportamiento estadísticamente estable y relativamente predecible dentro del análisis multivariado.

Sin embargo, la predictibilidad del Mg se ve limitada por características propias de su espectro en la región MIR. A diferencia del Ca, el Mg no presenta bandas de absorción directas y definidas; su concentración debe inferirse de manera indirecta a través de asociaciones con carbonatos, ciertos tipos de arcillas y otros componentes del suelo que afectan la reflectancia.

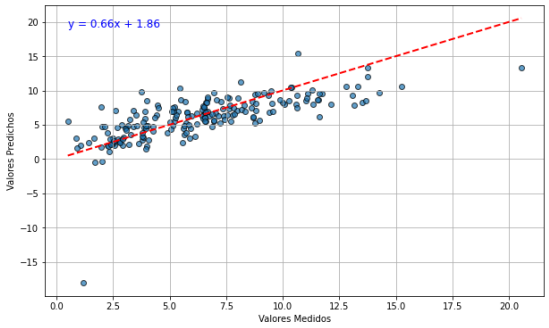
La heterogeneidad mineralógica y textural de los suelos puede enmascarar parcialmente esta señal, reduciendo la capacidad del PLSR para identificar correlaciones lineales precisas a lo largo de todo el rango de concentración. Además, factores como la presencia de distintos minerales portadores de Mg, la humedad del suelo y la distribución de partículas finas contribuyen a incrementar el ruido espectral, dificultando la calibración óptima (Stenberg y Rossel, 2010).

A pesar de estas limitaciones, la combinación de una distribución relativamente equilibrada en la base de datos y la asociación indirecta del Mg con componentes espectrales detectables permite construir modelos de desempeño moderado. Esto demuestra que la predictibilidad de nutrientes no solo depende de la fuerza de las bandas espectrales, sino también de la representatividad y uniformidad de la variable en la base de calibración. En consecuencia, aunque el Mg no alcanza la precisión observada para Ca, su comportamiento intermedio proporciona información valiosa sobre cómo las propiedades químicas y la distribución estadística interactúan para definir el éxito de la modelización espectral (Lelago y Bibiso, 2022; Stenberg y Rossel, 2010).



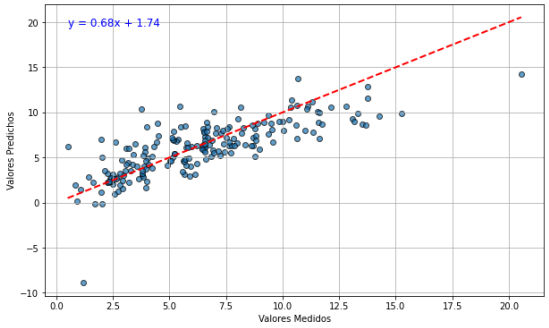
(a) Valores reales vs. predichos para Magnesio con SNV + Savitzky-Golay 2ª Derivada

- **RMSE**: 2.25
- **R**^{2}: 0.56



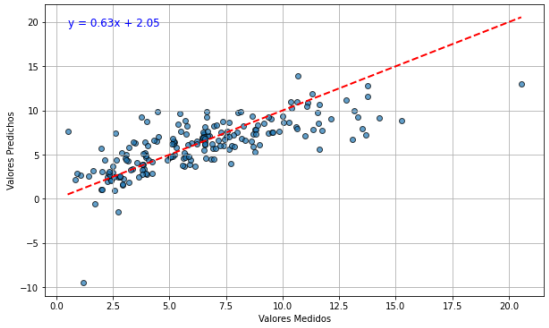
(b) Valores reales vs. predichos para Magnesio con Detrend

- **RMSE**: 2.3
- **R**^{2}: 0.54



(c) Valores reales vs. predichos para Magnesio con SNV

- **RMSE**: 2.37
- **R**^{2}: 0.51



(d) Valores reales vs. predichos para Magnesio con Savitzky-Golay 2ª Derivada + SNV

- **RMSE**: 2.33
- **R**^{2}: 0.52

Figura 28: Mejores modelos para la predicción del Magnesio.

Modelos menos efectivos en la predicción de Magnesio

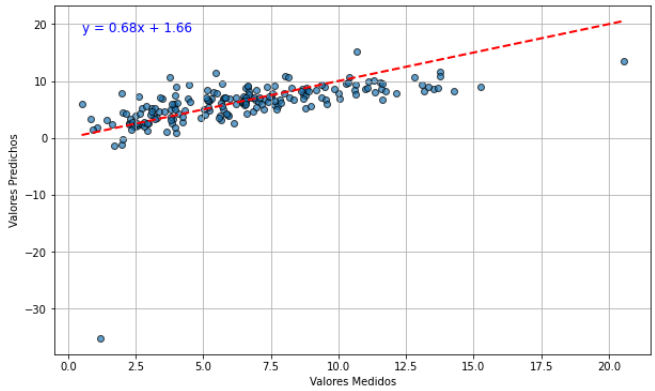


Figura 29: Valores reales vs. predichos para Magnesio con el pretratamiento Savitzky-Golay 2ª Derivada

- **RMSE**: 3.52
- **R^2**: -0.09

Los pretratamientos SNV y SNV combinados con la segunda derivada de Savitzky-Golay demostraron ser los más efectivos, con RMSE de 2.25-2.37 y R² de 0.51-0.56. Estos pretratamientos ayudan a mejorar la precisión del modelo de predicción para el magnesio en el suelo al eliminar variaciones sistemáticas y resaltar las características espectrales relevantes del Mg.

La aplicación del pretratamiento SNV (Standard Normal Variate) es efectiva para eliminar las variaciones sistemáticas debidas a diferencias en el tamaño de las partículas y la dispersión de la luz, normalizando los espectros y reduciendo el ruido no relacionado con las características químicas del magnesio. Al combinar SNV con la segunda derivada de Savitzky-Golay, se resalta aún más las características espectrales relevantes, mejorando así la precisión del modelo de predicción de Mg (Hati et al., 2022).

La combinación de SNV con derivadas permite una mejor identificación de las señales espectrales asociadas con el magnesio. La segunda derivada de Savitzky-Golay es particularmente útil para detectar y separar los picos espectrales relevantes de Mg, mejorando la capacidad del modelo para predecir la concentración de este nutriente en el suelo (Jakkan et al., 2023).

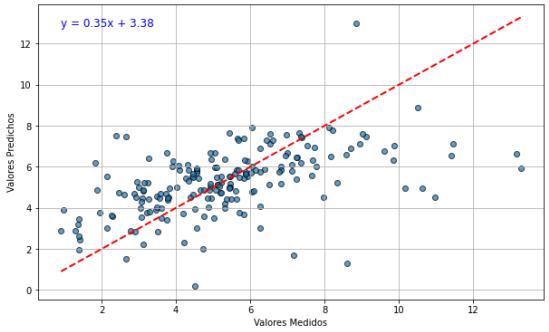
El uso de la segunda derivada de Savitzky-Golay sin SNV resultó en los peores resultados, con un RMSE de 3.52 y un R^2 de -0.09. Esto se debe a la amplificación del ruido que ocurre cuando se aplica la derivada sin un pretratamiento adecuado para normalizar y limpiar los datos espectrales. La segunda derivada puede intensificar tanto el ruido como las características no relacionadas, lo que afecta negativamente la precisión del modelo de predicción (Hati et al., 2022).

Mejores modelos para la predicción de Materia Orgánica

La Materia Orgánica (MO) se destaca en la literatura como uno de los componentes con mayor potencial de predicción espectral mediante modelos PLSR, mostrando un comportamiento especialmente favorable que comparte estrechamente con elementos como el calcio (Ca) y el magnesio (Mg). Esta condición se explica tanto por la naturaleza de sus datos de laboratorio como por sus características espectrales en la región del infrarrojo medio (MIR).

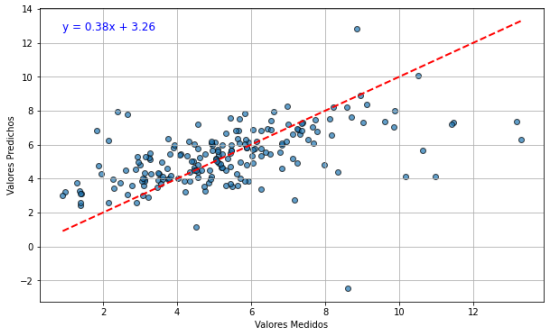
En este caso, la MO a pesar de su potencial como predictor, no pudo destacarse debido a limitaciones en la base de datos de calibración. Estudios como el realizado en suelos etíopes muestran que, para lograr resultados estables en la predicción de carbono orgánico (OC), es indispensable contar con un muestreo que cubra adecuadamente la variabilidad del parámetro y una proporción adecuada de muestras para calibración en torno al 20 %, siempre que estas muestras sean representativas En este estudio, sin embargo, la cobertura de concentraciones de MO no fue suficientemente homogénea ni suficientemente amplia, lo que impidió que el modelo capturara relaciones espectrales consistentes (Lelago y Bibiso, 2022).

Adicionalmente, la calidad del preprocesamiento de muestras afecta la precisión del modelo. La predicción de carbono orgánico mediante MIR puede verse comprometida si los suelos no son procesados adecuadamente (secado, tamizaje), ya que el ruido espectral aumenta y la señal se dispersa, aun cuando la composición molecular permita una buena detección. Esta combinación de falta de variabilidad representativa y posible inconsistencia en la preparación de muestras explica la baja capacidad predictiva de la MO en los modelos PLSR de este estudio (Żelazny y Šimon, 2022).



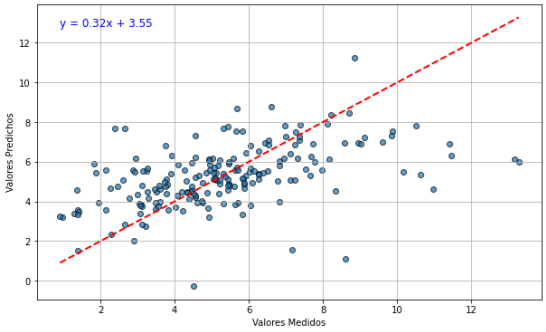
(a) Valores reales vs. predichos para Materia orgánica con Savitzky-Golay 2ª Derivada + SNV

- **RMSE**: 1.89
- **R**^{2}: 0.32



(b) Valores reales vs. predichos para Materia orgánica con SNV

- **RMSE**: 1.91
- **R**^{2}: 0.30

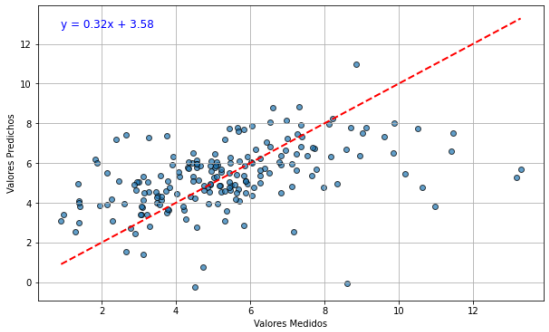


(c) Valores reales vs. predichos para Materia orgánica con Detrend

- **RMSE**: 1.98
- **R**^{2}: 0.24

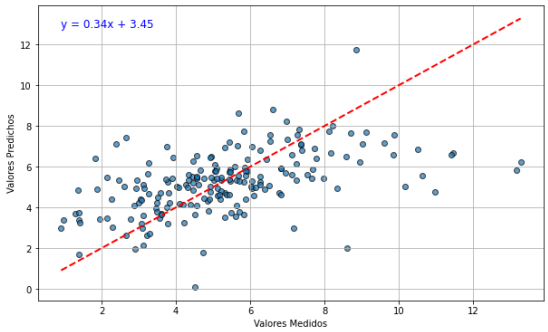
Figura 30: Mejores modelos para la predicción de la Materia orgánica.

Modelos menos efectivos para la predicción de Materia Orgánica



(a) Valores reales vs. predichos para Materia orgánica con Savitzky-Golay 2ª Derivada

- **RMSE**: 2.09
- **R**^{2}: 0.16



(b) Valores reales vs. predichos para Materia orgánica con Savitzky-Golay 1ª Derivada

- **RMSE**: 2.05
- **R**^{2}: 0.19

Figura 31: Peores modelos para la predicción de materia orgánica.

Los pretratamientos SNV y su combinación con la primera derivada de Savitzky-Golay demostraron ser los más efectivos, con RMSE de 1.89-1.91 y R² de 0.3-0.32. Estos pretratamientos son eficaces para mejorar la uniformidad de los datos espectrales y eliminar efectos no deseados, permitiendo que el modelo capture mejor las señales espectrales asociadas con la materia orgánica.

El pretratamiento SNV (Standard Normal Variate) ayuda a corregir las variaciones sistemáticas causadas por diferencias en el tamaño de las partículas y la dispersión de la luz. Este proceso normaliza los espectros, reduciendo el ruido y mejorando la consistencia de los datos espectrales. Al combinar SNV con la primera derivada de Savitzky-Golay, se resaltan las características espectrales relevantes y se eliminan tendencias lineales no deseadas, mejorando así la precisión del modelo de predicción de MO (Wu et al., 2024).

El uso de la segunda derivada de Savitzky-Golay sin SNV resultó en los peores resultados, con un RMSE de hasta 2.09 y R² de 0.16. Esto se debe a la amplificación del ruido que ocurre cuando se aplica la derivada sin un pretratamiento adecuado para normalizar los datos. La segunda derivada intensifica tanto el ruido como las características no relacionadas con la materia orgánica, afectando negativamente la precisión del modelo de predicción (Li et al., 2022).

La materia orgánica tiene una influencia significativa en varias propiedades del suelo, como la estructura, la capacidad de retención de agua y la fertilidad química. Estos efectos amplifican las señales espectrales asociadas, haciendo que la MO sea más detectable y predecible mediante técnicas espectroscópicas avanzadas (Li et al., 2022).

Los mapas de calor presentados (Figuras 32 y 33) ofrecen una visión global de la eficacia de los diferentes modelos y pretamientos empleados para predecir los nutrientes de interés. En la Figura 32, los valores de R^2 permiten identificar de manera rápida las combinaciones de pretamiento y nutriente que producen mejores ajustes, mientras que en la Figura 33 se observa la magnitud del error de predicción (RMSE) en cada caso.

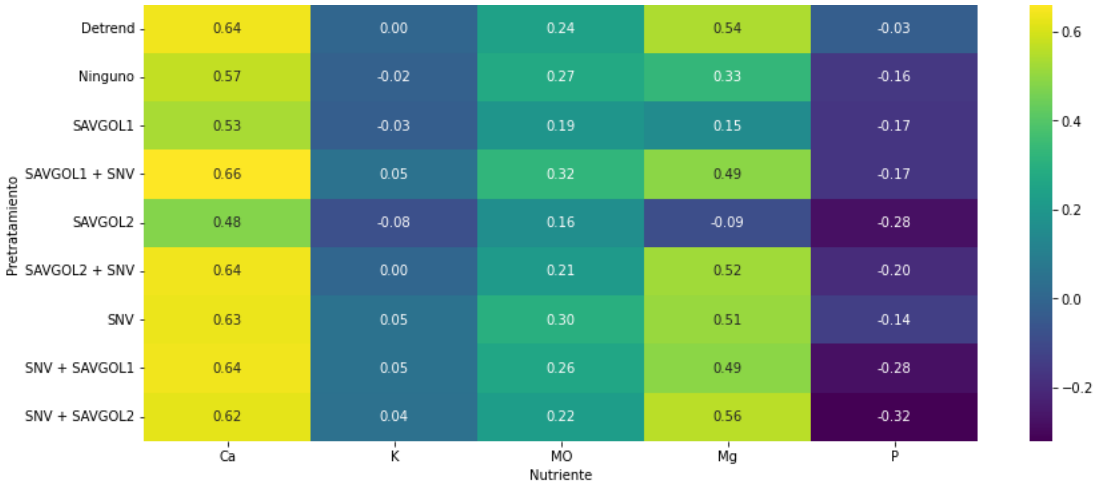


Figura 32: Mapa de calor de todos los valores de R^2 clasificados por nutriente y pretratamiento

La comparación de ambos mapas de calor revela que, para calcio y magnesio, las estrategias de pretamiento que combinan filtros de suavizado (SAVGO o SAVGO+SNV) tienden a ofrecer un mejor equilibrio entre precisión y estabilidad del modelo. En contraste, para fósforo y potasio, la mayoría de las combinaciones presentan ajustes mucho más discretos, lo que confirma la complejidad propia a la detección espectral de estos nutrientes.

Este panorama consolidado permite, por un lado, identificar los enfoques más prometedores para cada nutriente y, por otro, evidenciar la necesidad de ajustes o métodos adicionales, especialmente en el caso de fósforo y potasio para mejorar la capacidad predictiva de los modelos.

La visualización conjunta de R^2 y RMSE (figura 34) facilita la toma de decisiones sobre qué

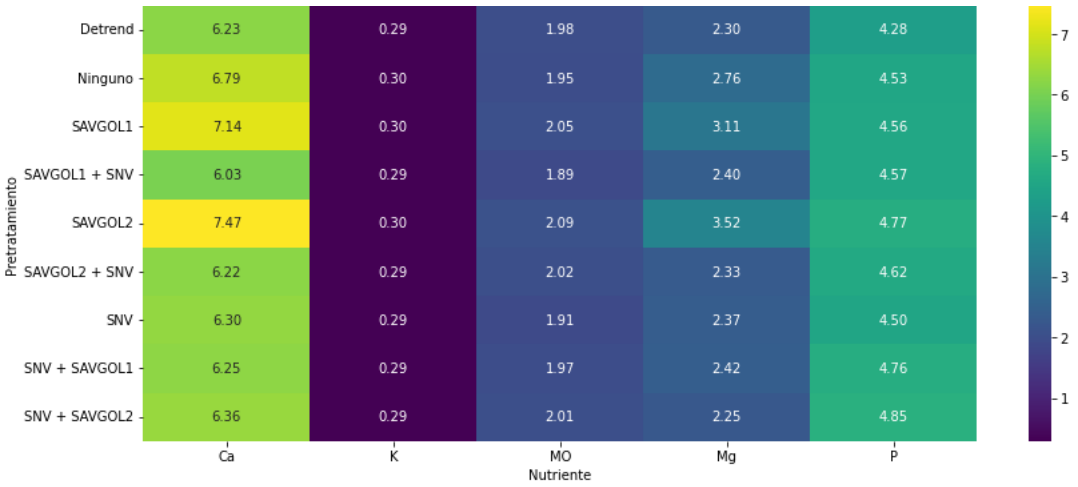


Figura 33: Mapa de calor de todos los valores de RMSE clasificados por nutriente y pretratamiento

configuraciones de pretamiento resultan más adecuadas para un objetivo de modelado específico (maximizar la precisión o minimizar el error), sirviendo como guía para el refinamiento de los métodos espectrales en futuros trabajos.

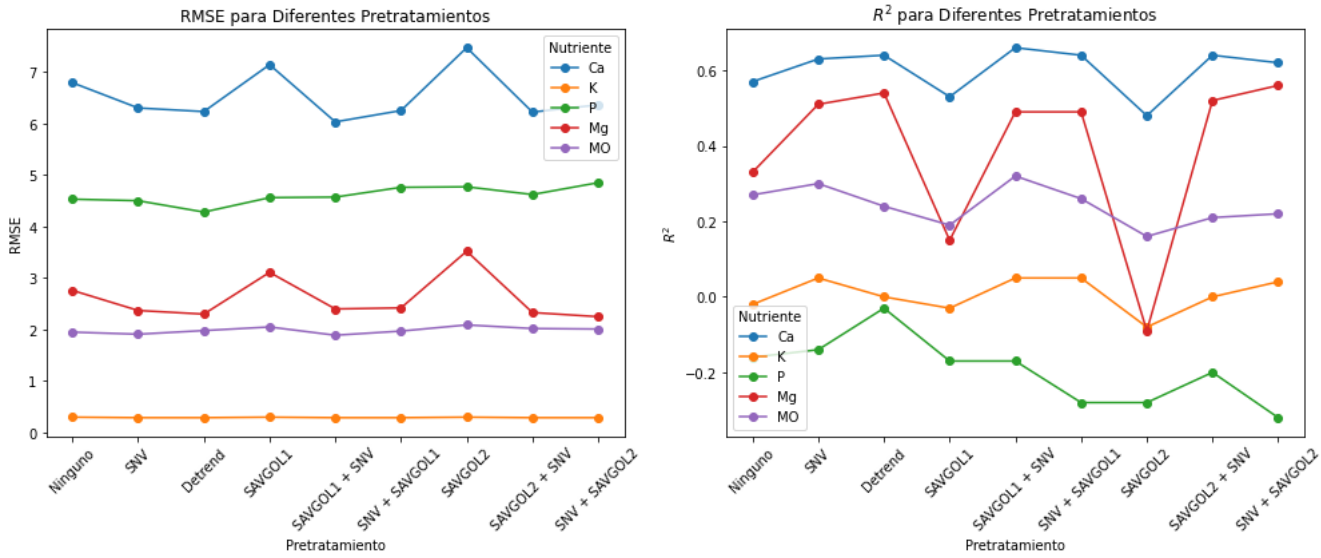


Figura 34: Resumen lineal de los parámetros de rendimiento R² y RMSE para todos los pretratamientos

Aunado a la interpretación basada en R² y RMSE, la incorporación del RPD (Ratio of Performance to Deviation) proporciona una perspectiva complementaria para evaluar la calidad de los

modelos. El RPD compara la desviación estándar de la variable objetivo con el RMSE de la predicción.

El autor (Chang, Laird, Mausbach, y Hurburgh, 2001) propone una clasificación del poder predictivo de la espectroscopía NIRS (y, por extensión, MIR) utilizando el valor del Ratio de Desviación de Predicción (RPD), dividiéndolo en tres categorías bien definidas según el desempeño del modelo.

Tabla 5: Clasificación del desempeño del modelo según RPD

RPD	**Categoría**	**Interpretación**
> 2	A	Alto rendimiento
1.4 – 2.0	B	Aceptable
< 1.4	C	Bajo rendimiento

Basado en la figura 35, los valores de RPD obtenidos para cada nutriente bajo las distintas estrategias de pretratamiento indican una variabilidad en el desempeño de los modelos. Por ejemplo, los modelos para calcio y magnesio presentan valores de RPD cercanos a 1.7 y 1.5 respectivamente, lo que los ubica en la **Categoría B** (RPD entre 1.4 y 2.0). Esta categoría también está asociada con valores del coeficiente de determinación (r^2) entre 0.50 y 0.80, rango que efectivamente se observa en varias combinaciones de Ca y Mg dentro del presente estudio. Esta etiqueta no garantiza que la predicción sea precisa o confiable para todos los usos. En realidad, estos modelos tienden a ser limitados en su exactitud y presentan errores que pueden ser significativos respecto al rango total de la variable medida. Por esta razón, su utilidad práctica se restringe principalmente a aplicaciones donde no se requiere una estimación exacta, sino más bien una orientación general.

Los modelos con clasificación aceptable, podrían ser útiles para monitorear tendencias, clasificar muestras en categorías amplias (bajo, medio, alto) o detectar valores atípicos que merezcan un análisis más detallado. Sin embargo, no son recomendables para decisiones críticas o aplicaciones que demanden alta precisión, como de fertilización precisas o control de nutrientes con tolerancias estrictas. En este sentido, la clasificación de “aceptable” refleja más la capacidad del modelo de capturar patrones generales que su capacidad de ofrecer predicciones exactas.

En contraste, los modelos desarrollados para potasio, fósforo y materia orgánica presentan valores de RPD entre 0.7 y 1.1, ubicándose en la **Categoría C** (RPD < 1.4). Esto sugiere que, para estas propiedades, el modelo no logra un desempeño predictivo confiable, ya que los valores de r^2 también suelen ser inferiores a 0.50, lo que limita la capacidad de estimación cuantitativa a partir de la información espectral.

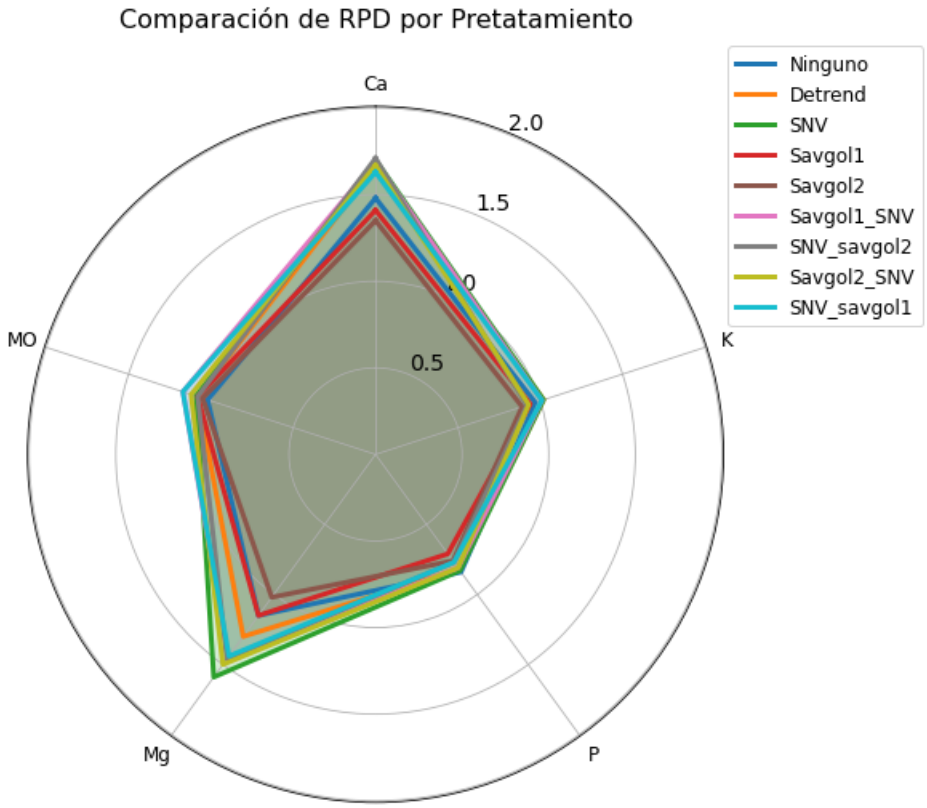


Figura 35: Valores de RPD clasificados por pretratamiento y nutriente

La tabla 6 respalda la idea de que los modelos basados en la señal espectral del calcio tienen un desempeño relativamente superior en comparación con los desarrollados para otros nutrientes, validando que una distribución de concentración más simétrica y amplia, como la observada en su histograma, proporciona al algoritmo PLSR una base de datos aceptable y consistente para identificar correlaciones lineales fiables con las características espectrales, resultando en una mejor capacidad predictiva, bajo el contexto de aplicación mencionado

Tabla 6: Top 5 combinaciones con mayores valores de RPD

Pretratamiento	**Nutriente**	**RPD**
SNV_savgol2	Ca	1.71
SNV	Ca	1.70
Savgol1_SNV	Ca	1.69
Savgol2_SNV	Ca	1.67
Detrend	Ca	1.66

Los resultados de la tabla 7 resaltan la complejidad ya mencionada en la detección espectral del fósforo, lo cual se traduce en modelos con un desempeño predictivo muy pobre para este nutriente, en comparación con otros como el calcio o el magnesio. Esta dificultad es directamente atribuible a la naturaleza de su distribución de concentración, altamente sesgada. Esta configuración limita severamente la capacidad del algoritmo PLSR para establecer relaciones aceptables y consistentes con las señales espectrales, resultando en la baja capacidad predictiva observada.

Tabla 7: Top 5 peores combinaciones según RPD

Pretratamiento	**Nutriente**	**RPD**
Savgol1	P	0.71
Savgol2	P	0.76
SNV_savgol2	P	0.77
SNV_savgol1	P	0.77
Savgol2_SNV	P	0.81

Es importante destacar que también se desarrollaron modelos utilizando la totalidad de los datos espectrales, sin aplicar el promediado cada 20 mediciones. Esta estrategia se implementó con el objetivo de evaluar el impacto que dicho promediado podría tener sobre el desempeño de los modelos. No obstante, los resultados obtenidos bajo esta condición fueron considerablemente menos favorables en comparación con los presentados previamente en esta sección. Los detalles y resultados de estos modelos se incluyen en el apartado de anexos para su consulta.

5. Conclusiones

- Los histogramas revelaron que los nutrientes que presentan distribuciones de concentración muy variadas como el Calcio (Ca), Materia Orgánica (MO) y Magnesio (Mg) muestran distribuciones más amplias y simétricas, mientras que Potasio (K) y Fósforo (P) exhiben un sesgo extremo a la izquierda, con la mayoría de los datos concentrados en valores bajos.

- En el Análisis de Componentes Principales de los datos de laboratorio, los largos vectores de carga de P y K no reflejan una variabilidad uniformemente distribuida, sino la desproporcionada influencia de unos pocos valores atípicos. Esto contrasta con la variabilidad más manejable de Ca y MO.

- Los resultados subrayan la importancia de seleccionar y optimizar los métodos de pretratamiento para cada elemento. Las combinaciones de pretratamientos como SNV y derivadas de Savitzky-Golay jugaron un papel clave en mejorar la precisión y fiabilidad de los modelos predictivos, especialmente para el calcio.

- La capacidad predictiva de los modelos PLSR está fuertemente correlacionada con la naturaleza de la distribución del nutriente. Nutrientes con distribuciones más equilibradas (Ca, Mg) tienden a generar modelos aceptables debido a la consistencia de la señal espectral a lo largo de su rango.

- Las distribuciones altamente sesgadas de K y P (con dominancia de valores bajos y valores atípicos extremos) representan un desafío considerable para el modelado PLSR, resultando en una baja capacidad predictiva a pesar de su aparente contribución a la varianza total en el PCA.

- El modelo predictivo para Calcio (Ca), aunque presentó el mejor desempeño con un RPD de 1.7 utilizando SNV combinado con Savitzky-Golay de segunda derivada, muestra limitaciones importantes en su precisión cuantitativa. Los valores de RMSE y r^2 indican que el error promedio es considerable respecto al rango de la variable, por lo que su aplicación requiere cautela y no es recomendable para estimaciones exactas de concentración de calcio. El modelo puede resultar útil en contextos donde no se requiera una alta exacti-

tud, como en el monitoreo rápido de tendencias, la clasificación general de muestras o la detección de valores atípicos que ameriten análisis más detallados. Esto sugiere que, aunque no sustituye métodos de laboratorio, puede ser un apoyo práctico para evaluaciones preliminares o de carácter exploratorio.

- Los modelos predictivos son altamente dependientes de las metodologías de extracción utilizadas para el calcio, fósforo, potasio, magnesio y materia orgánica. Cualquier variación en estos métodos o técnicas influirá significativamente en los resultados del modelo, destacando la importancia de estandarizar los procedimientos de extracción.

6. Recomendaciones

- Ampliar la recolección de muestras de suelo de diferentes lugares, no limitándose únicamente al Cantón de Nicoya. Esto permitiría enriquecer los modelos predictivos con un espectro más amplio de variabilidad del suelo, mejorando su representatividad y precisión para aplicaciones en diversas condiciones agrícolas y regiones geográficas.
- Incrementar el número de muestras para los conjuntos de validación y calibración. Aumentar la cantidad de muestras en estos conjuntos mejoraría significativamente la calidad predictiva de los modelos.
- Fomentar colaboraciones interdisciplinarias con expertos en otras áreas, como la ecología, la biología y la climatología, para enriquecer el enfoque y las metodologías utilizadas en la investigación del suelo. Estas colaboraciones pueden abrir nuevas perspectivas y métodos de análisis que mejoren la comprensión y gestión de los recursos del suelo.
- Explorar técnicas más avanzadas de aprendizaje automático y análisis de datos, como redes neuronales profundas o máquinas de soporte vectorial, PLSR locales, que podrían capturar complejidades no lineales e interacciones entre variables más efectivamente que los modelos tradicionales.

Referencias

Abdel-Fattah, M. K., Mohamed, E. S., Wagdi, E. M., Shahin, S. A., Aldosari, A. A., Lasaponara, R., y Alnaimy, M. A. (2021). Quantitative evaluation of soil quality using principal component analysis: The case study of El-Fayoum depression Egypt. *Sustainability (Switzerland)*, *13*(4), 1–19. doi: 10.3390/su13041824

Anderson, L. D., Liu, B., Balser, D. S., Bania, T. M., Haffner, L. M., Linville, D. J., ... Wenger, T. V. (2023, oct). Methods for Averaging Spectral Line Data. (1999). Descargado de http://arxiv.org/abs/2310.09076

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.-M., y McBratney, A. (2010, oct). Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *TrAC Trends in Analytical Chemistry*, *29*(9), 1073–1081. Descargado de http://dx.doi.org/10.1016/j.trac.2010.05.006https://linkinghub.elsevier.com/retrieve/pii/S0165993610001585 doi: 10.1016/j.trac.2010.05.006

Bertsch, F. (2012). *Guía para la interpretación de la fertilidad de los suelos de costa rica*. Editorial Universidad de Costa Rica. Descargado de https://www.researchgate.net/publication/279172745_Guia_para_la_interpretacion_de_la_fertilidad_de_los_suelos_de_Costa_Rica

Bertsch, F., Bejarano, J. A., y Corrales, M. (2005). Correlación entre las soluciones extractoras kcl-olsen modificado y mehlich 3, usadas en los laboratorios de suelos de costa rica. *Agronomía Costarricense*, *29*(3), 137–142. Descargado de https://www.redalyc.org/pdf/436/43626961016.pdf

Bonilla Segovia, J. S., Dávila Rojas, F. A., y Villa Quishpe, M. W. (2021, jan). Estudio del uso de técnicas de inteligencia artificial aplicadas para análisis de suelos para el sector agrícola. *RECIMUNDO*, *5*(1), 4–19. Descargado de http://recimundo.com/index.php/es/article/view/978 doi: 10.26820/recimundo/5.(1).enero.2021.4-19

Cabalceta, G., y Cordero, A. (1994). Niveles críticos de fósforo en ultisoles, inceptisoles, vertisoles y andisoles de costa rica. *Agronomía Costarricense*, *18*(2), 147–161. Descargado

de https://www.mag.go.cr/rev_agr/v18n02_147.pdf

Cao, H., Gu, Y., Hu, Y., Wang, X., Ding, W., Chen, G., ... He, H. (2023). Mid-infrared spectroscopy coupled with chemometrics for quantitative determination of biomaterial activity. *Optik*, *281*(August 2022), 170854. Descargado de https://doi.org/10.1016/j.ijleo.2023.170854 doi: 10.1016/j.ijleo.2023.170854

Chang, C.-W., Laird, D. A., Mausbach, M. J., y Hurburgh, C. R. (2001, mar). Near-Infrared Reflectance Spectroscopy–Principal Components Regression Analyses of Soil Properties. *Soil Science Society of America Journal*, *65*(2), 480–490. Descargado de https://acsess.onlinelibrary.wiley.com/doi/10.2136/sssaj2001.652480x doi: 10.2136/sssaj2001.652480x

Cheng, J.-H., y Sun, D.-W. (2017). Partial Least Squares Regression (PLSR) Applied to NIR and HSI Spectral Data Modeling to Predict Chemical Properties of Fish Muscle. *Food Engineering Reviews*, *9*(1), 36–49. doi: 10.1007/s12393-016-9147-1

Chinchilla, M., Mata, R., y Alvarado, A. (2011, jun). Andisoles, inceptisoles y entisoles de la subcuenca del río Pirrís, región de Los Santos, Talamanca, Costa Rica. *Agronomía Costarricense*, *35*(1), 83–107. Descargado de https://revistas.ucr.ac.cr/index.php/agrocost/article/view/6688 doi: 10.15517/rac.v35i1.6688

Corrales, M., Bertsch, F., y Bejarano, J. A. (2005). Los laboratorios de análisis de suelos y foliares en costa rica: Informe del comité de laboratorios de análisis de suelos, plantas y aguas. *Agronomía Costarricense*, *29*(3), 125–135. Descargado de https://www.redalyc.org/pdf/436/43626961015.pdf

Davies, A. M. C., y Fearn, T. (2004). TONY DAVIES COLUMN Back to basics : the principles of principal component analysis. *Tpny Davies Column*.

de Palaminy, L., Daher, C., y Moulherat, C. (2022, apr). Development of a non-destructive methodology using ATR-FTIR and chemometrics to discriminate wild silk species in heritage collections. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *270*, 120788. Descargado de https://doi.org/10.1016/j.saa.2021.120788https://linkinghub.elsevier.com/retrieve/pii/S1386142521013652 doi: 10.1016/j.saa.2021.120788

Ditta, A., Nawaz, H., Mahmood, T., Majeed, M. I., Tahir, M., Rashid, N., ... Byrne, H. J.

(2019). Principal components analysis of Raman spectral data for screening of Hepatitis C infection. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, *221*, 117173. Descargado de https://doi.org/10.1016/j.saa.2019.117173 doi: 10.1016/j.saa.2019.117173

Dotto, A. C., Dalmolin, R. S. D., ten Caten, A., y Grunwald, S. (2018). A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra. *Geoderma*, *314*(May 2017), 262–274. Descargado de https://doi.org/10.1016/j.geoderma.2017.11.006 doi: 10.1016/j.geoderma.2017.11.006

FAO. (2022). *A primer on soil analysis using visible and near-infrared (vis-NIR) and mid-infrared (MIR) spectroscopy*. Autor. Descargado de https://doi.org/10.4060/cb9005enhttp://www.fao.org/documents/card/en/c/cb9005en doi: 10.4060/cb9005en

Guerrero, A., y Bertsch, F. (2020). *Primer informe del ejercicio de intercomparación de la red latinoamericana de laboratorios de suelos - latsolan.* Informe técnico. Descargado de https://openknowledge.fao.org/server/api/core/bitstreams/4c0db9e3-6eae-4dee-aa1b-dd9565c5d861/content

Hati, K. M., Sinha, N. K., Mohanty, M., Jha, P., Londhe, S., Sila, A., ... Chaudhari, S. K. (2022, apr). Mid-Infrared Reflectance Spectroscopy for Estimation of Soil Properties of Alfisols from Eastern India. *Sustainability*, *14*(9), 4883. Descargado de https://www.mdpi.com/2071-1050/14/9/4883 doi: 10.3390/su14094883

Huang, H., Fang, Z., Xu, Y., Lu, G., Feng, C., Zeng, M., ... Zhao, Z. (2024). Stacking and ridge regression-based spectral ensemble preprocessing method and its application in near-infrared spectral analysis. *Talanta*, *276*(February), 126242. Descargado de https://doi.org/10.1016/j.talanta.2024.126242 doi: 10.1016/j.talanta.2024.126242

Hunter, A. H. (1975). New techniques and equipment for routine plant analytical procedure. En E. Bornemisza y A. Alvarado (Eds.), *Soil management in tropical america* (pp. 467–483). Raleigh, NC: North Carolina State University. Descargado de https://edepot.wur.nl/485014

Jakkan, D. A., Ghare, P., y Sakode, C. (2023). Multi-parameter Soil Property Prediction Incorpo-

rating Mid-infrared Spectroscopy and Dropout Sequential Artificial Neural Network. *Water, Air, and Soil Pollution*, *234*(11), 1–18. Descargado de https://doi.org/10.1007/s11270-023-06726-6 doi: 10.1007/s11270-023-06726-6

Ji, W., Adamchuk, V. I., Biswas, A., Dhawale, N. M., Sudarsan, B., Zhang, Y., ... Shi, Z. (2016, dec). Assessment of soil properties in situ using a prototype portable MIR spectrometer in two agricultural fields. *Biosystems Engineering*, *152*, 14–27. Descargado de http://dx.doi.org/10.1016/j.biosystemseng.2016.06.005https://linkinghub.elsevier.com/retrieve/pii/S1537511015304633 doi: 10.1016/j.biosystemseng.2016.06.005

Kandpal, L. M., Munnaf, M. A., Cruz, C., y Mouazen, A. M. (2022). Spectra Fusion of Mid-Infrared (MIR) and X-ray Fluorescence (XRF) Spectroscopy for Estimation of Selected Soil Fertility Attributes. *Sensors*, *22*(9). doi: 10.3390/s22093459

Karray, E., Elmannai, H., Toumi, E., Gharbia, M. H., Meshoul, S., Aichi, H., y Rabah, Z. B. (2023). Evaluating the Potentials of PLSR and SVR Models for Soil Properties Prediction Using Field Imaging, Laboratory VNIR Spectroscopy and Their Combination. *CMES - Computer Modeling in Engineering and Sciences*, *136*(2), 1399–1425. doi: 10.32604/cmes.2023.023164

Lelago, A., y Bibiso, M. (2022, mar). Performance of mid infrared spectroscopy to predict nutrients for agricultural soils in selected areas of Ethiopia. *Heliyon*, *8*(3), e09050. Descargado de https://linkinghub.elsevier.com/retrieve/pii/S2405844022003383 doi: 10.1016/j.heliyon.2022.e09050

Li, H., Wang, J., Zhang, J., Liu, T., Acquah, G. E., y Yuan, H. (2022). Combining Variable Selection and Multiple Linear Regression for Soil Organic Matter and Total Nitrogen Estimation by DRIFT-MIR Spectroscopy. *Agronomy*, *12*(3). doi: 10.3390/agronomy12030638

Mammadov, E., Denk, M., Mamedov, A. I., y Glaesser, C. (2024). Predicting Soil Properties for Agricultural Land in the Caucasus Mountains Using Mid-Infrared Spectroscopy. *Land*, *13*(2), 1–25. doi: 10.3390/land13020154

Martínez, F., y Gutierrez, R. (2021). *Recuperación de macronutrientes (n, p, k) mediante la bioestimulación de compost y guano de isla en suelos agrícolas contaminados por un plaguicida (superfuran)* (Tesis Doctoral, Universidad Andina del Cusco). Des-

cargado de http://repositorio.ucv.edu.pe/bitstream/handle/20.500.12692/47102/Gutierrez_RS-SD.pdf?sequence=1&isAllowed=y

Mata, R., Rosales, A., Sandoval, D., Vindas, E., y Alemán, B. (2020). *Mapa de suelos de costa rica (Órdenes).* Centro de Investigaciones Agronómicas, Universidad de Costa Rica. (Escala 1:200,000. San José, Costa Rica)

Metz, M., Abdelghafour, F., Roger, J. M., y Lesnoff, M. (2021). A novel robust PLS regression method inspired from boosting principles: RoBoost-PLSR. *Analytica Chimica Acta*, *1179*, 338823. Descargado de https://doi.org/10.1016/j.aca.2021.338823 doi: 10.1016/j.aca.2021.338823

Molina, E., y Bornemisza, E. (2006). Nivel crítico de zinc en suelos de costa rica. *Agronomía Costarricense*, *30*(2), 45–59. Descargado de https://www.redalyc.org/pdf/436/43630204.pdf

Molina, E., y Cabalceta, G. (1990). Correlación de diferentes soluciones extractoras en vertisoles y ultisoles de costa rica. *Agronomía Costarricense*, *14*(1), 37–44. Descargado de https://www.mag.go.cr/rev_agr/v14n01_037.pdf

Nath, D., Laik, R., Meena, V. S., Pramanick, B., y Singh, S. K. (2021, sep). Can mid-infrared (mid-IR) spectroscopy evaluate soil conditions by predicting soil biological properties? *Soil Security*, *4*(May), 100008. Descargado de https://doi.org/10.1016/j.soisec.2021.100008https://linkinghub.elsevier.com/retrieve/pii/S2667006221000058 doi: 10.1016/j.soisec.2021.100008

Oberholzer, S., Summerauer, L., Steffens, M., y Ifejika Speranza, C. (2023, jun). *Dataset variability and carbonate concentration influence the performance of local visible-near infrared spectral models* (n.º June). Descargado de https://egusphere.copernicus.org/preprints/2023/egusphere-2023-1087/ doi: 10.5194/egusphere-2023-1087

ONU. (2021). *Ods costa rica.* Descargado de https://ods.cr/

Perret, J., Villalobos Leandro, J. E., Abdalla Bolaños, K., Fuentes Fallas, C. L., Cuarezma Espinoza, K. M., Macas Amaya, E. N., ... Drewry, D. (2020, jul). Desarrollo de métodos de análisis de espectroscopia y algoritmos de aprendizaje automático para la evaluación de algunas propiedades del suelo en Costa Rica. *Agronomía Costarricense*, *44*(2), 139–154. Descargado de https://revistas.ucr.ac.cr/index.php/

agrocost/article/view/43108 doi: 10.15517/rac.v44i2.43108

Sadzawka, A., Carrasco, M., Grez, R., Mora, M., Flores, H., y Neaman, A. (2006). *Soil Methods Recommended for the Soils of Chile*. Instituto de Investigaciones Agropecuarias (INIA). Descargado de https://biblioteca.inia.cl/items/9ec1aab6-f9aa-4e4b-b2c5-28d1fed798c3

Seema, Ghosh, A., Mouli Hati, K., Kumar Sinha, N., Mridha, N., y Sahu, B. (2022, dec). Regional soil organic carbon prediction models based on a multivariate analysis of the Mid-infrared hyperspectral data in the middle Indo-Gangetic plains of India. *Infrared Physics Technology*, *127*(September), 104372. Descargado de https://doi.org/10.1016/j.infrared.2022.104372https://linkinghub.elsevier.com/retrieve/pii/S135044952200353X doi: 10.1016/j.infrared.2022.104372

Shan, P., Zhao, Y., Wang, Q., Ying, Y., y Peng, S. (2020). Principal component analysis or kernel principal component analysis based joint spectral subspace method for calibration transfer. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, *227*, 117653. Descargado de https://doi.org/10.1016/j.saa.2019.117653 doi: 10.1016/j.saa.2019.117653

Shepherd, K. D., y Walsh, M. G. (2002). Development of Reflectance Spectral Libraries for Characterization of Soil Properties. *Soil Science Society of America Journal*, *66*(3), 988. Descargado de https://www.soils.org/publications/sssaj/abstracts/66/3/988 doi: 10.2136/sssaj2002.0988

Shi, Z., Yin, J., Li, B., Sun, F., Miao, T., Cao, Y., ... Ji, W. (2023). Comparison of Depth-Specific Prediction of Soil Properties: MIR vs. Vis-NIR Spectroscopy. *Sensors*, *23*(13). doi: 10.3390/s23135967

Shin, J., Kim, D. C., Cho, Y., Yang, M., y Cho, W. J. (2024). A Preprocessing Technique Using Diffuse Reflectance Spectroscopy to Predict the Soil Properties of Paddy Fields in Korea. *Applied Sciences (Switzerland)*, *14*(11). doi: 10.3390/app14114673

Silalahi, D. D., Midi, H., Arasan, J., Mustafa, M. S., y Caliman, J.-P. (2018, jul). Robust generalized multiplicative scatter correction algorithm on pretreatment of near infrared spectral data. *Vibrational Spectroscopy*, *97*(May), 55–65. Descargado de https://linkinghub

.elsevier.com/retrieve/pii/S0924203117303508 doi: 10.1016/j.vibspec.2018.05.002

Staff, S. S. (2022). *Kellogg soil survey laboratory methods manual. soil survey investigations report no. 42, version 6.0.* Descargado de https://www.nrcs.usda.gov/resources/guides-and-instructions/kellogg-soil-survey-laboratory-methods-manual

Stenberg, B., y Rossel, R. V. (2010). Diffuse Reflectance Spectroscopy for High-Resolution Soil Sensing. En *Proximal soil sensing* (Vol. 35, pp. 29–47). Dordrecht: Springer Netherlands. Descargado de http://link.springer.com/10.1007/978-90-481-8859-8_3 doi: 10.1007/978-90-481-8859-8_3

Thomas, G. (1982). Exchangeable cations. En A. Page, R. Miller, y D. Keeney (Eds.), *Methods of soil analysis. part 2. chemical and microbiological properties* (2.ª ed., pp. 159–165). Madison, WI: American Society of Agronomy and Soil Science Society of America. Descargado de https://acsess.onlinelibrary.wiley.com/doi/10.2134/agronmonogr9.2.c8

Tsagkaris, A., Bechynska, K., Ntakoulas, D., Pasias, I., Weller, P., Proestos, C., y Hajslova, J. (2023, jun). Investigating the impact of spectral data pre-processing to assess honey botanical origin through Fourier transform infrared spectroscopy (FTIR). *Journal of Food Composition and Analysis*, *119*(January), 105276. Descargado de https://doi.org/10.1016/j.jfca.2023.105276https://linkinghub.elsevier.com/retrieve/pii/S0889157523001503 doi: 10.1016/j.jfca.2023.105276

Vestergaard, R.-J., Vasava, H., Aspinall, D., Chen, S., Gillespie, A., Adamchuk, V., y Biswas, A. (2021, oct). Evaluation of Optimized Preprocessing and Modeling Algorithms for Prediction of Soil Properties Using VIS-NIR Spectroscopy. *Sensors*, *21*(20), 6745. Descargado de https://www.mdpi.com/1424-8220/21/20/6745 doi: 10.3390/s21206745

Walkley, A., y Black, I. A. (1934). An examination of the degtjareff method for determining soil organic matter and a proposed modification of the chromic acid titration method. *Soil Science*, *37*, 29–38. doi: 10.1097/00010694-193401000-00003

Wu, M., Huang, Y., Zhao, X., Jin, J., y Ruan, Y. (2024). Effects of different spectral processing

methods on soil organic matter prediction based on VNIR-SWIR spectroscopy in karst areas, Southwest China. *Journal of Soils and Sediments*, *24*(2), 914–927. Descargado de https://doi.org/10.1007/s11368-023-03691-9 doi: 10.1007/s11368-023-03691-9

Żelazny, W. R., y Šimon, T. (2022, may). Calibration Spiking of MIR-DRIFTS Soil Spectra for Carbon Predictions Using PLSR Extensions and Log-Ratio Transformations. *Agriculture*, *12*(5), 682. Descargado de https://www.mdpi.com/2077-0472/12/5/682 doi: 10.3390/agriculture12050682

Zhang, J., y Mouazen, A. M. (2023). Fractional-order Savitzky–Golay filter for pre-treatment of on-line vis–NIR spectra to predict phosphorus in soil. *Infrared Physics and Technology*, *131*(April), 104720. Descargado de https://doi.org/10.1016/j.infrared.2023.104720 doi: 10.1016/j.infrared.2023.104720

Zhao, J., y Wan, S. (2023). Artificial Intelligence and Hyperspectral Modeling for Soil Management. En (pp. 67–91). Descargado de https://link.springer.com/10.1007/978-981-99-2828-6_4 doi: 10.1007/978-981-99-2828-6_4

Zimmermann, B., y Kohler, A. (2013). Optimizing savitzky-golay parameters for improving spectral resolution and quantification in infrared spectroscopy. *Applied Spectroscopy*, *67*(8), 892–902. doi: 10.1366/12-06723

7. Anexos

7.1. Código para el desarrollo de los modelos

Listing 1: Código Python para el modelo PLSR y cálculo de métricas

```
Autor: josea

Este script combina la optimizaci n del n mero de componentes latentes (
con el c lculo de modelos PLSR para predecir los nutrientes.
"""

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.cross_decomposition import PLSRegression
from sklearn.metrics import mean_squared_error, r2_score

# ------------------------------
# 1. Cargar y preparar los datos
# ------------------------------

# Cargar los datos espectrales y los datos objetivo (nutrientes) desde arc
# Nota: en este ejemplo se utiliza 'snv_savgol2.xlsx'. Puedes cambiar la r
spectral_df = pd.read_excel('C:/Users/josea/Desktop/SNV_savgol2mt.xlsx')
target_df = pd.read_excel('C:/Users/josea/Desktop/datos.xlsx')

# Asegurarse de que los nombres de columnas sean strings
spectral_df.columns = spectral_df.columns.astype(str)
```

```
target_df.columns = target_df.columns.astype(str)

# Dividir los datos en conjuntos de entrenamiento y prueba
X_train, X_test = train_test_split(spectral_df, test_size=0.2, random_stat
y_train, y_test = train_test_split(target_df, test_size=0.2, random_state=

# Fusionar los DataFrames de entrenamiento y prueba usando la columna 'ID'
train_df = pd.merge(X_train, y_train, on='ID')
test_df = pd.merge(X_test, y_test, on='ID')

# ------------------------------
# 2. Definir la funci n de evaluaci n con CV
# ------------------------------

def evaluate_pls(n_components, X, y, cv=5):
    """
    Eval a un modelo PLSRegression con un n mero dado de componentes usa
    Devuelve el RMSE promedio obtenido.
    """
    pls = PLSRegression(n_components=n_components)
    # Se utiliza el scoring negativo para el MSE
    mse_scores = cross_val_score(pls, X, y, scoring='neg_mean_squared_erro
    rmse_scores = np.sqrt(-mse_scores)
    return np.mean(rmse_scores)

# ------------------------------
# 3. Determinar el n mero  ptimo  de variables latentes para cada nutrien
# ------------------------------

nutrientes = ['Ca', 'K', 'P', 'Mg', 'MO']
```

```
max_components = 25  # Rango m ximo de componentes a evaluar
n_components_range = range(1, max_components + 1)

# Diccionarios para almacenar resultados de la optimizaci n
best_components = {}  # Nutriente: mejor n mero de componentes
rmse_all = {}         # Nutriente: lista de RMSE para cada valor de n_com

for nutrient in nutrientes:
    print(f"Evaluando nutriente: {nutrient}")
    # Preparar los datos: se eliminan 'ID' y la columna del nutriente
    X_train_nutrient = train_df.drop(columns=['ID', nutrient])
    y_train_nutrient = train_df[nutrient]

    rmse_values = []
    # Evaluar para cada cantidad de componentes
    for n in n_components_range:
        rmse = evaluate_pls(n, X_train_nutrient, y_train_nutrient, cv=5)
        rmse_values.append(rmse)

    # Guardar la evoluci n del RMSE para este nutriente
    rmse_all[nutrient] = rmse_values

    # Determinar el n mero de componentes que minimiza el RMSE
    best_n = list(n_components_range)[np.argmin(rmse_values)]
    best_components[nutrient] = best_n
    print(f"  El n mero ideal de variables latentes para {nutrient} es: {

# ------------------------------
# 4. Graficar la evoluci n del RMSE para cada nutriente
# ------------------------------
```

```
n_plots = len(nutrientes)
n_cols = 2
n_rows = int(np.ceil(n_plots / n_cols))

plt.figure(figsize=(n_cols * 6, n_rows * 4))
for i, nutrient in enumerate(nutrientes, start=1):
    plt.subplot(n_rows, n_cols, i)
    plt.plot(list(n_components_range), rmse_all[nutrient], marker='o', lin
    plt.xlabel('N mero de Variables Latentes')
    plt.ylabel('RMSE (CV)')
    plt.title(f'{nutrient} (Mejor: {best_components[nutrient]} comp.)')
    plt.grid(True)
plt.tight_layout()
plt.show()

print("Resumen de mejores componentes por nutriente:")
for nutrient in nutrientes:
    print(f"  {nutrient}: {best_components[nutrient]} componentes")

# ------------------------------
# 5. Entrenar, evaluar y visualizar el modelo PLSR con el n mero  ptimo
# ------------------------------

models = {}
results = {}

for nutrient in nutrientes:
    # Preparar datos de entrenamiento y prueba eliminando 'ID' y la column
    X_train_nutrient = train_df.drop(columns=['ID', nutrient])
```

```
y_train_nutrient = train_df[nutrient]
X_test_nutrient = test_df.drop(columns=['ID', nutrient])
y_test_nutrient = test_df[nutrient]

# Entrenar el modelo con el n mero  ptimo  de componentes para el nu
best_n = best_components[nutrient]
plsr = PLSRegression(n_components=best_n)
plsr.fit(X_train_nutrient, y_train_nutrient)
models[nutrient] = plsr

# Predicci n en el conjunto de prueba
y_pred_nutrient = plsr.predict(X_test_nutrient)

# Calcular m tricas de evaluaci n
rmse = mean_squared_error(y_test_nutrient, y_pred_nutrient, squared=F
r2 = r2_score(y_test_nutrient, y_pred_nutrient)

# Ajuste lineal sobre las predicciones para obtener la ecuaci n de r
coef = np.polyfit(y_test_nutrient, y_pred_nutrient.flatten(), 1)
equation = f"y = {coef[0]:.2f}x + {coef[1]:.2f}"

# Almacenar los resultados
results[nutrient] = {
    'rmse': rmse,
    'r2': r2,
    'y_test': y_test_nutrient,
    'y_pred': y_pred_nutrient,
    'equation': equation
}
```

```
# Imprimir los resultados para cada nutriente
for nutrient, res in results.items():
    print(f"\nResultados para {nutrient}:")
    print(f"  RMSE: {res['rmse']:.2f}")
    print(f"  R^2: {res['r2']:.2f}")
    print(f"  Ecuaci n: {res['equation']}")

# Visualizaci n gr fica: Valores reales vs Predichos con la ecuaci n de
for nutrient, res in results.items():
    y_test = res['y_test']
    y_pred = res['y_pred']

    plt.figure(figsize=(10, 6))
    plt.scatter(y_test, y_pred, edgecolor='k', alpha=0.7)
    plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()],
    plt.xlabel('Valores Reales')
    plt.ylabel('Valores Predichos')
    plt.title(f'Valores Reales vs Predichos para {nutrient}')
    plt.text(y_test.min(), y_test.max(), res['equation'], fontsize=12, co
    plt.grid(True)
    plt.show()
```

7.2. Imágenes varias

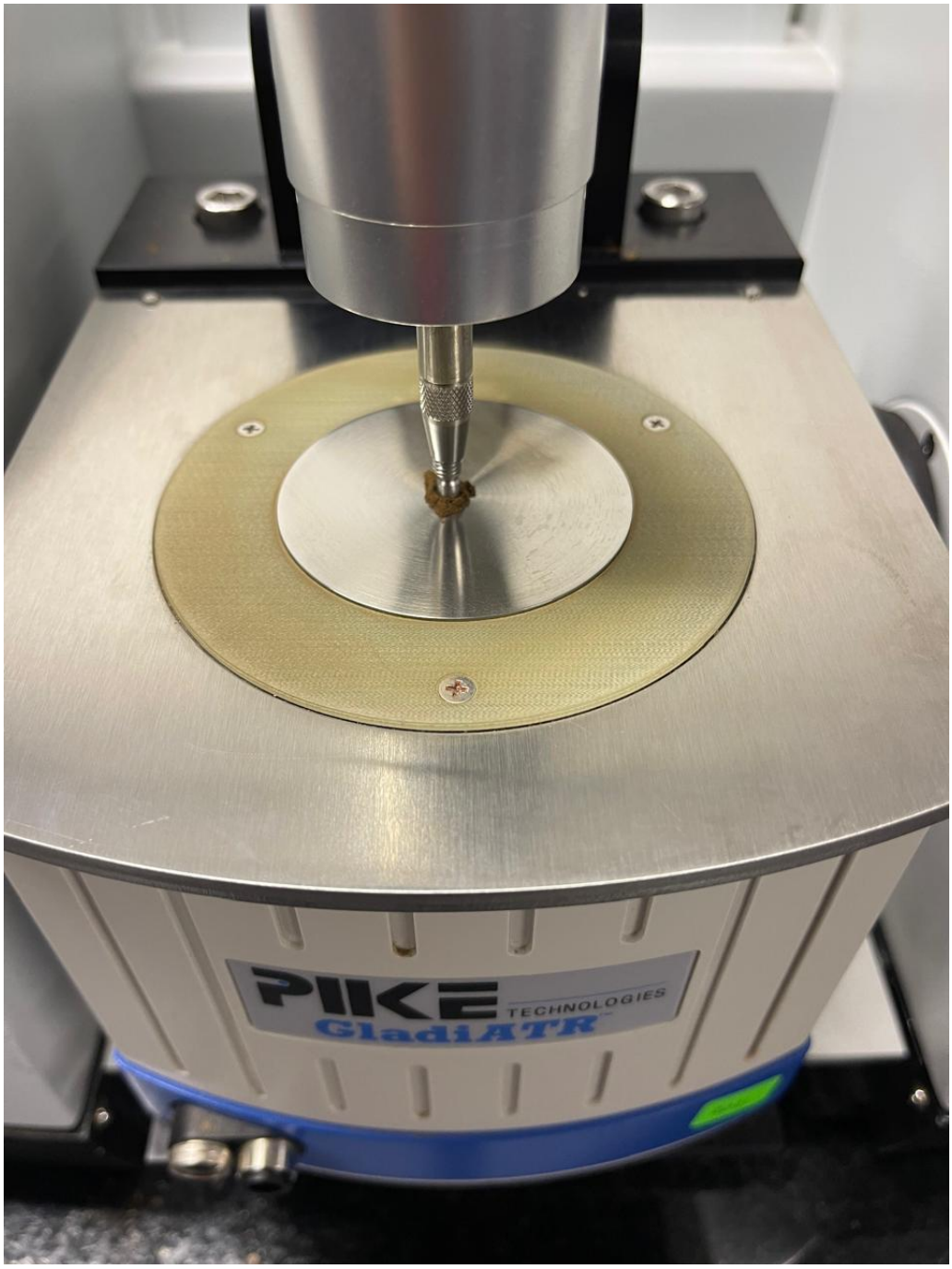


Figura 36: FTIR Perkin Elmer modelo Frontier, obteniendo muestra con la técnica de reflectancia atenuada ATR

Figura 37: Preparación de las muestras utilizadas

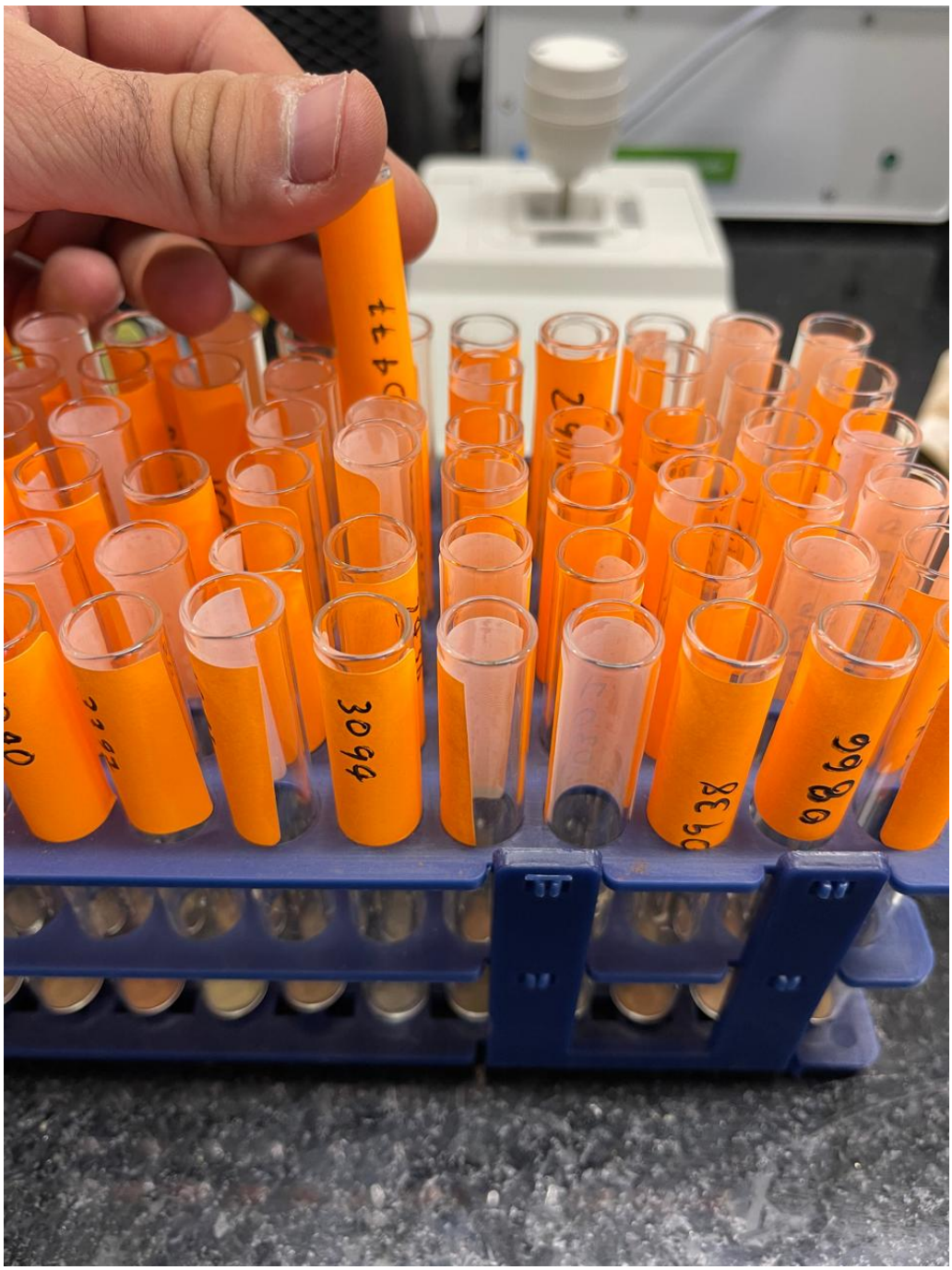
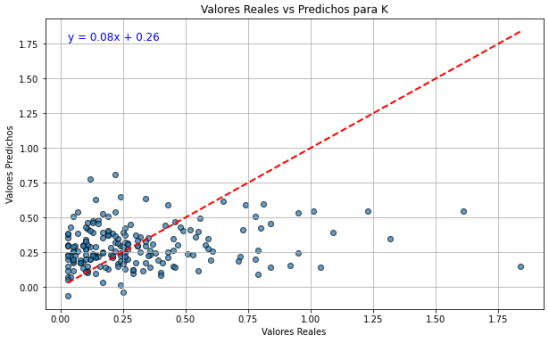


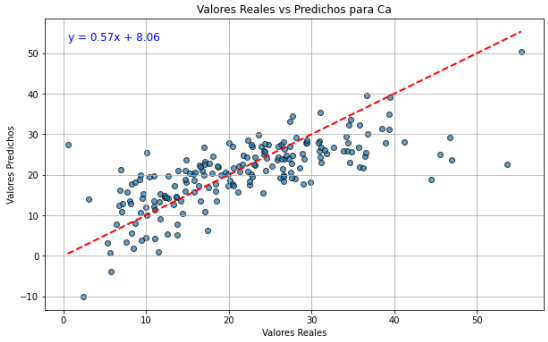
Figura 38: Preparación y transporte de las muestras utilizadas

Modelos con los datos sin promediar cada 20 mediciones



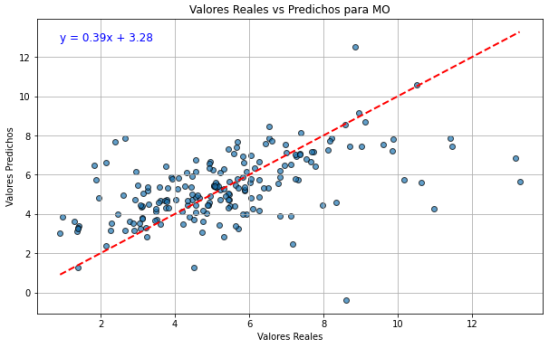
(a) Valores reales vs. predichos para Potasio con Detrend

- **RMSE**: 0.31
- **R**^{2}: -0.09



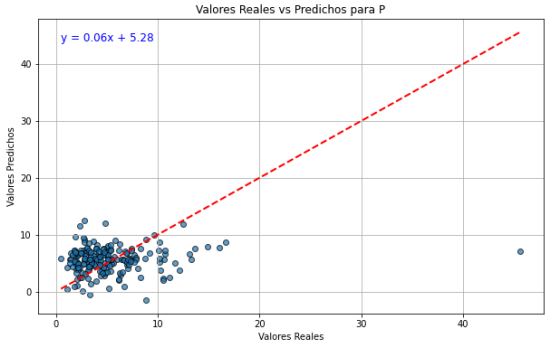
(b) Valores reales vs. predichos para Calcio con Savitzky-Golay 1ª Derivada

- **RMSE**: 7.25
- **R**^{2}: 0.5



(c) Valores reales vs. predichos para Materia orgánica con SNV

- **RMSE**: 1.99
- **R**^{2}: 0.24



(d) Valores reales vs. predichos para Fósforo con Savitzky-Golay 1ª Derivada + SNV

- **RMSE**: 4.55
- **R**^{2}: -0.17

Figura 39: Modelos varios para la predicción con los datos espectrales sin promediar.

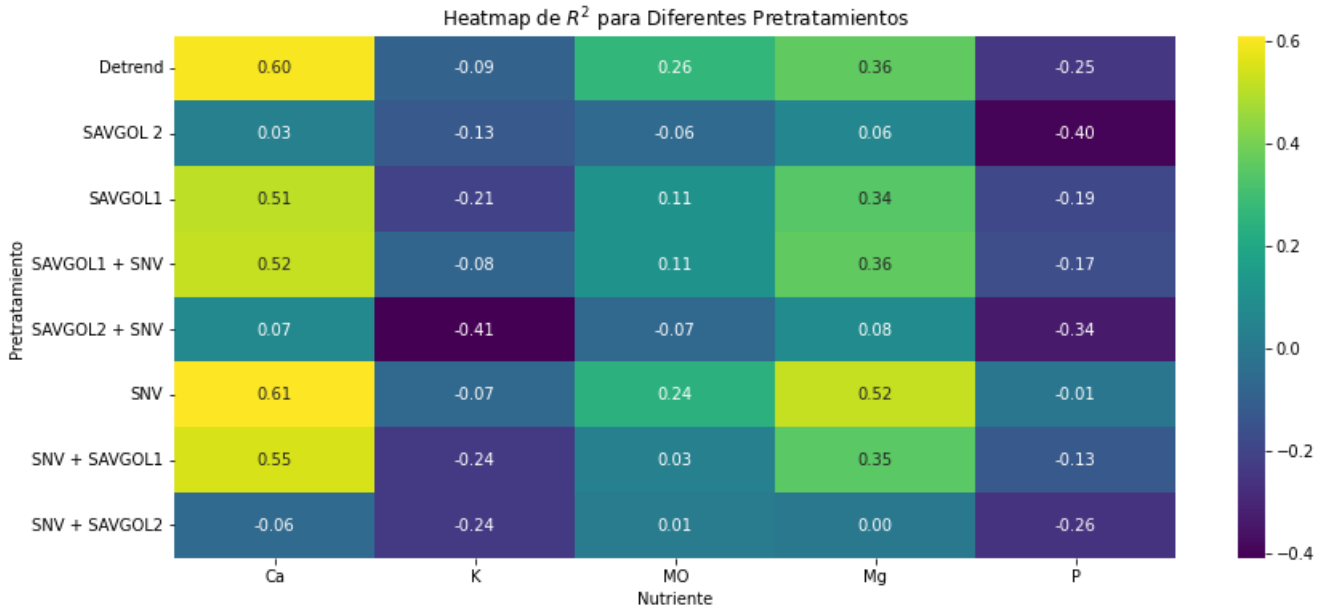


Figura 40: Mapa de calor para los valores de R^2 para los modelos con todos los datos

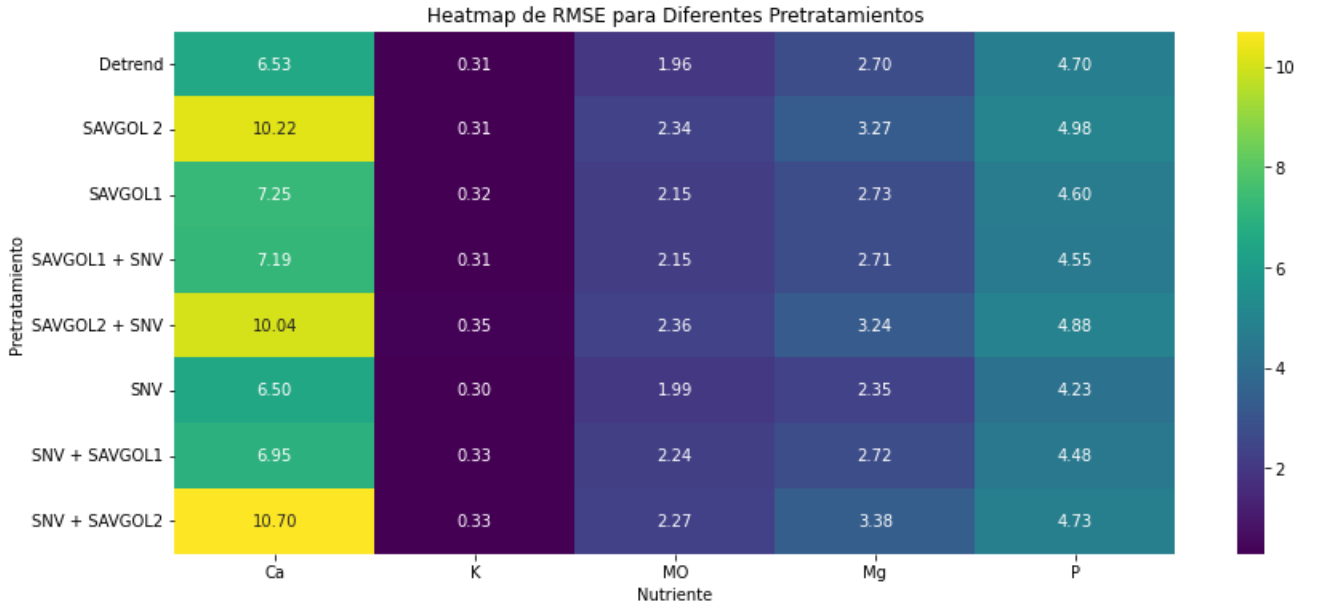


Figura 41: Mapa de calor para los valores de RMSE para los modelos con todos los datos