



**MODELACIÓN PREDICTIVA DEL CARBONO ORGÁNICO DEL SUELO CON
DATOS MULTIFUENTE Y APRENDIZAJE AUTOMÁTICO.**

RICARDO MONTOYA INFANTE

LUDWIG ESTEBAN ÁNGEL CORREDOR

PROYECTO APLICADO PARA OPTAR AL TÍTULO DE

MAGISTER EN CIENCIA DE DATOS

DIRECTOR

EDGAR LEONAIRO PENCUE

FACULTAD DE INGENIERÍA Y CIENCIAS

MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI, MAYO 11 DE 2026

TABLA DE CONTENIDO

INTRODUCCIÓN	11
1. Definición del problema	12
1.1 Planteamiento del problema	12
1.2 Formulación del problema	13
1.2.1 Sistematización	14
2.1 OBJETIVOS	15
2.1 Objetivo General	15
2.2 Objetivos Específicos	15
3. MARCO TEÓRICO Y ANTECEDENTES	16
3.1 Marco Teórico	16
3.1.1 Fundamentos del Carbono Orgánico del Suelo (COS) y su Relevancia.....	16
3.1.2 Factores Determinantes y Dinámica Espacial del COS	16
3.1.3 Adquisición de Datos para la Estimación del COS	17
3.1.4 Aprendizaje Automático para el Modelamiento del COS	17
3.1.5 Regresión por Vectores de Soporte: formulación, kernel polinómico y propiedades para n moderado.....	18
3.2 Antecedentes	19
4. METODOLOGÍA	21
4.1 Enfoque general del estudio	21
4.2 Fuente de datos y construcción de la base analítica	21
4.3 Preprocesamiento y estandarización del flujo analítico	21
4.4 Definición de escenarios de modelación	22
4.5 Familias de modelos evaluadas	23
4.6 Estrategia de búsqueda, screening y ajuste fino de hiperparámetros.	23

4.6.1 Screening y ranking multicriterio	23
4.6.2 Ajuste fino de hiperparámetros	24
4.6.3 Validación espacial	24
4.7 Contraste estadístico entre modelos	24
4.7.1 Evaluación de incertidumbre y análisis residual	25
4.7.2 Interpretación de los modelos y aporte de covariables	25
4.8 Selección del modelo final	25
4.8.1 Comparación entre alternativas competitivas	25
4.8.2 Criterios de selección del modelo óptimo	26
4.9 Modelación espacial y generación cartográfica	26
4.9.1 Protocolo de producción cartográfica	26
4.10 Adquisición y preprocesamiento de datos de percepción remota	27
4.10.1 Fuentes de datos espectrales	27
4.10.2 Índices espectrales derivados	28
4.11 Anexos del estudio	28
5. RESULTADOS	29
5.1 Resultados - Objetivo 1: Analizar las relaciones estadísticas entre las mediciones de COS y un conjunto consolidado de covariables ambientales para identificar los predictores más relevantes en la Granja San José.	29
5.1.1 Consolidación de la Base de Datos Analítica	29
5.1.2 Análisis Exploratorio y Caracterización de la Variable Objetivo (COS)	30
5.1.3 Análisis univariado	31
5.1.4 Análisis bivariado	35
5.1.5 Relevancia predictiva y diagnóstico de multicolinealidad de las covariables	37
5.1.6 Factor de Inflación de la Varianza (VIF)	38

5.1.7 Síntesis integradora del objetivo 1	40
5.2 Resultados - Objetivo 2: Comparar el desempeño predictivo de los algoritmos de aprendizaje automático para seleccionar el modelo óptimo de estimación de COS.	41
5.2.1 Modelos lineales	41
5.2.2 Modelos basados en árboles.....	42
5.2.3 Modelos probabilísticos	44
5.2.4 Modelos basados en funciones kernel.....	45
5.2.5 Modelos no paramétricos	46
5.2.6 Modelos Aditivos/GAM	47
5.2.7 Modelos basados en redes neuronales	48
5.2.8 Síntesis integradora del objetivo 2	49
5.3 Resultados - Objetivo 3: Evaluar el resultado de los diversos algoritmos de aprendizaje automático implementados, cuantificando la exactitud y la incertidumbre de las predicciones.	50
5.3.1 Escenario I sin exclusión de covariables	51
5.3.3 Escenario II sin exclusión de covariables	54
5.3.4 Escenario II con exclusión de covariables MO y N.....	55
5.3.5 Escenario III.....	57
5.3.6 Selección de modelos individuales definitivos para la predicción de COS por escenario	59
5.3.7 Interpretación de los modelos y aporte de las covariables.....	61
5.3.8 Reproducibilidad del modelo campeón	65
5.3.9 Síntesis integradora del objetivo 3	66
5.4 Resultados - Objetivo 4: Generar cartografía digital de alta resolución que represente la distribución espacial del COS y la incertidumbre asociada a la predicción en el área de estudio.	66

5.4.1 Mapa de predicción espacial del COS	67
5.4.2 Coherencia espacial de la superficie predictiva	69
5.4.3 Incertidumbre predictiva	70
5.4.4 Distribución espacial de residuales	72
5.4.5 Síntesis de la producción cartográfica	73
6. CONCLUSIONES Y TRABAJOS FUTUROS.....	75
REFERENCIAS BIBLIOGRÁFICAS	77

TABLA DE FIGURAS

Figura 1 Mapa de predicción espacial del COS (%) —Escenario II_excl, modelo SVR_poly, resolución 3 m. Coordenadas MAGNA-SIRGAS (EPSG:3116).	68
Figura 2 Distribución de las predicciones del COS sobre los 28 981 píxeles válidos — Escenario II_excl. $\mu = 8.35 \%$, $\sigma = 0.73 \%$, rango [4.70, 10.80].	69
Figura 3 Diagrama de dispersión observado vs. predicho en las 145 coordenadas de muestreo — Escenario II_excl. $R^2 = 0.8707$, $RMSE = 0.4326$, $MAE = 0.2212$, sesgo = -0.0077	70
Figura 4 Mapa de incertidumbre predictiva (σ) estimada por bootstrap (N = 100) — Escenario II_excl.....	71
Figura 5 Mapa de ancho del intervalo de confianza al 95 % (P97.5 – P2.5) — Escenario II_excl.	72
Figura 6 Distribución espacial de residuales (observado – predicho) en los 145 puntos de muestreo — Escenario II_excl. $\mu_{resid} = 0.0077$, $\sigma_{resid} = 0.4325$	73

TABLA DE TABLAS

Tabla 1 Comparación de estudios representativos sobre predicción del COS mediante ML.	19
Tabla 2 Definición de escenarios de evaluación de modelos predictivos	22
Tabla 3 Familias de modelos evaluadas y propósito metodológico	23
Tabla 4 Flujo operativo de la modelación espacial	27

Tabla 5 Índices espectrales derivados de PlanetScope (NIR) e imágenes RGB	28
Tabla 6 Tipo y función de variables usadas en el estudio	29
Tabla 7 Evaluación de calidad y estructura del conjunto de datos.....	31
Tabla 8 Variables cuantitativas edáficas incluidas en el estudio	32
Tabla 9 Variables cualitativas incluidas en el estudio.....	33
Tabla 10 Selección de variables a través del método SelectKBest	37
Tabla 11 Síntesis de los hallazgos del análisis exploratorio y su implicación para la modelación del carbono orgánico	39
Tabla 12 Síntesis de métricas de desempeño predictivo y validación de supuestos estadísticos por escenario de modelos lineales.	42
Tabla 13 Síntesis de métricas de desempeño predictivo y validación de supuestos estadísticos por escenario de modelos basados en árboles	42
Tabla 14 Síntesis de métricas de desempeño predictivo y validación de supuestos estadísticos por escenario de modelos probabilísticos	44
Tabla 15 Síntesis de métricas de desempeño predictivo y validación de supuestos estadísticos por escenario de modelos basados en funciones kernel	45
Tabla 16 Síntesis de métricas de desempeño predictivo y validación de supuestos estadísticos por escenario de modelos no paramétricos.....	46
Tabla 17 Síntesis de métricas de desempeño predictivo y validación de supuestos estadísticos por escenario de modelos aditivos/GAM	48
Tabla 18 Síntesis de métricas de desempeño predictivo y validación de supuestos estadísticos por escenario de modelos basados en redes neuronales.	49
Tabla 19 Evaluación multicriterio de robustez predictiva, cuantificación de la incertidumbre y jerarquización inferencial de modelos campeones por familia algorítmica (Escenario I).	51
Tabla 20 Evaluación multicriterio de robustez predictiva, cuantificación de la incertidumbre y jerarquización inferencial de modelos campeones por familia algorítmica (Escenario I con exclusión de covariables).	52
Tabla 21 Desempeño del modelo de stacking en el escenario I con exclusión de covariables MO y N.....	54

Tabla 22 Evaluación multicriterio de robustez predictiva, cuantificación de la incertidumbre y jerarquización inferencial de modelos campeones por familia algorítmica (Escenario II sin exclusión).	54
Tabla 23 Evaluación multicriterio del desempeño espacial, incertidumbre bootstrap, diagnóstico residual y agrupación inferencial de los modelos campeones por familia algorítmica (Escenario II con exclusión)	55
Tabla 24 Desempeño del modelo de stacking en el escenario II, con exclusión de covariables MO y N	57
Tabla 25 Evaluación multicriterio de robustez predictiva, cuantificación de la incertidumbre y jerarquización inferencial de modelos campeones por familia algorítmica (Escenario III).....	57
Tabla 26 Desempeño del modelo de stacking frente a modelos individuales en el escenario III..	58
Tabla 27 Selección de modelos individuales definitivos para la predicción de COS por escenario.	59
Tabla 28 Comparativa multidimensional de modelos stacking y modelos base: diagnóstico de error espacial, estabilidad de las predicciones y validación inferencial	60
Tabla 29 Contribución agregada de grupos de covariables al desempeño predictivo bajo esquemas de permutación y protocolos de exclusión de variables.....	61
Tabla 30 Contribución agregada de grupos de covariables al desempeño predictivo bajo esquemas de permutación y protocolos de exclusión de variables.....	62
Tabla 31 Diagnóstico de la arquitectura predictiva local: atribución de importancia y sentido de influencia de las covariables edafoclimáticas y sensores remotos mediante LIME.	63
Tabla 32 Evaluación de la redundancia y estabilidad estructural del espacio de características mediante RFECV para los modelos campeones seleccionados.	64
Tabla 33 Especificaciones para la reproducibilidad del modelo campeón	65
Tabla 34 Resumen de la producción cartográfica — Escenario II con exclusión (SVR)	74
Tabla 35 Resumen de la producción cartográfica — Escenario II_excl (SVR_poly)	74

TABLA DE ECUACIONES

Ecuación 1 Problema de optimización primal — Support Vector Regression (SVR).....	18
Ecuación 2 Solución dual kernelizada de la SVR	18

Ecuación 3 Índice global de ranking multicriterio	24
---	----

INTRODUCCIÓN

La degradación del suelo y el cambio climático impactan la estabilidad de los sistemas socioecológicos rurales. El carbono orgánico del suelo (COS) es un indicador crítico de la salud edáfica y un sumidero estratégico en el ciclo global del carbono. Según el IPCC [1], las emisiones netas de CO₂ derivadas del uso de la tierra, su cambio y la silvicultura representan aproximadamente una décima parte del total antropogénico global. En Colombia, la conversión de ecosistemas naturales a sistemas agropecuarios genera reducciones de COS que oscilan entre una quinta parte y la mitad en suelos minerales, con pérdidas superiores en suelos orgánicos altoandinos bajo manejo intensivo [2].

La cuantificación directa del COS a escala predial es costosa y técnicamente compleja, lo que limita la toma de decisiones en sistemas productivos de pequeña escala. Si bien existen avances en cartografía digital de suelos en Colombia —principalmente con modelos de regresión kriging [3],[4] y variables edafoclimáticas, persiste un vacío metodológico en la integración sistemática de covariables funcionales de la vegetación (relación C/N, índices espectrales de alta resolución) y en la comparación rigurosa de familias de algoritmos de aprendizaje automático bajo validación espacial en entornos tropicales [5].

Ante este vacío, la presente investigación evalúa cuál familia de algoritmos ofrece la mejor capacidad predictiva del COS sobre el conjunto de datos disponibles, integrando covariables multifuente bajo validación cruzada espacial. Se consolidó una base de datos de 145 muestras georreferenciadas (IES CINOC, Caldas) con imágenes PlanetScope (3 m) y repositorios edafoclimáticos (IGAC, IDEAM, WorldClim, SoilGrids). Para mitigar el desbalance dimensional y prevenir el sobreajuste, la evaluación se estructuró en tres escenarios de covariables de complejidad ascendente. Se contrastaron siete familias algorítmicas mediante validación cruzada espacial GroupKFold, contraste estadístico no paramétrico y diagnóstico de incertidumbre predictiva, con criterios de selección multicriterio orientados a la generalización espacial y a la mínima dependencia de covariables redundantes. El producto resultante es una cartografía digital de la distribución del COS y de su incertidumbre asociada, concebida como insumo para la planificación ambiental, la restauración ecológica y los mercados de carbono.

1. Definición del problema

1.1 Planteamiento del problema

El Carbono Orgánico del Suelo (COS) es un componente del ciclo del carbono y un indicador clave de la salud ecosistémica. Constituye el reservorio más grande de carbono, superando la cantidad encontrada en la atmósfera y la biomasa vegetal en conjunto, influyendo directamente en la regulación climática, fertilidad del suelo, estructura edáfica y retención de agua; aspectos vitales para la sostenibilidad agrícola y forestal [2], [6]. Pequeñas alteraciones en sus niveles repercuten en impactos considerables en la estabilidad del ecosistema y la dinámica climática, por esto, comprender y cuantificar su variabilidad espacial es importante para el diseño de estrategias de mitigación y adaptación al cambio climático y la gestión sostenible del suelo en diferentes escalas.

La cuantificación directa del carbono orgánico del suelo (COS) a nivel regional se ve obstaculizada por sus elevados costos y complejidad técnica. Esta carencia de datos de alta resolución espacial dificulta la toma de decisiones informadas en sistemas productivos de pequeña escala, limitando o dificultando el monitoreo ambiental preciso, la planificación territorial informada y la formulación de políticas públicas basadas en evidencia. La dinámica del COS presenta una complejidad inherente, regulada por la interacción de factores edafoclimáticos, topográficos y antrópicos. Es particularmente determinante el papel de la cobertura vegetal y las propiedades del material orgánico que esta suministra, elementos que suelen ser subestimados en los modelos predictivos aplicados a entornos tropicales como Colombia [7],[8].

Este panorama exige una transición paradigmática desde las representaciones cartográficas discretas hacia el modelado de procesos edáficos. Como sostienen Wilson (2024) y otros autores, la limitación de los sistemas vectoriales tradicionales ha sido superada por la adopción de estructuras rasterizadas de alta resolución, las cuales se alinean con las capacidades actuales de análisis espacial. Bajo el paradigma del mapeo predictivo de suelos, el uso de algoritmos de aprendizaje automático de conjunto (ensemble ML) permite integrar flujos masivos de datos geoespaciales —incluyendo sensores remotos y modelos climáticos— en marcos tridimensionales. Esta capacidad para modelar interacciones no lineales resulta determinante para vincular el COS con variables funcionales frecuentemente subestimadas en modelos estáticos, tales como la arquitectura del perfil edáfico y su dinámica hídrica [9].

La superioridad de los algoritmos de aprendizaje computacional frente a la regresión lineal múltiple ha sido ampliamente documentada en la estimación espacial del COS [3], [10]. Actualmente, la integración de aprendizaje profundo y minería de datos espaciales se ve potenciada por procesos automáticos de selección de variables, como RFECV y SelectKBest, junto con ingeniería de características multiescalar. Estos avances no solo incrementan la exactitud de los modelos, sino que también fortalecen su transparencia mediante enfoques de interpretabilidad como SHAP o LIME [11].

La implementación de redes neuronales profundas junto a sensores remotos de alta resolución permite descifrar interacciones complejas entre factores edafoclimáticos y el uso del suelo que determinan la estabilización del carbono en el perfil edáfico. Complementariamente, el empleo de modelos probabilísticos y métodos bayesianos [12] optimiza la estimación de la incertidumbre en la cartografía del COS, otorgando mayor rigor técnico a la toma de decisiones y a las políticas de mitigación. En este contexto, la ciencia de datos aplicada a la agroecología no solo facilita el procesamiento de bases de datos masivas, sino que genera herramientas escalables y reproducibles esenciales para la gestión territorial y el fortalecimiento de la resiliencia climática.

1.2 Formulación del problema

La estimación precisa y espacialmente explícita del Carbono Orgánico del Suelo (COS) resulta fundamental para una gestión sostenible de los agroecosistemas y para el diseño de estrategias de mitigación frente al cambio climático. No obstante, los métodos tradicionales de mapeo del COS presentan limitaciones significativas en términos de resolución espacial, escalabilidad y capacidad de representar la complejidad ambiental de entornos heterogéneos, como el Centro de Transferencia Tecnológica Granja San José. En este contexto, se plantea la necesidad de incorporar enfoques basados en ciencia de datos que permitan mejorar la precisión y robustez de las estimaciones espaciales del COS mediante el uso de información multifuente y modelos avanzados de predicción.

A partir de este planteamiento, surge la pregunta general para el proyecto:

¿De qué manera la inclusión de un conjunto diverso de covariables ambientales, analizadas mediante algoritmos de aprendizaje automático, contribuye a mejorar el rendimiento predictivo y la robustez de los modelos de distribución espacial del Carbono Orgánico del Suelo?

1.2.1 Sistematización

- Mediante el uso de análisis exploratorio de datos y algoritmos de selección de características, ¿cuál es el subconjunto óptimo de covariables (topográficas, espectrales, edáficas) que maximiza el poder predictivo y minimiza la redundancia para la modelación del COS?
- ¿Qué algoritmos de aprendizaje automático —como Random Forest, Support Vector Machines, Gradient Boosting y Redes Neuronales Artificiales— ofrecen un mejor desempeño predictivo y mayor capacidad de generalización en este contexto?
- ¿Qué mejoras cuantitativas y cualitativas presentan los mapas generados mediante estos modelos en comparación con aquellos que no integran sistemáticamente covariables funcionales?

2.1 OBJETIVOS

2.1 Objetivo General

Desarrollar un modelo predictivo mediante la aplicación de técnicas de aprendizaje automático, para estimar la concentración y distribución del Carbono Orgánico del Suelo (COS), integrando covariables ambientales.

2.2 Objetivos Específicos

- Analizar las relaciones estadísticas entre las mediciones de COS y un conjunto consolidado de covariables ambientales para identificar los predictores más relevantes en la Granja San José.
- Comparar el desempeño predictivo de los algoritmos de aprendizaje automático para seleccionar el modelo óptimo de estimación de COS.
- Evaluar el resultado de los diversos algoritmos de aprendizaje automático implementados, cuantificando la exactitud y la incertidumbre de las predicciones.
- Generar cartografía digital de alta resolución que represente la distribución espacial del COS y la incertidumbre asociada a la predicción en el área de estudio.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1 Marco Teórico

3.1.1 Fundamentos del Carbono Orgánico del Suelo (COS) y su Relevancia

La estimación del carbono orgánico del suelo (COS) se fundamenta en la evolución de los modelos de pedogénesis hacia el paradigma del mapeo digital de suelos (MDS) [13], el cual define al suelo como un producto de la interacción de factores ambientales y espaciales. Estos factores regulan los procesos biogeoquímicos de fijación, estabilidad y distribución del carbono en el perfil edáfico [14], [15]. Desde una perspectiva funcional, el COS es un pilar de la salud edáfica: actúa como agente de unión para la estabilidad de agregados y potencia la fertilidad química al elevar la capacidad de intercambio catiónico (CIC) [10].

Como componente central del ciclo biogeoquímico, el COS constituye el principal almacén de carbono terrestre, superando los depósitos atmosféricos y bióticos combinados [2], [6]. Su importancia funcional en la agricultura es crítica; niveles óptimos aseguran una productividad estable y resiliencia ante el estrés ambiental. Por ello, el monitoreo preciso del COS es fundamental para la gestión agrícola sustentable y la modelación de servicios ecosistémicos globales [3], [16].

3.1.2 Factores Determinantes y Dinámica Espacial del COS

La variabilidad geográfica del carbono orgánico resulta de un balance entre factores del entorno. La delimitación de patrones espaciales se basa en predictores como el uso del suelo, la topografía (elevación, pendiente, TWI) y condiciones climáticas [10], [17]. El clima influye mediante el control de la precipitación y la temperatura sobre la producción primaria neta, regulando la actividad microbiana y la descomposición de residuos [18], [14].

Las propiedades edáficas, especialmente la textura, son cruciales. Los suelos de textura fina (arcillosos) favorecen la protección física de la materia orgánica mediante la formación de complejos y la oclusión en agregados estables [19]. En contraste, suelos arenosos presentan menores niveles de COS debido a una mayor aireación y reducida capacidad de protección. El pH influye en la composición de las comunidades microbianas y en la solubilidad de la materia orgánica [19], [20]. La topografía modula los flujos hidrológicos y procesos de erosión/sedimentación que redistribuyen el carbono en el paisaje [21]. Las prácticas de manejo (labranza, rotación) impactan directamente los niveles de COS, donde la conversión de ecosistemas naturales a sistemas agrícolas intensivos suele derivar en pérdidas significativas [22], [23].

En el contexto colombiano, Burbano Orjuela [2] ha destacado el papel del COS frente al cambio climático en suelos tropicales, subrayando que la variabilidad climática y los regímenes de humedad propios de la zona intertropical generan gradientes de COS especialmente marcados en comparación con latitudes templadas. A escala subregional en Antioquia, Ortega-Monsalve et al. [24] documentaron la relación del COS con nitrógeno total y otras variables edáficas en sistemas ganaderos, evidenciando la heterogeneidad espacial característica de los paisajes colombianos.

3.1.3 Adquisición de Datos para la Estimación del COS

El modelado robusto requiere una arquitectura de datos integrada. Los datos de referencia (ground truth) procedentes de muestras georreferenciadas son indispensables para el entrenamiento y validación de algoritmos [17]. La representatividad del diseño de muestreo es determinante para capturar la variabilidad espacial [25].

La teledetección y las variables espectrales permiten monitorear la reflectancia de la superficie en relación con el contenido de materia orgánica mediante sensores multiespectrales como Sentinel-2 y Landsat 8/9 [26], [16]. Los índices de productividad vegetal (NDVI, SAVI) actúan como estimadores de la biomasa aérea y predictores de la acumulación de COS a largo plazo [27], [27]. En áreas de baja cobertura, índices como el de suelo desnudo (BSI) explotan la relación inversa entre el albedo y el contenido de carbono [16]. Asimismo, índices de humedad (NDMI) proporcionan información sobre el estado hídrico, factor regulador de la mineralización microbiana [28].

3.1.4 Aprendizaje Automático para el Modelamiento del COS

El machine learning (ML) ha transformado el MDS al capturar interdependencias no lineales complejas entre el COS y predictores ambientales [17]. Esta flexibilidad permite procesar volúmenes masivos de datos multidimensionales [9].

Modelos lineales: Ofrecen estructuras parsimoniosas e interpretables. En espectroscopía de suelos, la regresión por mínimos cuadrados parciales (PLSR) es un referente para manejar colinealidad en datos Vis-NIR [20]. Los enfoques lineales también sirven como líneas base robustas en mapeos regionales [21].

Modelos basados en árboles y ensambles: Capturan relaciones no lineales e interacciones complejas. Random Forest destaca por su capacidad para integrar información heterogénea de sensores remotos y datos de campo [12], [29].

Modelos aditivos y enfoques GAM: Permiten modelar relaciones no lineales suaves, conservando interpretabilidad. Son eficaces para representar simultáneamente relaciones graduales y no estacionariedad espacial [30].

Modelos no paramétricos: Como k-nearest neighbours, se adaptan a distribuciones locales complejas y estructuras heterogéneas sin formas funcionales fijas [31].

Modelos basados en funciones kernel: Las máquinas de soporte vectorial (SVM) proyectan datos a espacios de mayor dimensión para modelar estructuras no lineales de alta complejidad, especialmente en contextos espectrales[32] .

Modelos probabilísticos: La regresión por procesos gaussianos (GPR) y métodos bayesianos permiten cuantificar explícitamente la incertidumbre asociada a las predicciones [33]. Su formulación se desarrolla en la siguiente subsección.

Redes neuronales: Aprenden patrones de alta dimensionalidad y son útiles para maximizar la capacidad predictiva en sistemas complejos integrando percepción remota [20], [34].

3.1.5 Regresión por Vectores de Soporte: formulación, kernel polinómico y propiedades para n moderado.

La regresión por vectores de soporte (SVR) busca estimar una función $f(x) = w \cdot \phi(x) + b$ que aproxime las observaciones con un error inferior a un umbral ε y, simultáneamente, sea lo más plana posible en el espacio de características inducido por el mapeo no lineal $\phi(\cdot)$ [35]. El problema primal se formula como un programa cuadrático convexo con función de pérdida ε -insensible.

Ecuación 1 Problema de optimización primal — Support Vector Regression (SVR)

$$\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \cdot \sum_i (\xi_i + \xi_i^*), \quad s. a.: \quad y_i - w \cdot \phi(x_i) - b \leq \varepsilon + \xi_i; \quad w \cdot \phi(x_i) + b - y_i \leq \varepsilon + \xi_i^*; \quad \xi_i, \xi_i^* \geq 0$$

donde el hiperparámetro C controla el equilibrio entre la planitud de la solución (regularización) y la tolerancia al error de ajuste, mientras que las variables de holgura ξ_i, ξ_i^* permiten el incumplimiento del margen ε bajo penalización lineal [32]. La aplicación de la dualidad de Lagrange transforma el problema en una formulación equivalente que sólo depende de productos internos en el espacio de características, lo que habilita el truco del kernel:

Ecuación 2 Solución dual kernelizada de la SVR

$$f(x) = \sum_i (\alpha_i - \alpha_i^*) \cdot k(x_i, x) + b$$

3.2 Antecedentes

El mapeo digital del COS mediante ML ha experimentado una expansión metodológica acelerada en la última década. A continuación, se sintetizan críticamente los trabajos más relevantes para el presente estudio, identificando métodos, tamaños muestrales, métricas y limitaciones declaradas que justifican el enfoque adoptado.

Tabla 1 Comparación de estudios representativos sobre predicción del COS mediante ML.

Referencia	Área / País	Método principal	n muestras	Métrica principal	Limitación declarada
[3]Tadesse et al. (2026)	Etiopía	Regression Kriging	N.E.	R ² , RMSE	Escala local; covariables limitadas
[26]Etinosa et al. (2026)	EE. UU. semiárido	Teledetección + ML	N.E.	R ² , RMSE	Cobertura espectral limitada
[16]Beisekenov et al. (2025)	Asia Central	RF, SVM, ANN	>100	R ² , RMSE	Agricultura conservación; sin CV espacial
[36] et al. (2025)	Irán (Golestán)	ML + spline depth	N.E.	RMSE, R ²	Área geográfica reducida
[25]Bócoli et al. (2025)	Brasil tropical	RF, XGBoost, Kriging	Alta densidad	R ² , RMSE	Costo de muestreo intensivo
[22]Delgado-Vargas & Benavides (2024)	Nariño, Colombia	Estadística clásica	<50	Correlaciones	Sin predicción espacial; n pequeño
[37] et al. (2024)	Global	ML bottom-up	Grande	R ² , RMSE, Incertidumbre	Escala gruesa; datos climáticos
[8]Serrano-Agudelo et al. (2024)	Páramo Cortadera, Colombia	RS + correlación	N.E.	Coefficiente correlación	Sin modelo predictivo formal; sin CV
[12]Bernardini et al. (2024)	Multi-región	ML vs modelo de proceso	Variado	R ² , RMSE	Sin cuantificación de incertidumbre unificada
[38]Gouda et al. (2024)	Zonas áridas	ANN + RS	N.E.	R ² , RMSE	Entornos áridos; poca generalización
[39]Zheng et al. (2024)	Shanxi, China	GWR con términos suaves	N.E.	R ² , AICc	No estacionariedad; sin CV espacial explícita

[21]Guevara et al. (2018)	América Latina	Múltiples ML	Grande	R ² , RMSE por país	Resolución gruesa; heterogeneidad regional
------------------------------	----------------	--------------	--------	--------------------------------	--

Los estudios de la tabla permiten identificar tres brechas metodológicas recurrentes que el presente trabajo busca cubrir:

Primera: la mayoría de los estudios, incluso los más recientes, omite la validación cruzada espacial estricta, reportando métricas potencialmente optimistas. Solo Radočaj et al. [40], Milà et al. [41] y Mushagalusa et al. [42] abordan explícitamente este problema, ninguno en el contexto colombiano.

Segunda: la cuantificación formal de la incertidumbre predictiva es escasa. Poggio et al. [43] la implementan a escala global, y Wang et al. [18] reportan intervalos en su enfoque ascendente, pero los estudios a escala local o regional raramente publican mapas de incertidumbre como producto independiente.

Tercera: en Colombia, la predicción digital del COS sigue dominada por análisis de correlación y estadística clásica [22], [8], con ausencia de modelos ML con validación espacial rigurosa. Guevara et al. [21] es uno de los pocos antecedentes regional de escala LATAM con múltiples algoritmos, pero con resolución gruesa (≥ 1 km) y sin validación cruzada espacial explícita.

En conjunto, estos antecedentes justifican el desarrollo de un enfoque que integre: (i) la evaluación comparativa de familias algorítmicas heterogéneas de aprendizaje automático bajo criterios homogéneos de desempeño y generalización espacial; (ii) validación cruzada espacial estricta (GroupKFold con malla bivariada de cuantiles); y (iii) cartografía de incertidumbre al 95 % como producto cartográfico explícito, en un área de estudio colombiana con $n = 145$ muestras y 48 covariables multifuente.

4. METODOLOGÍA

4.1 Enfoque general del estudio

La presente investigación se desarrolla bajo un enfoque cuantitativo, predictivo y aplicado, orientado a identificar la estrategia de modelación más adecuada para la estimación del carbono orgánico del suelo (COS) bajo distintos escenarios de covariables y niveles de redundancia informacional. El diseño metodológico se estructura como un estudio comparativo de familias algorítmicas, transformaciones y configuraciones analíticas, con énfasis en la reproducibilidad del flujo de trabajo y la evaluación rigurosa de la capacidad de generalización espacial con $n = 145$ observaciones. La investigación se enmarca en los enfoques contemporáneos de cartografía digital de suelos (DSM), donde la predicción de propiedades edáficas exige determinar qué estrategias ofrecen mayor robustez, precisión e interpretabilidad bajo condiciones reales de limitación muestral [10], [13], [15].

4.2 Fuente de datos y construcción de la base analítica

La base analítica integra datos de laboratorio de suelos (variable respuesta: COS en porcentaje) con covariables derivadas de análisis espacial y percepción remota obtenidas en el Centro de Transferencia Tecnológica (CTT) Granja San José del IES CINOC y la plataforma Planet. El conjunto de predictores incluye variables fisicoquímicas, atributos topográficos a múltiples ventanas espaciales y métricas espectrales derivadas de imágenes PlanetScope (3 m) y ortomosaicos RGB vía UAV. Tras la integración, se verificó la integridad estructural, la coherencia de tipos de variable y la correspondencia espacial entre observaciones y covariables [9], [44].

4.3 Preprocesamiento y estandarización del flujo analítico

El preprocesamiento se implementó mediante una arquitectura modular basada en *Pipeline* y *ColumnTransformer* (scikit-learn), que integra en una sola secuencia reproducible la transformación de variables numéricas, su escalamiento y la codificación de variables categóricas mediante *OneHotEncoder*. Este esquema garantiza la coherencia entre las fases de entrenamiento y predicción, previene la fuga de información y asegura la aplicación idéntica de cada transformación sobre la grilla de predicción espacial.

Las variables numéricas fueron sometidas a cuatro estrategias de transformación —original, raíz cúbica (*cbrt*), Yeo–Johnson y normalización por cuantiles— combinadas con cuatro esquemas de escalamiento: sin escalado, *StandardScaler*, *MinMaxScaler* y *RobustScaler*. Se incorporó

opcionalmente reducción de dimensionalidad mediante PCA para mitigar efectos de redundancia o colinealidad [45].

4.4 Definición de escenarios de modelación

La comparación de modelos se organizó mediante escenarios definidos a partir de distintos conjuntos de covariables. El factor de inflación de la varianza (VIF) evidenció valores de 446.84 para materia orgánica (MO) y 34.5 para nitrógeno total (N), valores que exceden ampliamente el umbral convencional de $VIF > 10$ adoptado en este estudio como criterio de exclusión por redundancia estructural [17][46]. La exclusión de MO y N se asume como decisión metodológica orientada a reducir la dependencia informacional y evaluar con mayor rigor el aporte predictivo del resto de covariables, sin desconocer su relevancia edafológica [24].

Se definieron tres escenarios de modelación:

- Escenario I — Conjunto completo de covariables (físico-químicas, texturales, categóricas, topográficas y espectrales), evaluado con y sin MO y N.
- Escenario II — Conjunto de variables edáficas (físico-químicas, texturales, categóricas y topográficas), evaluado con y sin MO y N.
- Escenario III — Conjunto de variables espectrales exclusivamente; la condición de exclusión de MO y N no aplica.

Tabla 2 Definición de escenarios de evaluación de modelos predictivos

Escenario	Dataset	MO y N
I	Conjunto completo (edáfico + topográfico + espectral)	Incluye MO y N
I	Conjunto completo (edáfico + topográfico + espectral)	Excluye MO y N
II	Variables edáficas (físico-químicas, texturales, categóricas y topográficas)	Incluye MO y N
II	Variables edáficas (físico-químicas, texturales, categóricas y topográficas)	Excluye MO y N
III	Variables espectrales (índices RGB y multispectrales)	No aplica

Los escenarios con MO y N evalúan capacidad predictiva con covariables de alta proximidad funcional; los escenarios sin MO y N evalúan capacidad de predicción indirecta con menor redundancia.

4.5 Familias de modelos evaluadas

La selección de familias respondió a una pregunta metodológica concreta: qué estructura predictiva resulta más adecuada para estimar el COS bajo distintos niveles de información disponible, con $n = 145$ y escenarios de diferente grado de redundancia entre covariables.

Los criterios de selección de familias fueron: (i) representar distintos niveles de complejidad funcional; (ii) contrastar modelos con diferente sensibilidad a la colinealidad; (iii) incluir familias con distintos niveles de interpretabilidad; y (iv) incorporar alternativas capaces de producir estimaciones explícitas de incertidumbre.

Tabla 3 Familias de modelos evaluadas y propósito metodológico

Familia	*Propósito metodológico
Lineales	Servir como referencia parsimoniosa e interpretable
Árboles y ensambles	Capturar relaciones no lineales e interacciones
Aditivos / GAM	Representar no linealidad conservando interpretabilidad
No paramétricos	Modelar patrones locales de similitud
Kernels	Representar dependencias complejas en espacios transformados
Probabilísticos	Incorporar predicción e incertidumbre
Redes neuronales	Explorar relaciones altamente no lineales y de mayor complejidad funcional

* *Análisis detallado en el Anexo 3. Descripción de familias de modelos evaluadas para la estimación del carbono orgánico del suelo*

4.6 Estrategia de búsqueda, screening y ajuste fino de hiperparámetros.

4.6.1 Screening y ranking multicriterio

Se evaluaron comparativamente 48 arquitecturas, combinando algoritmos, transformaciones, esquemas de escalamiento y PCA opcional. El ranking multicriterio integró cuatro indicadores: R^2 medio (R^2_mean), RMSE medio ($RMSE_mean$), MAE medio (MAE_mean)

y desviación estándar del desempeño entre folds. El ordenamiento se realizó de forma descendente para R^2_{mean} y ascendente para las métricas de error. A cada configuración se le asignó un rango individual por criterio, integrado en un índice global calculado como el promedio aritmético de los rangos parciales (Ecuación 1). Dado el volumen de configuraciones evaluadas sobre el mismo conjunto de datos con $n = 145$, el riesgo de selección optimista es no trivial; los resultados del screening se interpretaron como evidencia orientativa, no como selección definitiva.

Ecuación 3 Índice global de ranking multicriterio

$$Rank_{\text{global}} = (Rank_{R^2} + Rank_{RMSE} + Rank_{MAE} + Rank_{STD}) / 4$$

4.6.2 Ajuste fino de hiperparámetros

Con base en el screening, las cinco configuraciones mejor posicionadas por familia fueron sometidas a ajuste fino mediante búsqueda en grilla (*GridSearchCV*) integrada en validación cruzada espacial (*GroupKFold*), evaluando cada combinación de hiperparámetros sobre particiones espaciales mutuamente excluyentes con métricas R^2 y RMSE. La selección de la configuración óptima se basó en la capacidad de generalización espacial observada, no en el ajuste local [47].

4.6.3 Validación espacial

La validación espacial explícita mediante *GroupKFold* reduce el riesgo de fuga de información que los grupos espaciales definidos no se mezclen entre entrenamiento y prueba, reduciendo la fuga de información asociada a la autocorrelación espacial y proporcionando estimaciones más realistas del desempeño en áreas no observadas [48], [47], [49].

Los grupos espaciales se definieron mediante discretización bivariada de coordenadas geográficas por cuantiles, construyendo una malla de 3×3 bloques con tamaño muestral relativamente balanceado. Este procedimiento no constituye un bloqueo geométrico estricto: los bloques delimitados por cuantiles pueden contener puntos cercanos entre bloques adyacentes, lo que puede permitir cierta fuga de información residual [50].

4.7 Contraste estadístico entre modelos

Tras la identificación de las configuraciones con desempeño superior bajo validación espacial, se empleó la prueba de Friedman como contraste global no paramétrico para evaluar la

igualdad de rendimiento entre los modelos [51]. Ante resultados significativos en dicho contraste, se procedió con un análisis post-hoc mediante pruebas de rangos con signo de Wilcoxon, aplicando el ajuste de Holm para mitigar la inflación del error de tipo I derivada de las comparaciones múltiples [52].

4.7.1 Evaluación de incertidumbre y análisis residual

Para los modelos determinísticos (Gradient Boosting, EBM, SVR, ARD, Extra Trees), la incertidumbre se estimó mediante remuestreo bootstrap con $N = 300$ réplicas, garantizando la estabilidad de los percentiles 2.5 y 97.5 que definen la banda de incertidumbre al 95%. Para el modelo probabilístico GPR, la varianza predictiva es endógena a la formulación bayesiana del proceso gaussiano, sin requerir procedimientos de remuestreo externo [33].

Para las redes neuronales se empleó Monte Carlo Dropout, que aprovecha la interpretación bayesiana aproximada del dropout para obtener una distribución empírica de predicciones bajo inferencia estocástica [53].

4.7.2 Interpretación de los modelos y aporte de covariables

El análisis de interpretabilidad se realizó mediante tres estrategias complementarias sobre el modelo campeón de cada familia: (i) SHAP Kernel sobre la matriz preprocesada, para descomponer la predicción en aportes aditivos locales y caracterizar patrones globales de dependencia [54];(ii) importancia por permutación sobre el pipeline completo ajustado, con 30 permutaciones por variable y reescalamiento como aporte relativo porcentual; [10] y (iii) eliminación recursiva de variables con validación cruzada espacial (RFECV-GroupKFold), para identificar subconjuntos parsimoniosos que conserven el desempeño bajo condiciones de generalización espacial.

4.8 Selección del modelo final

4.8.1 Comparación entre alternativas competitivas

Concluidas las fases de screening, ajuste fino, validación espacial y análisis de incertidumbre por familia, se compararon los modelos ganadores de cada familia mediante los mismos contrastes no paramétricos (Friedman + Wilcoxon-Holm) aplicados sobre los folds espaciales. En escenarios donde las configuraciones individuales no alcanzaron niveles de precisión considerados satisfactorios, se evaluó el ensamblaje por stacking, donde las predicciones

out-of-fold de los modelos base sirvieron como entrada a un meta-modelo seleccionado bajo el mismo esquema de validación espacial.

4.8.2 Criterios de selección del modelo óptimo

La selección del modelo óptimo se sustentó en un enfoque multicriterio que integró: R^2 espacial medio y su variabilidad entre folds, RMSE espacial medio, MAE espacial medio, métricas bootstrap con intervalos de confianza al 95%, comportamiento residual (media, desviación estándar, cobertura del 95%) y ancho de la banda predictiva. Se otorgó peso prioritario al desempeño bajo validación cruzada espacial, en tanto proporciona la estimación más exigente de la capacidad de generalización territorial. La consistencia estadística de las diferencias observadas fue evaluada mediante los contrastes de Friedman y Wilcoxon-Holm.

4.9 Modelación espacial y generación cartográfica

La generación de cartografía predictiva del COS constituye la fase operativa mediante la cual el modelo seleccionado —aquel que supere los criterios de desempeño, robustez frente al sobreajuste y consistencia estadística establecidos en las secciones anteriores— se aplica sobre la extensión continua del área de estudio para producir superficies espacialmente explícitas de estimación e incertidumbre. El escenario y la configuración exacta que llegarán a esta fase no se [55]determinan a priori, sino como resultado del proceso comparativo descrito. Esta sección define el protocolo de producción cartográfica aplicable al modelo finalmente seleccionado [15].

El marco conceptual se apoya en el modelo SCORPAN, según el cual cualquier propiedad del suelo puede expresarse como función de sus propiedades vecinas, el clima, los organismos, el relieve, el material parental, la edad y la posición espacial, extendiendo los factores clásicos de formación del suelo e incorporando explícitamente la dimensión espacial como predictor. La extrapolación opera bajo el supuesto de que las relaciones aprendidas a partir de las $n = 145$ observaciones puntuales se mantienen válidas en ubicaciones no observadas dentro del mismo dominio ambiental [33].

4.9.1 Protocolo de producción cartográfica

El protocolo se estructura en ocho etapas secuenciales, diseñadas para garantizar la trazabilidad y reproducibilidad de los productos:

Tabla 4 Flujo operativo de la modelación espacial

Etapa	Fase	Descripción
1	Selección del modelo	Identificación del modelo y escenario óptimo según criterios multicriterio
2	Capas ráster	Compilación, reproyección y recorte de covariables al CRS y resolución objetivo (3 m)
3	Grilla de predicción	Apilamiento multivariado y vectorización a formato tabular píxel × covariable
4	Preprocesamiento espacial	Aplicación de transformaciones y escaladores entrenados sobre la grilla
5	Predicción píxel a píxel	Inferencia del modelo sobre cada píxel válido del dominio
6	Incertidumbre	Bootstrap $N \geq 1.000$ (determinísticos) o varianza predictiva endógena (GPR)
7	Reconstrucción ráster	Conversión del vector de predicciones a GeoTIFF georreferenciado
8	Producción cartográfica	Generación de mapas de predicción media, incertidumbre e IC al 95%

Los tres productos cartográficos por escenario seleccionado son: (i) mapa de predicción media del COS en porcentaje, con paleta secuencial para discriminar variaciones graduales; (ii) mapa de incertidumbre predictiva (desviación estándar por píxel), que identifica zonas de alta y baja confianza; y (iii) mapa del ancho del intervalo de confianza al 95% (percentiles 2.5–97.5), como medida complementaria orientada a usuarios no técnicos y tomadores de decisiones.

4.10 Adquisición y preprocesamiento de datos de percepción remota

4.10.1 Fuentes de datos espectrales

El estudio adoptó una estrategia multisensor que combina dos plataformas: imágenes multiespectrales PlanetScope (3 m) y ortomosaicos RGB de alta resolución capturados mediante UAV operado por el CTT del IES CINOC. PlanetScope proporciona cuatro bandas espectrales — azul (455–515 nm), verde (500–590 nm), rojo (590–670 nm) e infrarrojo cercano (NIR, 780–860 nm)— corregidas atmosféricamente a reflectancia de superficie mediante el algoritmo 6SV2. Las imágenes corresponden al producto *PSScene*, obtenido mediante el Programa de Educación e Investigación de Planet Labs. Los ortomosaicos RGB del IES CINOC complementan la dimensionalidad espectral con índices de bandas visibles, particularmente útiles para discriminar variaciones finas de cobertura a escala predial [56].

Criterios de selección de escenas

Las imágenes PlanetScope fueron seleccionadas con base en: cobertura de nubes inferior al 10% sobre el área de interés, coincidencia temporal razonable con la campaña de muestreo de suelos,

ausencia de artefactos radiométricos visibles y calidad geométrica verificada mediante superposición con los puntos de muestreo georreferenciados.

4.10.2 Índices espectrales derivados

Los índices espectrales se calcularon píxel a píxel a partir de las bandas de reflectancia de superficie, y sus valores fueron muestreados espacialmente en las coordenadas de las 145 muestras de suelo mediante intersección punto-ráster (rasterio, geopandas).

Tabla 5 Índices espectrales derivados de PlanetScope (NIR) e imágenes RGB

Índice	Fuente	Nombre completo	Justificación
NDVI	PlanetScope	Normalized Difference Vegetation Index	Vigor, biomasa y actividad fotosintética
SAVI	PlanetScope	Soil Adjusted Vegetation Index	Cobertura vegetal con corrección por brillo del suelo
GNDVI	PlanetScope	Green NDVI	Sensibilidad a variaciones de clorofila y nitrógeno foliar
TVI	PlanetScope	Triangular Vegetation Index	Biomasa foliar; menor saturación en vegetación densa
NGRDI	RGB-UAV	Normalized Green-Red Difference Index	Verdor y condición vegetal sin banda NIR
VARI	RGB-UAV	Visible Atmospherically Resistant Index	Fracción de cobertura vegetal; robusto a efectos atmosféricos
ExG	RGB-UAV	Excess Green Index	Discriminación cobertura vegetal vs. suelo desnudo
TGI	RGB-UAV	Triangular Greenness Index	Contenido relativo de clorofila desde bandas RGB
GCC	RGB-UAV	Green Chromatic Coordinate	Proxy de cobertura vegetal y fenología

4.11 Anexos del estudio

El conjunto de anexos referenciado a lo largo del documento se encuentra disponible para consulta en línea, a través del repositorio del proyecto: ricardomontoyainfante/ModelacionCOS

5. RESULTADOS

5.1 Resultados - Objetivo 1: Analizar las relaciones estadísticas entre las mediciones de COS y un conjunto consolidado de covariables ambientales para identificar los predictores más relevantes en la Granja San José.

Este capítulo presenta los resultados de la primera fase del proyecto, centrada en la construcción de la base de datos analítica y la exploración de las relaciones estadísticas entre el Carbono Orgánico del Suelo (COS) y un conjunto de covariables geoambientales. El desarrollo de estas actividades da cumplimiento al primer objetivo específico: analizar las relaciones estadísticas entre las mediciones de COS y un conjunto consolidado de covariables ambientales para identificar los predictores más relevantes en la Granja San José.

5.1.1 Consolidación de la Base de Datos Analítica

La fase inicial de resultados se sustentó en la consolidación de una base de datos multifuente construida para la modelación del carbono orgánico del suelo (COS), mediante la integración de mediciones edáficas de laboratorio con covariables ambientales derivadas de distintas fuentes geoespaciales. La base final quedó conformada por 145 muestras de suelo georreferenciadas procedentes del Centro de Transferencia Tecnológica del IES CINOC, para las cuales se dispuso de determinaciones del COS y de un conjunto complementario de propiedades físico-químicas del suelo. Sobre esta estructura observacional se integraron, además, predictores topográficos, categóricos y espectrales, conformando un sistema analítico de 48 covariables potenciales para el área de estudio, cuyo detalle se presenta en la relación completa de variables y fuentes incluida en el Anexo 5.

Tabla 6 Tipo y función de variables usadas en el estudio

Tipo de variable	Variables incluidas	n	Función analítica en el estudio
Georreferenciación y soporte espacial	POINT_X, POINT_Y, asnm	3	Representación de la localización espacial de las observaciones y del gradiente altitudinal del área de estudio; soporte para el muestreo espacial de covariables y la estructuración de la validación espacial.

Edáficas	pH, N, MO, K, Ca, Mg, CIC, P, Fe, Mn, Zn, Cu, S, B, Al	1 5	Caracterización química del suelo asociada con fertilidad, acidez, reserva de nutrientes y dinámica del COS.
Edáficas texturales	Ar, L, A	3	Representación de la composición granulométrica del suelo y su relación con retención de agua, aireación y protección físico-química del carbono.
Categóricas edáficas y de cobertura	UCS_F, Tipo_Cober, Clasif, Relación C/N	4	Incorporación de diferencias estructurales asociadas a unidades de suelo, clase textural y tipo de cobertura.
Variable respuesta	C	1	Contenido de carbono orgánico del suelo a estimar.
Variables derivadas de MDE / topográficas	dem1, dem_aspect1, dem_aspect_slope1, slope_pt, aspect_pt, curvature_pt, TPI_w3_pt, TRI_w3_pt, Rough_w3_pt, TPI_w5_pt, TRI_w5_pt, Rough_w5_pt, TPI_w9_pt, TRI_w9_pt, Rough_w9_pt	1 5	Captura de gradientes geomorfológicos, orientación, pendiente, curvatura y métricas de posición y rugosidad del terreno en distintas ventanas espaciales.
Espectrales	NGRDI_pt, VARI_pt, ExG_pt, TGI_pt, GCC_pt, SAVI_pt, GNDVI_pt	7	Representación de vigor vegetativo, verdor, cobertura y condición biofísica de la superficie mediante índices RGB y multispectrales.

La incorporación de covariables espectrales se realizó mediante la extracción puntual de índices RGB y NIR en las coordenadas de muestreo, a través de procedimientos de intersección punto-ráster, con el propósito de asegurar la correspondencia espacial entre las observaciones edáficas y la información derivada de percepción remota. Este procedimiento permitió integrar al espacio analítico señales biofísicas asociadas a la cobertura y condición de la vegetación, potencialmente vinculadas con la variabilidad del COS. Desde el punto de vista de la integridad del conjunto de datos, la base consolidada no presentó valores ausentes, lo que permitió avanzar directamente hacia las etapas de exploración estadística y modelación sin recurrir a procedimientos de imputación.

5.1.2 Análisis Exploratorio y Caracterización de la Variable Objetivo (COS)

Evaluación de la calidad y consistencia del conjunto de datos

La evaluación de la calidad y consistencia del conjunto de datos se realizó con el propósito de verificar su integridad interna y garantizar condiciones adecuadas para la validez de los análisis posteriores. Como se sintetiza en la Tabla 7 Evaluación de calidad y estructura del conjunto de datos, la base analítica presentó un alto nivel de completitud, sin registrar valores ausentes en las covariables principales ni en la variable objetivo. Del mismo modo, se observaron distribuciones asimétricas y valores elevados en variables como fósforo, potasio y cationes intercambiables; sin

embargo, tales comportamientos fueron interpretados como manifestaciones plausibles de la heterogeneidad edafológica del área de estudio, más que como anomalías atribuibles a errores de medición o registro.

Tabla 7 Evaluación de calidad y estructura del conjunto de datos

Criterio evaluado	Resultado observado	Soporte complementario
Coherencia general de rangos y valores faltantes	Los valores observados fueron consistentes con la variabilidad esperada del sistema edáfico estudiado, sin registrar datos faltantes en las variables analizadas. El conjunto de datos reflejó suelos moderadamente ácidos, con contenidos relativamente altos de materia orgánica, carbono y capacidad de intercambio catiónico, así como predominio de texturas medias a finas.	Anexo 6. Indicadores estadísticos de variables cuantitativas. Anexo 7. Distribución y frecuencias de variables cualitativas.
Distribución de variables e identificación de valores extremos	Se identificaron distribuciones asimétricas en varias covariables y valores elevados compatibles con la heterogeneidad edafológica del sistema, sin evidencia de anomalías que justificaran su exclusión. En particular, nutrientes como fósforo y algunos cationes intercambiables exhibieron mayor dispersión y asimetría, mientras que el COS y variables estrechamente asociadas a la fracción orgánica del suelo mostraron un comportamiento comparativamente más consistente.	Anexo 8. Análisis de distribución y pruebas de normalidad.

5.1.3 Análisis univariado

Variables cuantitativas

Dentro de las variables cuantitativas, el análisis descriptivo mostró que el COS presentó una variabilidad apreciable dentro del área de estudio, evidenciando contrastes suficientes para sustentar un ejercicio de modelación predictiva. A su vez, variables como materia orgánica (MO), nitrógeno (N) y capacidad de intercambio catiónico (CIC) exhibieron comportamientos coherentes con su estrecha relación funcional con la fracción orgánica del suelo, mientras que nutrientes como fósforo (P), potasio (K), calcio (Ca), magnesio (Mg), zinc (Zn) y cobre (Cu) mostraron distribuciones más asimétricas y heterogéneas, reflejando una estructura edáfica espacialmente contrastante. En particular, el fósforo presentó una marcada concentración de valores bajos con algunos registros altos puntuales, y micronutrientes como el Zinc (Zn) evidenciaron colas derechas pronunciadas, lo que sugiere que parte de la señal cuantitativa del sistema se encuentra dominada por distribuciones no gaussianas y por la coexistencia de condiciones locales diferenciadas de fertilidad y manejo. El detalle de los estadísticos descriptivos, las medidas de forma y la

interpretación edafológica ampliada de las variables cuantitativas se presenta en el Anexo 6. Indicadores estadísticos de variables cuantitativas.

Desde la perspectiva textural, la fracción arcillosa mostró predominio dentro del conjunto analizado, mientras que las proporciones de arena y limo oscilaron dentro de rangos intermedios, configurando un sistema de texturas medias a finas. Este rasgo resulta analíticamente relevante, dado que la textura constituye uno de los principales mecanismos de control sobre la estabilización físico-química del carbono orgánico del suelo. De forma complementaria, la distribución altitudinal relativamente acotada del conjunto sugiere que, aunque la altitud introduce un gradiente ambiental relevante, su efecto debe interpretarse más como modulador del contexto climático y biológico local que como un factor de contraste extremo dentro del área de estudio.

Tabla 8 Variables cuantitativas edáficas incluidas en el estudio

Variable o bloque	Observaciones	Aspectos relevantes	Implicaciones para la modelación
Materia orgánica (MO)	Asociación estrecha con el COS	MO presentó media de 17.12 %, DE de 4.23 y rango de 4.4 a 26.8, mientras que el COS mostró media de 8.23 %, DE de 1.21 y rango de 4.6 a 11.0, con patrones distributivos coherentes entre ambas variables.	Predictor estructural clave por su proximidad funcional con la variable respuesta (El carbono representa una porción estructural de la Materia orgánica)
Nitrógeno (N)	Comportamiento coherente con la fracción orgánica	N presentó media de 0.59 %, DE de 0.11 y rango de 0.20 a 0.78, con baja dispersión relativa y distribución compatible con su estrecha asociación con MO y COS	Alta proximidad funcional con la respuesta y posible redundancia informacional
CIC	Relación consistente con el COS	CIC registró media de 29.35 cmol(+)/kg, DE de 4.69 y rango de 15 a 42, mostrando una variación estable y compatible con su vínculo con la fracción orgánica del suelo	Indicador de capacidad funcional del suelo y de estabilización del carbono
Textura	Predominio de arcilla y texturas medias a finas	La arcilla (A) presentó media de 57.79 %, frente a 18.8 % de arena y 23.35 % de limo; además, las clases franco arenoso y franco arcilloso arenoso concentraron cerca del 89 % de las muestras	Evidencia de posible control textural sobre la estabilización físico-química del COS
P, K, Ca y Mg	Alta dispersión y asimetría	P presentó media de 9.52 mg/kg, DE de 8.21, asimetría de 4.61 y curtosis de 36.89;	Señal indirecta y espacialmente heterogénea, posiblemente

		K media de 0.36 y asimetría de 1.76; Ca media de 0.98 y asimetría de 3.23; Mg media de 0.25 y asimetría de 3.37	influida por fertilidad localizada y manejo
Zn y Cu	Valores elevados puntuales y colas derechas	Zn presentó media de 8.7, DE de 12.3, máximo de 92.8, asimetría de 4.74 y curtosis de 24.84; Cu presentó media de 3.42, máximo de 19.1, asimetría de 3.37 y curtosis de 15.39	Señal complementaria, menos estructural que la observada en las covariables orgánicas
Altitud	Variación relativamente acotada	Altitud presentó media de 1921.82 msnm, DE de 43.8 y asimetría de -0.38, indicando dispersión moderada dentro del gradiente altitudinal del área de estudio	Gradiente ambiental secundario, con posible efecto modulador más que dominante

Variables cualitativas

En cuanto a las variables cualitativas, la estructura de frecuencias reveló un claro predominio de determinadas clases texturales, tipos de cobertura y unidades cartográficas de suelo, lo que permitió reconocer desde esta fase preliminar la existencia de contextos edáficos y de uso del suelo con peso diferencial dentro de la muestra. La clasificación textural estuvo dominada por las clases franco arenoso y franco arcilloso arenoso, mientras que, en términos de cobertura, los sistemas silvopastoriles y de bosque concentraron la mayor proporción de observaciones. Por su parte, la distribución de las unidades cartográficas de suelo mostró una participación importante de las clases GAe_{fl}, S_{lde} y SEe_{fl}. Más allá de su función descriptiva, estas frecuencias permiten anticipar que la variabilidad del COS no se distribuye sobre un sistema homogéneo, sino sobre una estructura espacial y funcionalmente estratificada, en la que la textura, la cobertura y la unidad de suelo constituyen dimensiones relevantes de diferenciación. El desarrollo detallado de las tablas de frecuencia, estadísticas por categoría, distribuciones y representaciones gráficas asociadas se presenta en el Anexo 7. Distribución y frecuencias de variables cualitativas.

Tabla 9 Variables cualitativas incluidas en el estudio

Variable cualitativa	Hallazgo principal	Evidencia cuantitativa relevante	Relevancia para la modelación
----------------------	--------------------	----------------------------------	-------------------------------

Clasificación textural	Predominaron texturas medias, con clara concentración en clases franco arenoso y franco arcilloso arenoso	FrancoArenoso representó 57.93 % de las muestras y FrancoArcillosoArenoso 31.03 %; en conjunto, ambas clases concentraron 88.96 % del total	Indica un sistema textural relativamente concentrado, pero con contraste suficiente para evaluar la influencia de la textura sobre el COS
COS según clasificación textural	El COS mostró diferencias entre clases texturales, con menores valores y mayor dispersión en la clase Franco	FrancoArenoso presentó media de C = 84.31; FrancoArcillosoArenoso, 83.49; Franco, 66.53, además de la mayor variabilidad (IQR = 22.50). Anexo 8. Análisis de distribución y pruebas de normalidad, apartado Carbono (C)	Sugiere que la textura modula la acumulación y estabilidad del carbono, con distinto grado de control según la clase textural
Tipo de cobertura	Predominaron los sistemas silvopastoriles, seguidos por bosque	Silvopastoril representó 54.48 % de las observaciones y Bosque 27.59 %; Agrícola y Pasturas tuvieron menor representación (11.72 % y 6.21 %, respectivamente)	Evidencia una estructura de cobertura no homogénea, relevante para interpretar el efecto del uso/cobertura sobre el COS
COS según tipo de cobertura	El COS presentó los valores más altos en sistemas silvopastoriles y los más bajos en bosque	Silvopastoril mostró media de C = 85.92 y mediana = 86.00; Pasturas media = 83.11; Agrícola media = 81.06; Bosque media = 75.35 y mayor dispersión (IQR = 19.25). Anexo 8. Análisis de distribución y pruebas de normalidad, apartado Carbono (C)	Sugiere que la cobertura del suelo modula el contenido de COS, aunque su efecto puede estar condicionado por textura y manejo local
Clase agrológica (UCS_F)	La muestra se concentró principalmente en las unidades GAefl, SIde y SEefl	GAefl representó 42.07 % de las muestras, SIde 30.34 % y SEefl 27.59 %	Confirma la existencia de estratificación edáfica dentro del área de estudio y justifica evaluar el comportamiento del COS según unidades de suelo
COS según clase agrológica	La clase GAefl presentó contenidos de COS más altos y más estables	GAefl mostró media de C = 84.67 e IQR = 10.0, frente a SIde (media = 81.18; IQR = 21.5) y SEefl (media = 79.78; IQR = 15.5) Anexo 8. Análisis de distribución y pruebas de normalidad, apartado Carbono (C)	Sugiere que las unidades cartográficas de suelo capturan diferencias estructurales relevantes en la variabilidad del COS

5.1.4 Análisis bivariado

A partir de la estructura observada en el análisis univariado, se realizó un análisis bivariado con el fin de examinar la relación entre el carbono orgánico del suelo (COS) y las covariables del estudio.

Asociación entre el COS y las variables cuantitativas

El análisis bivariado de las variables cuantitativas muestra que la variabilidad del carbono orgánico del suelo (COS) se encuentra mucho más estrechamente asociada con covariables edáficas vinculadas a la fracción orgánica y a la capacidad funcional del suelo que con variables de fertilidad puntual o de naturaleza mineral aislada. En las variables que mostraron un comportamiento compatible con el uso de correlación de Pearson (*Anexo 10. Análisis de Correlación*), la asociación más intensa se observó entre el COS y la materia orgánica (MO), con un coeficiente prácticamente perfecto ($r = 1.00$), seguida por el nitrógeno total (N), con una correlación muy alta ($r = 0.94$), y por la capacidad de intercambio catiónico (CIC), con una asociación positiva moderada ($r = 0.59$).

Otras covariables químicas mostraron asociaciones de menor magnitud o de signo inverso. El aluminio intercambiable presentó una correlación positiva relativamente alta con el COS ($r = 0.70$), lo que sugiere una coexistencia entre mayores contenidos de carbono y ambientes edáficos ácidos con presencia significativa de Al^{3+} , rasgo coherente con sistemas de montaña húmeda donde la acumulación de materia orgánica puede ocurrir bajo condiciones de acidez persistente. Por el contrario, el calcio mostró una asociación negativa moderada ($r = -0.43$), mientras que el pH presentó una correlación negativa débil a moderada ($r = -0.29$), resultado consistente con la lógica edafológica de suelos ácidos donde menores valores de pH y menor saturación de bases coexisten con una mayor conservación relativa de carbono orgánico. El magnesio, por su parte, exhibió una relación negativa débil ($r = -0.12$), lo que indica un aporte explicativo comparativamente marginal dentro del bloque edáfico considerado.

En las variables cuya estructura distributiva no justificó una aproximación paramétrica, el examen mediante correlación de Spearman mostró que fósforo ($\rho = 0.06$) y potasio ($\rho = 0.10$) mantuvieron asociaciones prácticamente nulas con el COS, lo que indica que su relación con la variable respuesta no sigue un patrón monotónico relevante dentro del conjunto analizado.

Variación del COS según variables cualitativas

El análisis del carbono orgánico del suelo (COS) en función de las variables cualitativas (*Anexo 8. Análisis de distribución y pruebas de normalidad, apartado Carbono -C-; Anexo 9.*

Variable objetivo y covariables edáficas) mostró que su distribución no se organiza sobre un sistema homogéneo, sino sobre una estructura edáfica funcionalmente estratificada por diferencias texturales (Anexo 9. Variable objetivo y covariables edáficas, apartado Clasificación textural), de cobertura (Anexo 9. Variable objetivo y covariables edáficas, apartado Cobertura de suelo y contenido de materia orgánica) y de unidad cartográfica de suelo (Anexo 9. Variable objetivo y covariables edáficas, apartado Clase agrológica). En el caso de la textura, aunque las clases franco arenoso y franco arcilloso arenoso concentraron la mayor proporción de observaciones, el COS exhibió diferencias apreciables entre categorías, con menores valores centrales y mayor dispersión en la clase Franco, frente a contenidos más altos y relativamente más estables en FrancoArenoso y FrancoArcillosoArenoso. Este patrón sugiere que la textura actúa como un regulador efectivo de la estabilización del carbono, al condicionar procesos de agregación, retención de humedad y protección fisicoquímica de la materia orgánica.

Un comportamiento igualmente relevante se observó respecto al tipo de cobertura (*Anexo 9. Variable objetivo y covariables edáficas, apartado Cobertura de suelo y contenido de materia orgánica*). Los sistemas silvopastoriles presentaron los valores medios y medianos más altos de COS, mientras que la cobertura bosque exhibió menores valores centrales y una mayor dispersión. Por su parte, la clase agrológica (*Anexo 9. Variable objetivo y covariables edáficas, apartado Clase agrológica*) introdujo una dimensión adicional de diferenciación al condensar atributos relativamente estables del medio pedogeomorfológico. La clase GAefl presentó mayores contenidos medios de carbono y una menor dispersión interna que SIde y SEefl, lo que sugiere que la unidad cartográfica capta una señal territorial estructural asociada a combinaciones particulares de relieve, material parental, drenaje y uso del suelo.

Con el propósito de complementar la lectura descriptiva del carbono orgánico del suelo (COS) entre categorías, se aplicaron contrastes inferenciales sobre las principales variables cualitativas del estudio que se pueden consultar en el Anexo 11. Análisis del carbono orgánico del suelo según clases texturales. Dado que en los tres factores evaluados se verificó heterogeneidad de varianzas mediante la prueba de Levene, la comparación entre grupos se realizó mediante ANOVA de Welch y contrastes *post hoc* de Games–Howell, con el fin de preservar validez estadística bajo condiciones de varianza desigual.

En la clasificación textural, los resultados mostraron diferencias estadísticamente significativas en el contenido de carbono entre clases (ANOVA de Welch: $F = 9.73$; $p < 0.001$),

confirmando que la textura constituye un factor explicativo relevante de la variabilidad del COS. El análisis *post hoc* indicó que la clase Franco presentó contenidos de carbono significativamente inferiores a Franco Arcilloso Arenoso ($\Delta\text{media} = -16.96$; $p = 0.00056$) y a Franco Arenoso ($\Delta\text{media} = -17.78$; $p = 0.00037$), mientras que entre estas dos últimas no se detectaron diferencias significativas ($\Delta\text{media} = -0.82$; $p = 0.62$). En el caso del tipo de cobertura, el contraste también evidenció diferencias significativas entre grupos (ANOVA de Welch: $F = 6.16$; $p = 0.002$), lo que confirma que la cobertura modula de manera significativa la variabilidad del COS. Las comparaciones pareadas mostraron diferencias estadísticamente significativas entre Bosque y Silvopastoril ($\Delta\text{media} = -10.57$; $p = 0.00006$), así como entre Bosque y Pasturas ($\Delta\text{media} = -7.76$; $p = 0.0379$), con mayores contenidos de carbono en estos dos últimos sistemas. No se observaron diferencias significativas entre los demás pares, lo que indica que el principal contraste estructural se da entre la cobertura boscosa y ciertos sistemas productivos, especialmente aquellos con arreglo silvopastoril.

5.1.5 Relevancia predictiva y diagnóstico de multicolinealidad de las covariables

Análisis de relevancia de variables predictoras

Con el objetivo de identificar las variables con mayor capacidad explicativa sobre el carbono orgánico del suelo, se aplicó un procedimiento de selección basado en la prueba F de regresión ($f_{\text{regression}}$) mediante el método SelectKBest de la biblioteca *scikit-learn*. Este método evalúa la relación lineal entre cada predictor y la variable objetivo, permitiendo ordenar las variables según su contribución potencial al modelo.

Tabla 10 Selección de variables a través del método SelectKBest

Variable	Score	Interpretación
MO	41789	Variable extremadamente dominante
N	1085	Muy fuerte relación con COS
AI	137	Relación moderada-alta
CIC	75	Importancia relevante
L	52	Relación significativa

A	47	Relación significativa
Ca	32	Relación moderada
Zn	13	Relación baja-moderada
pH	13	Relación moderada
TGI_pt	10	Índice espectral relevante

Los resultados del análisis de relevancia de variables muestran que la materia orgánica (MO) presenta la mayor capacidad explicativa respecto al carbono orgánico del suelo, con un valor estadístico F significativamente superior al resto de predictores. Este resultado es consistente desde el punto de vista edafológico, dado que el carbono orgánico constituye una fracción principal de la materia orgánica del suelo. Otras variables con alta relevancia incluyen el nitrógeno (N), la capacidad de intercambio catiónico (CIC) y elementos como Al y Ca, los cuales reflejan propiedades químicas y de fertilidad del suelo estrechamente relacionadas con la dinámica del carbono. Se observa además la presencia de índices espectrales derivados de sensores remotos, como TGI, SAVI y GNDVI, lo cual indica que ciertas características de la vegetación y del estado del suelo capturadas mediante teledetección pueden aportar información relevante para la estimación del carbono orgánico del suelo.

5.1.6 Factor de Inflación de la Varianza (VIF)

El diagnóstico de multicolinealidad evidenció que el sistema predictor no es estadísticamente homogéneo, sino que presenta bloques de redundancia de distinta magnitud. Las mayores tensiones se concentraron en materia orgánica (MO), nitrógeno (N), altitud y, de forma especialmente marcada, en las fracciones texturales Arena, Limo y Arcilla, cuyos valores de VIF indican colinealidad estructural severa. Variables como CIC, P, Fe, Mn, S y B mostraron niveles bajos de redundancia, lo que sugiere una contribución más estable dentro del espacio predictor.

La modelación del carbono orgánico del suelo (COS) demanda una distinción crítica entre la capacidad explicativa y la redundancia funcional de los predictores. El diagnóstico de multicolinealidad revela que la materia orgánica y el nitrógeno total presentan una identidad estructural con la variable respuesta que, de ser incorporada irrestrictamente, induciría configuraciones predictivas artificialmente favorecidas y eclipsaría el aporte de otras covariables. Por el contrario, la elevada colinealidad en las fracciones texturales se identifica como una

consecuencia de su dependencia composicional intrínseca en un sistema cerrado, lo que justifica su permanencia para salvaguardar la integridad de los mecanismos físicos de estabilización edáfica. Así, la exclusión metodológica de los predictores dominantes no constituye una depuración estadística estándar, sino una estrategia para incrementar la exigencia analítica del sistema, permitiendo evaluar con mayor rigor la potencia de señales espectrales, topográficas y edáficas menos evidentes.

Tabla 11 Síntesis de los hallazgos del análisis exploratorio y su implicación para la modelación del carbono orgánico

Dimensión	Hallazgo principal	Evidencia reportada en resultados	Implicación para la modelación del COS
Estructura de la base analítica	Se consolidó una base multifuente completa, sin valores ausentes, con 145 observaciones y 46 covariables potenciales.	Integración de variables edáficas, texturales, categóricas, topográficas y espectrales en una sola base analítica.	Proporciona una base consistente y metodológicamente robusta para el análisis exploratorio y la modelación posterior.
Variable objetivo (COS)	El COS presentó variabilidad suficiente para sustentar el ejercicio predictivo.	Diferencias observadas en medias, dispersión y distribución según textura, cobertura y unidades de suelo.	Confirma la viabilidad estadística del problema de predicción y la existencia de señal suficiente para el modelado.
Predictores edáficos dominantes	MO y N mostraron la relación más fuerte con el COS; CIC y Al aportaron señal complementaria relevante.	Correlaciones altas con COS y valores elevados en el análisis de relevancia predictiva.	Identifica el núcleo de covariables de mayor proximidad funcional con la variable respuesta.
Textura del suelo	La textura introdujo diferencias estadísticamente significativas en el contenido de COS.	ANOVA de Welch significativo y contrastes post hoc entre clases texturales.	Sugiere un control físico relevante sobre la estabilización y protección del carbono orgánico.
Cobertura del suelo	El tipo de cobertura también introdujo contrastes significativos en la distribución del COS.	Diferencias entre bosque, silvopastoril y pasturas en medias y dispersión.	Indica que el uso/cobertura del suelo modula funcionalmente los aportes y la conservación de carbono.
Variables químicas secundarias	Nutrientes como P, K y otros elementos presentaron relaciones débiles o espacialmente heterogéneas con el COS.	Correlaciones bajas o nulas y distribuciones asimétricas.	Su aporte al modelo sería complementario, contextual o dependiente de condiciones locales de fertilidad y manejo.

Variables espectrales	Algunos índices espectrales mostraron señal explicativa incipiente, pero subordinada frente al bloque edáfico.	Presencia de ciertos índices entre variables con relevancia preliminar.	Justifica su incorporación en escenarios posteriores, aunque no como predictores dominantes del sistema.
Multicolinealidad	MO y N exhibieron redundancia estructural severa (a nivel de porción orgánica); las fracciones texturales presentaron colinealidad composicional (porción inorgánica del suelo).	Valores extremos de VIF para MO y N.	Justifica la evaluación de escenarios con exclusión selectiva de MO y N para someter el problema a una condición predictiva más exigente. Tanto los valores de MO y N son valores que se suelen determinar químicamente en análisis básicos de suelos y no tienen una proporción constante (relación C/N en la MO).

5.1.7 Síntesis integradora del objetivo 1

La evaluación de supuestos estadísticos mostró que una proporción importante de las covariables del sistema edáfico presenta distribuciones no normales, asimetrías marcadas, colas derechas pronunciadas y valores extremos, particularmente en variables asociadas a nutrientes y cationes intercambiables. Desde el punto de vista sustantivo, el análisis exploratorio, bivariado, inferencial y de relevancia predictiva permitió establecer que la variabilidad del carbono orgánico del suelo (COS) está dominada por un núcleo de predictores edáficos de alta proximidad funcional, conformado principalmente por la materia orgánica (MO) y el nitrógeno total (N), seguidos por la capacidad de intercambio catiónico (CIC) y, de manera complementaria, por el aluminio intercambiable (Al). Variables como calcio y pH presentaron asociaciones inversas de magnitud moderada a débil, mientras que fósforo y potasio mostraron relaciones prácticamente nulas o de aporte marginal. En el plano de las variables cualitativas, la textura y el tipo de cobertura introdujeron contrastes estadísticamente significativos en la distribución del COS, lo que permitió reconocerlas como factores estructurales relevantes en la modulación territorial del carbono, mientras que la clase agrológica aportó una señal descriptiva coherente, aunque sin soporte inferencial concluyente al nivel de significancia adoptado. Se identifica de forma explícita que los predictores más relevantes del COS en la Granja San José son MO y N como núcleo dominante, CIC y Al como señal edáfica complementaria, y la textura y la cobertura como factores categóricos estructurales; a su vez, se establece que la interpretación de MO y N debe realizarse con cuidado debido a su redundancia severa, evidenciada por los valores extremos de VIF, lo que justificó la

definición de escenarios diferenciados de modelación con y sin estas covariables para evaluar con mayor rigor la capacidad explicativa del resto del sistema predictor.

Como complemento a la determinación del conjunto óptimo de covariables para la estimación del COS, se aplicaron herramientas de interpretabilidad y selección *ex post* orientadas a evaluar la contribución funcional y la estabilidad de los predictores en los modelos campeones. Para ello, se emplearon SHAP Kernel sobre la matriz preprocesada, importancia por permutación sobre el pipeline completo y RFECV bajo validación cruzada espacial. Estas técnicas no se asumieron como mecanismos de inferencia causal, sino como soportes analíticos para contrastar la relevancia relativa de las covariables, reducir redundancia y verificar la consistencia del subconjunto predictor frente a la capacidad de generalización espacial del modelo.

5.2 Resultados - Objetivo 2: Comparar el desempeño predictivo de los algoritmos de aprendizaje automático para seleccionar el modelo óptimo de estimación de COS.

En el presente capítulo se exponen los resultados correspondientes a los modelos campeones de cada familia de aprendizaje automático evaluada. Los modelos campeones por familia se comparan dentro de cada escenario; posteriormente se selecciona un modelo definitivo por escenario. Por tanto, no todos los campeones familiares son modelos finales para los escenarios.

En cada apartado se presenta la tabla comparativa de resultados de cada una de las familias de modelos evaluados, presentando el modelo campeón en cada uno de los escenarios planteados, además de su configuración y desempeño espacial. El desarrollo técnico detallado del proceso de evaluación se consignó en los anexos correspondientes a cada familia de modelos y el escenario de modelación. En los anexos se documenta las configuraciones analizadas durante las etapas de *screening* y ajuste fino de hiperparámetros, las métricas completas de desempeño obtenidas bajo validación cruzada espacial, la identificación del modelo campeón con su respectiva configuración óptima, así como los resultados de incertidumbre y diagnóstico residual asociados al ajuste final. Según la naturaleza del modelo, se incluyen los diagnósticos de supuestos y condiciones del ajuste según cada caso.

5.2.1 Modelos lineales

Campeones de la familia de modelos lineales

Tabla 12 Síntesis de métricas de desempeño predictivo y validación de supuestos estadísticos por escenario de modelos lineales.

Escenario	Modelo	Configuración	Desempeño espacial	Supuestos
Escenario I sin exclusión de covariables	ARD	original + minmax + PCA=False	$R^2 = 0.9992$; RMSE = 0.0332; MAE = 0.0273	• Normalidad residual: Cumple. • Homocedasticidad: Cumple. • Independencia de errores: Cumple. • Multicolinealidad: No cumple.
Escenario I con exclusión de MO y N	ARD	cbt + robust + PCA=False	$R^2 = 0.6057$; RMSE = 0.7555; MAE = 0.5211	• Normalidad residual: No cumple. • Homocedasticidad: No cumple. • Independencia de errores: Cumple. • Multicolinealidad: No cumple.
Escenario II sin exclusión de covariables	ARD	original + minmax + PCA=False	$R^2 = 0.9993$; RMSE = 0.0322; MAE = 0.0264	• Normalidad residual: Cumple. • Homocedasticidad: Cumple. • Independencia de errores: Cumple. • Multicolinealidad: No cumple.
Escenario II con exclusión de MO y N	BayesianRidge	cbt + standard + PCA=False	$R^2 = 0.6850$; RMSE = 0.6752; MAE = 0.4724	• Normalidad residual: No cumple. • Homocedasticidad: No cumple. • Independencia de errores: Cumple. • Multicolinealidad: No cumple.
Escenario III espectral	BayesianRidge	original + minmax + PCA=False	$R^2 = -0.0884$; RMSE = 1.2552; MAE = 0.9572	• Normalidad residual: No cumple. • Homocedasticidad: Cumple. • Independencia de errores: Cumple. • Multicolinealidad: No cumple.

En los modelos lineales, ARD domina los escenarios con inclusión de MO y N, con desempeños prácticamente perfectos y errores mínimos, lo que evidencia una fuerte dependencia predictiva de covariables orgánicas estructuralmente próximas al COS. Al excluir MO y N, el rendimiento disminuye de forma importante, aunque BayesianRidge en el escenario II mantiene el mejor equilibrio lineal bajo restricción de información, con R^2 espacial moderado y errores aceptables. Es importante tener en cuenta que ninguno de los modelos ganadores en la familia cumple supuestos de linealidad.

5.2.2 Modelos basados en árboles

Campeones de la familia de modelos basados en árboles

Tabla 13 Síntesis de métricas de desempeño predictivo y validación de supuestos estadísticos por escenario de modelos basados en árboles

Escenario	Modelo	Configuración	Desempeño espacial	Supuestos
-----------	--------	---------------	--------------------	-----------

Escenario I sin exclusión de covariables	Gradient Boosting Regressor, GBR	yeojohnson	+	$R^2 = 0.9770$; RMSE = 0.0287; MAE = 0.1503	- Control de sobreajuste: Cumple. - Profundidad del modelo: Cumple. - Número de estimadores: Cumple. - Estabilidad por validación espacial: Cumple.
		learning_rate=0.05 max_depth=3 n_estimators=300	+		
Escenario I con exclusión de MO y N	Extra Trees	yeojohnson	+	$R^2 = 0.5458$; RMSE = 0.8108; MAE = 0.6159	- Control de sobreajuste: Cumple parcialmente. - Profundidad del modelo: Cumple. - Submuestreo de variables: Cumple. - Número de estimadores: Cumple. - Estabilidad por validación espacial: Cumple parcialmente.
		max_depth=12 max_features=0.7 min_samples_leaf=2 n_estimators=300	+		
Escenario II sin exclusión de covariables	Gradient Boosting Regressor, GBR	yeojohnson	+	$R^2 = 0.9797$; RMSE = 0.1714; MAE = 0.0781	- Control de sobreajuste: Cumple. - Profundidad del modelo: Cumple. - Número de estimadores: Cumple. - Estabilidad por validación espacial: Cumple.
		learning_rate=0.03 max_depth=3 n_estimators=300	+		
Escenario II con exclusión de MO y N	Random Forest, RF	original	+	$R^2 = 0.5807$; RMSE = 0.7790; MAE = 0.5936	- Control de sobreajuste: Cumple. - Profundidad del modelo: Cumple. - Submuestreo de variables: Cumple. - Número de estimadores: Cumple. - Diversidad entre árboles: Cumple. - Estabilidad por validación espacial: Cumple parcialmente.
		max_depth=12 max_features=0.7 min_samples_leaf=4 n_estimators=300	+		
Escenario III espectral	CatBoost	yeojohnson + depth=6 + iterations=300 learning_rate=0.05	+	$R^2 = 0.0550$; RMSE = 1.1695; MAE = 0.9343	- Profundidad del modelo: Cumple. - Número de iteraciones: Cumple. - Tasa de aprendizaje: Cumple. - Estabilidad por validación espacial: Limitada por bajo desempeño

En la familia de árboles y ensambles, los mejores desempeños se concentran en los escenarios con inclusión de MO y N, donde Gradient Boosting alcanza ajustes espaciales muy altos y errores reducidos. Al excluir estas covariables, el desempeño cae de forma marcada, aunque Extra Trees y Random Forest conservan una capacidad predictiva moderada, lo que indica que parte de la señal del COS puede recuperarse mediante relaciones no lineales entre covariables edáficas, topográficas y categóricas. En el escenario III, CatBoost fue el mejor dentro de la familia, pero su

R^2 cercano a cero confirma que los predictores espectrales aislados no son suficientes para estimar robustamente el COS bajo modelos de árboles.

5.2.3 Modelos probabilísticos

Campeones de la familia de modelos probabilísticos

Tabla 14 Síntesis de métricas de desempeño predictivo y validación de supuestos estadísticos por escenario de modelos probabilísticos

Escenario	Modelo	Configuración	Desempeño espacial	Supuestos
Escenario I sin exclusión de covariables	GPR_Dot	original + standard	$R^2 = 0.9979$; RMSE = 0.0551; MAE = 0.0383	• Coherencia del kernel: Cumple. • Regularización / ruido del modelo: Cumple. • Calibración de incertidumbre: Cumple. • Estabilidad por validación espacial: Cumple. • Coherencia del kernel: Cumple. • Regularización / ruido del modelo: Cumple.
Escenario I con exclusión de MO y N	GPR_Matern	yeojohnson + standard + alpha = 1e-10	$R^2 = 0.4276$; RMSE = 0.9103; MAE = 0.6569	• Coherencia del kernel: Cumple parcialmente. • Regularización / ruido del modelo: Cumple parcialmente. • Estabilidad por validación espacial: Cumple parcialmente. • Coherencia del kernel: Cumple. • Regularización / ruido del modelo: Cumple. • Calibración de incertidumbre: Cumple parcialmente. • Estabilidad por validación espacial: Cumple parcialmente.
Escenario II sin exclusión de covariables	GPR_Dot	original + robust + alpha = 0.001	$R^2 = 0.9992$; RMSE = 0.0345; MAE = 0.0283	• Coherencia del kernel: Cumple. • Regularización / ruido del modelo: Cumple. • Calibración de incertidumbre: Cumple. • Estabilidad por validación espacial: Cumple. • Coherencia del kernel: Cumple. • Regularización / ruido del modelo: Cumple. • Calibración de incertidumbre: Cumple. • Estabilidad por validación espacial: Cumple.
Escenario II con exclusión de MO y N	GPR_Matern	yeojohnson + robust + alpha = 0.01	$R^2 = 0.6596$; RMSE = 0.7020; MAE = 0.5060	• Coherencia del kernel: Cumple. • Regularización / ruido del modelo: Cumple. • Calibración de incertidumbre: Cumple parcialmente. • Estabilidad por validación espacial: Cumple parcialmente. • Coherencia del kernel: Cumple. • Regularización / ruido del modelo: Cumple. • Calibración de incertidumbre: Cumple parcialmente. • Estabilidad por validación espacial: Cumple parcialmente.
Escenario III espectral	GPR_RBF	yeojohnson + standard + alpha = 1e-10	$R^2 = -0.0473$; RMSE = 1.2312; MAE = 0.9605	• Coherencia del kernel: Cumple. • Regularización / ruido del modelo: Cumple. • Calibración de incertidumbre: Cumple parcialmente. • Estabilidad por validación espacial: Limitada por bajo desempeño.

En la familia probabilística, los procesos gaussianos con kernel DotProduct fueron claramente superiores en los escenarios con inclusión de MO y N, alcanzando desempeños espaciales casi perfectos y niveles bajos de error e incertidumbre, lo que evidencia una fuerte capacidad para capturar relaciones estructurales dominadas por covariables orgánicas altamente informativas. Al excluir MO y N, el mejor desempeño se desplazó hacia GPR_Matern, con una reducción importante del ajuste pero con comportamiento todavía moderado, especialmente en el escenario II, donde la información edáfica remanente permitió conservar parte de la señal predictiva.

5.2.4 Modelos basados en funciones kernel

Campeones de la familia de modelos basados en funciones kernel

Tabla 15 Síntesis de métricas de desempeño predictivo y validación de supuestos estadísticos por escenario de modelos basados en funciones kernel

Escenario	Modelo	Configuración	Desempeño espacial	Supuestos
Escenario I sin exclusión de covariables	SVR_poly	cbrt + robust + degree = 2	$R^2 = 0.9484$; $RMSE = 0.2734$; $MAE = 0.1545$; 9 grupos espaciales	• Escalamiento de variables: Cumple. • Coherencia del kernel: Cumple. • Regularización del margen: Cumple. • Estabilidad por validación espacial: Cumple.
Escenario I con exclusión de MO y N	KR_poly	yeojohnson + minmax	$R^2 = 0.5250$; $RMSE = 0.8292$; $MAE = 0.6338$; 9 grupos espaciales	• Escalamiento de variables: Cumple. • Coherencia del kernel: Cumple parcialmente. • Regularización del margen: Cumple parcialmente. • Estabilidad por validación espacial: Cumple parcialmente.
Escenario II sin exclusión de covariables	NuSVR_rbf	original + minmax + C = 50, gamma = auto, nu = 0.75	$R^2 = 0.9728$; $RMSE = 0.1984$; $MAE = 0.0693$	• Escalamiento de variables: Cumple. • Coherencia del kernel: Cumple. • Regularización del margen: Cumple. • Estabilidad por validación espacial: Cumple.
Escenario II con exclusión de MO y N	SVR_poly	yeojohnson + standard + C = 1, coef0 = 1.0, degree =	$R^2 = 0.7298$; $RMSE = 0.6253$; $MAE = 0.4449$	• Escalamiento de variables: Cumple. • Coherencia del kernel: Cumple. • Regularización del margen: Cumple.

Escenario espectral	III	NuSVR_rbf	2, epsilon = 0.1, gamma = scale	$R^2 = 0.0195$; $RMSE = 1.1913$; $MAE = 0.9527$	- Estabilidad por validación espacial: Cumple parcialmente. - Escalamiento de variables: Cumple. - Coherencia del kernel: Cumple. - Regularización del margen: Cumple. - Estabilidad por validación espacial: Limitada por bajo desempeño.
			yeojohnson + standard + C = 1, gamma = scale, nu = 0.5		

En la familia kernel, el mejor equilibrio entre rendimiento y estabilidad se observó en el escenario II sin exclusión, donde NuSVR_rbf alcanzó un R^2 alto, errores bajos y una incertidumbre estrecha, lo que indica que la estructura no lineal inducida por kernel aprovechó eficazmente la señal edáfica completa. En el escenario I sin exclusión, SVR_poly también mostró desempeño robusto, aunque con mayor error e incertidumbre. Al excluir MO y N, la familia kernel conservó capacidad predictiva moderada, especialmente en el escenario II con SVR_poly, pero con pérdida clara de estabilidad e incremento de incertidumbre.

5.2.5 Modelos no paramétricos

Campeones de la familia de modelos no paramétricos

Tabla 16 Síntesis de métricas de desempeño predictivo y validación de supuestos estadísticos por escenario de modelos no paramétricos

Escenario	Modelo	Configuración	Desempeño espacial	Supuestos
Escenario I sin exclusión de covariables	KNN_distance	yeojohnson + standard	$R^2 = 0.4478$; $RMSE = 0.8940$; $MAE = 0.6638$	- Escalamiento de variables: Cumple. - Validez de la métrica de distancia: Cumple parcialmente. - Densidad local de observaciones: Cumple parcialmente. - Estabilidad por validación espacial: Cumple parcialmente.
		yeojohnson + standard + n_neighbors = 7 + p = 1		- Escalamiento de variables: Cumple. - Validez de la métrica de distancia: Cumple parcialmente. - Densidad local de observaciones: No cumple plenamente. - Estabilidad por validación espacial: Limitada por bajo desempeño.
Escenario I con exclusión de MO y N	KNN_distance		$R^2 = 0.2394$; $RMSE = 1.0493$; $MAE = 0.7915$	

Escenario II sin exclusión de covariables	KNN_uniform	yeojohnson	+	$R^2 = 0.4478$; RMSE = 0.8940; MAE = 0.6638		• Escalamiento de variables: Cumple. • Validez de la métrica de distancia: Cumple parcialmente. • Densidad local de observaciones: Cumple parcialmente. • Estabilidad por validación espacial: Cumple parcialmente.
		standard	+			
Escenario II con exclusión de MO y N	KNN_distance	n_neighbors = 5		$R^2 = 0.2394$; RMSE = 1.0493; MAE = 0.7915		• Escalamiento de variables: Cumple. • Validez de la métrica de distancia: Cumple parcialmente. • Densidad local de observaciones: No cumple plenamente. • Estabilidad por validación espacial: Limitada por bajo desempeño.
		+ p = 1				
Escenario III espectral	KNN_distance	yeojohnson	+	$R^2 = -0.0045$; RMSE = 1.2058; MAE = 0.9503		• Escalamiento de variables: Cumple. • Validez de la métrica de distancia: No cumple plenamente. • Densidad local de observaciones: No cumple plenamente. • Estabilidad por validación espacial: Limitada por desempeño nulo.
		minmax	+			
		n_neighbors = 11	=			
		+ p = 1				

En la familia no paramétrica, los modelos KNN presentaron desempeños bajos a moderados, con mejores resultados en los escenarios con inclusión de MO y N, aunque sin alcanzar el rendimiento de las familias lineal, probabilística, kernel o aditiva. La exclusión de MO y N redujo de forma importante la capacidad predictiva, lo que indica que la señal remanente no fue suficientemente estructurada en el espacio de distancias para sostener predicciones locales robustas. En el escenario III espectral, el R^2 cercano a cero y los errores elevados evidencian que la información espectral aislada no generó vecindades predictivamente consistentes. En términos metodológicos, esta familia funcionó como contraste no paramétrico local, pero su desempeño sugiere limitaciones asociadas a dimensionalidad, baja densidad efectiva de observaciones y sensibilidad a la métrica de distancia.

5.2.6 Modelos Aditivos/GAM

Campeones de la familia de modelos aditivos/GAM

Tabla 17 Síntesis de métricas de desempeño predictivo y validación de supuestos estadísticos por escenario de modelos aditivos/GAM

Escenario	Modelo	Configuración	Desempeño espacial	Supuestos
Escenario I sin exclusión de covariables	EBM	original + standard + interactions = 5 + max_bins = 128 + learning_rate = 0.05 + max_rounds = 300	$R^2 = 0.8890$; RMSE = 0.4008; MAE = 0.2587	• Aditividad estructural: Cumple. • Control de interacciones: Cumple. • Control de complejidad: Cumple. • Estabilidad por validación espacial: Cumple.
Escenario I con exclusión de MO y N	EBM	original + standard + interactions = 10 + max_bins = 256 + learning_rate = 0.03 + max_rounds = 400	$R^2 = 0.6076$; RMSE = 0.7536; MAE = 0.5701	• Aditividad estructural: Cumple. • Control de interacciones: Cumple. • Control de complejidad: Cumple. • Estabilidad por validación espacial: Cumple parcialmente.
Escenario II sin exclusión de covariables	EBM	original + robust + interactions = 10 + max_bins = 128 + learning_rate = 0.05 + max_rounds = 300	$R^2 = 0.9308$; RMSE = 0.3164; MAE = 0.1745	• Aditividad estructural: Cumple. • Control de interacciones: Cumple. • Control de complejidad: Cumple. • Estabilidad por validación espacial: Cumple.
Escenario II con exclusión de MO y N	EBM	original + robust + interactions = 10 + max_bins = 128 + learning_rate = 0.05 + max_rounds = 300	$R^2 = 0.6497$; RMSE = 0.7121; MAE = 0.4950	• Aditividad estructural: Cumple. • Control de interacciones: Cumple. • Control de complejidad: Cumple. • Estabilidad por validación espacial: Cumple parcialmente.
Escenario III espectral	EBM	yeojohnson + standard + interactions = 5 + max_bins = 128	$R^2 = 0.1995$; RMSE = 1.0764; MAE = 0.8593	• Aditividad estructural: Cumple. • Control de interacciones: Cumple. • Control de complejidad: Cumple. • Estabilidad por validación espacial: Limitada por bajo desempeño.

En la familia aditivos/GAM, EBM fue el modelo campeón en todos los escenarios, lo que muestra una buena capacidad para capturar relaciones no lineales moderadas manteniendo interpretabilidad estructural. El mejor desempeño se presentó en el escenario II sin exclusión, donde la combinación de covariables edáficas con MO y N permitió alcanzar un R^2 alto y errores bajos. Al excluir MO y N, el desempeño disminuyó de forma clara, aunque EBM conservó una capacidad predictiva moderada y relativamente estable, especialmente en el escenario II.

5.2.7 Modelos basados en redes neuronales

Campeones de la familia de modelos basados en redes neuronales

Tabla 18 Síntesis de métricas de desempeño predictivo y validación de supuestos estadísticos por escenario de modelos basados en redes neuronales.

Escenario	Modelo	Configuración	Desempeño espacial	Supuestos
Escenario I sin exclusión de covariables	Residual MLP	selu_net + selu + dropout = 0.15 + lr = 0.001 + capas 128–128	$R^2 = -3.4474$; RMSE = 2.5372; MAE = 2.0556	- Control de sobreajuste: Cumple parcialmente. - Regularización: Cumple. - Estabilidad de entrenamiento: No cumple plenamente. - Estabilidad por validación espacial: No cumple.
Escenario I con exclusión de MO y N	MLP shallow	original + robust + softplus + batchnorm = 1 + dropout = 0.15 + lr = 0.002	$R^2 = -0.7486$; RMSE = 1.5909; MAE = 1.3180	- Control de sobreajuste: Cumple parcialmente. - Regularización: Cumple. - Estabilidad de entrenamiento: Cumple parcialmente. - Estabilidad por validación espacial: No cumple.
Escenario II sin exclusión de covariables	MLP residual	elu + residual_mlp + dropout = 0.10 + lr = 0.0008 + n_layers = 3	$R^2 = -2.3090$; RMSE = 2.1885; MAE = 1.6816	- Control de sobreajuste: Cumple parcialmente. - Regularización: Cumple parcialmente. - Estabilidad de entrenamiento: No cumple plenamente. - Estabilidad por validación espacial: No cumple.
Escenario II con exclusión de MO y N	MLP residual	softplus + residual_mlp + dropout = 0.05 + lr = 0.0005 + n_layers = 2	$R^2 = -0.3152$; RMSE = 1.3798; MAE = 1.0582	- Control de sobreajuste: Cumple parcialmente. - Regularización: Cumple parcialmente. - Estabilidad de entrenamiento: Cumple parcialmente. - Estabilidad por validación espacial: No cumple.
Escenario III espectral	MLP deep	original + robust + relu + dropout = 0.20 + l2_reg = 0.003 + lr = 0.0005	$R^2 = -22.9536$; RMSE = 5.8883; MAE = 3.4711	- Control de sobreajuste: No cumple plenamente. - Regularización: Cumple. - Estabilidad de entrenamiento: No cumple. - Estabilidad por validación espacial: No cumple.

En la familia de redes neuronales, todos los modelos campeones presentaron R^2 espacial negativo, lo que indica que las arquitecturas evaluadas no lograron generalizar adecuadamente frente a la estructura espacial del COS. Aunque se implementaron mecanismos de regularización como dropout, batch normalization, activaciones no lineales y arquitecturas residuales, los errores fueron altos y la incertidumbre amplia, especialmente en los escenarios I sin exclusión, II sin exclusión y III espectral.

5.2.8 Síntesis integradora del objetivo 2

Los resultados mostraron que el desempeño varía según la composición del conjunto de covariables y la exclusión o inclusión de MO y N, evidenciando diferencias entre estructuras lineales, no lineales, probabilísticas, aditivas, no paramétricas y neuronales. Entre los modelos

destacados se identificaron BayesianRidge y ARD en la familia lineal, RF/GBR en árboles y ensambles, SVR_poly y NuSVR_rbf en modelos kernel, GPR_Dot y GPR_Matern en modelos probabilísticos, EBM en modelos aditivos/GAM, KNN_distance en modelos no paramétricos y MLP residual en redes neuronales.

Por escenario, los mejores desempeños se concentraron en modelos con distinta estructura funcional, lo que confirma que no existió una única familia dominante para todas las condiciones de modelación. En el Escenario I sin exclusión de covariables, los modelos con mayor desempeño se asociaron principalmente con estructuras altamente ajustadas a la señal edáfica disponible, destacándose ARD, GBR, SVR_poly, GPR_Dot y EBM como alternativas competitivas dentro de sus respectivas familias. En el Escenario I con exclusión de MO y N, el desempeño disminuyó de forma general, pero sobresalieron modelos capaces de aprovechar relaciones indirectas entre covariables, especialmente EBM, BayesianRidge/ARD, Extra Trees y variantes kernel. En el Escenario II sin exclusión, los modelos más destacados fueron aquellos que capturaron con alta precisión la relación entre covariables edáficas y COS, particularmente ARD, GBR, NuSVR_rbf y GPR_Dot. En el Escenario II con exclusión de MO y N, la competencia fue más equilibrada y metodológicamente exigente, destacándose SVR_poly, BayesianRidge, GPR_Matern y EBM como los modelos de mejor comportamiento relativo. En el Escenario III, basado exclusivamente en covariables espectrales, los desempeños fueron más limitados, aunque EBM se posicionó como una de las alternativas más consistentes dentro de un escenario de menor capacidad predictiva.

5.3 Resultados - Objetivo 3: Evaluar el resultado de los diversos algoritmos de aprendizaje automático implementados, cuantificando la exactitud y la incertidumbre de las predicciones.

Los resultados del Objetivo 3 permitieron evaluar la robustez predictiva de los modelos campeones identificados en el Objetivo 2, comparando los criterios de incertidumbre, estabilidad residual, las métricas de exactitud y comparación inferencial entre familias algorítmicas. Bajo este enfoque, la evaluación no se limitó a seleccionar el modelo con mayor R^2 espacial, sino que consideró simultáneamente la magnitud del error, la dispersión entre particiones espaciales, la amplitud de las bandas predictivas, la cobertura aproximada del 95 % y la consistencia estadística de las diferencias observadas entre modelos.

5.3.1 Escenario I sin exclusión de covariables

Evaluación multicriterio del desempeño espacial, incertidumbre bootstrap, diagnóstico residual y agrupación inferencial de los modelos campeones por familia en el Escenario I sin exclusión de covariables.

Tabla 19 Evaluación multicriterio de robustez predictiva, cuantificación de la incertidumbre y jerarquización inferencial de modelos campeones por familia algorítmica (Escenario I).

Familia	Modelo	Grupo Friedman	Friedman	Desempeño espacial	Incertidumbre bootstrap	Diagnóstico residual
Lineales	ARD	Superior	$\chi^2 = 28.0286$; $p = 0.0000928$	$R^2 = 0.9992$; RMSE = 0.0332; MAE = 0.0273	$R^2b = 0.999$; RMSEb = 0.037; MAEb = 0.030; banda 95 % = 0.037	Media residual = 0.001; DE residual = 0.032; cobertura 95 % = 0.952
Probabilísticos	GPR_Dot	Superior	$\chi^2 = 28.0286$; $p = 0.0000928$	$R^2 = 0.9979$; RMSE = 0.0551; MAE = 0.0383	$R^2b = 0.991$; RMSEb = 0.104; MAEb = 0.064; banda 95 % = 0.097	Media residual = 0.008; DE residual = 0.055; cobertura 95 % = 0.966
Árboles y ensambles	GBR	Alto	$\chi^2 = 28.0286$; $p = 0.0000928$	$R^2 = 0.9770$; RMSE = 0.0287; MAE = 0.1503	$R^2b = 0.960$; RMSEb = 0.236; MAEb = 0.117; banda 95 % = 0.187	Media residual = 0.012; DE residual = 0.180; cobertura 95 % = 0.952
Kernels	SVR_poly	Intermedio alto	$\chi^2 = 28.0286$; $p = 0.0000928$	$R^2 = 0.9484$; RMSE = 0.2734; MAE = 0.1545	$R^2b = 0.862$; RMSEb = 0.415; MAEb = 0.225; banda 95 % = 0.329	Media residual = -0.003; DE residual = 0.274; cobertura 95 % = 0.959
Aditivos/GAM	EBM	Intermedio alto	$\chi^2 = 28.0286$; $p = 0.0000928$	$R^2 = 0.8890$; RMSE = 0.4008; MAE = 0.2587	$R^2b = 0.848$; RMSEb = 0.467; MAEb = 0.320; banda 95 % = 0.504	Media residual = 0.024; DE residual = 0.401; cobertura 95 % = 0.938
No paramétricos	KNN_distance	Bajo	$\chi^2 = 28.0286$; $p = 0.0000928$	$R^2 = 0.4478$; RMSE = 0.8940; MAE = 0.6638	$R^2b = 0.408$; RMSEb = 0.922; MAEb = 0.688; banda 95 % = 1.003	Media residual = -0.212; DE residual = 0.872; cobertura 95 % = 0.917

Redes neuronales	SELU selu_net	/	Muy bajo	$\chi^2 = 28.0286$; p = 0.0000928	$R^2 = -3.4474$;	$R^2b = -14.238$;	Media residual =
					RMSE = 2.5372; MAE = 2.0556	RMSEb = 4.635; MAEb = 3.669; banda 95 % = 5.685	DE residual = 2.127; cobertura 95 % = 0.945

En el Escenario I sin exclusión, el contraste de Friedman fue significativo — $\chi^2 = 28.0286$; p = 0.0000928—, evidenciando diferencias globales entre modelos campeones. ARD y GPR_Dot conformaron el grupo superior, con R^2 espacial ≥ 0.9979 , $RMSE \leq 0.0551$ y bandas de incertidumbre reducidas. GBR mantuvo alto desempeño, pero con MAE mayor. SVR_poly y EBM se ubicaron en un nivel intermedio alto, mientras que KNN_distance y SELU/selu_net mostraron menor generalización espacial.

Escenario I con exclusión de covariables MO y N

Evaluación multicriterio del desempeño espacial, incertidumbre bootstrap, diagnóstico residual y agrupación inferencial de los modelos campeones por familia en el Escenario I con exclusión de covariables

Tabla 20 Evaluación multicriterio de robustez predictiva, cuantificación de la incertidumbre y jerarquización inferencial de modelos campeones por familia algorítmica (Escenario I con exclusión de covariables).

Familia	Modelo	Grupo Friedman	Friedman	Desempeño espacial	Incertidumbre bootstrap	Diagnóstico residual
Lineales	ARD	Superior	$\chi^2 = 14.7429$; p = 0.0224	$R^2 = 0.6057$;	$R^2b = -0.363$;	Media residual =
				RMSE = 0.7555; MAE = 0.5211	RMSEb = 1.170; MAEb = 0.690; banda 95 % = 1.177	DE residual = 0.758; cobertura 95 % = 0.966
Aditivos/GAM	EBM	Superior	$\chi^2 = 14.7429$; p = 0.0224	$R^2 = 0.6076$;	$R^2b = 0.564$;	Media residual =
				RMSE = 0.7536; MAE = 0.5701	RMSEb = 0.789; MAEb = 0.598; banda 95 % = 0.944	DE residual = 0.756; cobertura 95 % = 0.959
Árboles y ensambles	Extra Trees	Intermedio alto	$\chi^2 = 14.7429$; p = 0.0224	$R^2 = 0.5458$;	$R^2b = 0.472$;	Media residual =
				RMSE = 0.8108; MAE = 0.6159	RMSEb = 0.870; MAEb = 0.660; banda 95 % = 0.870	DE residual = 0.812; cobertura 95 % = 0.924
Probabilísticos	GPR_Matern	Intermedio alto	$\chi^2 = 14.7429$; p = 0.0224	$R^2 = 0.4276$;	$R^2b = 0.376$;	Media residual =
				RMSE =	RMSEb = 0.946;	DE residual = -0.063;

Kernels	KR_poly	Intermedio	$\chi^2 = 14.7429$; p = 0.0224	0.9103; MAE =	MAEb = 0.687;	residual = 0.911;
				0.6569	banda 95 % =	cobertura 95 % =
					0.983	0.938
No paramétricos	KNN_distance	Bajo	$\chi^2 = 14.7429$; p = 0.0224	$R^2 = 0.5250$;	$R^2b = -2.392 \times 10^{15}$;	Media residual =
				RMSE =	5.011×10^6 ;	-0.066; DE
				0.8292; MAE =	MAEb = 7.719×10^5 ;	residual = 0.829;
Redes neuronales	MLP shallow	Muy bajo	$\chi^2 = 14.7429$; p = 0.0224	0.6338	banda 95 % =	cobertura 95 % =
					0.900	0.945
				$R^2 = 0.2394$;	$R^2b = 0.186$;	Media residual =
				RMSE =	RMSEb = 1.081;	-0.242; DE
				1.0493; MAE =	MAEb = 0.825;	residual = 1.025;
				0.7915	banda 95 % =	cobertura 95 % =
					1.178	0.924
				$R^2 = -0.7486$;	$R^2b = -2.331$;	Media residual =
				RMSE =	RMSEb = 2.144;	0.926; DE
				1.5909; MAE =	MAEb = 1.642;	residual = 1.299;
				1.3180	banda 95 % =	cobertura 95 % =
					2.506	0.952

En el Escenario I con exclusión de MO y N, la prueba de Friedman fue significativa — $\chi^2 = 14.7429$; p = 0.0224—, indicando diferencias globales entre los modelos campeones. ARD y EBM quedaron en el grupo superior, aunque la lectura multicriterio favorece a EBM por presentar R^2 espacial similar a ARD, menor inestabilidad bootstrap y menor amplitud de banda predictiva. Extra Trees y GPR_Matern ocuparon un nivel intermedio alto, KR_poly mostró inestabilidad severa en bootstrap, y KNN_distance junto con MLP shallow evidenciaron menor capacidad de generalización espacial.

5.3.2 Modelo Stacking - Escenario I, con exclusión de covariables MO y N

En el escenario I con exclusión de MO y N, el *stacking* se incorporó como estrategia complementaria de segundo nivel para explorar opciones de generalización bajo una condición de información edáfica restringida. El meta-modelo ExtraTrees no superó al mejor modelo base del escenario: aunque registró errores medios competitivos, su R^2 espacial medio fue inferior al de EBM (0.3597 vs. 0.4545), por lo que no constituyó una mejora predictiva efectiva. El detalle completo del procedimiento y de las métricas asociadas se presenta en el *Anexo 50*.

Tabla 21 Desempeño del modelo de stacking en el escenario I con exclusión de covariables MO y N

Modelo	Tipo	R ² espacial (media)	RMSE espacial (media)	MAE espacial (media)	Interpretación
ExtraTrees	Stacking	0.3597	0.7194	0.5722	Desempeño intermedio, sin superar al mejor modelo base

5.3.3 Escenario II sin exclusión de covariables

Evaluación multicriterio del desempeño espacial, incertidumbre bootstrap, diagnóstico residual y agrupación inferencial de los modelos campeones por familia en el Escenario II sin exclusión de covariables

Tabla 22 Evaluación multicriterio de robustez predictiva, cuantificación de la incertidumbre y jerarquización inferencial de modelos campeones por familia algorítmica (Escenario II sin exclusión).

Familia	Modelo	Grupo Friedman	Friedman	Desempeño espacial	Incertidumbre bootstrap	Diagnóstico residual
Lineales	ARD	Superior	$\chi^2 = 23.4000$; $p = 0.0002830$	$R^2 = 0.9993$; RMSE = 0.0322; MAE = 0.0264	$R^{2b} = 0.999$; RMSEb = 0.037; MAEb = 0.030; banda 95 % = 0.037	Media residual = 0.000; DE residual = 0.032; cobertura 95 % = 0.952
Probabilísticos	GPR_Dot	Superior	$\chi^2 = 23.4000$; $p = 0.0002830$	$R^2 = 0.9992$; RMSE = 0.0345; MAE = 0.0283	$R^{2b} = 0.998$; RMSEb = 0.046; MAEb = 0.035; banda 95 % = 0.052	Media residual = 0.000; DE residual = 0.035; cobertura 95 % = 0.952
Árboles y ensambles	GBR	Intermedio alto	$\chi^2 = 23.4000$; $p = 0.0002830$	$R^2 = 0.9797$; RMSE = 0.1714; MAE = 0.0781	$R^{2b} = 0.967$; RMSEb = 0.215; MAEb = 0.105; banda 95 % = 0.170	Media residual = 0.013; DE residual = 0.171; cobertura 95 % = 0.952
Kernels	NuSVR_rbf	Intermedio alto	$\chi^2 = 23.4000$; $p = 0.0002830$	$R^2 = 0.9728$; RMSE = 0.1984; MAE = 0.0693	$R^{2b} = 0.955$; RMSEb = 0.238; MAEb = 0.098; banda 95 % = 0.139	Media residual = 0.000; DE residual = 0.199; cobertura 95 % = 0.986
Aditivos/GAM	EBM	Intermedio	$\chi^2 = 23.4000$; $p = 0.0002830$	$R^2 = 0.9308$; RMSE =	$R^{2b} = 0.906$; RMSEb = 0.362; MAEb =	Media residual = 0.001; DE residual = 0.317; cobertura 95 % = 0.959

No paramétricos	KNN_uniform	Bajo	χ^2 = 23.4000; p = 0.0002830	0.3164; MAE = 0.1745	0.225; banda 95 % = 0.391	Media residual = -0.212; DE residual = 0.872; cobertura 95 % = 0.917
				R^2 = 0.4478; RMSE = 0.8940; MAE = 0.6638	R^2b = 0.408; RMSEb = 0.922; MAEb = 0.688; banda 95 % = 1.003	
Redes neuronales	MLP residual	Muy bajo	χ^2 = 23.4000; p = 0.0002830	R^2 = -2.3090; RMSE = 2.1885; MAE = 1.6816	R^2b = -1.188; RMSEb = 1.725; MAEb = 1.276; banda 95 % = 4.189	Media residual = 1.124; DE residual = 1.884; cobertura 95 % = 0.931

En el Escenario II sin exclusión, la prueba de Friedman fue significativa — $\chi^2 = 23.4000$; p = 0.0002830—, indicando diferencias globales entre modelos campeones por familia. ARD y GPR_Dot conformaron el grupo superior, con R^2 espacial ≥ 0.9992 , $RMSE \leq 0.0345$, $MAE \leq 0.0283$ y bandas predictivas muy reducidas. GBR y NuSVR_rbf integraron un bloque intermedio alto, con R^2 superiores a 0.97 y errores bajos, mientras que EBM conservó un desempeño sólido pero inferior.

5.3.4 Escenario II con exclusión de covariables MO y N

Evaluación multicriterio del desempeño espacial, incertidumbre bootstrap, diagnóstico residual y agrupación inferencial de los modelos campeones por familia en el Escenario II con exclusión de covariables.

Tabla 23 Evaluación multicriterio del desempeño espacial, incertidumbre bootstrap, diagnóstico residual y agrupación inferencial de los modelos campeones por familia algorítmica (Escenario II con exclusión)

Familia	Modelo	Grupo Friedman	Friedman	Desempeño espacial	Incertidumbre bootstrap	Diagnóstico residual
Kernels	SVR_poly	Superior	χ^2 = 22.6286; p = 0.0009309	R^2 = 0.7298; RMSE = 0.6253; MAE = 0.4449	R^2b = 0.555; RMSEb = 0.762; MAEb = 0.543; banda 95 % = 0.906	Media residual = 0.014; DE residual = 0.627; cobertura 95 % = 0.966
				R^2 = 0.6850; RMSE = 0.6752; MAE = 0.4724	R^2b = 0.620; RMSEb = 0.729; MAEb = 0.532; banda 95 % = 0.655	Media residual = 0.016; DE residual = 0.675; cobertura 95 % = 0.959
Lineales	BayesianRidge	Superior	χ^2 = 22.6286; p = 0.0009309			

Probabilísticos	GPR_Matern	Intermedio alto	χ^2 = 22.6286; p = 0.0009309	R ² = 0.6596; R ² b = 0.572; Media residual = 0.004; DE residual = 0.7020; MAE = 0.5060	RMSEb = 0.778; MAEb = 0.577; banda 95 % = 0.855	95 % = 0.938
Aditivos/GAM	EBM	Intermedio alto	χ^2 = 22.6286; p = 0.0009309	R ² = 0.6497; R ² b = 0.602; Media residual = 0.015; DE residual = 0.7121; MAE = 0.4950	RMSEb = 0.751; MAEb = 0.549; banda 95 % = 0.936	95 % = 0.959
Árboles y ensambles	RF	Intermedio	χ^2 = 22.6286; p = 0.0009309	R ² = 0.5807; R ² b = 0.491; Media residual = 0.040; DE residual = 0.7790; MAE = 0.5936	RMSEb = 0.852; MAEb = 0.642; banda 95 % = 0.849	95 % = 0.917
No paramétricos	KNN_distance	Bajo	χ^2 = 22.6286; p = 0.0009309	R ² = 0.2394; R ² b = 0.186; Media residual = 0.242; DE residual = 1.0493; MAE = 0.7915	RMSEb = 1.081; MAEb = 0.825; banda 95 % = 1.178	95 % = 0.924
Redes neuronales	MLP residual	Muy bajo	χ^2 = 22.6286; p = 0.0009309	R ² = -0.3152; R ² b = -0.476; Media residual = 0.041; DE residual = 1.3798; MAE = 1.0582	RMSEb = 1.440; MAEb = 1.080; banda 95 % = 2.794	95 % = 0.945

En el Escenario II con exclusión de MO y N, la prueba de Friedman fue significativa — $\chi^2 = 22.6286$; $p = 0.0009309$ —, evidenciando diferencias globales entre los modelos campeones por familia. Bajo este escenario, SVR_poly se posiciona como el mejor modelo, al presentar el mejor desempeño espacial del conjunto evaluado, con el mayor R² medio (0.7298) y los menores errores promedio (RMSE = 0.6253; MAE = 0.4449). Aunque BayesianRidge mostró una banda predictiva más estrecha y mayor parsimonia, su desempeño espacial fue inferior y, además, mantuvo limitaciones asociadas al incumplimiento de supuestos lineales, lo que reduce su conveniencia como modelo final en un escenario caracterizado por relaciones potencialmente no lineales entre covariables edáficas, topográficas y COS.

Modelo Stacking - Escenario II, con exclusión de covariables MO y N

En el escenario II con exclusión de MO y N, el modelo de *stacking* con ElasticNet mostró un desempeño intermedio (R² espacial medio = 0.5241), superior al bloque inferior del escenario, pero insuficiente para superar a los mejores modelos individuales. En particular, permaneció por debajo de SVR_poly, BayesianRidge, GPR_Matern, EBM y RF, lo que indica que la integración de predictores heterogéneos no produjo una mejora efectiva de generalización espacial frente a las alternativas base más sólidas.

Tabla 24 Desempeño del modelo de stacking en el escenario II, con exclusión de covariables MO y N

Modelo	Tipo	R ² espacial medio	DE R ² espacial	RMSE espacial medio	DE RMSE espacial	MAE espacial medio	DE MAE espacial	R ² global	RMS E global	MAE global
ElasticNet	Stacking	0.5241	0.3629	0.5942	0.1693	0.4534	0.1007	0.7129	0.6446	0.4663

5.3.5 Escenario III

Evaluación multicriterio del desempeño espacial, incertidumbre bootstrap, diagnóstico residual y agrupación inferencial de los modelos campeones por familia en el Escenario III.

Tabla 25 Evaluación multicriterio de robustez predictiva, cuantificación de la incertidumbre y jerarquización inferencial de modelos campeones por familia algorítmica (Escenario III).

Familia	Modelo	Grupo Friedman	Friedman	Desempeño espacial	Incertidumbre bootstrap	Diagnóstico residual
Aditivos/GAM	EBM	Superior	$\chi^2 = 14.9143$; p = 0.020934	R ² = 0.1995; RMSE = 1.0764; MAE = 0.8593	R ² b = 0.177; RMSEb = 1.089; MAEb = 0.865; banda 95 % = 1.392	Media residual = -0.051; DE residual = 1.079; cobertura 95 % = 0.945
Árboles y ensambles	CatBoost	Intermedio	$\chi^2 = 14.9143$; p = 0.020934	R ² = 0.0550; RMSE = 1.1695; MAE = 0.9343	R ² b = 0.062; RMSEb = 1.161; MAEb = 0.915; banda 95 % = 1.241	Media residual = -0.115; DE residual = 1.168; cobertura 95 % = 0.945
Kernels	NuSVR_rbf	Intermedio alto	$\chi^2 = 14.9143$; p = 0.020934	R ² = 0.0195; RMSE = 1.1913; MAE = 0.9527	R ² b = 0.024; RMSEb = 1.184; MAEb = 0.950; banda 95 % = 1.103	Media residual = -0.016; DE residual = 1.195; cobertura 95 % = 0.938
No paramétricos	KNN_distance	Intermedio	$\chi^2 = 14.9143$; p = 0.020934	R ² = -0.0045; RMSE = 1.2058; MAE = 0.9503	R ² b = -0.004; RMSEb = 1.201; MAEb = 0.949; banda 95 % = 1.291	Media residual = -0.141; DE residual = 1.202; cobertura 95 % = 0.938
Probabilísticos	GPR_RBF	Intermedio alto	$\chi^2 = 14.9143$; p = 0.020934	R ² = -0.0473; RMSE = 1.2312; MAE = 0.9605	R ² b = -0.064; RMSEb = 1.238; MAEb = 0.969;	Media residual = -0.119; DE residual = 1.230; cobertura 95 % = 0.931

Lineales	BayesianRidge	Intermedio bajo	χ^2 = 14.9143; p = 0.020934	R ² = -0.0884; RMSE = 1.2552; MAE = 0.9572	banda 95 % = 1.341 R ² b = -0.052; RMSEb = 1.231; MAEb = 0.945; banda 95 % = 0.656	Media residual = - 0.058; DE residual = 1.230; cobertura 95 % = 0.945
Redes neuronales	MLP deep	Muy bajo	χ^2 = 14.9143; p = 0.020934	R ² = -22.9536; RMSE = 5.8883; MAE = 3.4711	banda 95 % = 4.500 R ² b = -17.857; RMSEb = 5.066; MAEb = 3.403; banda 95 % = 4.500	Media residual = - 0.372; DE residual = 5.897; cobertura 95 % = 0.959

En el Escenario III, la prueba de Friedman fue significativa — $\chi^2 = 14.9143$; p = 0.020934—, lo que indica diferencias globales entre los modelos campeones evaluados con covariables espectrales. EBM ocupó el grupo superior y alcanzó el mayor R² espacial del escenario; sin embargo, su valor de R² = 0.1995 y RMSE = 1.0764 evidencian una capacidad predictiva absoluta limitada. CatBoost, NuSVR_rbf, KNN_distance, GPR_RBF y BayesianRidge mostraron desempeños bajos o marginales, mientras que MLP deep presentó el comportamiento menos favorable.

Modelo Stacking - Escenario III

En el escenario III, el modelo de *stacking* con SVR_rbf como meta-modelo mostró un desempeño globalmente limitado, con R² espacial medio = -0.3016, RMSE espacial medio = 1.0375 y MAE espacial medio = 0.8559, además de una variabilidad elevada entre folds (DE R² = 0.6135). Estos resultados indican que la integración de modelos base no logró compensar la debilidad estructural del escenario espectral, caracterizado por una señal predictiva insuficiente y una baja capacidad de generalización espacial. Aunque el ensamble alcanzó un R² global de 0.1423, su desempeño espacial siguió siendo inferior al de los mejores modelos individuales del escenario, particularmente EBM, por lo que no constituyó una mejora efectiva.

Tabla 26 Desempeño del modelo de stacking frente a modelos individuales en el escenario III

Modelo	Tipo	R ² espacial medio	DE R ² espacial	RMSE espacial medio	DE RMSE espacial	MAE espacial medio	DE MAE espacial	R ² global	RMS E global	MAE global
SVR_rbf_meta	Stacking	-0.3016	0.6135	1.0375	0.2988	0.8559	0.2611	0.1423	1.1142	0.8829

5.3.6 Selección de modelos individuales definitivos para la predicción de COS por escenario

Tabla 27 Selección de modelos individuales definitivos para la predicción de COS por escenario.

Escenario / condición	Familia / modelo campeón	Desempeño, incertidumbre y diagnóstico	Contraste inferencial global / grupo	Justificación
Escenario I / Sin exclusión de covariables	Probabilísticos – GPR_Dot	R ² espacial medio = 0.9979; RMSE = 0.0551; MAE = 0.0383; R ² bootstrap = 0.991; RMSE bootstrap = 0.104; MAE bootstrap = 0.064; media residual = 0.008; DE residual = 0.055; cobertura 95 % = 0.966; banda 95 % = 0.097.	Grupo superior. Posición global 2; ranking medio = 2.0. Friedman global: $\chi^2 = 13.5600$; p = 0.003570.	Se mantiene como modelo seleccionado porque combina desempeño espacial casi máximo, incertidumbre reducida y formulación probabilística explícita. Aunque modelos lineales presentaron ajuste similar, su incumplimiento de supuestos limita su valor inferencial; GPR_Dot conserva mayor solidez predictiva y probabilística.
Escenario I / Con exclusión de covariables MO y N	GAM aditivos – EBM	R ² espacial medio = 0.4949; RMSE = 0.7450; MAE = 0.5735; R ² bootstrap = 0.564; RMSE bootstrap = 0.789; MAE bootstrap = 0.598; media residual = -0.050; DE residual = 0.812; cobertura 95 % = 0.924; banda 95 % = 0.870.	Grupo inferior. Posición global 4; ranking medio = 3.6. Friedman global: $\chi^2 = 13.5600$; p = 0.003570.	Se mantiene EBM porque, al excluir MO y N, ofrece la lectura funcional más rica y menos redundante. Su valor no reside únicamente en el R ² , sino en la capacidad de redistribuir la señal predictiva hacia covariables edáficas indirectas, topográficas, espectrales y categóricas bajo una estructura interpretable.
Escenario II / Sin exclusión de covariables	Probabilísticos – GPR_Dot	R ² espacial medio = 0.9992; RMSE = 0.0345; MAE = 0.0283; R ² bootstrap = 0.998; RMSE bootstrap = 0.046; MAE bootstrap = 0.035; media residual $\approx$ 0.000; DE residual = 0.035; cobertura 95 % = 0.952; banda 95 % = 0.052.	Grupo superior. Posición global 1; ranking medio = 1.0. Friedman global: $\chi^2 = 13.5600$; p = 0.003570.	Se mantiene como mejor modelo global del estudio. Su desempeño prácticamente máximo se explica por la alta concentración de señal edáfica directa, principalmente MO y N. Frente a alternativas lineales, GPR_Dot es más defendible por incorporar incertidumbre predictiva explícita.
Escenario II / Con exclusión de covariables	Kernels – SVR_poly	R ² espacial medio = 0.7298; RMSE = 0.6253; MAE = 0.4449; R ² bootstrap = 0.555; RMSE bootstrap = 0.762;	Grupo superior. Friedman del escenario: $\chi^2 =$	Se selecciona SVR_poly porque presentó el mejor desempeño espacial del escenario y pertenece al grupo superior de Friedman. Aunque BayesianRidge mostró mayor R ²

covariables	MAE bootstrap = 0.543; 22.6286; p = bootstrap y menor banda predictiva, su
MO y N	media residual = 0.014; DE 0.0009309. pertenencia a la familia lineal y el
	residual = 0.627; cobertura incumplimiento de supuestos clásicos
	95 % = 0.966; banda 95 % = limitan su uso como modelo final.
	0.906. SVR_poly ofrece una alternativa no lineal
	regularizada, robusta para relaciones
	edáficas complejas sin exigir supuestos
	paramétricos clásicos.

La comparación global mediante Friedman evidenció diferencias estadísticamente significativas entre los modelos definitivos comparados, $\chi^2 = 13.5600$; $p = 0.003570$. Sin embargo, las comparaciones pareadas con ajuste de Holm no identificaron diferencias significativas entre pares específicos, lo que sugiere que la jerarquía debe interpretarse de forma conjunta con las magnitudes de R^2 , RMSE, MAE, incertidumbre bootstrap y diagnóstico residual. Bajo este criterio, los modelos GPR_Dot de los escenarios sin exclusión conformaron el bloque superior, SVR_poly ocupó una posición intermedia y EBM (no incluido en el cuadro) del escenario I con exclusión quedó como alternativa inferior en la comparación conjunta, aunque sigue siendo el modelo más defendible dentro de su escenario específico.

Tabla 28 Comparativa multidimensional de modelos stacking y modelos base: diagnóstico de error espacial, estabilidad de las predicciones y validación inferencial

Escenario	Modelo stacking	Desempeño espacial	Incertidumbre / estabilidad	Diagnóstico residual / global	Comparación frente al mejor modelo base
Escenario I con exclusión de MO y N	ExtraTrees_meta	R ² espacial medio = 0.3597; RMSE espacial medio = 0.7194; MAE espacial medio = 0.5722	DE R ² = 0.3756; RMSE = 0.2776; DE MAE = 0.2064	R ² global = 0.5788; RMSE global = 0.7808; MAE global = 0.5724	No superó al mejor modelo base. BayesianRidge_base presentó R ² espacial medio = 0.4597, RMSE = 0.6803 y MAE = 0.5269
Escenario II con exclusión de MO y N	ElasticNet_meta	R ² espacial medio = 0.5241; RMSE espacial medio = 0.5942; MAE espacial medio = 0.4534	DE R ² = 0.3629; RMSE = 0.1693; DE MAE = 0.1007	R ² global = 0.7129; RMSE global = 0.6446; MAE global = 0.4663	No superó al mejor modelo base en R ² espacial. NuSVR_rbf presentó R ² espacial medio = 0.5710, RMSE = 0.5781 y MAE = 0.4380
Escenario III espectral	SVR_rbf_meta	R ² espacial medio = 0.3016; RMSE espacial medio =	DE R ² = 0.6135; RMSE =	R ² global = 0.1423; RMSE global =	Mejoró parcialmente frente a varios modelos

1.0375; MAE 0.2988; DE 1.1142; MAE base, pero mantuvo R^2
espacial medio = MAE = 0.2611 global = 0.8829 espacial negativo
0.8559

Los modelos *stacking* fueron metodológicamente útiles como contraste de segundo nivel, pero no demostraron superioridad sistemática frente a los mejores modelos base. El caso más favorable fue el Escenario II con exclusión de MO y N, donde ElasticNet_meta alcanzó el mejor equilibrio entre R^2 espacial, error y estabilidad; sin embargo, no superó el desempeño de NuSVR_rbf. En los escenarios I con exclusión y III espectral, los ensambles mostraron limitaciones claras, lo que indica que la complementariedad entre familias fue insuficiente para justificar su selección como modelos definitivos.

5.3.7 Interpretación de los modelos y aporte de las covariables

Evaluación del subconjunto de covariables mediante permutación

Aporte relativo agregado de las familias de covariables según importancia por permutación en los modelos campeones por escenario

Tabla 29 Contribución agregada de grupos de covariables al desempeño predictivo bajo esquemas de permutación y protocolos de exclusión de variables.

Escenario / modelo campeón	Edáficas (%)	Topográficas (%)	Espectrales (%)	Categóricas (%)	Aporte dominante
Escenario I sin exclusión / GPR_Dot	99.493	0.095	0.107	0.305	Edáficas
Escenario I con exclusión MO y N / EBM	74.60	16.63	6.94	1.83	Edáficas + topográficas
Escenario II sin exclusión / GPR_Dot	99.703	0.000	No aplica	0.297	Edáficas
Escenario II con exclusión MO y N / SVR_poly	93.516	0.000	No aplica	6.484	Edáficas

La importancia por permutación evidencia que la estimación del COS estuvo dominada por covariables edáficas en todos los escenarios donde estas estuvieron disponibles. En los escenarios sin exclusión, el aporte edáfico superó el 99 %, lo que confirma que el desempeño casi perfecto de GPR_Dot estuvo condicionado por variables de alta proximidad funcional con el COS, especialmente MO y N. Al excluir estas covariables en el Escenario I, el aporte edáfico disminuyó a 74.60 % y emergió una contribución topográfica relevante de 16.63 %, seguida por las covariables espectrales con 6.94 % y categóricas con 1.83 %.

Evaluación del subconjunto de covariables mediante SHAP

Evaluación funcional del subconjunto de covariables mediante SHAP en los modelos campeones por escenario.

Tabla 30 Contribución agregada de grupos de covariables al desempeño predictivo bajo esquemas de permutación y protocolos de exclusión de variables.

Escenario / modelo campeón	Estructura funcional identificada mediante SHAP	Covariables con mayor soporte SHAP
Escenario I sin exclusión / GPR_Dot	Dependencia funcional altamente concentrada en covariables edáficas directas.	MO como covariable dominante; aportes secundarios de N, variables de cobertura, algunos índices espectrales y atributos topográficos.
Escenario I con exclusión de MO y N / EBM	Redistribución funcional de la señal hacia covariables edáficas indirectas, topográficas, espectrales y categóricas.	Al, A, K, pH, CIC, P, SAVI_pt y variables categóricas de textura y cobertura.
Escenario II sin exclusión / GPR_Dot	Estructura funcional edáfica casi exclusiva.	MO como eje explicativo principal; aportes marginales de variables físico-químicas y categóricas.
Escenario II con exclusión de MO y N / SVR_poly	Estructura funcional edáfico-química depurada, dominada por variables asociadas con acidez, intercambio catiónico, textura y bases intercambiables.	Al, CIC, K, pH, A, Mn, Cu, cobertura y unidad cartográfica.
Escenario III espectral / EBM	Dependencia restringida al bloque espectral.	Índices espectrales RGB y multiespectrales disponibles en el escenario.

El análisis SHAP permitió examinar la estructura funcional interna de los modelos campeones y diferenciar entre escenarios de alta precisión por disponibilidad de covariables orgánicas directas y escenarios de mayor exigencia metodológica por reconstrucción indirecta de la señal del COS. En los escenarios sin exclusión, SHAP confirmó una dependencia dominante de MO, lo que explica los desempeños casi unitarios pero limita la interpretación independiente de otras covariables. Los escenarios con exclusión de MO y N mostraron una redistribución de las contribuciones hacia variables edáficas indirectas y ambientales, especialmente Al, CIC, K, pH, textura, Mn, cobertura e índices espectrales, lo que resulta más informativo desde el punto de vista edafológico. En el escenario espectral, SHAP permitió caracterizar la contribución relativa de los índices disponibles, aunque la baja exactitud del modelo impide atribuirles suficiencia predictiva.

Evaluación del subconjunto de covariables mediante LIME

Evaluación local del aporte de covariables mediante LIME en los modelos campeones por escenario.

Tabla 31 Diagnóstico de la arquitectura predictiva local: atribución de importancia y sentido de influencia de las covariables edafoclimáticas y sensores remotos mediante LIME.

Escenario / modelo campeón	Naturaleza de la explicación LIME	Covariables o condiciones locales con mayor contribución	Sentido local del efecto
Escenario I sin exclusión / GPR_Dot	Explicación local dominada por una covariable orgánica directa.	MO en el intervalo $0.14 < MO \leq 0.71$; aportes menores de NGRDI_pt, K, pH, UCS_F_SEef1, GNDVI_pt, S, Mn y ExG_pt.	MO aportó el mayor efecto positivo. La ausencia de cobertura agrícola favoreció la predicción; algunas ausencias categóricas y ciertos índices espectrales o nutrientes generaron efectos negativos moderados.
Escenario I con exclusión de MO y N / EBM	Explicación local redistribuida hacia covariables edáficas indirectas, de cobertura y relieve.	$Al \leq -0.59$, $CIC \leq -0.50$, $Ca > 0.19$, $S > 0.72$, $L > 0.59$, $A \leq -0.61$, $Tipo_Cober \leq 1.00$, $SAVI_pt \leq -0.15$, $Clasif \leq 2.00$, $Rough_w5_pt \leq -0.69$, $Mn > 0.57$, $Fe \leq -0.66$ y un valor intermedio de Ar.	Predominaron contribuciones negativas acumuladas que redujeron la predicción local del COS. El efecto local se explicó por la combinación de acidez, capacidad de intercambio, bases, textura, cobertura, señal espectral y rugosidad.
Escenario II sin exclusión / GPR_Dot	Explicación local altamente concentrada en MO dentro del conjunto edáfico.	MO en el intervalo $0.00 < MO \leq 0.48$; presencia de cobertura silvopastoril; pH relativamente bajo; presencia de Clasif_FrancoArenoso.	MO presentó el mayor efecto positivo. Las variables categóricas de textura y cobertura, así como pH, tuvieron efectos positivos secundarios.
Escenario II con exclusión de MO y N / SVR_poly	Explicación local dominada por un núcleo edáfico-químico indirecto, con predominio de efectos negativos sobre la predicción local del COS.	$Al \leq -0.55$, $CIC \leq -0.69$, $A \leq -0.58$, $Ca > 0.66$, $L > 0.66$, $Cu \leq -0.78$, $Mn > 0.69$, $Clasif_Franco > 0.00$, K en el intervalo $-0.09 < K \leq 0.58$, presencia de Tipo_Cober_Bosque, Ar intermedio y Mg bajo-intermedio.	La predicción local fue reducida principalmente por Al bajo, CIC baja, A baja, Ca alto, L alto, Cu bajo, Mn alto y presencia de Clasif_Franco. Los únicos efectos positivos relevantes correspondieron a K intermedio, ausencia de pasturas y Mg bajo-intermedio.

La evaluación mediante LIM E complementa la lectura global de SHAP y permutación al mostrar cómo se construye una predicción particular en cada modelo campeón. En los escenarios sin exclusión, las explicaciones locales confirman una dependencia dominante de MO, lo que indica que la alta exactitud de GPR_Dot está fuertemente condicionada por la presencia de una covariable orgánica directamente asociada al COS. En los escenarios con exclusión de MO y N,

LIME muestra una arquitectura local más distribuida, donde la predicción se explica por combinaciones de Al, CIC, textura, K, pH, micronutrientes, cobertura, clasificación y señal espectral.

Evaluación del subconjunto de covariables mediante RFECV

La evaluación mediante RFECV permitió diferenciar entre un óptimo estrictamente predictivo y un óptimo metodológico-interpretativo del subconjunto de covariables. En los escenarios sin exclusión, la selección recursiva confirmó la dominancia de MO y, en menor medida, de N, lo cual explica el alto desempeño de los modelos, pero también evidencia una dependencia de covariables directamente asociadas con la variable respuesta.

Tabla 32 Evaluación de la redundancia y estabilidad estructural del espacio de características mediante RFECV para los modelos campeones seleccionados.

Escenario / condición	Modelo campeón asociado	Covariables con selección muy alta (≥ 0.80)	Covariables con selección intermedia (0.60–0.79)
Escenario I / Sin exclusión de covariables	GPR_Dot	MO, N, GCC_pt, Tipo_Cober_Agricola	—
Escenario I / Con exclusión de MO y N	EBM	Al, A, K, pH	CIC, Clasif_ArenoFrancoso, Clasif_FrancoArcillosoArenoso, P, SAVI_pt
Escenario II / Sin exclusión de covariables	GPR_Dot	MO, Tipo_Cober_Pasturas, Clasif_ArenoFrancoso, Clasif_FrancoArcillosoArenoso, Fe	Clasif_FrancoArenoso, P, Tipo_Cober_Agricola, UCS_F_SIde
Escenario II / Con exclusión de MO y N	SVR_poly	Al, CIC, A, K, Tipo_Cober_Bosque	UCS_F_SEefl
Escenario III / Espectral	EBM	Tipo_Cober_Silvopastoril, Clasif_ArenoFrancoso, Clasif_Franco, Clasif_FrancoArcillosoArenoso, Clasif_FrancoArenoso, Tipo_Cober_Pasturas	UCS_F_SEefl

Los escenarios con exclusión de MO y N revelaron subconjuntos más parsimoniosos y edafológicamente informativos, especialmente el Escenario II con exclusión, donde A, Al, CIC, K y Mn conformaron el núcleo predictor más robusto y menos redundante. El Escenario III mostró que, sin información edáfica directa, la selección se desplazó hacia variables categóricas

territoriales, confirmando que la señal espectral aislada no fue suficiente para sostener una predicción estructuralmente robusta del COS.

5.3.8 Reproducibilidad del modelo campeón

Tabla 33 Especificaciones para la reproducibilidad del modelo campeón

Componente	Especificación para reproducibilidad
Entorno computacional	Python en Google Colab
Librerías principales	pandas, numpy, scikit-learn, scipy, statsmodels, geopandas, rasterio, shap, tensorflow/keras, interpret, xgboost, lightgbm, catboost, ngboost y pygam.
Base analítica	145 muestras georreferenciadas, integrando covariables edáficas, texturales, categóricas, topográficas y espectrales.
Escenarios de modelación	Escenario I: conjunto completo; Escenario II: conjunto edáfico/topográfico; Escenario III: conjunto espectral; con variantes con y sin MO y N cuando aplica.
Criterio de exclusión de covariables	MO y N se excluyeron en escenarios específicos por alta redundancia estructural y proximidad funcional con el COS.
Preprocesamiento del flujo analítico	Arquitectura estandarizada mediante Pipeline y ColumnTransformer, incorporando transformación de variables, escalamiento, codificación categórica y PCA opcional.
Transformaciones y escaladores evaluados	Transformaciones: original, raíz cúbica, Yeo–Johnson y normalización por cuantiles. Escaladores: sin escalado, StandardScaler, MinMaxScaler y RobustScaler.
Validación espacial	Validación cruzada espacial mediante GroupKFold, con grupos definidos a partir de coordenadas geográficas mediante agrupamiento o malla espacial.
Búsqueda y ajuste de modelos	Screening inicial de configuraciones candidatas y ajuste fino de hiperparámetros sobre los mejores modelos por familia.
Métricas de evaluación	R ² , RMSE y MAE estimados bajo validación espacial.
Incertidumbre y diagnóstico residual	Bootstrap con 300 réplicas para modelos determinísticos, incertidumbre endógena en GPR, Monte Carlo Dropout en redes neuronales, media residual, desviación estándar, cobertura aproximada del 95 % y revisión de supuestos cuando aplica.
Contraste estadístico	Prueba de Friedman global y comparaciones pareadas mediante Wilcoxon con ajuste de Holm.
Modelo final seleccionado	SVR_poly en escenario II con exclusión de MO y N.
Configuración final del modelo	Yeo–Johnson + StandardScaler; kernel polinomial grado 2; C=1, coef0=1.0, epsilon=0.1, gamma='scale'.
Productos y soporte de réplica	SVR_poly: https://colab.research.google.com/drive/18sx3Lj9J4zd7xAfdpudur1z4sKGGOCic?usp=drive_link Estimación espacial SVR_poly: https://drive.google.com/file/d/1EVh_4QaPa4F9fgMb3f1akz9vDZqZwLcd/view?usp=drive_link Permutación, SHAP y LIME: https://drive.google.com/file/d/1zN0yByfid9cKN5haJ22jLJUxVL_5kNEO/view?usp=sharing

REFECV:

https://drive.google.com/file/d/1GyfYbnDvR6AqOM8vpPOs3pG5dLzSXXr4/view?usp=drive_link

5.3.9 Síntesis integradora del objetivo 3

Los resultados mostraron que el desempeño predictivo del COS estuvo fuertemente condicionado por la composición del conjunto de covariables, con mejores métricas en los escenarios que incluyeron MO y N, y una reducción esperada del ajuste en los escenarios con exclusión de estas variables, los cuales representaron una condición más exigente para evaluar la capacidad predictiva indirecta. Dentro de los modelos más relevantes se destacaron ARD y BayesianRidge en la familia lineal, GBR/RF en árboles y ensambles, GPR_Dot/GPR_Matern en probabilísticos, EBM en aditivos/GAM, KNN_distance en no paramétricos, MLP residual en redes neuronales y, especialmente, SVR_poly dentro de los modelos kernel. La comparación entre métricas espaciales, bandas de incertidumbre y diagnóstico residual permitió evidenciar que no todos los modelos con buen ajuste puntual mantenían la misma estabilidad bajo validación espacial. En el Escenario II con exclusión de MO y N, el modelo SVR_poly se consolidó como la alternativa más relevante desde el criterio predictivo y explicativo, al alcanzar el mayor desempeño espacial entre los modelos campeones evaluados, con $R^2 = 0.7298$, $RMSE = 0.6253$ y $MAE = 0.4449$. Aunque modelos como BayesianRidge, GPR_Matern y EBM mostraron comportamientos competitivos y ventajas parciales en incertidumbre, parsimonia o estabilidad, SVR_poly ofreció la mejor capacidad de generalización espacial bajo una condición de información restringida. Su selección se justifica porque, al excluir las covariables de mayor proximidad funcional con el COS, el modelo kernel polinomial logró representar de manera más eficiente relaciones no lineales entre las covariables remanentes y la variable respuesta, manteniendo además un diagnóstico residual aceptable y una incertidumbre bootstrap compatible con un desempeño robusto para la estimación predictiva del carbono orgánico del suelo.

5.4 Resultados - Objetivo 4: Generar cartografía digital de alta resolución que represente la distribución espacial del COS y la incertidumbre asociada a la predicción en el área de estudio.

El Objetivo 3 concluyó con la selección del modelo SVR con kernel polinomial de grado 2 (SVR_poly) como modelo campeón definitivo del Escenario II con exclusión de MO y N. Este modelo alcanzó el mayor R^2 espacial medio del escenario (0.7298), con $RMSE = 0.6253$ y $MAE =$

0.4449, ubicándose en el grupo superior de la prueba de Friedman ($\chi^2 = 22.6286$; $p = 0.0009$). Su selección se fundamentó en la capacidad de representar relaciones no lineales entre covariables edáficas remanentes y el COS, sin requerir el cumplimiento de supuestos paramétricos clásicos que limitaron alternativas como BayesianRidge. La evaluación mediante SHAP, LIME y RFECV identificó a Al, CIC, K, pH, A, Mn y Cu como el núcleo predictor del modelo, conformando una estructura funcional edáfico-química coherente con los procesos de retención de carbono en suelos andinos.

El presente objetivo materializa la transición del modelo validado puntualmente hacia su aplicación espacial exhaustiva. Para ello, se construyó un cubo de datos de 29 bandas que integró las covariables edáficas fisicoquímicas (excluyendo MO y N), regionalizadas mediante interpolación IDW ($p = 2$), y los derivados topográficos del DEM, remuestreados por interpolación bilineal, sobre la grilla maestra de 3 m de resolución (176×267 píxeles, EPSG:3116). El modelo SVR_poly fue re-entrenado localmente sobre las 145 muestras, extrayendo los valores de las covariables directamente del cubo en las coordenadas georreferenciadas de muestreo, garantizando la alineación perfecta entre los features de entrenamiento y las bandas de predicción. La incertidumbre predictiva se estimó mediante bootstrap con $N = 100$ réplicas.

5.4.1 Mapa de predicción espacial del COS

La superficie de predicción del COS (Figura 1) presentó un rango de valores entre 4.70 % y 10.80 %, con una media de 8.35 % y una desviación estándar de 0.73 %. La mediana (8.50 %) resultó ligeramente superior a la media, indicando una asimetría negativa moderada consistente con la distribución observada en las muestras de laboratorio. A diferencia de los escenarios exclusivamente espectrales, donde el rango predicho se comprimió hacia la media (6.25–9.62 % en el Escenario III), el modelo SVR_poly preservó la variabilidad original del COS (4.60–11.00 % observado en laboratorio), confirmando que las covariables edáficas interpoladas retuvieron suficiente señal predictiva para capturar los extremos de la distribución.

Espacialmente, se identificaron contenidos elevados de COS (> 9.0 %) en el sector nororiental y en parches del centro-sur de la granja, zonas asociadas a coberturas de bosque y arreglos silvopastoriles con mayor aporte de hojarasca al horizonte superficial. Los valores más bajos (< 6.0 %) se localizaron en el extremo suroccidental y en áreas puntuales del borde occidental, coincidentes con pasturas abiertas de menor retención de carbono. La transición gradual entre zonas

refleja la influencia de los gradientes edáficos interpolados, particularmente Al, CIC y K, que el análisis SHAP del Objetivo 3 identificó como los predictores de mayor contribución funcional.

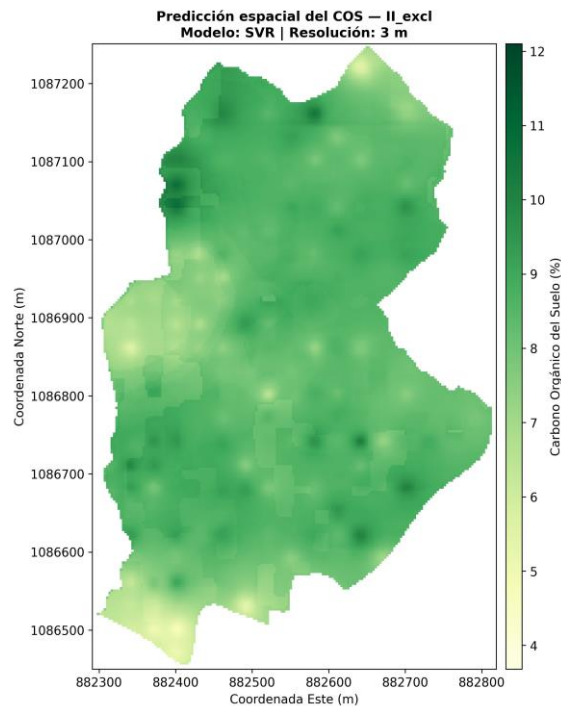


Figura 1 Mapa de predicción espacial del COS (%) —Escenario II_excl, modelo SVR_poly, resolución 3 m. Coordenadas MAGNA-SIRGAS (EPSG:3116).

La distribución de las predicciones sobre los 28 981 píxeles válidos Figura 2 exhibió una concentración modal en el intervalo 8.3–8.9 %, con colas extendidas hacia ambos extremos. La asimetría negativa leve ($\text{media} = 8.35 < \text{mediana} = 8.50$) sugiere que una fracción mayor del territorio presentó contenidos de COS por encima del promedio, coherente con la predominancia de coberturas de bosque y silvopastoril en el área de estudio.

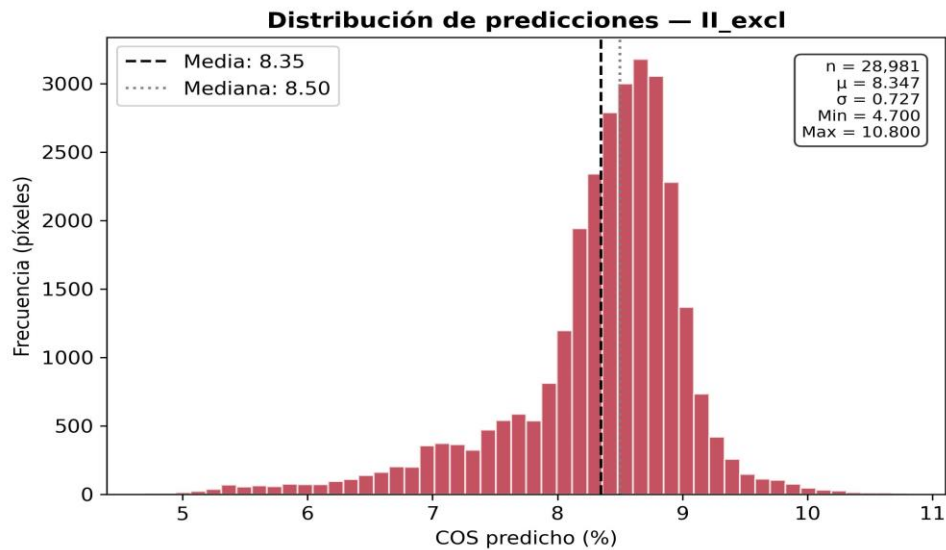


Figura 2 Distribución de las predicciones del COS sobre los 28 981 píxeles válidos — Escenario II_excl. $\mu = 8.35 \%$, $\sigma = 0.73 \%$, rango [4.70, 10.80].

5.4.2 Coherencia espacial de la superficie predictiva

La superposición de las predicciones del mapa sobre las 145 coordenadas de muestreo arrojó un $R^2 = 0.8707$, $RMSE = 0.4326$, $MAE = 0.2212$ y sesgo = -0.0077 (Figura 3). Estas métricas corresponden a una evaluación de coherencia interna, dado que el modelo fue re-entrenado sobre la totalidad de las muestras; por lo tanto, no constituyen una validación independiente, función que fue cubierta por la validación cruzada espacial GroupKFold del Objetivo 3 ($R^2 = 0.7298$, $RMSE = 0.6253$). La diferencia entre ambos valores refleja la reducción de varianza esperada al evaluar sobre datos de entrenamiento respecto a datos de validación, y su magnitud (R^2 de 0.73 a 0.87) se sitúa dentro de márgenes razonables que descartan sobreajuste severo.

El gráfico de observado versus predicho mostró una concentración de puntos a lo largo de la línea 1:1 en todo el rango, con dispersión moderada en el segmento inferior ($COS < 6 \%$), donde la menor densidad de observaciones reduce la capacidad del modelo para capturar la variabilidad local. Se identificaron dos observaciones en el extremo superior ($COS \text{ observado} > 10 \%$) con predicciones ligeramente inferiores, posiblemente asociadas a microhábitats con acumulación orgánica puntual no representada por las covariables a la resolución de 3 m.

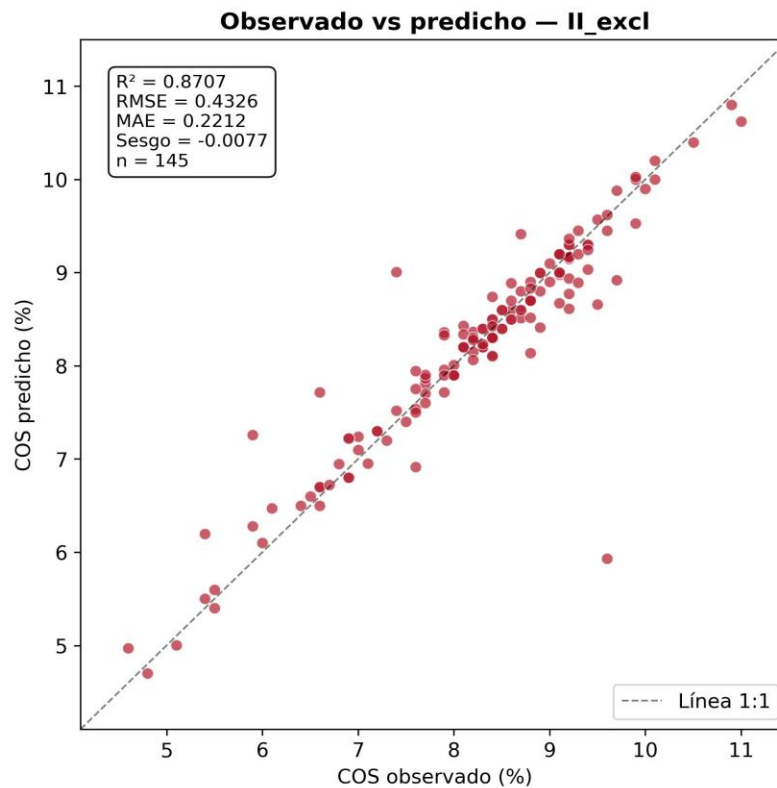


Figura 3 Diagrama de dispersión observado vs. predicho en las 145 coordenadas de muestreo — Escenario II_excl.
 $R^2 = 0.8707$, $RMSE = 0.4326$, $MAE = 0.2212$, sesgo = -0.0077 .

5.4.3 Incertidumbre predictiva

El mapa de desviación estándar predictiva (Figura 4) estimada por bootstrap ($N = 100$) mostró valores de σ predominantemente inferiores a 0.3 en la mayor parte del área de estudio, con focos puntuales de mayor incertidumbre ($\sigma > 0.6$) distribuidos de forma dispersa. Estos focos coincidieron con la ubicación de puntos de muestreo donde el bootstrap detectó mayor sensibilidad a la composición del set de entrenamiento. La incertidumbre media fue baja y espacialmente homogénea, lo que indica estabilidad del modelo SVR_poly frente a perturbaciones del conjunto de datos. Este resultado es consistente con la banda predictiva del 95 % reportada en el Objetivo 3 (0.906), que ubicó al modelo dentro de un rango de incertidumbre aceptable para la predicción espacial del COS a escala predial.

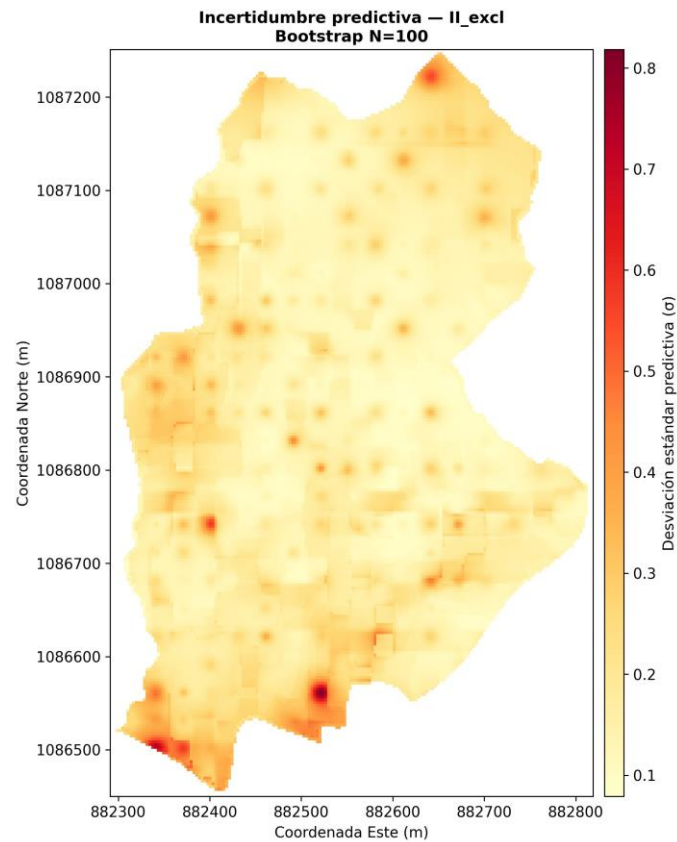


Figura 4 Mapa de incertidumbre predictiva (σ) estimada por bootstrap ($N = 100$) — Escenario II_excl.

El mapa de ancho del intervalo de confianza al 95 % (Figura 5) presentó valores entre 0.5 y 3.0 en la mayor parte del dominio, con un foco aislado superior a 3.0 en el extremo nororiental. La distribución espacial de la incertidumbre no exhibió un gradiente centro-periferia, lo cual descarta un efecto de borde generalizado y sugiere que los puntos de mayor incertidumbre corresponden a combinaciones locales de covariables edáficas menos representadas en el conjunto de entrenamiento.

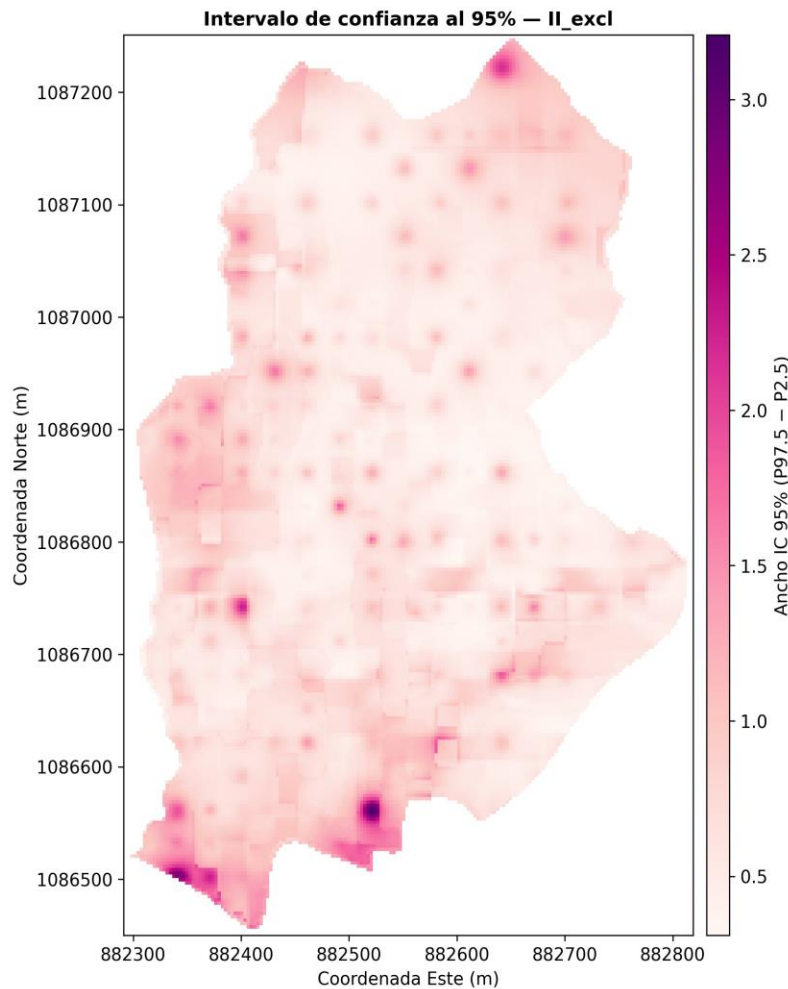


Figura 5 Mapa de ancho del intervalo de confianza al 95 % (P97.5 – P2.5) — Escenario II_excl.

5.4.4 Distribución espacial de residuales

La distribución espacial de los residuales en los 145 puntos de muestreo (Figura 6) mostró una media prácticamente nula ($\mu = 0.0077$) y una desviación estándar de 0.4325, sin evidencia de agrupamiento geográfico de errores de signo común. Los residuales positivos y negativos se distribuyeron de manera intercalada a lo largo del área de estudio, lo que valida el cumplimiento de los supuestos de independencia residual verificados en el Objetivo 3 (cobertura aproximada del 95 % = 0.966). Se identificó un único residual atípico negativo de magnitud -3.0 en el extremo suroccidental, correspondiente a una observación con COS elevado en una zona de baja densidad muestral.

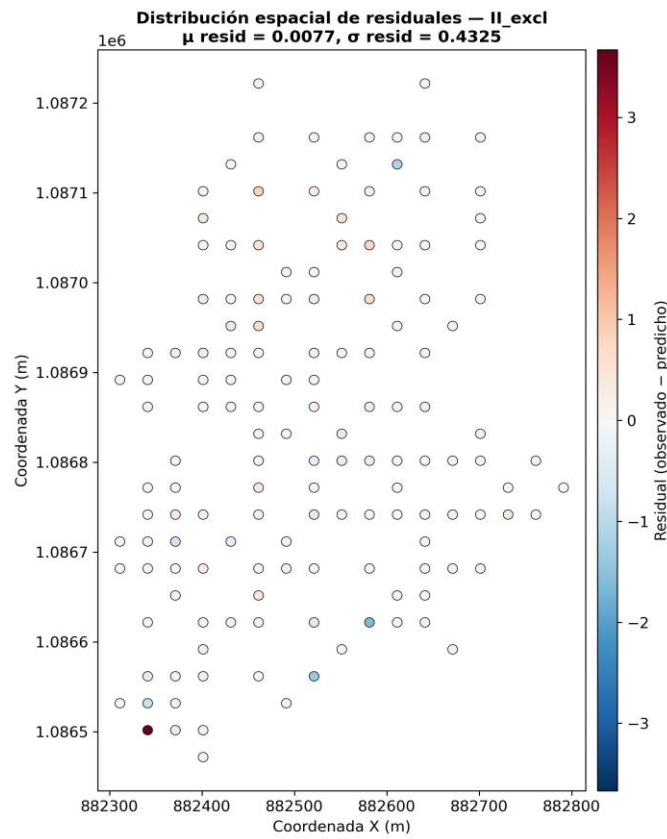


Figura 6 Distribución espacial de residuales (observado – predicho) en los 145 puntos de muestreo — Escenario II_excl. μ resid = 0.0077, σ resid = 0.4325.

5.4.5 Síntesis de la producción cartográfica

La producción cartográfica del Escenario II con exclusión de MO y N demostró que el modelo SVR_poly, seleccionado en el Objetivo 3 como campeón bajo criterios de desempeño espacial, estabilidad bootstrap y cumplimiento de supuestos, generó una superficie predictiva del COS espacialmente coherente y con incertidumbre controlada sobre el área de la Granja San José. La consistencia entre las métricas de coherencia interna ($R^2 = 0.87$) y las de validación cruzada espacial ($R^2 = 0.73$ reportado en el Objetivo 3) descarta un sobreajuste severo y confirma que la transición del modelo puntual a la predicción exhaustiva no degradó su capacidad predictiva. La cartografía resultante constituye un insumo operativo para la zonificación territorial del COS a escala predial, donde las covariables edáficas fisicoquímicas indirectas (Al, CIC, K, pH, textura), espacializadas mediante interpolación geoestadística, proporcionaron suficiente señal para sostener una predicción estructuralmente robusta sin depender de la materia orgánica ni del nitrógeno total.

Tabla 34 Resumen de la producción cartográfica — Escenario II con exclusión (SVR)

Indicador	Valor
Modelo campeón	SVR_poly (grado 2)
Covariables	29 bandas (edáficas + topográficas)
Resolución espacial	3 m (EPSG:3116)
Píxeles válidos	28 981
COS predicho — media $\pm \sigma$	8.35 $\pm$ 0.73 %
COS predicho — rango	[4.70, 10.80] %
R ²	0.8707
RMSE	0.4326
MAE	0.2212
Sesgo	−0.0077
σ residual	0.4325
Incertidumbre (bootstrap)	N = 100 réplicas

Tabla 35 Resumen de la producción cartográfica — Escenario II_excl (SVR_poly)

6. CONCLUSIONES Y TRABAJOS FUTUROS

El análisis estadístico de las covariables evidenció que la variabilidad del COS en la Granja San José está principalmente estructurada por variables edáficas asociadas a la fracción orgánica del suelo, la capacidad de intercambio catiónico, la textura y la acidez, mientras que las covariables topográficas y espectrales aportan información complementaria para representar la heterogeneidad espacial del sistema.

La comparación entre familias de modelos mostró que no existe una única estructura algorítmica superior en todos los escenarios, sino que el desempeño depende del conjunto de covariables disponible, del grado de redundancia informacional y de la capacidad del modelo para generalizar bajo validación espacial.

La evaluación multicriterio permitió establecer que el modelo SVR_poly en el escenario II con exclusión de MO y N ofreció el mejor equilibrio entre desempeño predictivo, estabilidad espacial, control de incertidumbre y menor dependencia de covariables directamente asociadas al COS, por lo que fue seleccionado como modelo más adecuado para la estimación indirecta del carbono orgánico del suelo.

La cartografía generada permitió representar espacialmente la distribución del COS y su incertidumbre asociada, evidenciando que la aplicación del modelo seleccionado es útil para identificar patrones territoriales de acumulación, zonas de mayor variabilidad predictiva y áreas prioritarias para monitoreo o validación adicional.

La Ciencia de Datos constituyó el eje decisivo del trabajo al transformar datos edáficos, topográficos y espectrales heterogéneos en un sistema predictivo reproducible, validado espacialmente y acompañado de incertidumbre. Su aporte central fue convertir la estimación del COS en una arquitectura analítica robusta, capaz de integrar modelación, evaluación multicriterio y cartografía digital para generar conocimiento territorial verificable y útil para la gestión sostenible del suelo.

Trabajos futuros

Ampliar el tamaño muestral y la cobertura ambiental del estudio, incorporando nuevas observaciones en unidades de paisaje y coberturas subrepresentadas, con el fin de fortalecer la capacidad de generalización del modelo y reducir la incertidumbre espacial.

Realizar validación externa en predios o áreas distintas al dominio original de entrenamiento, para establecer la transferibilidad real del modelo y diferenciar entre desempeño local y capacidad de generalización regional.

Incorporar covariables adicionales de tipo climático, hidrológico y de manejo, incluyendo variables asociadas a humedad del suelo, historia de uso, fertilización e intensidad de disturbio, con el propósito de capturar dimensiones del COS no completamente representadas en el presente estudio.

Desarrollar enfoques de modelación multitemporal que permitan analizar la dinámica temporal del COS, integrando nuevas campañas de muestreo y series históricas de percepción remota para avanzar desde una cartografía estática hacia esquemas de monitoreo.

Extender la arquitectura metodológica desarrollada hacia otras propiedades edáficas de interés agronómico y ambiental, de modo que el trabajo evolucione desde la modelación de una variable específica hacia una plataforma integral de cartografía digital de suelos.

Evaluar la utilidad operativa de la cartografía generada en procesos de planificación predial, restauración ecológica, gestión del suelo y estimación de reservas de carbono, para consolidar la transición del resultado analítico hacia aplicaciones territoriales concretas.

REFERENCIAS BIBLIOGRÁFICAS

- [1] IPCC, “IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.”, jul. 2023. doi: 10.59327/IPCC/AR6-9789291691647.
- [2] H. Burbano Orjuela, “El carbono orgánico del suelo y su papel frente al cambio climático”, *Revista de Ciencias Agrícolas*, vol. 35, núm. 1, p. 82, jun. 2018, doi: 10.22267/rcia.183501.85.
- [3] T. L. Jiru, A. W. Moomen, M. K. Domfeh, y O. Appiah, “Spatial prediction of soil organic carbon stocks using regression kriging to support land degradation neutrality (SDG 15.3.1) in Negele Arsi District, Ethiopia”, *Sustainable Futures*, vol. 11, jun. 2026, doi: 10.1016/j.sftr.2026.101796.
- [4] G. A. Araujo-Carrillo *et al.*, “Colombia: Soil Organic Carbon Sequestration Potential National Map Executive summary”. [En línea]. Disponible en: <http://54.229.242.119/GloSIS/>
- [5] A. Hoyos-Sanclemente, J. C. Menjivar-Flores, y G. Rueda-Saa, “Soil organic carbon in agricultural soils of an inter-Andean valley in Colombia: understanding the effects of environmental and geographic variables”, *Environ. Monit. Assess.*, vol. 197, núm. 6, p. 697, may 2025, doi: 10.1007/s10661-025-14123-1.
- [6] T. Hengl *et al.*, “SoilGrids1km — Global Soil Information Based on Automated Mapping”, *PLoS One*, vol. 9, núm. 8, p. e105992, ago. 2014, doi: 10.1371/JOURNAL.PONE.0105992.
- [7] A. S. Moreno Muñoz, Á. I. Guzmán Alvis, y I. F. Benavides Martínez, “A random forest model to predict soil organic carbon storage in mangroves from Southern Colombian Pacific coast”, *Estuar. Coast. Shelf Sci.*, vol. 299, abr. 2024, doi: 10.1016/j.ecss.2024.108674.
- [8] P. C. Serrano-Agudelo, L. J. Martínez-Martínez, P. A. Serrano-Cely, A. Bolívar-Gamboa, y D. F. Moreno-Pérez, “Determining of variables influencing soil organic carbon content of PNR-Cortadera paramo (Colombia) via remote sensing”, *Revista Colombiana de Ciencias Hortícolas*, vol. 18, núm. 2, may 2024, doi: 10.17584/RCCH.2024V18I2.17464.
- [9] J. P. Wilson, *Spatial Data Science*. 380 New York Street, Redlands, California 92373-8100: Esri Press, 2024.

- [10] M. Wiesmeier *et al.*, “Soil organic carbon storage as a key function of soils - A review of drivers and indicators at various scales”, *Geoderma*, vol. 333, pp. 149–162, ene. 2019, doi: 10.1016/j.geoderma.2018.07.026.
- [11] A. Fisher, C. Rudin, y F. Dominici, “All Models are Wrong, but Many are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously”, 2019. [En línea]. Disponible en: <http://jmlr.org/papers/v20/18-760.html>.
- [12] L. G. Bernardini *et al.*, “Learning vs. understanding: When does artificial intelligence outperform process-based modeling in soil organic carbon prediction?”, *N. Biotechnol.*, vol. 81, pp. 20–31, jul. 2024, doi: 10.1016/j.nbt.2024.03.001.
- [13] B. Malone, D. Arrouays, L. Poggio, B. Minasny, y A. B. McBratney, “Digital soil mapping: Evolution, current state and future directions of the science”, *Encyclopedia of Soils in the Environment, Second Edition*, pp. V4-684-V4-695, ene. 2023, doi: 10.1016/B978-0-12-822974-3.00130-0.
- [14] C. A. Sierra, A. M. Hoyt, Y. He, y S. E. Trumbore, “Soil Organic Matter Persistence as a Stochastic Process: Age and Transit Time Distributions of Carbon in Soils”, *Global Biogeochem. Cycles*, vol. 32, núm. 10, pp. 1574–1588, oct. 2018, doi: 10.1029/2018GB005950.
- [15] A. B. McBratney, M. L. Mendonça Santos, y B. Minasny, “On digital soil mapping”, *Geoderma*, vol. 117, núm. 1–2, pp. 3–52, nov. 2003, doi: 10.1016/S0016-7061(03)00223-4.
- [16] N. Beisekenov, W. Banakinaou, A. D. Ajayi, H. Hasegawa, y A. Tadao, “Remote sensing-based soil organic carbon monitoring using advanced machine learning techniques under conservation agriculture systems”, *Smart Agricultural Technology*, vol. 11, ago. 2025, doi: 10.1016/j.atech.2025.101036.
- [17] A. M. J. C. Wadoux, B. Minasny, y A. B. McBratney, “Machine learning for digital soil mapping: Applications, challenges and suggested solutions”, el 1 de noviembre de 2020, *Elsevier B.V.* doi: 10.1016/j.earscirev.2020.103359.
- [18] Q. Wang *et al.*, “Assessing spatiotemporal variations of soil organic carbon and its vulnerability to climate change: A bottom-up machine learning approach”, *Climate Smart Agriculture*, vol. 1, núm. 2, nov. 2024, doi: 10.1016/j.csag.2024.100025.
- [19] R. B. Jackson, K. Lajtha, S. E. Crow, G. Hugelius, M. G. Kramer, y G. Piñeiro, “The Ecology of Soil Carbon: Pools, Vulnerabilities, and Biotic and Abiotic Controls”, *Annu.*

- Rev. Ecol. Evol. Syst.*, vol. 48, pp. 419–445, nov. 2017, doi: 10.1146/ANNUREV-ECOLSYS-112414-054234.
- [20] Y. Tekin, Z. Tümsavas, y A. M. Mouazen, “Comparing the artificial neural network with parcial least squares for prediction of soil organic carbon and pH at different moisture content levels using visible and near-infrared spectroscopy”, *Rev. Bras. Cienc. Solo*, vol. 38, núm. 6, pp. 1794–1804, 2014, doi: 10.1590/S0100-06832014000600014.
- [21] M. Guevara *et al.*, “No silver bullet for digital soil mapping: Country-specific soil organic carbon estimates across Latin America”, *SOIL*, vol. 4, núm. 3, pp. 173–193, ago. 2018, doi: 10.5194/soil-4-173-2018.
- [22] I. A. Delgado-Vargas y N. B. Franco, “Soil organic carbon storage in different agroforestry systems associated with coffee in Nariño, Colombia”, *Agronomia Mesoamericana*, vol. 35, 2024, doi: 10.15517/am.2024.59765.
- [23] L. B. Guo y R. M. Gifford, “Soil carbon stocks and land use change: A meta analysis”, *Glob. Chang. Biol.*, vol. 8, núm. 4, pp. 345–360, 2002, doi: 10.1046/j.1354-1013.2002.00486.x.
- [24] M. Ortega-Monsalve, M. Medina-Sierra, y M. F. Cerón-Muñoz, “Relación del carbono orgánico y el nitrógeno total con otras variables del suelo en producciones ganaderas del departamento de Antioquia, Colombia”, *Revista U.D.C.A Actualidad & Divulgación Científica*, vol. 28, núm. 2, dic. 2025, doi: 10.31910/rudca.v28.n2.2025.2553.
- [25] F. A. Bócoli *et al.*, “Can environmental variables, high sampling density and machine learning deliver detailed maps of soil organic carbon and carbon stock in tropical regions?”, *Catena (Amst)*, vol. 249, feb. 2025, doi: 10.1016/j.catena.2025.108718.
- [26] I. Etinosa, J. DeMarco, y P. Crossley, “Remote sensing-based modeling of soil organic carbon in wet meadow and sagebrush ecosystems in semi-arid landscapes”, *J. Arid Environ.*, vol. 233, feb. 2026, doi: 10.1016/j.jaridenv.2025.105545.
- [27] A. Mulakaledu, B. Swathi, M. M. Jadhav, S. M. Shukri, V. Bakka, y P. Jangir, “Satellite Image–Based Ecosystem Monitoring with Sustainable Agriculture Analysis Using Machine Learning Model”, *Remote Sens. Earth Syst. Sci.*, vol. 7, núm. 4, pp. 764–773, dic. 2024, doi: 10.1007/s41976-024-00157-5.
- [28] R. W. Aslam *et al.*, “Machine Learning-Based Wetland Vulnerability Assessment in the Sindh Province Ramsar Site Using Remote Sensing Data”, *Remote Sens. (Basel)*, vol. 16, núm. 5, mar. 2024, doi: 10.3390/rs16050928.

- [29] L. Breiman, “Random Forests”, 2001.
- [30] J. Roca-Pardiñas, C. Cadarso-Suárez, y W. González-Manteiga, “8 Bootstrap Methods for Testing Interactions in GAMs”.
- [31] S. Theodoridis, “Learning in Parametric Modeling: Basic Concepts and Directions”, en *Machine Learning*, Elsevier, 2020, pp. 67–120. doi: 10.1016/B978-0-12-818803-3.00012-X.
- [32] P. Mathur, “Support vector machine, an important supervised learning category”, en *Artificial Intelligence for Computational Fluid Dynamics*, Elsevier, 2026, pp. 237–260. doi: 10.1016/B978-0-443-29118-0.00005-0.
- [33] H. Omre, T. M. Fjeldstad, y O. B. Forberg, “Bayesian Spatial Modelling”, *Springer Books*, pp. 3–16, 2024, doi: 10.1007/978-3-031-65418-3_2.
- [34] Md. M. Hossain, M. J. Simu, N. Sakib, K. M. Safin Kamal, y A. W. Reza, “Deep Learning-Based Soil Image Classification for Sustainable Agriculture: A Case Study on Crop Recommendation”, 2025, pp. 151–165. doi: 10.1007/978-981-96-6297-5_12.
- [35] V. Cherkassky y Y. Ma, “Comparison of model selection for regression”, *Neural Comput.*, vol. 15, núm. 7, pp. 1691–1714, jul. 2003, doi: 10.1162/089976603321891864.
- [36] M. Emami, F. Khormali, M. R. Pahlavan-Rad, y S. Ebrahimi, “Predicting the spatial distribution of organic carbon in soil by combining machine learning algorithms and spline depth function in a part of Golestan Province, Iran”, *Soil Tillage Res.*, vol. 251, sep. 2025, doi: 10.1016/j.still.2025.106530.
- [37] N. Wang, Q. Wu, Y. Gui, Q. Hu, y W. Li, “Cross-Modal Segmentation Network for Winter Wheat Mapping in Complex Terrain Using Remote-Sensing Multi-Temporal Images and DEM Data”, *Remote Sens. (Basel)*, vol. 16, núm. 10, may 2024, doi: 10.3390/rs16101775.
- [38] M. Gouda *et al.*, “Integration of remote sensing and artificial neural networks for prediction of soil organic carbon in arid zones”, *Front. Environ. Sci.*, vol. 12, 2024, doi: 10.3389/fenvs.2024.1448601.
- [39] Y. Zheng, X. Zhao, X. Li, H. Chen, C. Li, y C. Zhang, “Mapping soil organic carbon density via geographically weighted regression with smooth terms: A case study in Shanxi Province”, *Ecol. Indic.*, vol. 166, sep. 2024, doi: 10.1016/j.ecolind.2024.112588.
- [40] D. Radočaj, I. Plaščak, y M. Jurišić, “A Comparative Assessment of Regular and Spatial Cross-Validation in Subfield Machine Learning Prediction of Maize Yield from Sentinel-2

- Phenology”, *Eng* 2025, Vol. 6, Page 270, vol. 6, núm. 10, p. 270, oct. 2025, doi: 10.3390/eng6100270.
- [41] C. Milà, M. Ludwig, E. Pebesma, C. Tonne, y H. Meyer, “Random forests with spatial proxies for environmental modelling: Opportunities and pitfalls”, *Geosci. Model Dev.*, vol. 17, núm. 15, pp. 6007–6033, ago. 2024, doi: 10.5194/GMD-17-6007-2024.
- [42] C. A. Mushagalusa, A. B. Fandohan, y R. Glèlè Kakaï, “Random forest and spatial cross-validation performance in predicting species abundance distributions”, *Environmental Systems Research*, vol. 13, núm. 1, dic. 2024, doi: 10.1186/s40068-024-00352-9.
- [43] L. Poggio *et al.*, “SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty”, *SOIL*, vol. 7, núm. 1, pp. 217–240, jun. 2021, doi: 10.5194/SOIL-7-217-2021.
- [44] T. Behrens y R. A. Viscarra Rossel, “On the interpretability of predictors in spatial data science: the information horizon”, *Sci. Rep.*, vol. 10, núm. 1, dic. 2020, doi: 10.1038/s41598-020-73773-y.
- [45] M. Riani, A. C. Atkinson, y A. Corbellini, “Automatic robust Box–Cox and extended Yeo–Johnson transformations in regression”, *Stat. Methods Appl.*, vol. 32, núm. 1, pp. 75–102, mar. 2023, doi: 10.1007/s10260-022-00640-7.
- [46] R. Tibshirani, “Regression Shrinkage and Selection Via the Lasso”, *Journal of the Royal Statistical Society. Series B: Methodological*, vol. 58, núm. 1, pp. 267–288, ene. 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.
- [47] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, y A. Brenning, “Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data”, *Ecol. Modell.*, vol. 406, pp. 109–120, ago. 2019, doi: 10.1016/j.ecolmodel.2019.06.002.
- [48] H. Li, X. Wang, y Y. Zhang, “Satellite-based precipitation retrieval using cross-temporal attention and multiscale spatial representation”, *Advances in Space Research*, abr. 2026, doi: 10.1016/j.asr.2026.04.109.
- [49] H. Meyer y E. Pebesma, “Predicting into unknown space? Estimating the area of applicability of spatial prediction models”, *Methods Ecol. Evol.*, vol. 12, núm. 9, pp. 1620–1633, sep. 2021, doi: 10.1111/2041-210X.13650.

- [50] V. A. Filla, A. P. Coelho, A. D. Ferroni, A. S. R. de S. Bahia, y J. Marques Júnior, “Estimation of clay content by magnetic susceptibility in tropical soils using linear and nonlinear models”, *Geoderma*, vol. 403, p. 115371, dic. 2021, doi: 10.1016/j.geoderma.2021.115371.
- [51] J. Derrac, S. García, D. Molina, y F. Herrera, “A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms”, *Swarm Evol. Comput.*, vol. 1, núm. 1, pp. 3–18, mar. 2011, doi: 10.1016/j.swevo.2011.02.002.
- [52] S. García, A. Fernández, J. Luengo, y F. Herrera, “Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power”, *Inf. Sci. (N. Y.)*, vol. 180, núm. 10, pp. 2044–2064, may 2010, doi: 10.1016/j.ins.2009.12.010.
- [53] Y. Gal y Z. Ghahramani, “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning”, oct. 2016, [En línea]. Disponible en: <http://arxiv.org/abs/1506.02142>
- [54] S. Lundberg y S.-I. Lee, “A Unified Approach to Interpreting Model Predictions”, nov. 2017, [En línea]. Disponible en: <http://arxiv.org/abs/1705.07874>
- [55] R. Amundson, “Factors of soil formation in the 21st century”, *Geoderma*, vol. 391, 2021, doi: 10.1016/j.geoderma.2021.114960.
- [56] D. P. Roy, H. Huang, R. Houborg, y V. S. Martins, “A global analysis of the temporal availability of PlanetScope high spatial resolution multi-spectral imagery”, *Remote Sens. Environ.*, vol. 264, p. 112586, oct. 2021, doi: 10.1016/j.rse.2021.112586.