UNERG AGRO-Científica
2025 6(1): 478-504
https://unerg.edu.ve/unerg-agro-cientifica/
ISSN: 2665-0061 (Online)

APLICACIÓN DE LA TECNICA BOSQUES ALEATORIOS (*RANDOM FOREST*) EN LA ESTIMACION DEL CARBONO ORGÁNICO DEL SUELO EN PAISAJES DE MONTAÑA

Ángel Rafael Valera Valera[1*]

[1] Universidad Rómulo Gallegos, Centro de Investigación y Extensión en Suelos y Aguas (CIESA-UNERG), San Juan de los Morros, Estado Guárico, Venezuela, e-mail: angelvalera@unerg.edu.ve. ORCID: https://orcid.org/0000-0001-5500-1332

*Autor de correspondencia

Recibido: 15/03/2025; **Aceptado**: 15/06/2025; **Publicado**: 30/06/2025

RESUMEN

La estimación de la reserva de carbono orgánico del suelo (COS) es de gran importancia para el conocimiento del grado de fertilidad y del aporte de nutrientes para el sustento de las plantas, para evaluar la potencialidad del ecosistema como sumidero de carbono, y para ejercer acciones contra los efectos del cambio climático. Para estimar la reserva de COS se requiere el conocimiento de la variación espacial de variables edafológicas asociadas con el contenido relativo del carbono orgánico (%CO), el espesor de los horizontes (Esp A), la densidad aparente de los suelos (Dap) y el contenido de esqueleto grueso (%EG). Con la finalidad de estimar la reserva de CO en la capa superficial de suelos de paisajes de montaña en un sector de la Cordillera de la Costa Central en Venezuela, se realizaron evaluaciones utilizando el algoritmo de aprendizaje automatizado denominado bosques aleatorios (RF, *random forest*) en una superficie de 6.760 hectáreas. Con este método de máquina de aprendizaje o *machine learning* se evaluaron las variables relacionadas con la predicción del COS utilizando los datos de 130 sitios de muestreo y variables auxiliares derivadas de modelos digitales de elevación e imágenes de satélite de 15m de resolución espacial, para una representación a escala 1:50.000. El entrenamiento y la generación de los modelos RF arrojaron coeficientes de determinación (R^2) de 0,963, 0,948, 0,932 y 0,946, para las variables %CO, Esp A, Dap y %EG, respectivamente; y la validación realizada con un conjunto de datos independientes arrojó un coeficiente de concordancia (CC) de 0,813, 0,593, 0,665, 0,855 para las variables consideradas, lo cual corresponde con una moderada a alta consistencia de los modelos base asociados con la estimación del COS. Los resultados indicaron que el método RF es capaz de estimar la distribución espacial de la reserva de COS, con variaciones desde 6,5 a 130 t ha^{-1} entre sectores con problemas de erosión y zonas con cobertura boscosa, lo que indicó que el método proporciona una apropiada estimación de la variabilidad espacial de la reserva de carbono orgánico en la capa superficial de los suelos de áreas montañosas.

Palabras clave: Bosques Aleatorios, Cartografía Digital de Suelos, Carbono Orgánico del Suelo, Paisajes de Montaña, Teledetección.

APPLICATION OF THE RANDOM FOREST TECHNIQUE FOR SOIL ORGANIC CARBON ESTIMATION IN MOUNTAIN LANDSCAPES

ABSTRACT

The estimation of the soil organic carbon stock (SOC) is of great importance for the knowledge of the degree of fertility and nutrient supply for plant sustenance, for assessing the potential of the ecosystem as a carbon sink, and for taking action

against the effects of climate change. To estimate the SOC pool requires knowledge of the spatial variation of soil variables associated with relative organic carbon content (%OC), A-horizon thickness (A-ht), soil bulk density (BD) and coarse skeleton content (%SC). In order to estimate the OC stock in the topsoil of mountain landscapes in a sector of the Cordillera de la Costa Central in Venezuela, evaluations were carried out using the automated learning algorithm called random forest (RF) on an area of 6,760 hectares. With this machine learning method, variables related to SOC prediction were evaluated using data from 130 sampling sites and auxiliary variables derived from digital elevation models and satellite images of 15 m spatial resolution, for a 1:50,000 scale representation. The training and generation of the RF models yielded coefficients of determination (R^2) of 0.963, 0.948, 0.932 and 0.946, for the variables %OC, A-ht, Bd and %SC, respectively; and the validation performed with an independent data set yielded a coefficient of concordance (CC) of 0.813, 0.593, 0.665, 0.855 for the variables considered, which corresponds to a moderate to high consistency of the base models associated with the estimation of SOC. The results indicated that the RF method is able to estimate the spatial distribution of the COS stock, with variations from 6.5 to 130 t ha^{-1} between sectors with erosion problems and areas with forest cover, which indicated that the method provides an appropriate estimate of the spatial variability of the organic carbon stock in the topsoil of mountainous areas.

Keywords: Random Forests, Digital Soil Mapping, Soil Organic Carbon, Mountain Landscapes, Remote Sensing.

INTRODUCCIÓN

La materia orgánica juega un papel muy importante en la determinación de la fertilidad y la productividad de los suelos, y por consiguiente en todas las características más importantes de los suelos productivos, además de los efectos ambientales por medio del secuestro de carbono orgánico del suelo (COS), en cuyo proceso el carbono se fija desde la atmósfera a través de las plantas o los residuos orgánicos y se almacena en el suelo. El COS representa más del 55% de la materia orgánica, por lo que desempeña un papel fundamental en el manejo sostenible de la tierra y en su productividad (Lal, 2004); además, controla diversas propiedades del suelo como la estructura, la capacidad de retención de agua, la capacidad de intercambio catiónico, la relación carbono-nitrógeno y otras (USDA-NRCS, 1995), influye en las propiedades físicas, químicas y biológicas del suelo (Odebiri *et al*., 2020)y desempeña un papel vital en la calidad y la salud del suelo (Lal, 2004; Kingsley *et al*., 2021).

El COS es un indicador importante de la calidad del suelo y determina directamente la fertilidad del suelo, por lo que es necesario comprender su distribución espacial y los factores que controlan su variabilidad, para un manejo eficiente y sostenible de los nutrientes del suelo (Kingsley *et al.,* 2020). Además, un suelo con un contenido óptimo de CO puede absorber y almacenar agua y ponerla a disposición de los cultivos en condiciones de sequía (FAO, 2017), lo que mejora la capacidad de adaptación de la agricultura a los impactos del cambio climático y, en consecuencia, aumenta su resiliencia (Kidemo *et al*., 2023).

La evaluación del carbono del suelo tanto en las tierras de cultivo como en áreas boscosas, es útil para el secuestro de carbono y el manejo sostenible del suelo. Sin embargo, las intervenciones antropogénicas severas en las tierras de cultivo, principalmente en terrenos planos

(Huang, 2022), y en zonas sometidas a intensos procesos erosivos, crea incertidumbre en la obtención de información precisa del suelo con datos de muestra limitados. Además, cuando la estimación espacial del COS se produce en diferentes zonas litológicas, sobre distintas coberturas del suelo, diversas áreas agrícolas y zonas climáticas contrastantes, es muy importante que la elaboración de mapas de CO deba tratarse con la máxima atención, con la finalidad de mejorar el manejo del suelo y la evaluación medioambiental (Zeraatpishe, 2012).

La distribución espacial del carbono orgánico del suelo es muy heterogénea, por las variaciones edafoclimáticas y el efecto del uso y cobertura de la tierra, lo cual afecta la forma en que el ecosistema reacciona al grado de intervención por la pérdida de cobertura vegetal (Kakhani *et al*, 2023). Sin embargo, la gran variabilidad del COS debido a la influencia de los factores formadores del suelo y a un bajo número de observaciones de campo, es capaz de limitar la certeza de las estimaciones espaciales del SOC (Wang *et al*., 2022; Zeraatpisheh *et al*, 2023). A tal efecto, la obtención de datos representativos de la variabilidad espacial del COS, es uno de los problemas que dificultan la generación de información representativa para grandes áreas (Kidemo *et al*., 2023).

En general, el conocimiento de la variación espacial del carbono orgánico del suelo es esencial para el manejo sostenible de los suelos, la mitigación del cambio climático, la planificación agrícola y la conservación de la biodiversidad. El estudio de la reserva de COS es fundamental para garantizar la sostenibilidad ambiental, mitigar el cambio climático, conservar la biodiversidad y promover prácticas agroambientales más sostenibles en estas áreas clave para el mantenimiento de la biodiversidad. Además, la estimación del contenido de COS de zonas de montaña también es crucial para evaluar el cambio climático, comprender la vulnerabilidad de estos ecosistemas, desarrollar estrategias de adaptación y mitigación, comprender las conexiones entre los diversos ecosistemas y evaluar los efectos del cambio climático en todo el sistema.

A pesar de su importancia, la estimación precisa y oportuna del COS presenta desafíos significativos, especialmente en paisajes de montaña. La distribución espacial del carbono orgánico del suelo es inherentemente heterogénea, influenciada por variaciones edafoclimáticas y los efectos del uso y cobertura de la tierra. Esta variabilidad se acentúa en las regiones montañosas debido a la complejidad topográfica, la diversidad de factores ambientales y la presencia de procesos geomorfológicos dinámicos (Vela *et al.,* 2012; Khanal *et al*., 2025; Hoyos-Sanclemente *et al*., 2025).

Para la evaluación de la reserva de COS se han empleado desde modelos estadísticos lineales hasta modelos de aprendizaje automático (*machine learning*), en el cual los modelos no lineales han demostrado una mayor eficiencia para explicar la compleja relación suelo-ambiente (Huang, 2022). En el marco de la cartografía digital de suelos (DSM, *Digital Soil Mapping*), las

propiedades del suelo (incluyendo el COS) de lugares no visitados se estiman mediante algoritmos estadísticos y matemáticos que relacionan la variable de salida con un gran número de variables medioambientales (Lagacherie y McBratney, 2006; McBratney *et al*., 2003). Las técnicas de aprendizaje automático (ML) más comunes empleadas en DSM para predecir el COS incluyen el bosque aleatorio (RF, *random forest*), k-vecinos *más cercanos* (kNN), máquina de vectores de soporte (SVM, *support vector machine*), redes neuronales artificiales, y el árbol Cubista (Cu) de regresión, entre otros (Zeraatpisheh et al., 2023; *Adhikari et al., 2020; Garosi et al.*, *2022; Lamichhane et al.*, 2019; Lamichhane *et al.*, 2022). Esto se debe principalmente a su excelente rendimiento predictivo y a su capacidad para modelar la compleja relación entre las variables dependientes y las variables independientes proporcionadas (Dangeti, 2017).

Recientemente, también se han explorado varios tipos de covariables ambientales en la cartografía del carbono del suelo, basados en la integración de técnicas de sensores remotos, modelos digitales de elevación y la obtención de variables auxiliares con alta resolución espacial. Al respecto, las covariables climáticas y el material parental juegan un papel importante en el carbono del suelo a escala regional, mientras que, a escala local, la variabilidad del carbono del suelo a menudo depende de la topografía, el manejo agrícola y las propiedades del suelo (Huang, 2022).

Dentro del campo de la inteligencia artificial, el aprendizaje automático emplea algoritmos con la capacidad de identificar patrones en datos masivos y elaborar predicciones, permitiendo la realización de tareas específicas de forma autónoma, sin necesidad de ser programados (Molla *et al,* 2022). En estas técnicas de aprendizaje, RF destaca como un método de clasificación y regresión basado en la agregación de muchos árboles de decisión. El método de bosques aleatorios fue descrito por primera vez por (Breiman, 2001) y más recientemente, varios estudios han demostrado que es una de las mejores técnicas de aprendizaje automático disponibles en la actualidad como apoyo a la cartografía digital de suelos. Su formulación matemática detallada se ha dado a conocer en diversos trabajos de investigación (Vaysse y Lagacherie 2015; Olson *et al.,* 2017; Nussbaum et al. 2018). El modelo RF suele ser un método de aprendizaje automático muy exitoso utilizando datos continuos y discretos en estudios de DSM (Wadoux *et al*., 2020), y solamente requiere dos pasos fundamentales para la construcción del modelo: ii) la creación de numerosos árboles de decisión utilizando embolsado o empaquetado aleatorio (*bootstrap*) con reemplazo a partir de los datos originales, y 2) la validación interna utilizando un conjunto de datos independientes, no empleados en el *bootstrap* (Breiman, 2001; Liaw y Wiener, 2002; Zeraatpisheh *et al,* 2023).

En esta investigación se utilizó la técnica de bosques aleatorios (*Random Forest*) con la finalidad de estimar la reserva del COS en la capa superficial, en un sector

de la cuenca del rio Caramacate dominado por paisajes de montaña, con suelos de alta variabilidad, influenciada por el uso de la tierra dominante basado en ganadería extensiva y la ocurrencia de movimientos en masa. Para lograr este objetivo, se aplicó un método de inferencia inductivo y se desarrolló una secuencia metodológica con las siguientes actividades i) Generación de modelos de predicción de propiedades del suelo (%CO, Esp A, Dap, %EG) a partir de variables ambientales derivadas de un modelo digital de elevación (MDE) e imágenes satelitales; ii) Evaluación de la exactitud y la confiabilidad de estos modelos de predicción; y iii) Estimación de la distribución espacial de la reserva de carbono orgánico en la capa superficial del suelo, mediante la integración de los modelos generados en la zona de estudio.

MATERIALES Y MÉTODOS

Área de Estudio

La investigación se realizó en un sector de la cuenca alta del río Guárico, específicamente en la cuenca del río Caramacate, la cual está ubicada entre los municipios Santos Michelena y San Sebastián de los Reyes del estado Aragua (Venezuela), entre las coordenadas geográficas 9,55° a 10,09° Norte y -67,12° a -67,03° Oeste, (Figura 1). La cuenca del río Caramacate representa el 8,5% de la cuenca alta del río Guárico, de la cual es tributaria. El área muestra (piloto) utilizada para el presente estudio corresponde a una superficie de 6.760 ha, cuyo paisaje está dominado por laderas de montaña con pendientes del 40%.

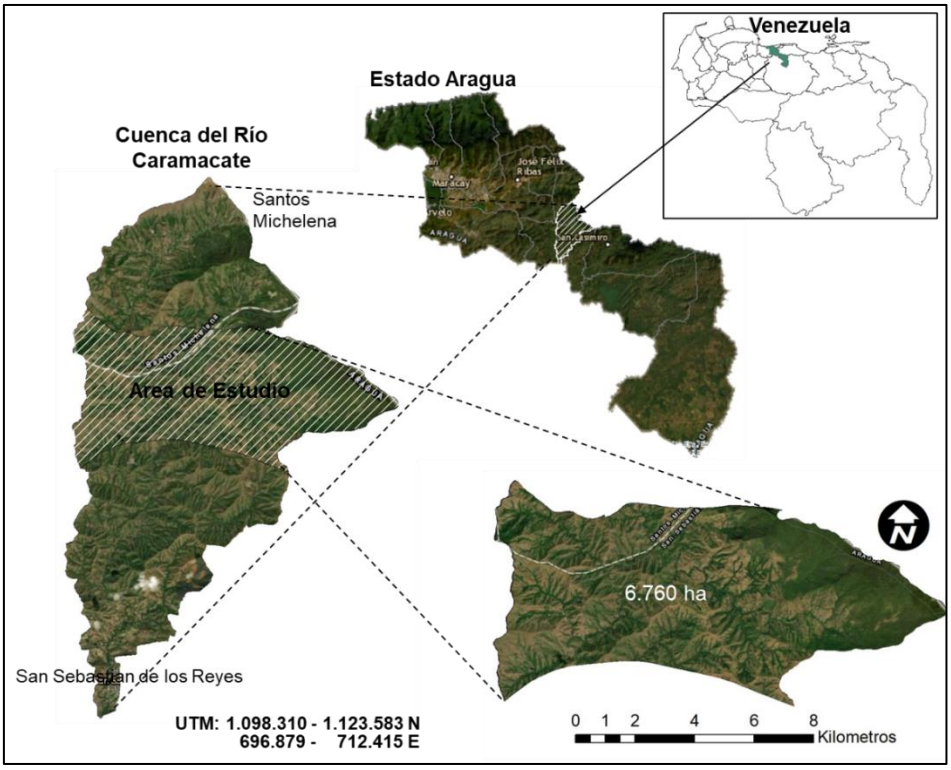

Figura 1. Ubicación relativa del área de estudio dentro de la cuenca del río Guárico, estado Aragua, Venezuela.

La geología está representada por rocas metavolcánicas y basaltos de la formación El Caño-El Chino, por metalavas máficas de la formación El Carmen (Urbani y Rodríguez 2004), sedimentos aluviales acarreados por el río Caramacate y los mantos coluvio-aluviales del Grupo Villa de Cura. La zona presenta una altitud comprendida entre 334 a 1.405 msnm, con una precipitación media anual que oscila entre 1.100 y 1.400 mm y una temperatura media anual varía entre 22 y 26 ºC. La vegetación herbácea ocupa más del 50% de la cobertura del sector, como producto de la deforestación y las quemas para el uso de la ganadería en condiciones extensivas. Los suelos en su mayoría son Entisoles, Inceptisoles y Alfisoles, cuya variabilidad se ha incrementado por el uso de la tierra dominante y la incidencia de los movimientos en masa (Pineda *et al*., 2011; Valera, 2018).

Datos de suelo

El conjunto de datos en la zona de estudio seleccionada está conformado por 130 muestras superficiales derivadas de perfiles de suelo, ubicados en diferentes posiciones geomorfológicas de laderas de paisajes de montaña y valles, de la cuenca del rio Caramacate (Valera, 2018). Los sitios de muestreo corresponden a observaciones del horizonte superficial de calicatas, cortes, y barrenos agrológicos, cuya distribución se presenta en la Figura 2, donde los círculos (color rojo) corresponden a los perfiles empleados en la generación y validación de los modelos.

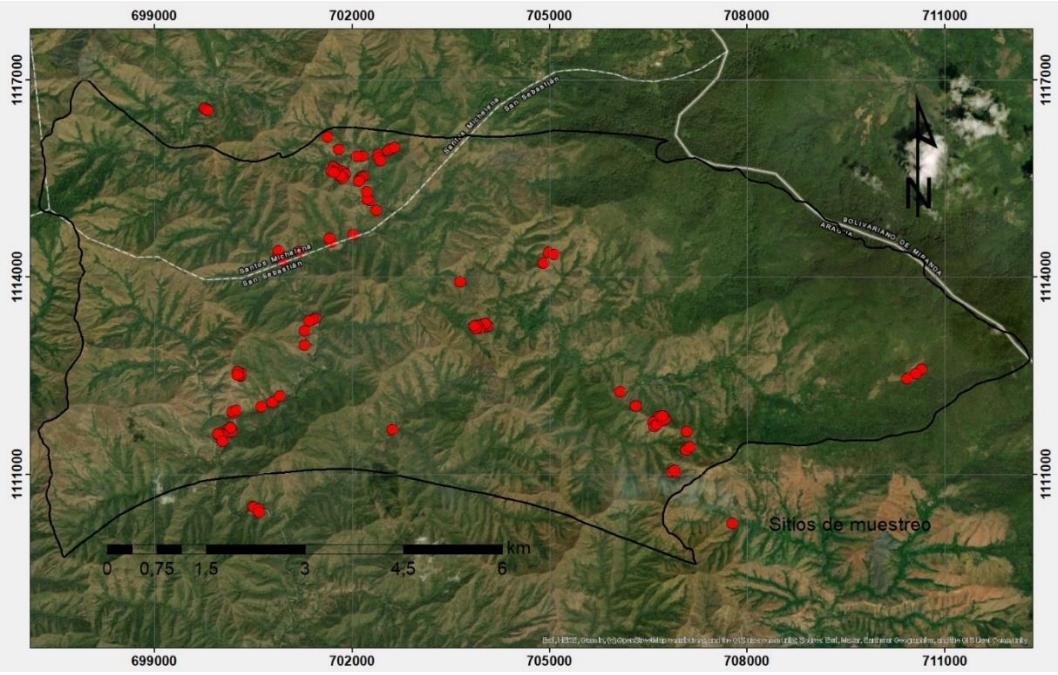

Figura 2. Distribución de los sitios de muestreo en el sector de estudio de la cuenca del río Caramacate.

Las variables del suelo seleccionadas para esta investigación, por su vinculación directa con la reserva de COS,

fueron: el porcentaje de carbono orgánico (%**CO**), el espesor del horizonte A (**Esp A**, en cm), la densidad aparente (**Dap**, en g cm^{-3}) y el contenido de esqueleto grueso (%**EG**). La variable morfológica Esp A se obtuvo en campo al momento del muestreo; la Dap se obtuvo mediante muestreo con toma muestra tipo *Uhland* con el método del cilindro metálico; el %EG se determinó mediante técnicas de tamizado en base a peso, y el %CO se realizó por el método de combustión húmeda (Walkley y Black, 1934) modificado, en el cual el suelo se oxida con una solución de dicromato de potasio estandarizada, utilizando el calor producido por la dilución de ácido sulfúrico concentrado, en la solución crómica (Walkley y Black, 1934), con lectura a 650 nm en espectrofotómetro de luz visible (Valera, 2018).

Variables auxiliares

Las variables auxiliares empleadas en este estudio se seleccionaron en concordancia con el modelo de factores de formación de suelos de Jenny (1941) y el modelo geoespacial multivariado ***scorpan***, formulado por McBratney *et al.* (2003) para la predicción de propiedades del suelo. Este modelo establece las relaciones entre las propiedades del suelo y los factores de formación espacialmente referenciados, incluyendo otros datos de suelo (**s**), clima (**c**), organismos (**o**), relieve (**r**), material parental (**p**), tiempo (**a**) y localización espacial (**n**). Las fuentes de datos digitales utilizadas para generar estas covariables ambientales incluyeron un modelo digital de elevación (MDE) de 15 m de resolución espacial (Figura 3), a partir del cual se derivaron los diversos parámetros morfométricos (variables auxiliares o variables topográficas) tales como: altitud (Alt) (Zevenbergen y Thorne, 1987; Burrough y McDonnell, 1998), gradiente (Pend) y orientación (Asp) de la pendiente (Moore *et al.*, 1991), posición relativa (PR) (Verbrugge, 2006), perfil de curvatura (Perfil_C, curvatura vertical o longitudinal), y plano de curvatura (Plano_C, curvatura horizontal o transversal) (Moore *et al*., 1993), área de captación (Area_C) (Tarboton *et al.* 1991), el índice topográfico de humedad (ITH) (Wilson y Gallant, 2000), el índice del potencial de escorrentía (ISP) y el índice de transporte de sedimentos (LS), los cuales fueron generadas con el programa SAGA v. 2.1.4. (Conrad *et al*., 2015).

También se utilizó un mapa de precipitación media anual (PP) realizado por kriging ordinario (Pineda *et al.,* 2011), y el índice de vegetación de diferencia normalizada (NDVI) calculado a partir de las bandas roja e infrarroja de una imagen Spot 5 multiespectral 658-330 (Rouse, 1974). Estas variables auxiliares fueron empleadas como parámetros de correlación ambiental para la estimación de las propiedades específicas del suelo en áreas no muestreadas.

Análisis estadístico

Los datos de las variables de suelo y las variables auxiliares se sometieron a un análisis exploratorio (AED) con apoyo del paquete estadístico SPSS® (IBM® *Statistics*, versión 20), para determinar los estadísticos descriptivos, como: media, mediana, varianza, coeficiente de variación, valores máximos y mínimos, y los índices de asimetría y curtosis. Para detectar la presencia de valores atípicos se utilizó la metodología de cercas externas e internas de Tukey (1977), y

posteriormente se realizó la prueba de normalidad de Kolmogorov-Smirnov, para evaluar la distribución de los datos.

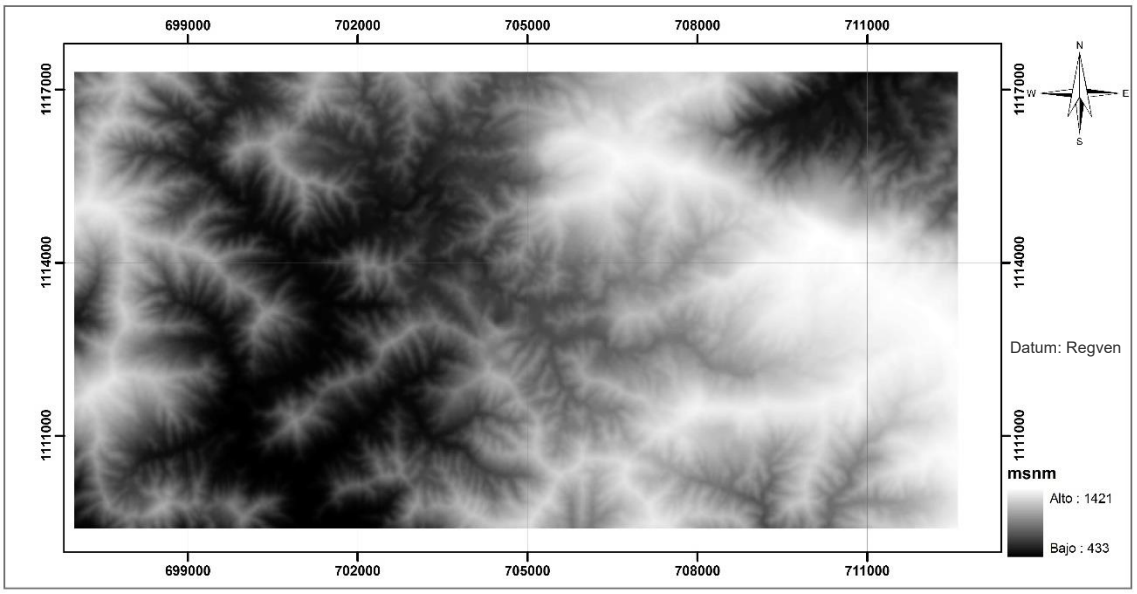

Figura 3. Modelo digital de elevación del área de referencia espacial de la cuenca del río Caramacate. **Fuente:** Valera (2018).

Técnica de predicción basada en bosques aleatorios (RF, *Random Forest*)

Para la predicción y evaluación de la distribución espacial del COS, se utilizó el algoritmo de aprendizaje automático de bosques aleatorios (RF, Random Forest) (Breiman, 2001). El método RF es un algoritmo de aprendizaje automático (*machine learning*) que se ha utilizado con éxito para predecir el carbono orgánico del suelo en varias regiones y condiciones de suelo. *Random Forest* es un algoritmo de aprendizaje automático de tipo ensamble, lo cual significa que combina varios algoritmos de aprendizaje individuales para producir una mejor precisión en las predicciones. El algoritmo crea varios árboles de decisión y combina las predicciones hechas por cada árbol para producir una predicción final.

RF consiste en un conjunto de clasificación aleatoria y árboles de regresión, cuyo algoritmo de hace crecer diferentes árboles seleccionando al azar y repetidamente variables predictoras y casos de entrenamiento para desarrollar una población aleatoria de árboles. El funcionamiento del algoritmo RF implica dos pasos fundamentales: i) Creación de árboles de decisión, en la cual se generan numerosos árboles de decisión utilizando un método de remuestreo llamado *bootstrap aggregation* (*bagging*) con reemplazo a partir de los datos originales. Para cada árbol, se selecciona aleatoriamente un subconjunto de variables predictoras en cada nodo para la división, lo que introduce aleatoriedad y reduce la correlación entre los árboles;

y ii) Validación interna, donde se utiliza un conjunto de datos independientes, no empleados en el *bootstrap* (conocidos como datos *out-of-bag* u OOB), para la validación interna del modelo. La salida final de RF es el promedio de las predicciones de los árboles individuales. (Breiman, 2001). La eficiencia de RF es notable, especialmente cuando el número de descriptores es muy grande, ya que puede manejar simultáneamente variables categóricas y continuas, así como relaciones complejas de alto orden, incluyendo la no linealidad y los efectos de interacción entre factores.

En esta investigación, se utilizó el software ArcGIS Pro® para llevar a cabo la predicción de los modelos RF de los atributos edafológicos en consideración, utilizando las variables auxiliares más importantes del sector. Para la predicción de las propiedades del suelo se utilizaron mapas en formato ráster de la precipitación anual media (PP), el índice de vegetación de diferencia normalizada (NDVI), y los atributos morfométricos derivados de un MDE de 15 m de resolución espacial de la zona estudiada.

Generación de modelos de predicción de variables del suelo

La generación de modelos de propiedades del suelo (%CO, Esp A, Dap, %EG) con el programa Arc GIS Pro® se realizó en tres fases: i) Entrenamiento con un 75% de los datos, hasta obtener la mejor combinación de variables predictoras para el ajuste de los parámetros del modelo para minimizar el error de predicción, las variables predictoras de mayor importancia, y el mejor ajuste del coeficiente de determinación (R^2); ii) Entrenamiento con las variables de mayor importancia, y generación de los mapas de propiedades del suelo en formato *raster*, conjuntamente con el grado de incertidumbre de cada modelo de predicción; y iii) Validación de los modelos de predicción con un conjunto de datos independientes equivalente al 25% del total.

El entrenamiento del *Random Forest* y la posterior combinación de sus resultados para inferir el COS total es un proceso de inferencia a partir de datos, lo cual es la esencia del razonamiento inductivo. El método se caracteriza por partir de observaciones o datos específicos (en este caso, los datos de variables del suelo utilizados para entrenar el algoritmo *Random Forest* y generar mapas individuales) para luego construir una generalización a través de un mapa final de COS). El algoritmo *Random Forest*, al aprender patrones de los datos de entrada para predecir las propiedades del suelo, opera de manera inductiva, partiendo de lo particular (cada propiedad del suelo por separado) a lo general (el contenido total de COS).

Evaluación de la bondad de ajuste de modelos de predicción variables del suelo

En la evaluación de la bondad de ajuste y precisión de los modelos (fase de calibración de modelos o entrenamiento) de los modelos de predicción se utilizaron los índices del diagnóstico de regresión generado en el proceso de entrenamiento. Al predecir una variable continua, el valor observado para cada

una de las entidades de prueba se compara con las predicciones de dichas entidades en función del modelo entrenado, y se obtienen los siguientes índices: el coeficiente de determinación (R^2), valor *P* y el error estándar (SE, *Standard Error)* del valor estimado, más la raíz del error cuadrático medio (RMSE, *Root Mean Square Error*). El coeficiente R^2 describe el grado de colinealidad entre los datos observados (medidos) y estimados. Un valor del error estándar de la predicción más bajo indica una estimación más precisa del modelo, y el RMSE evalúa la precisión de la predicción y sus valores deben ser lo más pequeño posible. Las ecuaciones que definen los índices R^2 y RMSE son:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}}\sum_{i=1}^{n}(yi - \hat{y})^2 \quad (2)$$

donde: $\boldsymbol{y_i}$ es el valor observado o medido en un sitio determinado ***i;*** $\boldsymbol{\hat{y}_i}$ es el valor estimado o inferido de *y*; $\boldsymbol{\bar{y}_i}$ es el valor medio de y; y ***n*** es el número de observaciones.

Validación de los modelos de predicción de variables del suelo

Para la evaluación de la confiabilidad de los modelos de predicción de COS (validación) se utilizaron seis (6) índices: 1) error medio (ME, *Mean Error*), 2) error absoluto medio (MAE, *Mean Absolute Error*), 3) raíz del error cuadrático medio (RMSE), 4) error medio estandarizado (SME), 5) raíz del error cuadrático medio estandarizado (SRMSE), y 6) coeficiente de concordancia (CC). Los índices, ME, MAE y RMSE contribuyen al análisis de los resultados indicando el error en los valores de la propiedad del suelo de interés (Hengl *et al*., 2004). El ME evalúa el error sistemático e indica la presencia de subestimación (-) o sobrestimación del modelo (+). El MAE expresa el tamaño del error producido por el modelo en comparación con el valor real (Chicco *et al.,* 2021). RMSE evalúa la precisión de la predicción y mide la cantidad de error que hay entre los conjuntos de datos medidos y estimados. El índice RMSE y ME fueron estandarizados con la desviación estándar de los residuos de los datos. El SME indica la variabilidad de las predicciones, cuyas estimaciones serán más adecuadas si sus valores están más cerca de cero. El índice SRMSE es más preciso mientras más se aproxime al valor ideal de la unidad (1), si el error estandarizado de la raíz cuadrada media es mayor que 1, existe subestimación de la variabilidad en sus predicciones, y si es menor que 1, se sobreestima la variabilidad en sus predicciones. El índice CC refleja el grado en el cual las observaciones son estimadas en forma correcta por el modelo. No es una medida de correlación como tal, sino una medida del grado en el cual las predicciones del modelo están libres de errores. Este índice CC se emplea como una medida estandarizada del grado de error de predicción del modelo y varía entre 0 y 1. Un valor calculado de 1 indica una concordancia perfecta entre los valores medidos y los predichos, y 0 indica que no hay concordancia en absoluto (Willmott *et al*., 2012). Las ecuaciones complementarias de ME, MAE y CC son:

$$ME = \frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i) \quad (3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|yi - \hat{y}i| \quad (4)$$

$$CC = 1 - \frac{n.(RMSE)^2}{PE} \quad (5)$$

$$PE = \sum_{j=1}^{n}(|\hat{y}i - \bar{y}| + |yi - \overline{y}|)^2 \quad (6)$$

donde ***n*** es el número de observaciones y ***PE*** es el error potencial de la varianza; **$\bar{y}$** es el valor medio observado; ***$\hat{y}_i$*** y ***y_i*** son los valores estimados y medidos, respectivamente.

Estimación de la reserva de carbono orgánico del suelo

La reserva de COS expresada en t ha^{-1} se calculó utilizando la siguiente fórmula (Penman *et al.*, 2003); FAO, 2017):

$$\textbf{\textit{COS}}\ (t\ ha^{-1}) = CO \times Dap \times p \times (1 - \%EG) \times 10 \quad (7)$$

donde ***CO*** es el contenido de CO en gkg $^{-1}$, ***Dap*** es la densidad aparente del suelo en t m^{-3}, ***p*** es el espesor o profundidad del suelo en metros (m), y ***EG*** es la corrección por presencia de material grueso en los suelos (> 2 mm) expresado como fracción decimal, y el factor 10 se utiliza para la conversión de unidades. En la estimación del COS se emplearon los mapas raster de 15 m de resolución espacial de cada una de las variables edafológicas de la ecuación 7, empleando la calculadora *raster* de la herramienta SIG ArcGIS Pro®.

RESULTADOS Y DISCUSIÓN

Estadísticos descriptivos

Los resultados de los estadísticos descriptivos de las propiedades edáficas empleadas en la generación y validación de los modelos de predicción, se presentan en el Cuadro 1.

Cuadro 1. Estadísticos descriptivos de las variables edáficas empleadas en la generación y validación de los modelos de predicción.

Variable	N	Min	Max	Media	Mediana	S	Varianza	CV (%)
CO (%)	130	0,59	3,65	1,77	1,71	0,41	0,24	23,2
Esp A (cm)	130	5,00	38,0	16,1	14,0	7,30	53,6	45,6
Dap (g cm^{-3})	100	1,11	1,70	1,36	1,36	0,12	0,002	9,1
EG (%)	130	0,00	73,0	26,9	22,4	19,3	373,0	71,7

S: Desviación estándar, CV: Coeficiente de variación, CO: Carbono orgánico, Esp A: Espesor A, EG: Esqueleto grueso, Dap: Densidad aparente.

En el conjunto de datos edáficos destacan los siguientes aspectos: a) las altas variaciones del espesor del contenido de fragmentos gruesos en la superficie; b) las moderadas variaciones del espesor del horizonte superficial de los suelos, y c) la baja variabilidad del contenido de carbono orgánico y la densidad aparente de los suelos, con CV del 23,2% y 9,1% respectivamente.

La correlación lineal entre las propiedades edáficas y las variables

auxiliares derivadas de MDE e imágenes satelitales se presenta en el Cuadro 2, en el cual se destacaron las siguientes tendencias entre propiedades edáficas y variables ambientales: a) el espesor del horizonte superficial expresa una ligera correlación positiva con el índice de humedad y el área de captación, y correlación negativa con la pendiente del terreno; b) el contenido de esqueleto grueso refleja una ligera correspondencia positiva con la altitud, la pendiente del terreno y la precipitación; y la densidad aparente correlaciona negativamente con el índice topográfico de humedad y la orientación de la pendiente; y c) el carbono orgánico está relacionado directamente con la altitud, la precipitación y el índice de vegetación.

Cuadro 2. Coeficientes de correlación lineal entre las variables edáficas y las variables auxiliares.

Variable Edáfica	Variables Auxiliares									
	Alt	**Pend**	**Asp**	**ITH**	**Area C**	**Perfil C**	**Plano C**	**PR**	**NDVI**	**PP**
%CO	0,13	0,03	0,12	0,02	0,01	-0,07	0,05	-0,11	**0,23	0,13
Esp A	0,02	*-0,19	-0,04	**0,29	**0,24	-0,08	0,09	-0,01	0,05	0,04
Dap (g cm^{-3})	0,05	*0,17	**-0,23	**-0,23	**-0,20	*0,17	-0,05	0,10	*0,17	-0,00
%EG	**0,29	**0,23	-0,15	**-0,28	*-0,21	0,02	-0,12	-0,04	-0,06	**0,36

**La correlación es significativa al nivel 0,01. *La correlación es significante al nivel 0,05.

Alt: Altitud, Pend: Pendiente, Asp: Orientación, Area_C: Área de captación, Perfil_C: Perfil de curvatura, Plano_C: Plano de curvatura, PR: Posición relativa, PP: Precipitación.

Las correlaciones observadas entre los dos conjuntos de variables no son muy elevadas, aunque presentan niveles significativos estadísticamente, sin embargo, todos los valores del estadístico lineal son inferiores a 0,36 y en promedio no superan al valor de r= 0,20. Esta situación, donde las relaciones lineales son débiles a pesar de la significancia estadística, es un indicativo de la alta complejidad de las relaciones suelo-ambiente en paisajes montañosos, sugiriendo que estas relaciones no son explícitamente lineales. Esta característica del conjunto de datos justifica la elección de un método de aprendizaje automático como *Random Forest*, que es capaz de modelar relaciones complejas y no lineales, a diferencia de los modelos lineales tradicionales que probablemente tendrían un rendimiento predictivo insatisfactorio en este entorno (Valera, 2018).

Entrenamiento y generación de modelos de predicción de propiedades del suelo con técnicas RF

Los resultados de las evaluaciones para la obtención de los modelos de predicción de las propiedades de los suelos con entrenamiento del 75% de los datos, permitió la obtención de la combinación de variables predictoras más apropiadas para el ajuste de los parámetros del

modelo. Las características generales para la generación de los modelos de predicción de las propiedades del suelo se indican en el Cuadro 3. El entrenamiento se alcanzó con 500 árboles a una profundidad variable que osciló entre 7 y 20 árboles, con 4 variables aleatorias para el proceso de generación de los modelos de las variables %CO, Esp A, Dap y %EG.

Cuadro 3. Características de los datos de entrenamiento de los modelos de predicción

Criterios	Valor
Número de árboles	500
Tamaño de hoja	1
Profundidad de los árboles	7-20
Profundidad media del árbol	13
Cantidad de datos de entrenamiento disponibles por árbol (%)	100
Numero de variables auxiliares	12
Número de variables muestreadas aleatoriamente	4
Cantidad de datos utilizados para el entrenamiento (%)	75
Cantidad de datos utilizados para validación (%)	25

Los modelos de las variables consideradas (%CO, Esp A, Dap, %EG) se generaron con las covariables más importantes para cada una, las cuales se señalan en la figura 3. Para el contenido de carbono orgánico, las variables más importantes fueron el índice de vegetación, el área de captación y el plano de curvatura, lo cual significa que las zonas con presencia de vegetación boscosa, un área de captación y una curvatura horizontal tendrá contenidos de CO distintas, capaces de explicar la variación espacial de dicha propiedad.

En cuanto al espesor del horizonte superficial del suelo, las variables auxiliares de mayor importancia son la orientación de las laderas, la altitud, la pendiente y la precipitación, aunado a la curvatura del terreno y el índice de vegetación. La variación de Esp A está relacionada con las características de relieve, la forma del terreno y la dominancia o poca existencia de vegetación, y expresa valores que obedecen a la ocurrencia de procesos de erosión de materiales sólidos y sedimentos, y a movimientos en masa, característicos de la zona evaluada. Este impacto se acelera por las pendientes existentes y se activa con el factor precipitación, generando desde suelos muy delgados en las partes altas de las laderas, a suelos moderadamente profundos en las partes bajas (pie de ladera, terrazas de valles intramontanos).

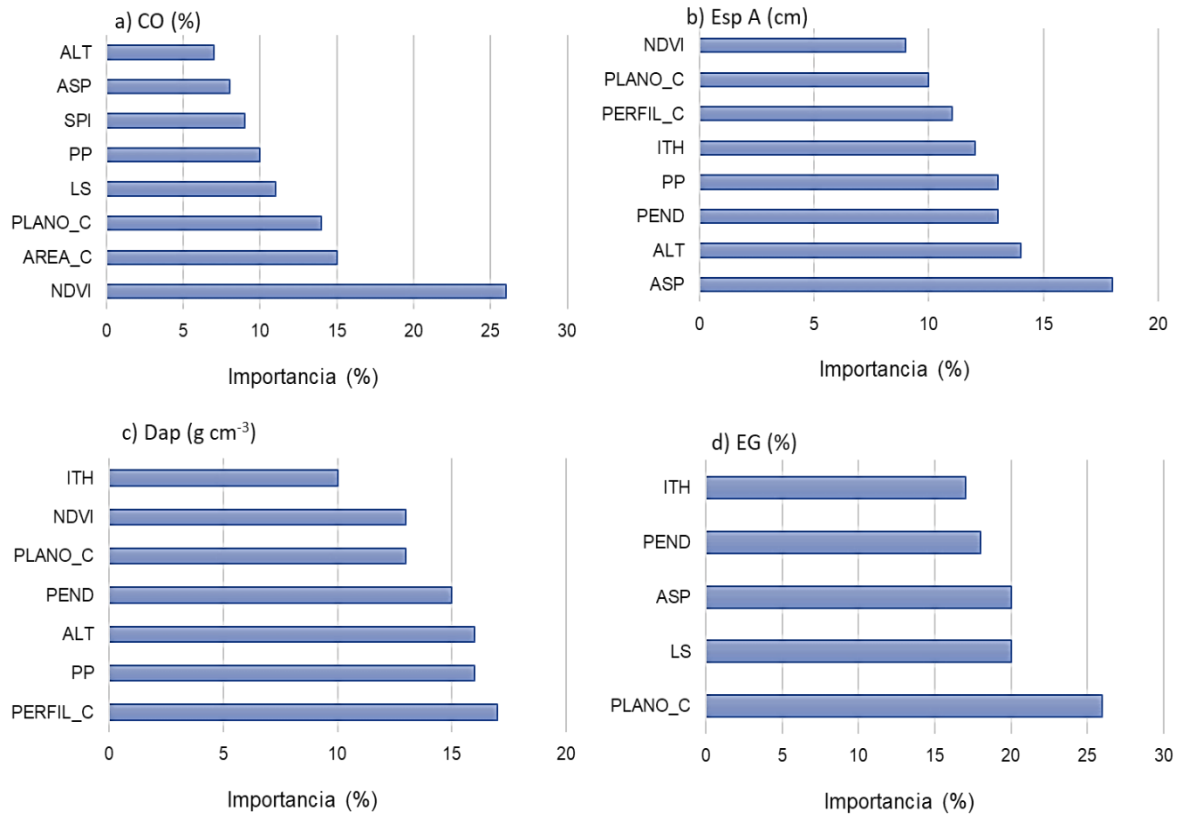

Figura 3. Importancia de las variables empleadas en el entrenamiento y generación de los modelos de predicción de propiedades del suelo.

Esas variables auxiliares también son consideradas de gran importancia en la generación del modelo de densidad aparente del suelo, pero con un mayor peso dado por la curvatura vertical del terreno y la altitud, con influencia de la precipitación. También la variable EG resultó influenciada por la curvatura horizontal del terreno y el factor LS, relacionado con el transporte de sedimentos.

Evaluación de los modelos de predicción de variables edáficas

La evaluación de la bondad de ajuste de los modelos estimados de las propiedades del suelo generó coeficientes de determinación que explican más del 93% de la variabilidad de los suelos. Los resultados del entrenamiento se resumen en el Cuadro 4, donde también se puede observar el bajo nivel de incertidumbre obtenido con los índices RMSE y el error estándar de los datos. Los mayores errores están dados por las variables Esp A y %EG, los cuales presentaron una mayor varianza y los más altos valores del coeficiente de variación.

Estos altos valores de R^2 durante la fase de entrenamiento indican que los modelos RF fueron altamente efectivos para explicar la variabilidad observada en los datos de calibración para todas las

propiedades del suelo consideradas. Un R^2 superior a 0,90 para todas las variables es un indicador de un ajuste muy fuerte del modelo a los datos de entrenamiento, lo que demuestra la capacidad de RF para capturar las complejas interacciones entre las propiedades edáficas y las covariables ambientales en el área de estudio. Este rendimiento superior en la fase de entrenamiento es un prerrequisito para la capacidad de generalización del modelo a datos no vistos, que se evalúa en la fase de validación.

La figura 4 corrobora las aseveraciones anteriores, mediante la representación de los valores estimados en función de los valores observados.

Cuadro 4. Indicadores de la bondad de ajuste de los modelos de predicción de variables del suelo con la técnica RF.

Índice	Variable del suelo			
	CO (%)	Esp A (cm)	Dap (g cm^{-3})	EG (%)
R^2	0,963	0,948	0,932	0,946
Valor-*p*	0,000	0,000	0,000	0,000
SE	0,013	0,017	0,021	0,016
RMSE	0,181	2,734	0,050	6,430

R^2: Coeficiente de determinación, SE: Error estándar, RMSE: Raíz del error cuadrático medio, Valor-*p*: nivel de significación (p<0,05)

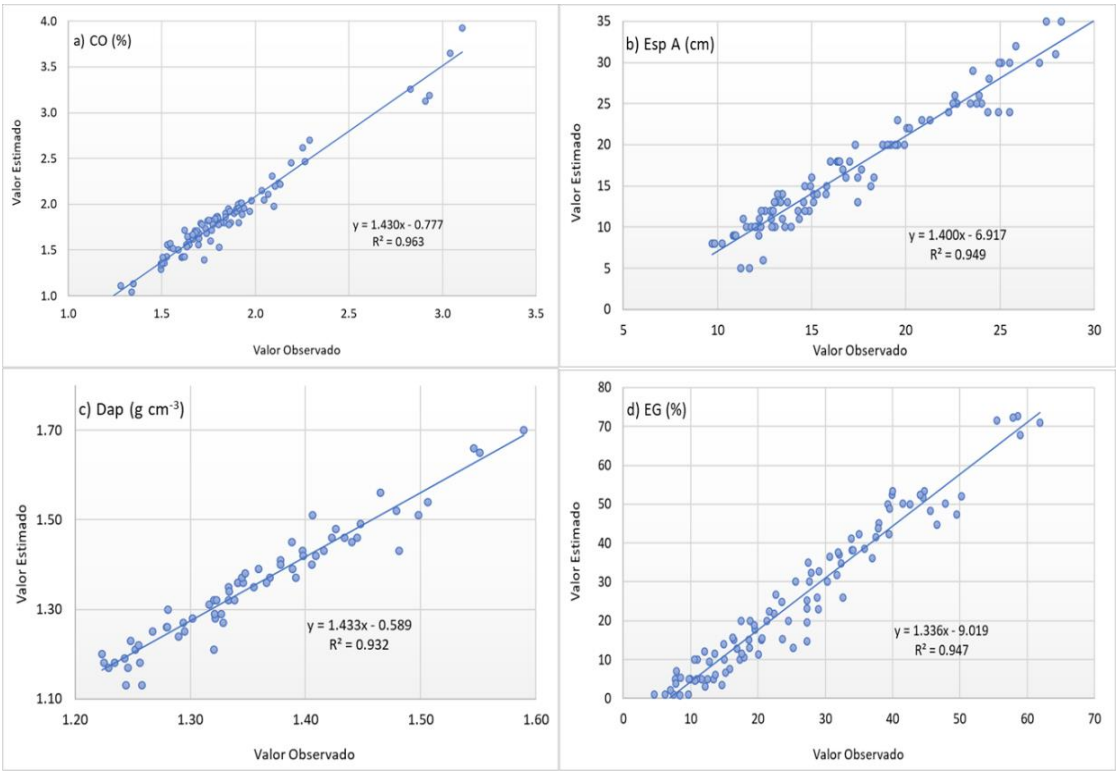

Figura 4. Calibración de los modelos de predicción espacial de variables del suelo.

Estos resultados indican la potencialidad del método RF, mostrando altos coeficientes de determinación y errores muy cercanos a cero, capaz de considerar la variabilidad presente en los suelos del sector evaluado dentro de la cuenca del río Caramacate.

Modelos de predicción espacial de propiedades del suelo con las técnicas RF

La aplicación de las técnicas de *Random Forest* permitió la generación de mapas de distribución espacial para el %CO, Esp A, Da y %EG en la cuenca del río Caramacate. Los resultados de los modelos de las propiedades del suelo en formato ráster se presentan en la figura 5. Los mapas de %CO y Esp A mostraron valores más altos en las zonas con mayor cobertura boscosa y menor alteración, como las áreas al Nor-este de la cuenca, donde la altitud y la precipitación son mayores, favoreciendo la acumulación de materia orgánica y el desarrollo de horizontes superficiales de mayor espesor. Por el contrario, en las zonas con problemas de erosión y uso intensivo para ganadería extensiva, especialmente en las laderas de alta pendiente en la zona central y occidental de la cuenca, se visualiza que el %CO y el Esp A son menores, reflejando la pérdida de suelo superficial y materia orgánica.

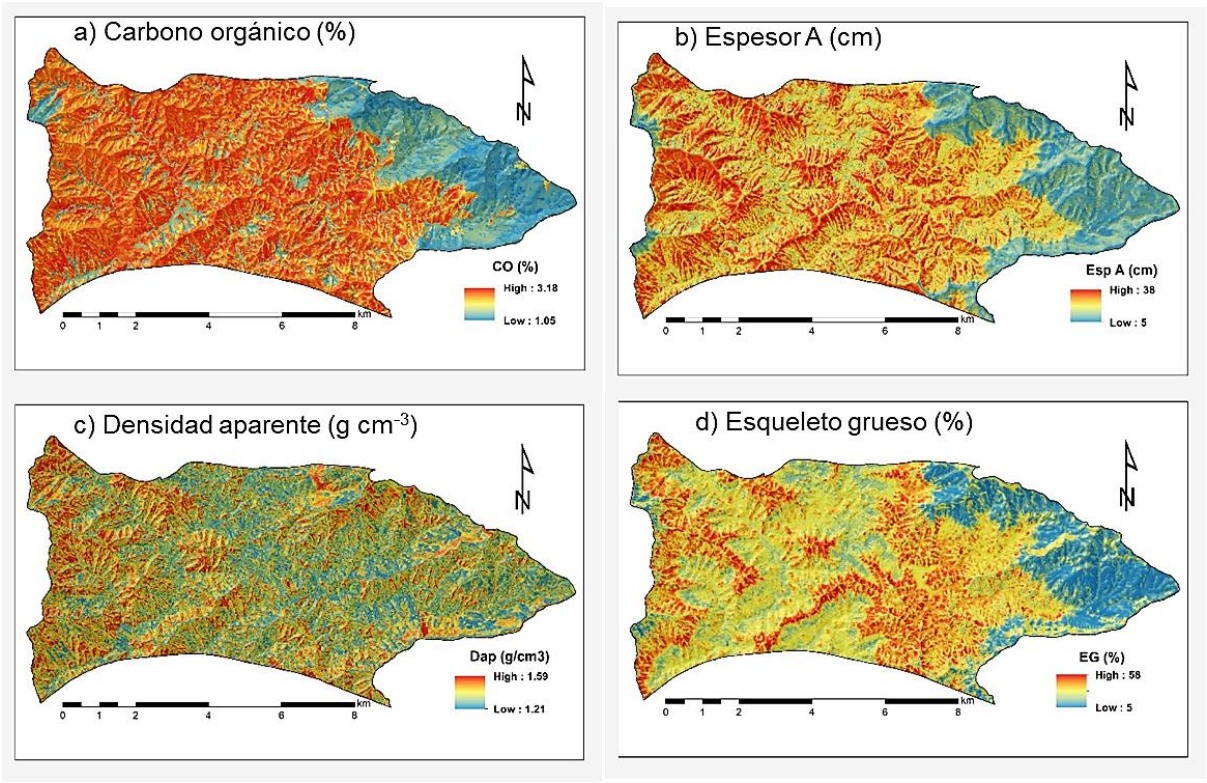

Figura 5. Modelos de predicción de variables edáficas relacionadas con la estimación del carbono orgánico en la capa superficial del suelo.

En cuanto a la densidad aparente (Dap) y el contenido de esqueleto grueso (%EG), los mapas reflejan la influencia del material parental y los procesos de

degradación, ya que las zonas con mayor %EG corresponden a áreas con horizontes más delgados, donde la erosión ha expuesto el material parental, mientras que la Dap varía en función de la compactación del suelo generada por el uso de la tierra, y el bajo contenido de materia orgánica. En tal sentido, la capacidad de RF para generar estas representaciones continuas y detalladas de las propiedades del suelo fue fundamental para comprender la heterogeneidad espacial, facilitando la interpretabilidad del modelo de clasificación resultante.

Validación de los modelos de predicción RF

Los índices de evaluación de la exactitud de los modelos RF obtenidos con el conjunto de datos independientes para la validación se señalan en el Cuadro 5. Tanto MAE como ME y RMSE, reflejan valores indicativos de ciertos errores a nivel de las propiedades edáficas, lo cual evidencia una relación directa entre el conjunto de datos empleado en la calibración de los modelos y los datos utilizados para la validación.

Cuadro 5. Evaluación de la exactitud de los modelos de predicción de variables del suelo con RF en el área estudiada.

Índice	Variable del suelo*			
	CO (%)	Esp A (cm)	Dap (g cm^{-3})	EG (%)
ME	0,006	1,697	-0,015	1,441
MAE	0,091	2,487	0,073	8,535
RMSE	0,114	2,886	0,089	10,55
SME	0,049	0,715	-0,167	0,136
SRMSE	0,999	0,991	0,997	0,992
CC	0,813	0,593	0,665	0,855

*Validación con datos independientes: 30 perfiles de suelo. RMSE: Raíz del error cuadrático medio, SRMSE: Raíz del error cuadrático medio estandarizado, ME: Error medio, SME: Error medio estandarizado, MAE: Error medio absoluto, CC: Coeficiente de concordancia.

En el caso del MAE y el ME, la mayoría de los valores obtenidos son cercanos a cero lo que indica una predicción imparcial no sesgada o con muy poco sesgo, aunque la excepción está dada por contenido de fragmentos gruesos (EG), donde el error sistemático es mayor, debido posiblemente a la influencia de procesos de arrastre y pérdida de parte de los suelos en los relieves de ladera de alta pendiente de los paisajes montañosos. Los valores de RMSE son ligeramente pequeños, lo cual sugiere que la evaluación de la predicción de las propiedades del suelo en la zona de interés es confiable, aun con ciertas diferencias generadas por el %EG en diversas zonas de la cuenca del río Caramacate. Además, los valores de la SRMSE en todas las variables edáficas son cercanos a uno, lo que indica que la varianza del error de predicción es una

evaluación realista de la precisión observada.

Los coeficientes de concordancia (CC) obtenidos para las variables fueron: 0,813 para %CO, 0,593 para Esp A, 0,665 para Dap y 0,855 para %EG. Estos valores indican una consistencia de moderada a alta de los modelos base asociados con la estimación del COS. El CC para %CO y %EG es particularmente alto (superior a 0,81), lo que sugiere que el modelo RF es muy preciso en la predicción de estas propiedades. Este rendimiento es consistente con la baja variabilidad observada para %CO en los estadísticos descriptivos (CV del 23,2%), lo que facilita su predicción. Sin embargo, el CC para Esp A (0,593) y Dap (0,665) es moderadamente menor.

Esta diferencia en el rendimiento predictivo puede atribuirse a la mayor variabilidad inherente de estas propiedades, especialmente el Esp A (CV del 45,6%) y el %EG (CV del 71,7%). Esta alta variabilidad representa un desafío persistente para la cartografía, incluso con algoritmos avanzados como RF, ya que la heterogeneidad a microescala puede no ser completamente capturada por las covariables ambientales a 15m de resolución.

La confiabilidad de las predicciones también puede verse influenciada por factores como el material parental, la topografía y la susceptibilidad a movimientos en masa, como se observó en estudios previos en la misma cuenca (Pineda, 2008; Valera, 2015). Esto sugiere que las características geológicas aunado a las altas pendientes del terreno y los procesos erosivos pueden introducir una incertidumbre adicional en la predicción, modulando la capacidad predictiva del modelo RF. A pesar de estas variaciones, la capacidad de RF para generar modelos con una consistencia general de moderada a alta en un ambiente tan complejo es un avance significativo para la cartografía digital de suelos. Pero en general, el grado de acuerdos entre los valores medidos y estimados para un conjunto de datos independientes es superior al 59% de los casos, con un promedio global de 73% para las variables consideradas.

Esto significa que, mediante la información de los datos edáficos y auxiliares obtenidos con la aplicación del algoritmo RF se puede obtener una estimación aceptable en la zona de estudio, aun utilizando un conjunto de datos limitados. Estos resultados demostraron ser superiores a los obtenidos en la zona con la aplicación de métodos de regresión lineal múltiple, regresión *kriging*, *Fuzzy c-means* (FCM) y FKCN (*Fuzzy Kohonen Clustering Network*) en estudios previos (Valera, 2018).

Modelo digital de la reserva de carbono orgánico del suelo

La aplicación de los modelos RF para predecir el %CO, Esp A, Dap y %EG, y su posterior integración mediante la fórmula de Penman ***et al.*** (2003) y FAO (2017), permitió la estimación de la reserva de COS en la capa superficial de los suelos de la cuenca del río Caramacate. Los resultados indicaron que la reserva de COS en el área de estudio varía significativamente, con valores que

oscilan entre 6,5 y 130 t ha^{-1} (Figura 5).

Esta amplia variación espacial en la reserva de COS refleja directamente la heterogeneidad del paisaje y la influencia de los factores formadores del suelo y los procesos de degradación. Las zonas con problemas de erosión, típicamente asociadas con laderas de alta pendiente y uso de ganadería extensiva, muestran las reservas más bajas de COS. Esto se debe a la pérdida de la capa superficial del suelo, rica en materia orgánica, por la acción de la erosión hídrica y los movimientos en masa. En contraste, las zonas con cobertura boscosa, especialmente en las partes más altas y menos perturbadas de la cuenca, exhiben las reservas más elevadas de COS. La presencia de bosques favorece la acumulación de materia orgánica a través de la biomasa vegetal y la reducción de la erosión, lo que permite una mayor estabilidad del carbono en el suelo.

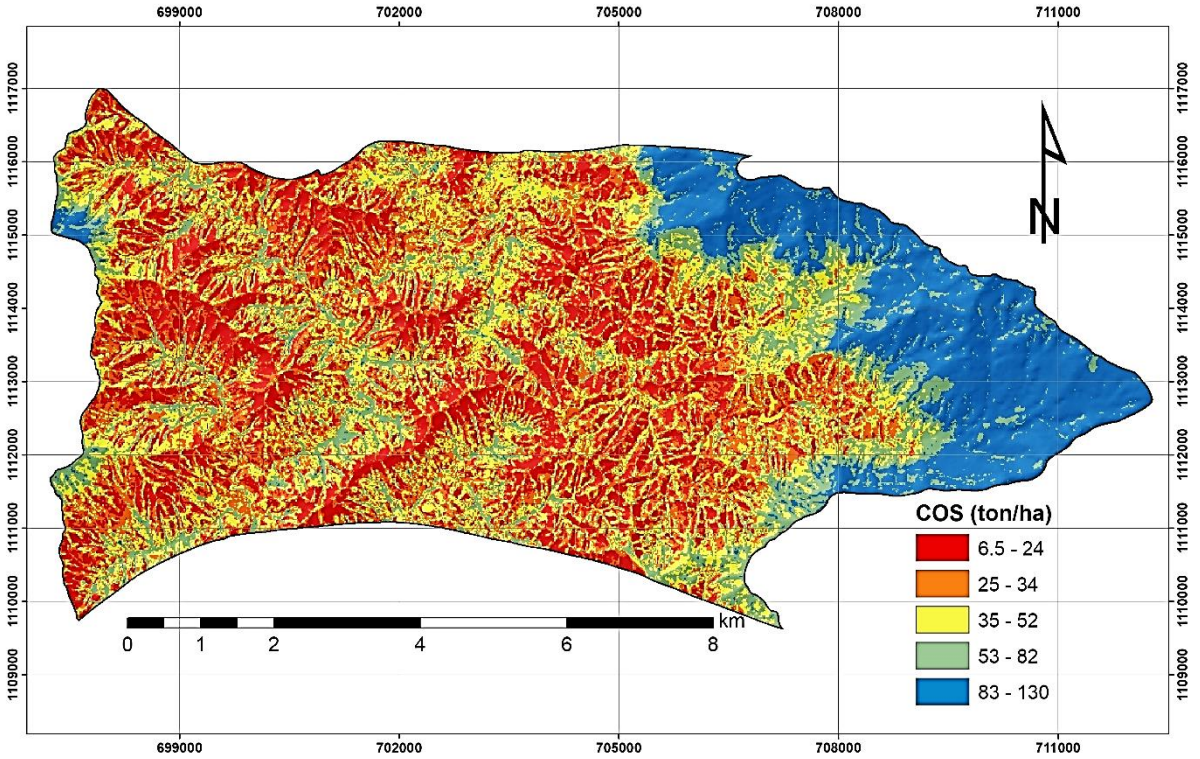

Figura 7. Modelo digital de la reserva de carbono orgánico del suelo en paisajes de la Cuenca del rio Caramacate.

El modelo digital de la reserva de COS ha sido capaz de proporcionar una representación espacial detallada que permite identificar áreas críticas con baja acumulación de carbono, así como zonas con alto potencial de secuestro. Esta información es crucial para la planificación del uso de la tierra y la implementación de estrategias de manejo sostenible, como la reforestación, medidas de conservación y el uso de materiales orgánicos, que puedan contribuir a aumentar el COS y mejorar la resiliencia de los ecosistemas

(Thompson y Kolka, 2005).

Los resultados obtenidos expresaron una correspondencia lógica con otras investigaciones que reflejaron que el área con mayor elevación y cubierta forestal contiene valores más altos de existencias de COS, donde también ocurre la mayor precipitación (Chen et al., 2021). Además, de forma similar se ha encontrado siempre un mayor contenido de materia orgánica en el suelo a mayor altitud que a menor altitud (Chen *et al.,* 2022). Los resultados de esta investigación también se corresponden con los estudios de Ramesh *et al*. (2019), quienes mostraron que, aunque los efectos del uso de la tierra sobre el COS son complejos, una mayor perturbación del suelo conduce a una mayor pérdida de COS y a una menor acumulación. Además, los bosques con mayor cobertura de copas y menor perturbación son capaces de proporcionar mejores condiciones para conservar el suelo y promover una mayor acumulación de materia orgánica en la capa superficial (Xiao *et al*., 2015), como ha sucedido en la cuenca del rio Caramacate. El modelo obtenido enfatiza la importancia de diversos factores donde destacan el clima, el relieve y la cobertura vegetal (Khanal *et al.,* 2023).

Un aporte final de la importancia de la presente investigación está referida al contraste con algunos estudios relacionados con el COS, a nivel nacional, en América Latina y a nivel global. La aplicación de la técnica de Bosques Aleatorios para la estimación del COS en la cuenca del río Caramacate representa una contribución vital para abordar la deficiencia crónica de información edáfica detallada en Venezuela. El país ha experimentado una pérdida significativa de ecosistemas naturales (más de 4 millones de hectáreas en casi cuatro décadas) y enfrenta riesgos de desertificación, especialmente en regiones transformadas para usos agrícolas (MapBiomas Venezuela, 2024). La cuenca del río Caramacate, con sus intensos procesos de degradación ambiental y escasez de datos, es un ejemplo paradigmático de esta problemática. Los resultados de este estudio no solo llenan un vacío de conocimiento local, sino que también ofrecen una metodología replicable para otras cuencas montañosas venezolanas que enfrentan desafíos similares de degradación y falta de información, lo que es crucial para la planificación y gestión de los recursos naturales a nivel local y nacional.

El enfoque de este estudio en el carbono orgánico del suelo en paisajes montañosos degradados por la actividad agrícola y los movimientos en masa, complementa los esfuerzos regionales en América Latina. Mientras que algunos países de la región, como Colombia, han centrado sus esfuerzos en la protección de grandes sumideros de carbono forestales en la Amazonía y la mitigación de la deforestación (Green Climate Fund., 2023), o como Costa Rica, que ha implementado programas de Pago por Servicios Ambientales (PSA) para la conservación forestal y el aumento de las reservas de carbono (MINAE, 2017). Este

estudio se enfoca en la dinámica del COS a nivel de suelo en un contexto de montaña con uso agropecuario extensivo.

La variabilidad espacial del COS en paisajes de montaña, influenciada por la topografía, el uso de la tierra y la degradación, es una característica común en muchas regiones andinas y centroamericanas (CCAD, 2010). Este estudio subraya la necesidad de soluciones adaptadas a las particularidades de cada región, reconociendo que la gestión del COS en suelos agrícolas de montaña, a menudo degradados, es un componente distinto y crucial de la mitigación del cambio climático y la gestión sostenible de la tierra que no siempre recibe la misma atención que los sumideros forestales. La metodología aplicada en Caramacate puede servir de modelo para la evaluación detallada del COS en otros paisajes agrícolas de montaña en América Latina, contribuyendo a una comprensión más holística de los balances de carbono a nivel regional.

A nivel mundial, el algoritmo *Random Forest* ha sido ampliamente reconocido como una de las técnicas de aprendizaje automático más exitosas para la cartografía digital de suelos, superando a menudo a los modelos lineales tradicionales. La precisión promedio de predicción de RF en estudios globales puede superar el 85%. (Sharma *et al.,* 2025). Los coeficientes de determinación (R^2) obtenidos en este estudio para el entrenamiento de los modelos RF (%CO: 0,963; Esp A: 0,948; Dap: 0,932; %EG: 0,946) son excepcionalmente altos y se encuentran entre los más elevados reportados en la literatura para la predicción de propiedades del suelo. Por ejemplo, estudios recientes que utilizan modelos RF con zonificación climática para estimar la densidad de COS (0-20 cm) en China reportaron un R^2 de 0,55 (Dong *et al.,* 2025), lo que resalta el rendimiento superior del modelo en este estudio.

A pesar de los desafíos globales en la cartografía del COS, como la necesidad de armonizar datos *in-situ* y la disponibilidad de imágenes de alta resolución (FAO, 2017), la metodología de este estudio demostró que RF puede lograr una precisión excepcional incluso en entornos montañosos complejos con datos de muestreo limitados. Aunque existen avances en otras técnicas de *machine learning*, como el aprendizaje por transferencia con redes neuronales convolucionales (CNN) que pueden mejorar la precisión en escenarios con datos limitados (Han *et al.,* 2025), la robustez y eficacia de RF, como se evidencia en este trabajo, mantienen su relevancia. Este estudio se alinea con el imperativo global de mejorar la precisión de los datos de COS para desarrollar estrategias de adaptación y mitigación del cambio climático (FAO, 2016), posicionándose como un caso de estudio relevante para regiones con características similares en todo el mundo.

CONCLUSIONES

La aplicación del algoritmo de aprendizaje automático *Random Forest* demostró ser una técnica robusta y eficaz para la generación de modelos de predicción espacial de propiedades clave

del suelo, como el porcentaje de carbono orgánico (%CO), el espesor del horizonte A (Esp A), la densidad aparente (Dap) y el contenido de esqueleto grueso (%EG), en el complejo paisaje montañoso de la cuenca del río Caramacate. Los altos coeficientes de determinación (R^2) obtenidos durante la fase de entrenamiento (superiores a 0,93 para todas las variables) y los satisfactorios coeficientes de concordancia (CC) en la validación (0,813 para %CO, 0,593 para Esp A, 0,665 para Dap y 0,855 para %EG) demuestraron la fuerte capacidad predictiva del modelo en este entorno.

El estudio ha logrado estimar con éxito la distribución espacial de las reservas de COS en la capa superficial de los suelos, revelando una significativa variabilidad en el paisaje, con valores que oscilan entre 6,5 y 130 t ha^{-1}. Esta variación está directamente relacionada con los procesos de degradación del suelo, el uso de la tierra y la cobertura vegetal, mostrando menores reservas en áreas erosionadas y mayores en zonas boscosas. El rendimiento del modelo, aunque generalmente alto, mostró una consistencia moderadamente menor para variables como el espesor del horizonte superficial y la densidad aparente, lo cual se atribuye a la alta variabilidad de estas propiedades en un ambiente tan dinámico y complejo como el de la cuenca del río Caramacate.

Esta investigación cumple con sus objetivos al proporcionar una alternativa cuantitativa y eficiente para la generación de información de suelos en ambientes montañosos con escasez de datos, contribuyendo a cerrar la brecha de información edáfica en zonas de difícil acceso. Los mapas detallados de COS generados son herramientas valiosas para la planificación del uso sostenible de la tierra, el control de la erosión y el diseño de estrategias de mitigación del cambio climático a nivel de la cuenca estudiada.

Para futuras investigaciones, se sugiere explorar la integración de covariables más dinámicas, como los cambios en el uso de la tierra a lo largo del tiempo, y evaluar la aplicabilidad de modelos de aprendizaje profundo (como las redes neuronales convolucionales) que han mostrado prometedoras mejoras en la precisión, especialmente con datos limitados. Asimismo, la replicación de esta metodología en otras cuencas montañosas venezolanas con características similares podría fortalecer la base de datos de COS a nivel local, regional y nacional.

AGRADECIMIENTO

El autor manifiesta su agradecimiento al Consejo de Desarrollo Científico y Humanístico (CDCH) de la Facultad de Agronomía de la Universidad Central de Venezuela, y al Centro de Investigación y Extensión en Suelos y Aguas de la Universidad Rómulo Gallegos (CIESA UNERG) por el apoyo logístico para el desarrollo de la investigación.

REFERENCIAS

Adhikari, K., Hartemink, A.E., Minasny, B., Bou Kheir, R., Greve, M.B., and Greve, M.H. (2020). Digital mapping of soil organic carbon contents and stocks in Denmark. PLOS ONE, 9(8), e105519.

Biau, G., and Scornet, E. (2016). *A random forest guided tour.* TEST, 25, 197–227. https://doi.org/10.1007/s11749-016-0481-7

Breiman, L. (2001). *Random forests.* Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324

Burrough, P. A., and McDonnell, R.A. (1998). Principles of geographical information systems. Oxford University Press.

CCAD. (2010). Estrategia Regional de Cambio Climático. Comisión Centroamericana de Ambiente y Desarrollo - CCAD Sistema de la Integración Centroamericana – SICA. Noviembre 2010. 93p.

Chen, X., Wang, D., Chen, J., Wang, C., and Shen, M. (2021). Soil organic carbon stocks in forest ecosystems of eastern China: Spatial patterns and environmental controls. Journal of Environmental Management, 289, 112573.

Chen, Y., Feng, X., Fu, B., and Lü, Y. (2022). Elevation-dependent soil organic carbon

Chicco, D., Warrens, M. J., and Jurman, G. (2021). *The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation.* PeerJ Computer Science, 7, e623. https://doi.org/10.7717/peerj-cs.623

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L. and Böhner, J. (2015). System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geoscientific

Dangeti, P. (2017). Statistics for machine learning. Packt Publishing Ltd, 21 jul. 2017 - 442 p.

Dong, Y., Zhang, G., and Lu, J. (2025). Spatial prediction of soil organic carbon density using machine learning and climate zoning in China. Environmental Science and Pollution Research, 32(5), 7890–7901.

FAO. (2017). Carbono orgánico del suelo - el potencial oculto. Editorial FAO. Roma, Italia. 90p. 978-92-5-309681-7. https://openknowledge.fao.org/handle/20.500.14283/i6937es

FAO. (2016). Estado Mundial del Recurso Suelo. Resumen Técnico preparado por el Grupo Técnico Intergubernamental del Suelo. Organización de las Naciones Unidas para la Alimentación y Agricultura y Grupo Técnico Intergubernamental del Suelo, Roma, Italia. ISBN 978-92-5-308960-4

Garosi, Y., Sheklabadi, M., and Pourghasemi, H.R. (2022). Digital mapping of soil organic carbon using ensemble learning models in a semi-arid region of Iran. Catena, 212, 106077.

Green Climate Fund. (2023). Funding Proposal FP203: Heritage Colombia (HECO): Maximizing the Contributions of Sustainably Managed Landscapes in Colombia for Achievement of Climate Goals Colombia | World Wildlife Fund, Inc. (WWF) | Decision B.35/05, 11 April 2023. 186 p.

Han, J., Wu, M., Qi, Y., Li, X., Chen, X., Wang, J., Zhu, J., and Li, Q. (2025). *A soil organic carbon mapping method*

based on transfer learning without the use of exogenous data. Frontiers in Environmental Science, 13. https://doi.org/10.3389/fenvs.2025.1580085

Hengl, T., Heuvelink, G. B., and Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma, 120(1–2), 75–93.

Hoyos-Sanclemente, A., Menjivar-Flores, J. C., y Rueda-Saa, G. (2025). *Soil organic carbon in agricultural soils of an inter-Andean valley in Colombia: Understanding the effects of environmental and geographic variables.* Environmental Monitoring and Assessment, 197, 697. https://doi.org/10.1007/s10661-025-14123-1

Huang, H. (2022). *A review on digital mapping of soil carbon in cropland: Progress, challenge, and prospect.* Environmental Research Letters, 17, 123004. https://doi.org/10.1088/1748-9326/aca41e

Jenny, H. (1941). Factors of Soil Formation: A System of Quantitative Pedology. New York: McGraw-Hill Book Company. 281 p.

Kakhani, N., Gläßle, T., Taghizadeh-Mehrjardi, R., Kebonye, N. M., and Scholten, T. (2023). *Exploring the "individual treatment effects" (ITE) of vegetation with causal inference on soil organic carbon prediction in Germany.* Paper presented at the EGU General Assembly 2023, Vienna, Austria. https://doi.org/10.5194/egusphere-egu23-1083

Khanal, S., Nolan, R.H., Medlyn, B.E., and Boer, M.M. (2023). *Mapping soil organic carbon stocks in Nepal's forests.* Scientific Reports, 13(1). https://doi.org/10.1038/s41598-023-34247-z

Kidemo, K. H., Gashu, D., and Nega, F. (2023). *Spatial variability of soil organic carbon in agricultural landscapes: A review.*Frontiers in Environmental Science, 11, 1205374. https://doi.org/10.3389/fenvs.2023.1205374

Kingsley, J., Isong Isong, A., Ndiye, M. K., Chapman, A. P., Okon, A. E., and Ahado, S. K. (2021). *Soil organic carbon prediction with terrain derivatives using geostatistics and sequential Gaussian simulation.* Journal of the Saudi Society of Agricultural Sciences, 20, 379–389. https://doi.org/10.1016/j.jssas.2021.04.005

Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. Science, 304(5677), 1623–1627.

Lamichhane, S., Kumar, L., and Adhikari, K. (2022). Comparing machine learning and geostatistical approaches for mapping soil organic carbon stocks at different scales. Remote Sensing, 14(3), 598.

Lamichhane, S., Kumar, L., and Wilson, B. (2019). Digital mapping of soil organic carbon at multiple depths using machine learning in Baneh region, Iran. Geoderma, 352, 64–77.

Loayza, NV, Sevilla, V., Olivera, C., Guevara, M., Olmedo, G., Vargas, R., Oyonarte, C., & Jiménez, W. (2020). Mapeo digital de carbono orgánico en suelos de Ecuador. *Revista Ecosistemas*, *29* (2).

MapBiomas Venezuela (2024). MapBiomas – Colección [2.0] de la Serie Anual de Mapas de Cobertura y Uso del Suelo de Venezuela (1985–2023), consultada el 28 de mayo de 2025 a través del enlace: https://plataforma.venezuela.mapbiomas.org/cobertura

McBratney, A. B., Mendonça Santos, M.L., and Minasny, B. (2003). On digital soil mapping. Geoderma, 117(1–2), 3–52.

MINAE. (2017). Emission Reductions Program to the FCPF Carbon Fund Costa Rica July 17th, 2017 Government of Costa Rica Ministry of Environment and Energy. 174P.

Molla, A., Zhang, W., Zuo, S., Ren, Y., and Han, J. (2022). *A machine learning and geostatistical hybrid method to improve spatial prediction accuracy of soil potentially toxic elements.* Research Square. https://doi.org/10.21203/rs.3.rs-1306764/v1

Moore, I D., Grayson, R.B., and Ladson, A.R. (1991). Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. Hydrological Processes, 5(1), 3–30.

Moore, I.D., Gessler, P.E., Nielsen, G.A., and Peterson, G.A. (1993). Soil attribute prediction using terrain analysis. Soil Science Society of America Journal, 57(2), 443–452.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., and Papritz, A. (2018). Evaluation of digital soil mapping approaches with large sets of environmental covariates. Soil, 4(1), 1–22.

Odebiri, O., Mutanga, O., Odindi, J., Peerbhay, K., Dovey, S., and Ismail, R. (2020). *Estimating soil organic carbon stocks under commercial forestry using topo-climate variables in KwaZulu-Natal, South Africa.* South African Journal of Science, 116(3–4), 1–8.

Olson, R.S., La Cava, W., Orzechowski, P., Urbanowicz, R.J., and Moore, J.H. (2017). PMLB: A large benchmark suite for machine learning evaluation and comparison. BioData Mining, 10(1), 36.

Penman, J., Gytarsky, M., Hiraishi, T., Krug, T., Kruger, D., Pipatti, R. and Wagner, F. (2003). Good practice guidance for land use, land-use change and forestry. IPCC.

Pineda, M.C. (2008). Determinación de áreas susceptibles a movimientos en masa y erosión laminar en unidades de paisaje de la subcuenca del río Caramacate, localizada en la Cuenca Alta del Río Guárico. Trabajo de Ascenso (Asistente). Facultad de Agronomía, UCV. Maracay. 313 p.

Pineda, M.C.; Elizalde, G. y Viloria, J. (2011). Determinación de áreas susceptibles a deslizamientos en un sector de la Cordillera de la Costa Central de Venezuela. Interciencia 36(5): 370-377.

Pouladi, N., Gholizadeh, A., Khosravi, V., and Borůvka, L. (2023). *Digital mapping of soil organic carbon using remote*

sensing data: A systematic review. Catena, 232, 107409. https://doi.org/10.1016/j.catena.2023.107409

Ramesh, T., Bolan, N.S., Kirkham, M.B., Wijesekara, H., Kanchikerimath, M., and Srinivasa Rao, C. (2019). Soil organic carbon dynamics: Impact of land use changes and management practices: A review. Advances in Agronomy, 156, 1–107.

Rouse, J.W. (1974). Monitoring the vernal advancement of retrogradation of natural vegetation. NASA/GSFC Final Report, 371.

Sharma, A., Liu, X., and Yang, X. (2025). Global assessment of random forest performance for digital soil mapping. Geoderma, 430, 116348.

Tarboton, D.G., Bras, R.L., and Rodriguez-Iturbe, I. (1991). On the extraction of channel networks from digital elevation data. Hydrological Processes, 5(1), 81–100.

Thompson, J.A., and Kolka, R.K. (2005). Soil carbon storage estimation in a forested watershed using quantitative soil-landscape modeling. Soil Science Society of America Journal, 69(4), 1086–1093.

Urbani, F.; J. A. Rodríguez. (2004). Atlas geológico de la Cordillera de la Costa, Venezuela. Mapas a escala 1:25.000. Versión Digital. Edic. Fundación Geos, UCV. Caracas.

USDA-NRCS (1995). Soil survey laboratory information manual. United States Department of Agriculture-Natural Resources Conservation Service. USDA.

Valera, A. (2015). Inventario de suelos y paisajes con apoyo de técnicas de cartografía digital en áreas montañosas. Caso Cuenca del Río Caramacate, Estado Aragua. Tesis de doctorado en Ciencias del Suelo. Universidad Central de Venezuela. Postgrado en Ciencias del Suelo. Maracay, Estado Aragua, Venezuela. 263 p. https://doi: 10.13140/RG.2.1.1714.3920

Valera, A. (2018). Geomorfometría y Edafometría. Cartografía Digital de Paisajes y Suelos con Técnicas de Inteligencia Artificial. Editorial Académica Española. Mauritius. 317p. ISBN: 978-620-2-12102-6.

Vaysse, K., and Lagacherie, P. (2015). *Regional evaluating digital soil mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France).* Geoderma Regional, 4, 20–30. https://doi.org/10.1016/j.geodrs.2014.11.003

Vela, G., Blanco, J., y Rodríguez, M. (2012). Niveles de carbono orgánico total en el suelo de conservación del Distrito Federal, centro de México. Investigaciones Geográficas, Boletín del Instituto de Geografía, UNAM, 77, 20–35.

Verbrugge, L. (2006). Depth of Soil in the Goss-Gasconade-Rock outcrop complex in Callaway County, Missouri Using the Soil Land Inference Model (SoLIM). A thesis presented to the department of geology and geography in candidacy for the degree of Master of

Science. Northwest Missouri State University. 76 p.

Walkley, A. and Black, A. (1934). An examination of the method for determining soil organic matter and proposed modification of the chromic acid titration method. Soil Science 37: 29-38.

Wadoux, M.J., Minasny, B., and McBratney, A.B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews, Volume 210,103359, ISSN 0012-8252, https://doi.org/10.1016/j.earscirev.2020.103359.

Wang, Y., Zhang, J., and Zhang, Z. (2022). Challenges in mapping soil organic carbon in mountainous regions: A review. Geoderma, 412, 115715.

Willmott, C.J., Robeson, S.M., and Matsuura, K. (2012). *A refined index of model performance.* International Journal of Climatology, 32, 2088–2094. https://doi.org/10.1002/joc.2419

Wilson, J.P., and Gallant, J.C. (2000). Digital Terrain analysis: Principles and applications. New York, NY. John Wiley. 127p.

Xiao, S., Zhang, W., Ye, Y., Zhao, J., and Wang, K. (2015). Soil organic carbon under different forest types in Southern China. Geoderma, 249–250, 129–138.

Zeraatpisheh, M., Ayoubi, S., Jafari, A., and Finke, P. (2023). Digital mapping of soil organic carbon using machine learning: A review. Catena, 232, 107409.

Zevenbergen, L.W., and Thorne, C.R. (1987). Quantitative analysis of land surface topography. Earth Surface Processes and Landforms, 12(1), 47–56.