Iranian Journal of Soil and Water Research
Homepage: http://ijswr.ut.ac.ir
ISNN: 2423-7833

Comparative Assessment of Selected Machine Learning Algorithms for Soil Organic Carbon Prediction and Mapping (Case Study: East of Khuzestan Province)

Farideh Karimi Dehkordi[1] | Saeid Hojati[2]✉ | Asim Biswas[3]

1. Department of Soil Science, Faculty of Agriculture, Shahid Chamran University of Ahvaz, Khuzestan, Iran. E-mail: farideh.karimi444@gmail.com
2. Corresponding Author, Department of Soil Science, Faculty of Agriculture, Shahid Chamran University of Ahvaz, Khuzestan, Iran. E-mail: s.hojati@scu.ac.ir
3. School of Environmental Science, Ontario Agricultural College, University of Guelph, Guelph N1G2W1, Ontario, Canada. Email: biswas@uoguelph.ca

Article Info

Article type: Research Article

Article history:

Received: Aug. 9, 2026

Revised: Aug. 24, 2026

Accepted: Sep. 9, 2026

Published online: Nov. 23, 2026

Keywords:
Random Forest,
XGBoost,
Regression Kriging,
Auxiliary Variables,
Uncertainty Analysis.

ABSTRACT

Given the spatial variability of soil organic carbon (SOC) as a key indicator of soil quality, its precise regional mapping is essential. This study compared the performance of random forest (RF), regression kriging (RK), and extreme gradient boosting (XGBoost) algorithms, and identified the most influential factors controlling SOC distribution in the Karun-3 watershed. Accordingly, 158 composite surface soil samples (0–5 cm) were collected using a conditioned Latin Hypercube Sampling (cLHS) design. After preparation, standard laboratory methods were used to measure several physical and chemical soil properties, including SOC, sand, silt, clay, electrical conductivity (EC), pH, calcite equivalent carbonate (CCE), and bulk density (BD). Additionally, a set of 31 auxiliary variables — comprising topographic, spectral, and climatic indices, as well as soil characteristics —was used to model the spatial variability of SOC. These auxiliary variables were pre-processed by removing those with near zero variance, and those highly intercorrelated ($r \geq 0.8$) with variance inflated factor (VIF) greater than 10, The XGBoost algorithm outperformed the others, demonstrating the highest coefficient of determination ($R2 = 0.628$), the lowest error metrics (RMSE = 0.473 and MAE = 0.316), and the highest ratio of performance to interquartile range (RPIQ = 1.98). BD, CCE, Carbonate Index (Carb.Ind), Normalized Difference Moisture Index (NDMI), and Length-Slope Factor (LS-Factor) contributed most significantly to the distribution of SOC. Analysis of SOC distribution maps indicated that the northern and northwestern parts of the area contain the highest levels, while the central, southern, and southeastern parts have the lowest. Evaluation of the patterns presented in uncertainty maps revealed that the values of predictive maps are less reliable in parts of the northwestern, eastern, and southern study area. Overall, it can be concluded that the simultaneous application of predictive and uncertainty maps is necessary to understand both the predicted values and their reliability.

Cite this article: Karimi Dehkordi, F., Hojati, S., & Biswas, A (2026). Comparative Assessment of Selected Machine Learning Algorithms for Soil Organic Carbon Prediction and Mapping (Case Study: East of Khuzestan Province), *Iranian Journal of Soil and Water Research*, 57 (9), 2133-2159. https://doi.org/10.22059/ijswr.2026.420209.670196

© The Author(s).

Publisher: University of Tehran Press.

DOI: https://doi.org/10.22059/ijswr.2026.420209.670196

EXTENDED ABSTRACT

Introduction

Soil organic carbon (SOC) is a key indicator of soil quality and a major component of the global carbon cycle. Accurate spatial prediction of SOC supports sustainable land management, enhances soil conservation planning, and contributes to climate-change mitigation by identifying areas with high carbon sequestration potential. However, SOC typically exhibits strong spatial heterogeneity, driven by interacting environmental factors such as topography, moisture regime, vegetation cover, and land use. Conventional soil mapping approaches are often labor-intensive, time-consuming, and expensive. This has motivated the rapid development of Digital Soil Mapping (DSM) as a cost-effective and scalable alternative.

The Karun 3 Dam watershed (eastern Khuzestan Province, Iran) provides a challenging environment for SOC mapping due to its complex terrain and diverse land uses, which can create non-linear and multi-factor controls on SOC distribution. In such landscapes, advanced predictive models are often required to capture the complex relationships between SOC and environmental covariates. Therefore, the main objective of this study was to evaluate and compare the predictive performance of three distinct modeling approaches—Random Forest (RF), Regression Kriging (RK), and Extreme Gradient Boosting (XGBoost)—for spatial prediction of SOC in the Karun 3 watershed. We hypothesized that the gradient boosting framework, particularly XGBoost, would outperform the other approaches by more effectively learning non-linear relationships and interactions between SOC and environmental covariates under heterogeneous environmental conditions.

Materials and Methods

The study was conducted in the Karun 3 Dam watershed located in eastern Khuzestan Province, Iran. The watershed is characterized by rugged topography and heterogeneous land-use patterns, which together contribute to high spatial variability in soil properties, including SOC. To ensure the sample set accurately represented the multivariate distribution of key environmental variables, including topography, geology, soil types, and vegetation cover, a Conditioned Latin Hypercube Sampling (cLHS) strategy was implemented. This was done using the TEUI 460 add-in widget in ArCGIS 10.8.2. For designing the proposed sampling points with this plugin, all input layers were converted to raster format with identical pixel sizes. Subsequently, a total of 158 composite surface soil samples were collected from a depth of 0–5 cm.

Soil samples were air-dried, passed through a 2 mm sieve, and prepared for the measurement of physical and chemical properties. These properties included soil organic carbon (SOC) determined by hot digestion oxidation; soil pH measured in a saturated soil paste using a pH meter; electrical conductivity (EC) measured in the saturated paste extract with an EC meter; soil texture determined by the hydrometer method; bulk density (ρb) measured using a cylinder; and calcium carbonate equivalents (CCE) of the samples determined by back-titration. The land use type and geographical location of the sampling sites were recorded using a Global Positioning System (GPS).

A suite of 55 environmental covariates was derived from Landsat 9 imagery and a 10 m digital elevation model (DEM) using ENVI 4.7, ArcGIS 10.8.2, and SAGA GIS 6.4.0 softwares. These covariates included topographic, spectral, climatic, and soil features, all commonly linked to SOC variability. To enhance model stability and eliminate redundant variables, we assessed multicollinearity by considering correlation coefficients among variables ($r \geq 0.8$) and the variance inflation factor (VIF > 10), after first removing variables with near-zero variance.

We then followed a predictive modeling framework and compared three approaches: Random Forest (RF), an ensemble learning method based on recursive partitioning; Extreme Gradient Boosting (XGBoost), a gradient-boosted tree framework designed for high predictive accuracy and efficient learning of complex non-linear patterns; and Regression Kriging (RK), a hybrid approach combining a multiple linear regression (MLR) component with ordinary kriging of regression residuals to incorporate spatial autocorrelation not captured by covariates.

Model performance analysis was evaluated using an independent validation strategy, where 70% of samples (n = 111) were used for model training and 30% (n = 47) for independent validation. Splitting data into training and testing subsets was performed employing a cLHS approach to ensure a more balanced and spatial coverage of both the train and test sample. Predictive accuracy assessment was caried out using the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and ratio of performance to interquartile range (RPIQ) metrics. To further examine model robustness, bootstrap resampling was performed within the caret package in R.

To prepare a map of the spatial distribution of SOC using RF and XGBoost models, the final and optimized model for each method was applied to every pixel of the region using the full set of environmental

variables to produce the SOC prediction map. For developing the SOC distribution map using the RK approach, the final SOC distribution map consisted of two separate steps. In the first step, a multiple linear regression model was fitted to the sampling points and applied to all raster layers of the region to generate the predicted SOC trend map. In the second step, using the optimum variogram fitted to the residuals of the regression model and ordinary kriging, the residual values were estimated and interpolated, then applied to all pixels in the study area. Finally, the final SOC distribution map was produced by the sum of the trend map and the kriged residuals.

To quantify the uncertainty of predictions using RF and XGBoost models, we employed the 90% prediction interval (PI) width To compute this index for tree-based methods, we first calculated, for each pixel, the distribution of predictions obtained from different trees (in the RF model) or from bootstrapped models (in the XGBoost model). The distance between the 5th and 95th percentiles of this distribution was then computed and used as the uncertainty measure. However, uncertainty in the RK method was derived from two separate sources. First, the variance of the trend estimate, attributable to uncertainty in the regression model coefficients, was computed. This value was then added to the variance derived from residual kriging.

Results

Model comparison revealed a clear ranking in predictive performance. XGBoost provided the most accurate SOC predictions, achieving R^2=0.628, RMSE = 0.473, MAE = 0.316, and RPIQ = 1.980. Random Forest showed moderate performance with R^2=0.545, RMSE = 0.509, and RPIQ = 1.841. Regression Kriging exhibited the lowest predictive accuracy (R^2=0.434, RMSE = 0.534), indicating limited added value from residual spatial interpolation under the conditions of this study. The superior performance of XGBoost suggests that gradient boosting is well-suited for SOC prediction in heterogeneous watersheds where complex, non-linear relationships and interactions between environmental covariates and SOC are expected. The weaker performance of RK may reflect limited spatial autocorrelation remaining in the residuals after accounting for covariates, or potential challenges in fitting a stable variogram structure under strong environmental heterogeneity.

Variable-importance analysis highlighted BD, CCE, Carbonate Index (Carb.Ind), Normalized Difference Moisture Index (NDMI), and Lthe ength Slope Factor (Ls-Factor) as the most influential predictors contributing to the spatial distriction of SOC in this study. The importance of BD, CCE, NDMI and LS-Factor underscores the role of soil physical condition and moisture status in regulating SOC patterns within the watershed.

The resulting SOC maps indicated higher SOC concentrations in the northern and northwestern parts of the watershed. In contrast, the central areas showed the lowest prediction uncertainty, suggesting stronger model confidence in this zone. Spatial uncertainty patterns further emphasize the need for careful sampling design and potentially additional covariates or sampling intensification in areas where model uncertainty remains high.

Conclusions

This study demonstrates XGBoost as the most effective approach for regional SOC mapping within the complex Karun 3 watershed, outperforming Random Forest (RF) and Regression Kriging (RK). The practical value of XGBoost lies in its ability to accurately capture the non-linear, multi-factor environmental controls on SOC, thereby improving the reliability of digital soil mapping outputs. The generated SOC prediction and uncertainty maps can assist land managers and decision-makers in identifying high-potential carbon sinks and prioritizing soil conservation efforts. Future research should integrate additional high-resolution covariates and explore deep learning models to further enhance spatial prediction accuracy and reduce uncertainty in complex watersheds.

Funding

The authors would like to thank the Deputy of Research Technology and Social Relationships at Shahid Chamran University of Ahvaz (**SCU.AS1402.365**) for their financial support of this research.

Authorship contribution

First Author: Formal analysis, methodology, software, writing — original draft,

Second Author: Conceptualization, data curation, funding acquisition, writing — review and editing, methodology, validation.

Third Author: Conceptualization, methodology, software, writing-review and editing.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used **editGPT** for grammatical and typographical errors.

Data availability statement

Data available on request from the authors.

Acknowledgements

The authors would like to thank anonymous referees for their constructive comments.

The authors would like to thank anonymous reviewers for their valuable suggestions in manuscript revision.

Ethical considerations

The authors avoided data fabrication, falsification, and plagiarism, and any form of misconduct.

Conflict of interest

The authors declare no conflict of interest.

مجله تحقیقات آب و خاک ایران، دوره ۵۷، شماره ۹ شاپا: ۷۸۳۳-۲۴۲۳
Homepage: http://ijswr.ut.ac.ir

مقایسه برخی مدل‌های یادگیری ماشین در برآورد و نقشه‌برداری رقومی کربن آلی خاک (مطالعه موردی: شرق استان خوزستان)

فریده کریمی دهکردی[۱] | سعید حجتی[۲]✉ | آسیم بیسواس[۳]

۱. گروه علوم خاک، دانشکده کشاورزی، دانشگاه شهید چمران اهواز، خوزستان، ایران. رایانامه: farideh.karimi444@gmail.com
۲. ن نویسنده مسئول، گروه علوم خاک، دانشکده کشاورزی، دانشگاه شهید چمران اهواز، خوزستان، ایران. رایانامه: s.hojati@scu.ac.ir
۳. گروه علوم محیط زیست، دانشگاه گوئلف، آنتاریو، کانادا. رایانامه: biswas@uoguelph.ca

اطلاعات مقاله

نوع مقاله: مقالهٔ پژوهشی

تاریخ دریافت: ۱۴۰۵/۵/۱۸
تاریخ بازنگری: ۱۴۰۵/۶/۲
تاریخ پذیرش: ۱۴۰۵/۶/۱۰
تاریخ انتشار: ۱۴۰۵/۹/۱

واژه‌های کلیدی:
جنگل تصادفی، گرادیان تقویت شده، رگرسیون کریجینگ، متغیرهای کمکی، عدم قطعیت.

چکیده

با توجه به تغییرپذیری مکانی بالای کربن آلی خاک (SOC) به‌عنوان شاخصی کلیدی در ارزیابی کیفیت خاک، پهنه‌بندی دقیق این ویژگی در مقیاس منطقه‌ای ضروری است. هدف این پژوهش مقایسه کارایی مدل‌های جنگل تصادفی (RF)، رگرسیون کریجینگ(RK) و گرادیان تقویت شده(XGBoost) و تعیین مهمترین عومل کنترل کننده SOC در حوضه آبخیز سد کارون ۳ بود. بر این اساس، ۱۵۸ نمونه خاک سطحی مرکب (۰ تا ۵ سانتی‌متر) با روش ابرمکعب لاتین مشروط (cLHS) برداشت شد. پس از آماده‌سازی نمونه‌ها، برخی ویژگی‌های فیزیکی و شیمیایی خاک شامل فراوانی ذرات شن، سیلت و رس، قابلیت هدایت الکتریکی (EC)، pH، کربنات کلسیم معادل (CCE) و جرم مخصوص ظاهری (BD) به روش‌های استاندارد آزمایشگاهی اندازه‌گیری شدند. همچنین، مجموعه‌ای از ۳۱ متغیر کمکی اعم از شاخص‌های توپوگرافی، طیفی، اقلیمی و ویژگی‌های خاک پس از حذف متغیرهای با واریانس نزدیک به صفر، دارای همبستگی بالا ($r \geq ۰/۸$) و با مقدار $VIF > ۱۰$، برای مدل‌سازی تغییرپذیری مکانی SOC استفاده شدند. نتایج نشان داد که مدل XGBoost با بیشترین ضریب تبیین ($R2 = ۰/۶۲۸$)، کمترین میزان خطا ($RMSE = ۰/۴۷۳$ و $MAE = ۰/۳۱۶$) و بالاترین نسبت عملکرد به بازه چارکی ($RPIQ = ۱/۹۸$) کارایی بهتری نسبت به دو مدل دیگر داشت. متغیرهای BD، CCE، شاخص کربنات (Carb.Ind)، شاخص رطوبت تفاوت نرمال شده(NDMI) و فاکتور طول شیب (S-Factor) بیشترین اهمیت را در توزیع SOC نشان دادند. نقشه‌های پراکنش SOC در منطقه نشان داد که شمال و شمال غربی منطقه بیشترین مقادیر و مناطق مرکزی، جنوبی و جنوب شرقی کمترین مقادیر SOC را دارند. نقشه‌های عدم قطعیت نشان داد که پیش‌بینی‌های ارائه شده در بخش‌هایی از شمال‌غرب، شرق و جنوب منطقه، عدم‌قطعیت بیشتری برخوردارند. به‌طورکلی، می‌توان نتیجه‌گیری نمود که استفاده همزمان از نقشه‌های پیش‌بینی و عدم قطعیت آن‌ها علاوه برای آگاهی از مقدار برآوردشدهٔ SOC، برای ارزیابی میزان اعتماد به نتایج ضروری است.

استناد: کریمی دهکردی؛ فریده، حجتی؛ سعید، بیسواس؛ آسیم.(۱۴۰۵) مقایسه برخی مدل¬های یادگیری ماشین در برآورد و نقشه¬برداری رقومی کربن آلی خاک (مطالعه موردی: شرق استان خوزستان) ، *مجله تحقیقات آب و خاک ایران*، ۵۷ (۹)، ۲۱۵۹-۲۱۳۳.
https://doi.org/10.22059/ijswr.2026.420209.670196

ناشر: مؤسسه انتشارات دانشگاه تهران.
© نویسندگان.
DOI: https://doi.org/10.22059/ijswr.2026.420209.670196

مقدمه

خاک یکی از بزرگ‌ترین منابع کربن آلی در زیست‌کره بوده و نقشی کلیدی در چرخه جهانی کربن ایفا می‌کند (Batjes, 2014). در مقیاس محلی، مدیریت کربن آلی خاک از اهمیت ویژه‌ای برخوردار است، زیرا مستقیماً بر خواص فیزیکی، شیمیایی و بیولوژیکی خاک تأثیر می‌گذارد. در مقیاس جهانی نیز مدیریت مناسب کربن خاک به دلیل نقش کلیدی آن در کاهش غلظت گازهای گلخانه‌ای اتمسفر و مقابله با تخریب زمین، حیاتی و زیست‌محیطی است (Stavi et al., 2015). کربن آلی خاک علاوه بر اثرگذاری بر کاهش یا افزایش دی‌اکسیدکربن اتمسفری، یکی از مهم‌ترین شاخص‌های کیفیت خاک به شمار می‌رود و با حاصلخیزی، ظرفیت نگهداشت آب و پایداری تولید در سامانه‌های کشاورزی ارتباط مستقیم دارد (Wiesmeier et al., 2019; Zhang et al., 2022). شناسایی چارچوب‌بندی نقش کربن خاک برای مقابله با چالش‌های جهانی، درک پویایی و ویژگی‌های کربن کل خاک، همراه با تعامل آن‌ها با عوامل خاک–محیط‌زیست الزامی می‌باشد. از این رو، برآورد دقیق کربن آلی خاک و شناخت الگوی تغییرات مکانی آن برای مدیریت پایدار خاک، برنامه‌ریزی کاربری اراضی و ارزیابی ظرفیت ترسیب کربن در مقیاس حوضه و منطقه ضروری است.

تعیین توزیع مکانی ویژگی‌های خاک با وضوح مناسب از طریق روش‌های سنتی نقشه‌برداری، دشوار است، بنابراین، توسعه یک روش و فرآیند دقیق‌تر و قابل اطمینان‌تر برای پیش‌بینی ویژگی‌های خاک بسیار مورد نیاز است (Jena et al., 2023; Moharana et al., 2022). نقشه‌برداری رقومی خاک با اتکا بر رابطه خاک–محیط و بهره‌گیری از متغیرهای کمکی مانند توپوگرافی، اقلیم و داده‌های سنجش از دور، امکان پیش‌بینی ویژگی‌های خاک در نقاط نمونه‌برداری‌نشده را فراهم می‌کند (McBratney et al., 2003; Keskin et al., 2018). از جمله کارکردهای اصلی نقشه‌برداری رقومی خاک می‌توان به تهیه نقشه‌های اولیه، بازنگری و به‌روزرسانی بانک‌های اطلاعاتی موجود، ارائه تفسیرهای ویژه از ویژگی‌های خاک و همچنین ارزیابی مخاطرات زیست‌محیطی اشاره کرد (Carre et al., 2007).

فرآیند مدل‌سازی در نقشه‌برداری رقومی خاک از طریق برقراری ارتباط میان ویژگی‌های خاک و متغیرهای محیطیِ کم‌هزینه (نظیر شاخص‌های حاصل از سنجش از دور و پارامترهای ژئومورفومتری) انجام می‌شود. در این فرآیند، نه تنها روابط آماری بین متغیرها مدل‌سازی می‌گردد، بلکه ساختار خودهمبستگی مکانی داده‌ها نیز جهت بهبود دقت برآوردها لحاظ می‌شود؛ بر این اساس، کارایی نهایی مدل‌ها مستقیماً به کمّیت و کیفیت داده‌های میدانی و همچنین قدرت تبیین متغیرهای محیطی وابسته است (Stoorvogel et al., 2009). نقشه‌برداری رقومی خاک به طور روزافزون از مدل‌های یادگیری ماشین برای آشکارسازی روابط میان ویژگی‌های خاک و متغیرهای گوناگون در نواحی مختلف بهره می‌گیرد. در این میان، انتخاب مدل مناسب برای توسعه مدل، نقش تعیین‌کننده‌ای در بهینه‌سازی نتایج متناسب با داده‌های در دسترس دارد (Khaledian and Miller., 2020).

تاکنون رویکردهای متعددی جهت برآورد و نقشه‌برداری رقومی کربن آلی خاک توسعه یافته‌اند؛ این روش‌ها طیف وسیعی از مدل‌های زمین‌آماری نظیر رگرسیون کریجینگ، مدل‌های خطی، ساختارهای هوشمند غیرخطی مانند شبکه‌های عصبی مصنوعی، ماشین بردار پشتیبان و نیز رویکردهای مبتنی بر درخت تصمیم همچون کوبیست و جنگل تصادفی را در بر می‌گیرند. در سال‌های اخیر، در بسیاری از مطالعات نقشه‌برداری رقومی خاک به دلیل توانایی مدل‌های یادگیری ماشین در مدل‌سازی روابط پیچیده و غیرخطی کربن آلی خاک استفاده می‌شود. مدل‌هایی مانند جنگل تصادفی (RF) و گرادیان تقویت شده (XGBoost) می‌توانند تعاملات بین متغیرها و پاسخ‌های غیرخطی را بدون نیاز به فرض‌های کلاسیک مدل‌های پارامتریک استخراج کنند و در بسیاری از کاربردهای محیطی عملکرد قابل قبولی نشان داده‌اند (Wadoux et al., 2020; Kasraei et al., 2024). بااین‌حال، شواهد جدید نشان می‌دهد که برتری یک مدل در یک منطقه، الزاماً در محیطی دیگر تکرار نمی‌شود و عملکرد مدل‌ها به‌شدت تابع ویژگی‌های محیطی، تراکم داده‌ها و دامنه تعمیم‌پذیری آن‌هاست (Tziolas et al., 2024; Dadgar et al., 2024). علاوه بر این، یکی از مسائل مهم در مطالعات نقشه‌برداری رقومی خاک‌ها، ارزیابی دامنه اعتبار پیش‌بینی در بخش‌هایی از منطقه است که از نظر فضای ویژگی شباهت کمتری به داده‌های آموزشی دارند. بر این اساس، Hateffard et al., (2024) نشان دادند که کاهش دقت در برون‌یابی مکانی بعضاً می‌تواند قابل توجه باشد. این نکته به‌ویژه برای حوضه‌های کوهستانی و ناهمگن اهمیت بیشتری دارد، زیرا تغییرات شدید شیب، جهت دامنه، رطوبت موضعی و فرآیندهای انتقال جرم، الگوهای مکانی SOC را پیچیده‌تر می‌سازند. در چنین شرایطی، نقشه نهایی تنها زمانی از نظر مدیریتی ارزشمند است که همراه با

تفسیر میزان عدم‌قطعیت آن ارائه شود.

در سال‌های اخیر، توجه به عدم‌قطعیت پیش‌بینی‌ها به یکی از محورهای اصلی مطالعات نقشه‌برداری رقومی خاک‌ها تبدیل شده است. Lilburne et al., (2024) نشان دادند که برآوردهای عدم‌قطعیت در محصولات بزرگ‌مقیاس خاک همیشه به‌خوبی کالیبره نیستند و لازم است با ارزیابی مستقل و تحلیل‌های مکانی تکمیل شوند. Rohmer et al., (2024) نیز نشان دادند که عوامل مؤثر بر مقادیر پیش‌بینی یک متغیر لزوماً همان عواملی نیستند که اعتمادپذیری پیش‌بینی‌ها را کنترل می‌کنند. بنابراین، تحلیل هم‌زمان مقدار پیش‌بینی متغیر هدف و عدم‌قطعیت آن، برای تفسیر کاربردی نتایج ضروری است.

استان خوزستان و به‌ویژه بخش شرقی آن (حوزه آبخیز کارون ۳)، به دلیل تنوع ارتفاعی و تغییرات اقلیمی–توپوگرافی، می‌تواند الگوهای مکانی پیچیده‌ای از کربن آلی خاک را نشان دهد. اگرچه در سال‌های گذشته در زمینه نقشه‌برداری رقومی خاک در کشور پیشرفت‌های فراوانی صورت پذیرفته است (Mousavi et al., 2020; Matinfat et al., 2021; Rahmati et al., 2024)، با این حال، علی‌رغم این پیشرفت‌ها، هنوز شکاف‌های تحقیقاتی قابل‌توجهی وجود دارد که ضرورت انجام تحقیق حاضر را توجیه می‌کند. نخست این‌که، بیشتر مطالعات انجام شده به‌ویژه در استان خوزستان بر دشت‌ها و اراضی کشاورزی متمرکز بوده‌اند و دقت مدل‌ها در مناطق با توپوگرافی نسبتاً پیچیده به‌ندرت ارزیابی شده است. دوم آن‌که، بسیاری از این مطالعات تنها به ارائه نقشه پیش‌بینی ویژگی‌های خاک اکتفا کرده و تحلیل عدم‌قطعیت مکانی در آن‌ها انجام نشده است. بر این اساس، این مطالعه با اهداف: (الف) ارزیابی و مقایسه عملکرد سه مدل جنگل تصادفی (RF)، گرادیان تقویت‌شده (XGBoost) و رگرسیون کریجینگ (RK) در برآورد کربن آلی خاک، (ب) شناسایی و تحلیل اهمیت متغیرهای محیطیِ کنترل‌کننده کربن آلی خاک، و (ج) تهیه نقشه رقومی پراکنش کربن آلی خاک و تحلیل عدم قطعیت آن در بخشی از اراضی حوزه آبخیز کارون ۳ انجام شد. نتایج حاصل از این مطالعه و نقشه‌های رقومی تولیدشده، می‌توانند ابزاری کاربردی برای برای شناسایی کانون‌های تخریب خاک، اولویت‌بندی عملیات حفاظت خاک در بالادست سد کارون ۳ و همچنین تدوین برنامه‌های مدیریت پایدار اراضی و ارتقای ظرفیت ترسیب کربن در مقیاس حوضه‌ای مورد استفاده واقع شود.

مواد و روش‌ها

منطقه مورد مطالعه

منطقهٔ مطالعاتی در این پژوهش در شرق استان خوزستان و حوزهٔ آبخیز سد کارون ۳ واقع شده است که از شمال شرقی به استان چهارمحال و بختیاری، از جنوب شرقی به استان کهکیلویه و بویراحمد، از غرب و شمال غربی به شهرستان ایذه و از جنوب به شهرستان باغملک محدود می‌شود (شکل ۱). ارتفاع منطقه ازسطح دریا از ۸۷۸ تا ۱۹۴۶متر و مساحت منطقه مورد مطالعه نیز حدود ۵۲/۴۱۶ کیلومتر مربع می‌باشد. بر اساس آمار ایستگاه هواشناسی سینوپتیک ایذه (نزدیک‌ترین ایستگاه هواشناسی به منطقه مطالعاتی)، متوسط دمای سالیانه هوا ۱۴ درجه سانتی‌گراد است و میزان رطوبت نسبی از ۲۹/۶ تا ۶۳/۳ درصد متغیر می‌باشد. همچنین، میانگین بارندگی ۱۰ ساله منطقه حدود ۶۰۰ میلی‌متر گزارش شده‌است. بر اساس سیستم طبقه‌بندی آمریکایی، خاک‌های منطقه دارای رژیم رطوبتی زریک[1] و رژیم حرارتی ترمیک[2]می‌باشند (Banaei, 1998). از نظر زمین‌شناسی نیز سازندهای متعددی در حوزه آبخیز سد کارون ۳ شناسایی شده‌اند که عمدتاً به دوران‌های دوم و سوم زمین‌شناسی (شکل ۲)، یعنی مزوزوئیک و سنوزوئیک تعلق دارند. در میان آن‌ها، سازند آسماری–جهرم به‌عنوان قدیمی‌ترین سازند شناخته می‌شود و سازندهای گچساران، آغاجاری، بختیاری و همچنین رسوبات دوره چهارم زمین‌شناسی به‌عنوان جوان‌ترین سازندها در بخش‌های مختلف این حوزه مشاهده می‌شوند (Talebian et al., 1999). در این منطقه دو نوع کاربری اراضی شامل جنگل تخریب شده و کشاورزی مشاهده می‌شود و بیش‌ترین کاربری این منطقه کشاورزی است.

نمونه‌برداری و اندازه‌گیری‌های آزمایشگاهی

نمونه‌برداری با استفاده از روش ابرمکعب لاتین شرطی (cLHS) در نیمه دوم مهرماه سال ۱۴۰۲ انجام و در هر نقطه از عمق ۰ تا ۵

1. Xeric
2. Thermic

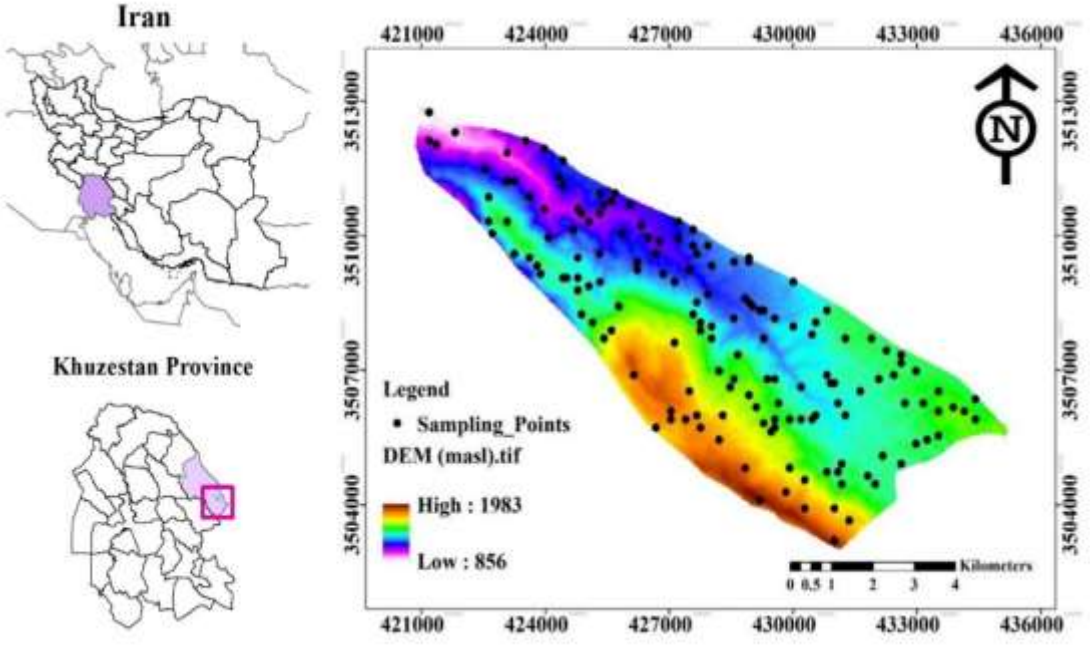
سانتی‌متر به صورت مرکب برداشت شد؛ بدین صورت که در هر نقطه ابتدا یک مربع به ابعاد ۱ متر درنظر گرفته‌شد و سپس از رئوس و مرکز مربع مزبور نمونه‌برداری انجام و نمونه‌ها با یکدیگر ترکیب و به آزمایشگاه منتقل شدند. لازم به توضیح است که در مطالعه حاضر از مدل رقومی ارتفاع منطقه با قدرت تفکیک مکانی ۹۰ متر، نقشهٔ خاک منطقه (در سطح رده)، نقشه شاخص پوشش گیاهی تفاوت نرمال شده (NDVI) و نقشهٔ زمین‌شناسی منطقه به عنوان متغیرهای محیطی ورودی برای تعیین موقعیت نقاط نمونه‌برداری در روش cLHS استفاده شد (شکل ۲). مدل رقومی ارتفاع به‌عنوان نماینده شرایط توپوگرافی و کنترل‌کننده فرآیندهایی نظیر شدت جریان و تجمع رواناب، فرسایش، رسوب‌گذاری و تغییرات رطوبت خاک در نظر گرفته شد. نقشه خاک، با هدف توجه به تغییرات کلاس‌های خاک و ویژگی‌های ذاتی آن‌ها استفاده گردید. نقشه زمین‌شناسی منطقه نیز با هدف بررسی اثر مواد مادری بر ویژگی‌های خاک به کار رفت. همچنین، از NDVI به دلیل ارتباط شناخته‌شده آن با تراکم، وضعیت و فعالیت پوشش گیاهی استفاده شد. در این پژوهش، به منظور پیاده‌سازی و اجرای الگوریتم cLHS در طراحی نقاط نمونه‌برداری از افزونه‌ی TEUI نسخه ۴۶۰ در محیط نرم افزار ArcGIS استفاده گردید. لازم به توضیح است برای استفاده از افزونه‌ی فوق در طراحی نقاط پیشنهادی برای نمونه‌برداری از خاک‌ها تمامی لایه‌های ورودی به فرمت رستری با اندازه پیکسل‌های مشابه تبدیل شدند.

نمونه‌های خاک بعد از هوا خشک کردن و عبور از الک ۲ میلی‌متر برای اندازه گیری خصوصیات فیزیکی و شیمیایی مورد نظر شامل اندازه‌گیری کربن آلی نمونه‌ها به روش اکسیداسیون تر، تعیین واکنش خاک (pH) در گل اشباع با دستگاه pH متر، اندازه‌گیری قابلیت هدایت الکتریکی در عصاره گل اشباع با دستگاه EC متر، تعیین بافت خاک به روش هیدرومتری ا اندازه‌گیری جرم مخصوص ظاهری (ρb) به وسیله سیلندر و کربنات کلسیم معادل نمونه‌ها نیز به روش تیتراسیون برگشتی، به‌کار گرفته شدند (Pansu and Gautheyrou, 2003). نمونه‌برداری با استفاده از استوانه‌های نمونه‌برداری با ابعاد حدوداً ۱۰ (قطر) در ۵ (ضخامت) سانتی‌متری انجام و در مجموع تعداد ۱۵۸ نمونه خاک سطحی برداشت و نوع کاربری فعلی و موقعیت جغرافیایی آن‌ها در محل‌های نمونه‌برداری با استفاده از سیستم موقعیت‌یاب جهانی (GPS) ثبت شد (شکل ۱).

آماده‌سازی، پیش‌پردازش و استخراج متغیرهای محیطی

در این پژوهش برای استخراج شاخص‌های طیفی سنجش از دوری از تصاویر ماهواره لندست ۹ استفاده شد که در شهریورماه سال ۱۴۰۲ از تارنمای سازمان زمین شناسی آمریکا (https://earthexplorer.usgs.gov) بارگیری شده بود. بعد از بارگیری، تصاویر ماهواره‌ای در محیط نرم‌افزار ENVI نسخه ۴/۷ فراخوانی شد. در مرحله پیش‌پردازش، ابتدا تصاویر ماهواره‌ای بر اساس مختصات جغرافیایی و مرز منطقه مورد مطالعه برش داده شدند. سپس، با استفاده از باندهای طیفی تصاویرِ پردازش‌شده، انواع شاخص‌های پوشش گیاهی به‌منظور استفاده به عنوان متغیرهای محیطی در مدل‌سازی تهیه شدند. همچنین، انواع شاخص‌های ژئومورفومتریک با استفاده از نقشه مدل رقومی ارتفاعی تهیه شده توسط سازمان نقشه‌برداری کشور با قدرت تفکیک مکانی ۱۰ متر در محیط نرم افزار SAGA GIS نسخه ۶/۴ تهیه گردید. در مجموع ۵۵ ویژگی برای مدل‌سازی پیش بینی کربن آلی خاک استفاده گردید. بعد از آماده‌سازی همهٔ متغیرهای کمکی (هم‌اندازه‌سازی پیکسل‌ها در لایه‌های اطلاعاتی، حذف پیکسل‌های گم‌شده، اصلاح نویز در تصاویر و هم‌پوشانی داده‌ها)، لایه‌های مختلف اعم از لایه‌های رستری و لایه‌ی نقاط نمونه‌برداری را روی هم انداخته و ماتریسی که شامل مختصات نقاط نمونه‌برداری شده، مقادیر متغیر هدف و متغیرهای محیطی در نقاط نمونه‌برداری شده است، تهیه و مدل‌های آماری بر داده‌ها برازش داده شد (McMilian et al., 2019).

مدل‌سازی و پیش‌بینی کربن آلی خاک، با استفاده از شاخص‌های طیفی به‌دست آمده از تصاویر ماهوارهٔ لندست ۹ و ویژگی‌های خاک (مطالعات صحرایی و آزمایشگاهی)، فاکتورهای اقلیمی و همچنین شاخص‌های توپوگرافی مشتق شده از مدل رقومی ارتفاع (DEM)(شکل ۲) انجام شد. لازم به توضیح است پس از پیش‌پردازش‌های اولیه و تعیین مناسب‌ترین متغیرها برای انجام مدل‌سازی با استفاده از مدل‌های جنگل تصادفی، رگرسیون کریجینگ و XGBoost ، بهترین مدل پیش‌بینی‌کننده کربن آلی را که دارای کمترین خطا و بیشترین ضریب تبیین بود، انتخاب و برای تهیه نقشه پراکنش کربن آلی از آن استفاده شد.



شکل ۱. نقشه مدل رقومی ارتفاع (DEM) منطقه مورد مطالعه که پراکنش نقاط نمونه‌برداری را در منطقه مورد مطالعه نمایش می‌دهد.

انتخاب متغیرهای محیطی

پیش از استفاده در مدل‌ها، تمام متغیرهای محیطی از نظر چندخطی بودن مورد بررسی قرار گرفت. عامل تورم واریانس (VIF) برای ارزیابی میزان هم‌خطی بین داده‌ها (معادلهٔ ۱) استفاده شد. بر این اساس، متغیرهای کمکی که همبستگی بالایی با یکدیگر نشان دادند ($r \geq 0.8$) و مقادیر $VIF \geq 10$ داشتند در فرآیند توسعه مدل حذف شدند.

$$VIF = \frac{1}{(1 - R_j^2)}$$ رابطه ۱)

در معادله ۱، R^2j ضریب تبیینی است که از رابطه رگرسیونی بین متغیر مورد نظر (ام) و سایر متغیرهای پیش‌بینی کننده به‌دست می‌آید (Hocking, 2013).

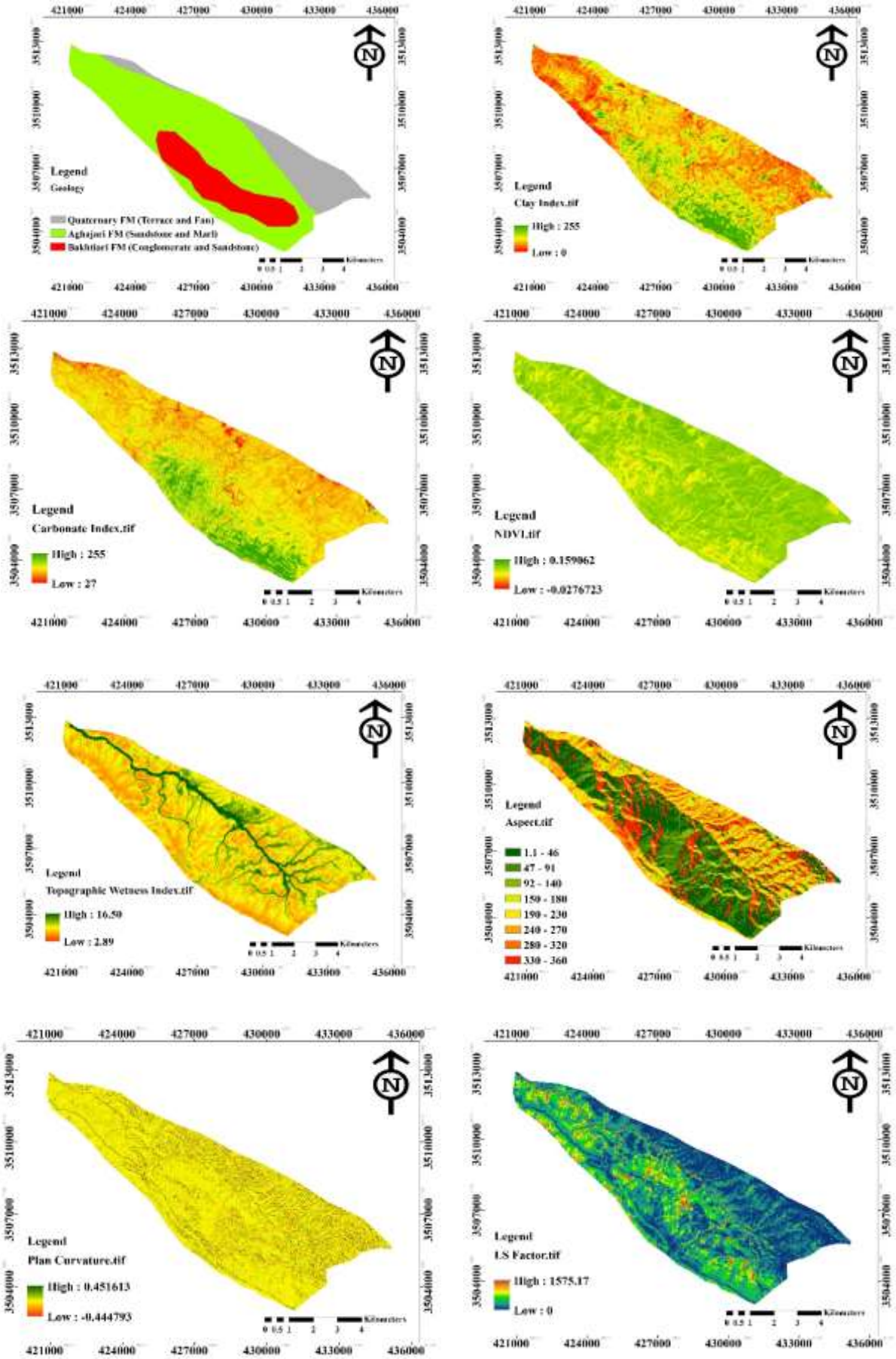
محاسبه عامل تورم واریانس (VIF) به صورت چند مرحله‌ای (Iterative) انجام شد. به‌گونه‌ای که در هر مرحله متغیری که دارای بیشترین مقدار VIF بود حذف گردید. این فرآیند تا زمانی که تمامی متغیرها دارای مقادیر VIF کمتر از ۱۰ باشند، انجام گردید (Hojati et al., 2024).

مدل‌های مورد استفاده

الف- مدل جنگل تصادفی

مدل جنگل تصادفی، یک روش یادگیری ماشین مبتنی بر مجموعه‌ای از درخت‌های تصمیم است که با استفاده از تکنیک انتخاب نمونه‌برداری با جایگذاری و انتخاب تصادفی زیرمجموعه‌ای از ویژگی‌ها، از بیش‌برازش جلوگیری می‌کند (Brieman, 2001). این مدل به دلیل ماهیت غیرپارامتریک خود، برخلاف مدل‌های آماری کلاسیک، نیازی به برقراری پیش‌فرض‌هایی چون توزیع نرمال داده‌ها، خطی بودن روابط بین متغیرها و هم‌سانی واریانس خطاها ندارد. این ویژگی به مدل اجازه می‌دهد تا تعاملات غیرخطی و پیچیده بین متغیرهای محیطی عددی و طبقه‌بندی‌شده را با دقت بالاتری مدل‌سازی نماید (Catani et al., 2013).

مدل جنگل تصادفی به سه پارامتر اصلی نیاز دارد که شامل تعداد متغیرهایی است که به‌طور تصادفی در هر تقسیم گره (mtry) با مقدار پیش‌فرض یک‌سوم تعداد کل متغیرها برای رگرسیون انتخاب می‌شوند، تعداد درختان (ntree) با مقدار پیش‌فرض ۵۰۰ و اندازه گره پایانی (nodesize) با مقدار استاندارد ۵ برای مدل‌های رگرسیونی است (Liaw and Wiener, 2002).



شکل ۲. نقشه تعدادی از متغیرهای محیطی تهیه شده در این پژوهش

برای تعیین اهمیت نسبی متغیرهای استفاده شده در فرآیند مدل‌سازی، افزایش میزان خطای استاندارد برای هر درخت با و بدون پیش‌بینی کننده‌های مربوطه محاسبه گردید. سپس مقدار میانگین تفاوت بین موارد با و بدون پیش‌بینی کننده‌ها برای همه درختان به صورت انحراف استاندارد تفاوت‌ها مشخص شد. رتبه‌های برتر در نمودار خروجی، اهمیت نسبی متغیرهای استفاده شده در فرآیند مدل‌سازی را نشان می‌دهند (Biau and Scornet, 2016; Kaya et al., 2022).

ب- مدل رگرسیون کریجینگ

یکی از روش‌های درون‌یابی ترکیبی، رگرسیون کریجینگ نام دارد. در این روش، ابتدا با استفاده از اطلاعات کمکی یک مدل رگرسیونی برازش می‌شود و سپس از کریجینگ ساده با میانگین معلوم برای درون‌یابی باقی‌مانده‌های حاصل از مدل رگرسیونی استفاده می‌شود. این ویژگی امکان استفاده از روش‌های رگرسیونی با پیچیدگی دلخواه، از جمله مدل‌های خطی تعمیم‌یافته، را فراهم می‌کند (Hengl et al., 2007). در پژوهش حاضر، ابتدا بخش نظام‌مند توزیع محتوای SOC با استفاده از رگرسیون خطی چندمتغیره (MLR) و با استفاده از تعداد ۳۱ متغیر محیطی مدل‌سازی گردید. همچنین، برای مدل‌سازی توزیع باقی‌مانده‌های حاصل از مدل رگرسیونی در الگوریتم رگرسیون کریجینگ، ابتدا چندین مدل واریوگرامی شامل Circular، Gaussian، Exponential، Spherical، Matérn و Pentaspherical بر داده‌های آزمایشی (باقیماندها) برازش داده شد. همچنین، میزان وابستگی مکانی باقی‌مانده‌ها در مدل‌های واریوگرامی مختلف با استفاده از نسبت اثر قطعه‌ای به سقف محاسبه و مقادیر وابستگی مکانی کمتر از ۲۵ درصد، ۲۵ تا ۷۵ درصد و بیش از ۷۵ درصد به‌ترتیب به عنوان کلاس‌های وابستگی مکانی ضعیف، متوسط و قوی در نظرگرفته شدند (Hojati, 2019; De Caires et al., 2025).

ج- مدل XGBoost

مدل XGBoost یک نسخه پیشرفته و بهینه‌سازی شده از مدل‌های "گرادیان تقویت شده" (Gradient Boosting) است. یکی از مهم‌ترین ویژگی‌های این مدل، توانایی بالای آن در کنترل خطای بیش‌برازش است؛ به این معنا که مدل اجازه نمی‌دهد محاسبات بیش از حد به داده‌های آموزشی وابسته شوند و در نتیجه، نقشه نهایی برای مناطق جدید دقیق‌تر خواهد بود. این کار از طریق افزودن پارامترهای کنترلی به تابع اصلی مدل انجام می‌شود که از پیچیده شدن بی دلیل درخت‌های تصمیم‌گیری جلوگیری می‌کند و از نظر سرعت و قدرت محاسباتی، بسیار هوشمندانه عمل می‌کند. این مدل با استفاده از محاسبات ریاضی دقیق‌تر (بسط مرتبه دوم)، مسیر رسیدن به کمترین خطا را سریع‌تر پیدا می‌کند. همچنین، برخلاف بسیاری از مدل‌های قدیمی که متغیرها را یکی‌یکی بررسی می‌کردند، می‌تواند داده‌ها را به صورت موازی پردازش کند که این موضوع زمان اجرای مدل را به‌ویژه در پروژه‌های با تعداد زیاد متغیرهای محیطی بسیار کاهش می‌دهد. علاوه بر این، مدل XGBoost در مواجهه با داده‌های ناقص یا گمشده در جدول داده‌ها، بسیار منعطف است و می‌تواند به طور خودکار بهترین مسیر را برای تحلیل آن‌ها پیدا کند (Chen and Guestrin., 2016). در این پژوهش، مدل XGBoost با تعداد ۳۶۱ درخت تقویتی پیاده‌سازی گردید. فرآیند اهمیت متغیرها در این مدل با استفاده از تابع پیش‌فرض این الگوریتم مبتنی بر معیارهای افزایش اطلاعات (Gain)، میزان پوشش‌دهی نمونه‌ها (Cover) و فرکانس تکرارها (Frequency) محاسبه شد.

ارزیابی اعتبار مدل برای مدلسازی کربن آلی خاک

برای ارزیابی دقت در روش‌های مدل‌سازی مورد استفاده، از آماره‌های ضریب تبیین (R^2)، میانگین قدر مطلق خطا (MAE)، خطای میانگین (ME)، نسبت عملکرد مدل به دامنه بین چارکی (RPIQ) و ریشه میانگین مربعات خطا (RMSE) استفاده شد:

$$R^2 = \left(\frac{\sum_{i=1}^{n}(Oi - \bar{O})(Pi - \bar{P})}{\sqrt{\sum_{i=1}^{n}(Oi - \bar{O})^2}\sqrt{\sum_{i=1}^{n}(Pi - \bar{P})^2}} \right)^2 \quad \text{رابطه ۲)}$$

$$RMSE = \sqrt[2]{\frac{1}{n}\sum_{i=1}^{n}(Pi - Oi)^2} \quad \text{رابطه ۳)}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Pi - Oi| \quad \text{رابطه ۴)}$$

$$ME = \frac{1}{n}\sum_{i=1}^{n}(Pi - Oi) \quad \text{رابطه ۵)}$$

$$IQ = Q_3 - Q_1 \quad \text{رابطه ۶)}$$

$$RPIQ = IQ / RMSE \quad \text{رابطه ۷)}$$

که در رابطه‌های ۲ تا ۵، n برابر تعداد نمونه‌های خاک Pi و Oi به‌ترتیب برابر مقادیر پیش‌بینی شده و اندازه‌گیری شده کربن آلی خاک در نقطه i، $\bar{P}$ برابر میانگین مقادیر پیش‌بینی شده و $\bar{O}$ میانگین مقادیر اندازه‌گیری شده می‌باشد (رحمتی و همکاران، ۱۴۰۳؛ Wang et al., 2021). در رابطه‌های ۶ و ۷، IQ دامنه بین چارکی داده‌های مشاهده شده است که از اختلاف مقادیر چارک سوم از چارک اول در متغیر هدف (کربن آلی خاک) به دست می‌آید. به طورکلی، هر چه مقادیر R^2 و RPIQ بزرگتر باشند نشان دهنده عملکرد بهتر مدل است. حال آن‌که، مدل با مقادیر (MAE، ME و RMSE) کوچکتر عملکرد بهتری را دارا خواهد بود.

تهیه نقشه‌های پراکنش کربن آلی خاک

به‌منظور تهیه نقشه پراکنش مکانی کربن آلی خاک با استفاده از مدل‌های RF و XGBOOST مدل نهایی و بهینه شده هر کدام از مدل‌های مزبور بر روی تمامی پیکسل‌های منطقه با استفاده از مجموعه کامل متغیرهای محیطی اعمال گردید تا نقشهٔ پیش‌بینی کربن آلی تهیه شود. برای تهیه نقشهٔ پراکنش کربن آلی در روش رگرسیون کریجینگ، تهیه نقشه نهایی پراکنش کربن آلی شامل دو مرحله مجزا بود. در گام اول، ابتدا مدل رگرسیونی خطی چند متغیرهٔ برازش داده شده بر نقاط نمونه‌برداری شده به تمامی لایه‌های رستری منطقه اعمال گردید تا نقشه پیش بینی شدهٔ روند (Trend) کربن آلی تهیه گردد. سپس و در گام دوم، با استفاده از بهترین مدل واریوگرامی برازش داده شده بر باقیمانده‌های مدل رگرسیونی و با استفاده از روش کریجینگ معمولی مقادیر باقیمانده‌های مدل برآورد و درون‌یابی شده و سپس به تمامی پیکسل‌های منطقه مطالعاتی اعمال گردید. در نهایت، نقشهٔ نهایی پراکنش کربن آلی از مجموع جبری نقشه‌های روند و باقیمانده‌های کریجینگ شده تولید گردید.

تهیهٔ نقشه‌های عدم قطعیت پیش بینی کربن آلی

با توجه به تغییرات غیرقابل پیش‌بینی خاک و همچنین فرآیندهای پیچیده خاک، معمولاً خطاهایی در خروجی مدل‌ها ایجاد می‌شود. لذا در پیش‌بینی ویژگی‌های خاک همواره درجات مختلفی از عدم قطعیت وجود دارد که لازم است در پروژه‌های نقشه‌برداری رقومی ویژگی‌های خاک نمایش داده شوند. از این رو، برای کمی سازی عدم قطعیت پیش‌بینی‌ها در استفاده از مدل‌های RF و XGBOOST از معیار عرض فاصله پیش‌بینی ۹۰ درصد استفاده شد. برای محاسبه این شاخص ابتدا برای هر پیکسل توزیع پیش‌بینی‌های حاصل از درخت‌های مختلف (در مدل RF) و یا مدل‌های بوت استراپ شده (در مدل XGBoost) محاسبه گردید و سپس فاصله بین صدک ۵ و ۹۵ این توزیع محاسبه و به عنوان معیار عدم قطعیت پیش‌بینی‌ها برای روش‌های مبتنی بر درخت تصمیم استفاده گردید.

میزان عدم قطعیت در روش RK از دو منبع مجزا محاسبه گردید. در ابتدا واریانس برآورد روند که ناشی از عدم قطعیت ضرایب مدل رگرسیونی است، محاسبه گردید و سپس این مقدار با واریانس حاصل از کریجینگ باقیمانده‌ها جمع جبری شد. در نهایت، فاصلهٔ پیش‌بینی ۹۰ درصد در این روش از رابطهٔ شمارهٔ ۸ به‌دست آمد:

$$PI_{90} = 1.645 \pm 1.645 \times SD \quad \text{رابطه ۸)}$$

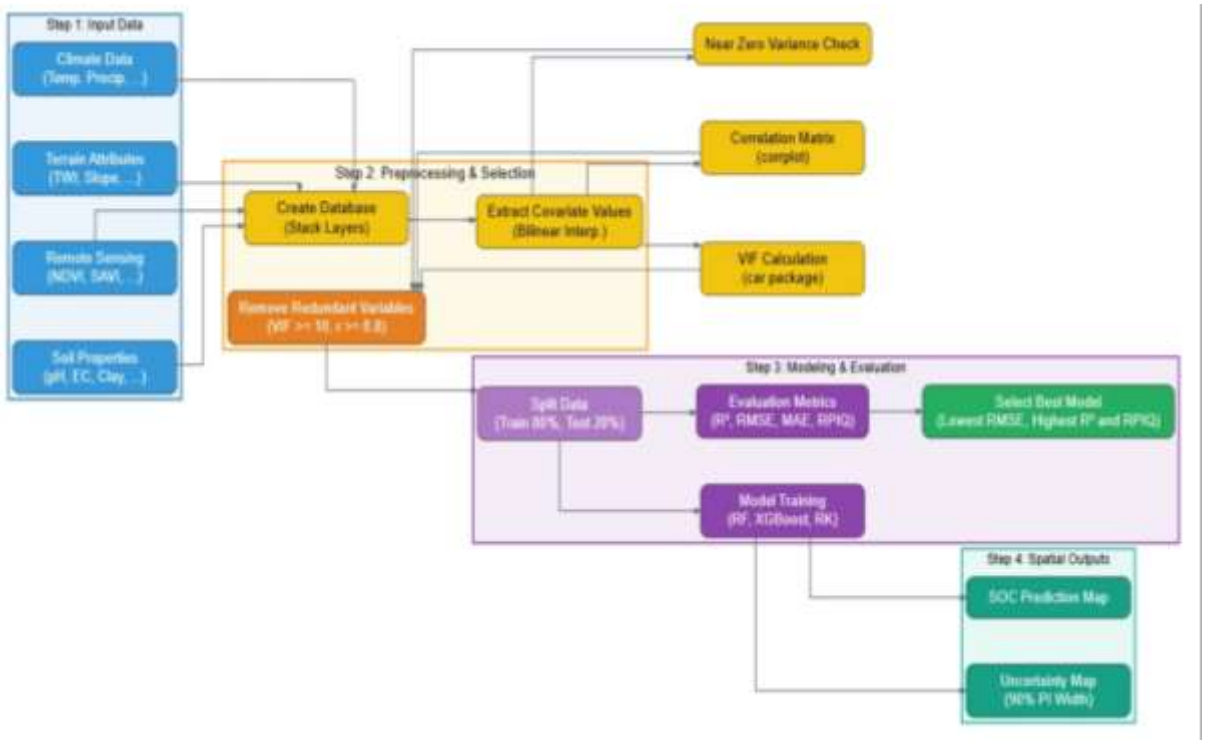
نرم‌افزارهای مورد استفاده

برای انجام این پژوهش اعم از آماده‌سازی داده‌ها، انجام تجزیه و تحلیل‌های آماری، اجرای مدل‌های مختلف، تهیه نمودارها و نقشه‌های مختلف از بسته‌های مختلفی در محیط نرم‌افزار R (R Core Team, 2023) استفاده گردید که فهرست آن‌ها در جدول ۱ ارائه شده است.

جدول ۱. بسته‌های استفاده شده در محیط نرم افزار R برای اجرای این پژوهش

ردیف	بسته	کاربرد	منبع
	clhs	**تعیین نقاط نمونه‌برداری و تقسیم بندی داده‌های آموزشی و آزمون**	**Roudier et al., (2012)**
۱	**readxl**	فراخوانی فایل‌های **Excel**	**Wickham and Bryan (2026)**
۲	**dplyrd**	مدیریت داده‌ها و محاسبه آماره‌های توصیفی	**Wickham et al., (2026)**
۳	**sp**	تبدیل داده‌ها به **SpatialPoints**	**Pebesma and Bivand (2005)**
۴	**MASS**	شناسایی نقاط فاقد داده (داده‌های گم‌شده)	**Venables and Ripley (2002)**
۵	**usdm**	شناسایی و حذف متغیرهای دارای واریانس صفر	**Naimi (2023)**
۶	**car**	محاسبه مقادیر **VIF**	**Fox and Weisberg (2019)**
۷	**corrplot**	ترسیم نمودار ماتریس همبستگی	**Wei and Simko (2021)**
۸	**randomForest**	اجرای مدل جنگل تصادفی	**Liaw and Wiener, (2002)**
۹	**xgboost**	اجرای مدل **XGBoost**	**Chen and Guestrin (2016)**
۱۰	**gstat**	محاسبه واریوگرام و اجرای مدل رگرسیون کریجینگ	**Pebesma, (2004)**
۱۱	**caret**	ارزیابی کارایی مدل‌های یادگیری ماشین	**Kuhn, (2008)**
۱۲	**quantregForest**	برآورد عدم قطعیت در مدل‌های جنگل تصادفی و **XGBoost**	**Meinshausen (2006)**
۱۳	**raster**	تهیهٔ نقشه‌های پیش‌بینی و عدم قطعیت پراکنش کربن آلی	**Hijmans, (2023)**
۱۴	**ggplot2**	ترسیم نمودارهای تعاملی	**Wickham, (2016)**

شکل ۳ روندنمای مراحل مختلف انجام این پژوهش را نشان می‌دهد.



شکل ۳. روند نمای مراحل مختلف تهیه نقشه رقومی کربن آلی خاک در منطقه مطالعاتی

یافته‌های پژوهش

انتخاب متغیرهای محیطی مؤثر

جدول ۲ و شکل ۴ به‌ترتیب مقادیر VIF و ضرایب همبستگی بین متغیرهای محیطی را بعد از حذف متغیرهای اضافی نشان می‌دهد. بر این اساس و پس از غربالگری اولیه تعداد ۳۱ متغیر محیطی برای فرآیند مدل‌سازی انتخاب شدند (جدول ۱). در مجموع، طی این فرآیند تعداد ۲۱ متغیر محیطی شامل SI2 ، OSAVI، SI3، SAVI، B4، NDVI، NDWI، SI5، B5، Gyp.Ind، SI1، B3، B6، WVP، BI، B7،

SI، SI6، Slope، Precipitation و Elevation به دلیل هم‌خطی بالا حذف شدند.

جدول شماره ۳ خلاصه توصیف آماری تعدادی از متغیرهای محیطی را بعد از حذف متغیرهای اضافی را نشان می‌دهد. توصیف ویژگی‌های آماری خاک و درک میزان پراکندگی آن‌ها، اولین گام در تحلیل‌های فضایی و مدل‌سازی‌های رقومی خاک است. لازم به ذکر است که طبقه‌بندی مقادیر ضریب تغییرات با استفاده از طبقه‌بندی وایلدینگ (Wilding., 1985) انجام گردید. طبق این دسته‌بندی، متغیرها به سه کلاس تغییرپذیری کم (با ضریب تغییرات کمتر از ۱۵ درصد)، تغییرپذیری متوسط (با ضریب تغییرات بین ۱۵ تا ۳۵ درصد) و تغییرپذیری زیاد (با ضریب تغییرات بالاتر از ۳۵ درصد) تقسیم‌بندی می‌شوند. بر اساس جدول ۴، متغیرهای pH خاک با ضریب تغییرات ۴/۶۵ درصد، جرم مخصوص ظاهری با ۶/۵۱ درصد، کربنات کلسیم معادل با ۹/۰۶ درصد و سیلت با ۱۴/۴۷ درصد در کلاس تغییرپذیری کم قرار گرفته‌اند.

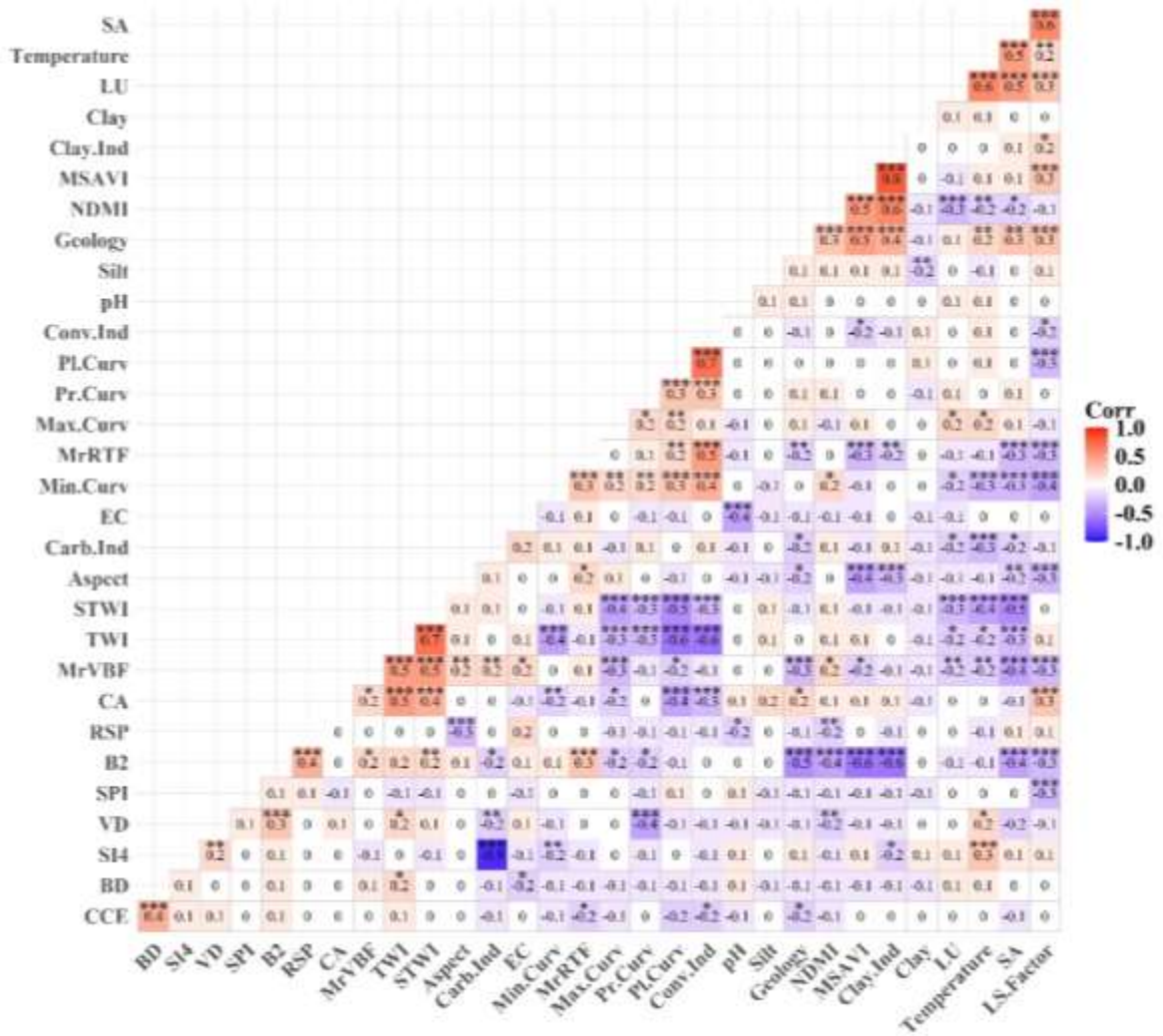
جدول ۲. متغیرهای نهایی انتخاب‌شده پس از تحلیل VIF

ردیف	علامت و نام متغیر به فارسی	نام لاتین	دسته بندی	VIF
۱	**SI4** (شاخص شوری ۴)	**Salinity Index 4**	تصاویر ماهواره‌ای	۷/۵۸
۲	**Carb.Ind** (شاخص کربنات)	**Carbonate Index**	تصاویر ماهواره‌ای	۷/۳۷
۳	**MSAVI** (شاخص پوشش گیاهی اصلاح شده)	**Modified Soil-Adjusted Vegetation Index**	تصاویر ماهواره‌ای	۵/۸۱
۴	**Clay.Ind** (شاخص رس)	**Clay Index**	تصاویر ماهواره‌ای	۴/۸۷
۵	**B2** (باند سبز لندست)	**Landsat 9 Green Band (Band 2)**	تصاویر ماهواره‌ای	۴/۳۴
۶	**NDMI** (شاخص رطوبت تفاوت نرمال شده)	**Normalized Difference Moisture Index**	تصاویر ماهواره‌ای	۲/۹۴
۷	**Temperature** (دما)	**Temperature**	اقلیمی	۲/۷۰
۸	**LU** (کاربری اراضی)	**Land Use**	موجودات زنده	۲/۴۰
۹	**Geology** (زمین‌شناسی)	**Geology**	مواد مادری	۱/۹۷
۱۰	**LS.Factor** (فاکتور طول و تندی شیب)	**Slope Length Factor**	توپوگرافی	۵/۵۸
۱۱	**TWI** (شاخص خیسی توپوگرافی)	**Topographic Wetness Index**	توپوگرافی	۵/۰۴
۱۲	**Conv.Ind** (شاخص تقعر/تحدب)	**Convergence Index**	توپوگرافی	۴/۴۸
۱۳	**STWI** (شاخص خیسی اصلاح شده)	**SAGA Topographic Wetness Index**	توپوگرافی	۳/۶۶
۱۴	**Pl.Curv** (انحنای تخت)	**Plan Curvature**	توپوگرافی	۳/۳۱
۱۵	**MrRTF** (شاخص مسطح بودن خطالراس)	**Multi-resolution Ridge Top Flatness**	توپوگرافی	۲/۱۸
۱۶	**SA** (مساحت شیب اراضی)	**Slope Area**	توپوگرافی	۵/۶۱
۱۷	**Min.Curv** (حداقل انحنا)	**Minimum Curvature**	توپوگرافی	۲/۱۷
۱۸	**MrVBF** (شاخص مسطح بودن کف دره)	**Multi-resolution Valley Bottom Flatness**	توپوگرافی	۲/۹۰
۱۹	**RSP** (موقعیت نسبی شیب)	**Relative Slope Position**	توپوگرافی	۱/۹۲
۲۰	**CA** (مساحت حوضه آبریز)	**Catchment Area**	توپوگرافی	۱/۸۴
۲۱	**V** (عمق دره)	**Valley Depth**	توپوگرافی	۱/۸۳
۲۲	**Aspect** (جهت شیب)	**Aspect**	توپوگرافی	۱/۷۴
۲۳	**Pr.Curv** (انحنای پروفیل)	**Profile Curvature**	توپوگرافی	۱/۷۰
۲۴	**SPI** (شاخص قدرت جریان)	**Stream Power Index**	توپوگرافی	۱/۳۴
۲۵	**Max.Curv** (حداکثر انحنا)	**Maximum Curvature**	توپوگرافی	۱/۵۸
۲۶	**EC** (قابلیت هدایت الکتریکی)	**Electrical Conductivity**	ویژگی‌های خاک	۱/۵۸
۲۷	**BD** (جرم مخصوص ظاهری)	**Bulk Density**	ویژگی‌های خاک	۱/۵۷
۲۸	**pH** (واکنش خاک)	**pH**	ویژگی‌های خاک	۱/۴۷
۲۹	**Clay** (فراوانی رس)	**Clay content**	ویژگی‌های خاک	۱/۳۱
۳۰	**Silt** (فراوانی سیلت)	**Silt content**	ویژگی‌های خاک	۱/۳۱
۳۱	**CCE** (کربنات کلسیم معادل)	**Calcium Carbonate Equivalent**	ویژگی‌های خاک	۱/۶۱

پایین بودن ضریب تغییرات pH نشان‌دهنده یکنواختی شرایط شیمیایی و وجود شرایط بافری قوی در کل منطقه است. به اعتقاد Boruvka et al., 2005، واکنش خاک به دلیل ماهیت لگاریتمی و کنترل شدید توسط کانی‌های مادری و کربنات‌ها، معمولاً کمترین میزان تغییرات مکانی را در بین ویژگی‌های خاک نشان می‌دهد. شاخص‌های شوری (SI4)،کربنات ورس، باند سبز تصاویر لندست (B2)،

و دما نیز از تغییرپذیری کمی برخوردار بودند. این متغیرها به دلیل ماهیت پایدار در مقیاس‌های خرد و کلان، نوسانات کمی نشان می‌دهند (Wang et al., 2009). میانگین آهک در اراضی منطقه معادل ۴۱/۸۸ درصد است که نشان‌دهنده ماهیت کاملاً آهکی مواد مادری خاک‌های منطقه است. تغییرپذیری بسیار پایین آهک (۹/۰۶ درصد) نیز می‌تواند مدلول آهکی بودن مواد مادری خاک‌ها در منطقه مطالعاتی باشد (مرادی نسب و همکاران، ۱۳۹۹؛ Tassinari et al., 2016). جرم مخصوص ظاهری با میانگین ۱/۳۳ گرم بر سانتی‌متر مکعب و ضریب تغییرات (۶/۵۱، درصد) پایداری ساختار فیزیکی خاک را بازگو می‌کند. از آنجا که نوسانات آن بیشتر تحت تأثیر بافت خاک و میزان ماده آلی است، یکنواخت بودن نسبی بافت اراضی مانع از تغییرات شدید این شاخص فیزیکی در منطقه شده است (Chen et al., 2019).

متغیرهای فراوانی رس، شاخص پوشش گیاهی اصلاح‌شده (MSAVI)، شاخص رطوبت (NDMI) و شاخص‌های خیسی (TWI, STWI) و موقعیت نسبی شیب (RSP) در گروه با تغییرپذیری مکانی متوسط قرار دارند. این متغیرها به شدت تحت تأثیر مدیریت اراضی، تغییر کاربری و رطوبت مقطعی خاک‌ها هستند. قرارگیری آن‌ها در کلاس متوسط، بیانگر تأثیرپذیری آن‌ها از فرآیندهای هیدرولوژیک و مدیریتی است که در طول زمان نوسانات معقولی دارند (Corwin and Lesch., 2005). محتوای رس با ضریب تغییرات ۱۹/۲۹ درصد تنها ویژگی فیزیکی خاک در این بخش است که در کلاس تغییرپذیری متوسط قرار می‌گیرد. بررسی‌ها حاکی از آن است که توزیع رس در خاک بیش از آنکه تحت تأثیر فرآیندهای سریع مدیریتی باشد، متأثر از فرآیندهای کند خاک‌سازی، هوازدگی خاک در طول شیب است. از این رو، نوسانات آن در حد متوسط ارزیابی می‌شود که با یافته‌های Wilding (1985) و Wang et al., (2009) مطابقت دارد. ایشان بیان داشتند که توزیع بافت خاک به صورت ذاتی و درازمدت در کنترل پستی و بلندی و هیدرولوژی منطقه قرار دارد.



شکل ۴. ماتریس همبستگی بین متغیرهای کمکی باقیمانده بعد از حذف متغیرهای اضافی. مخفف‌های ارائه شده در شکل، در جدول شماره ۲ معرفی شده‌اند.

هدایت الکتریکی خاک با ضریب تغییرات ۳۹/۵۱ درصد در کلاس تغییرپذیری زیاد قرار دارد. مقادیر بالای ضریب تغییرات در این ویژگی نشان‌دهنده ناهمگونی شدید آن در سطح منطقه و تأثیرپذیری زیاد از عواملی نظیر شرایط رطوبتی، وضعیت زهکشی موضعی، میزان تبخیر و شست‌وشو در امتداد پستی و بلندی‌ها است. افزون بر این، از آن‌جایی‌که کشاورزی یکی از کاربری‌های عمده منطقه را تشکیل می‌دهد، بخشی از این تغییرپذیری مکانی را می‌توان به فعالیت‌های مدیریتی و زراعی نسبت داد؛ به‌طوری‌که تفاوت در نوع کشت، شیوه آبیاری، میزان مصرف اصلاح‌کننده‌ها و نیز مدیریت بقایای گیاهی و شوری، می‌تواند سبب تغییرات قابل‌توجه در هدایت الکتریکی خاک در نقاط مختلف منطقه شود. Corwin and Lesch., (2005) نیز شوری خاک را از ویژگی‌های شیمیایی پویایی دانستند که به دلیل حساسیت زیاد به فرآیندهای هیدرولوژیک و مدیریتی، تغییرپذیری مکانی بالایی دارد.

مقادیر بسیار بالای ضریب تغییرات متغیرهایی مانند انحنای ماکزیمم (Max.Curv) و انحنای تخت (Pl.Curv) به این دلیل است که میانگین این متغیرها به صفر بسیار نزدیک است و از این رو، محاسبهٔ ضریب تغییرات برای چنین متغیرهایی منجر به اعداد بسیار بزرگی می‌شود. این ویژگی ذاتی شاخص‌های هندسی است و نشان‌دهنده پیچیدگی مورفولوژیک سطح زمین است. همان‌طور که Boruvka et al., (2005) نیز به آن اشاره می‌کند، چنین نوسانات شدیدی در توپوگرافی، عامل اصلی تفاوت در توزیع مکانی ویژگی‌های فیزیکی و شیمیایی خاک در مقیاس‌های کوچک است.

جدول ۳. توصیف آماری برخی ویژگی‌های خاک و متغیرهای محیطی استفاده شده در این پژوهش

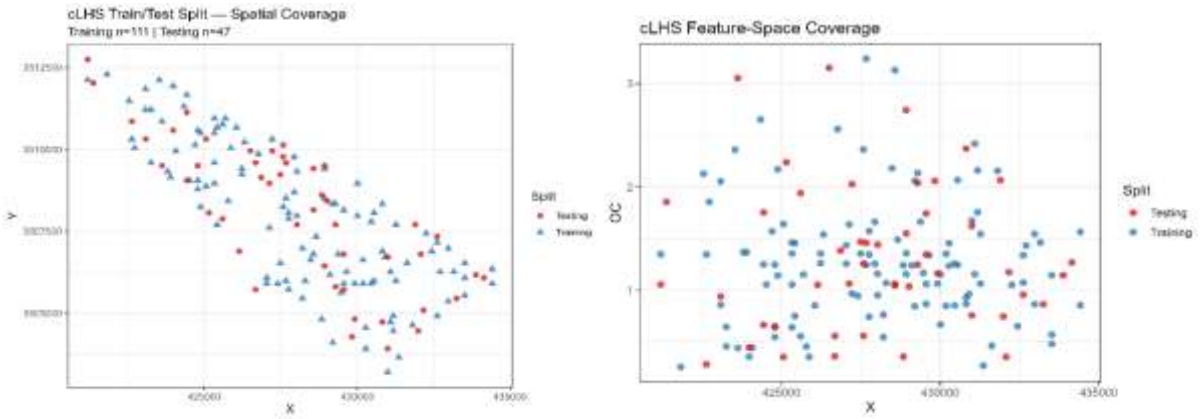
ردیف	نام متغیر	حداقل	حداکثر	میانگین	میانه	انحراف معیار	ضریب تغییرات**
۱	SOC	۰/۲۵۶	۳/۲۱۴	۱/۲۴۴	۱/۱۵۵	۰/۵۸۲	۴۶/۷۶
۲	BD	۱/۱۴۰	۱/۵۵	۱/۳۳	۱/۳۲	۰/۰۸۶	۶/۵۱
۳	pH	۵/۲۵۰	۸/۴۵	۷/۶۹	۷/۷۳	۰/۳۵	۴/۶۵
۴	Clay	۷/۷۵۰	۲۴/۲۵	۱۴/۸۴	۱۴/۰۰	۲/۸۶۰	۱۹/۲۹
۵	Silt	۲۳/۲۵	۶۲	۴۷/۰۷	۴۷/۸۷	۶/۸۱۰	۱۴/۴۷
۶	CCE	۳۴/۵۰	۵۳/۵۰	۴۱/۸۸	۴۱/۰۰	۳/۷۹۰	۹/۰۶
۷	EC	۰/۳۵	۱/۹۶	۰/۹۳	۰/۸۴	۰/۳۶۶	۳۹/۵۱
۸	SI4	۰/۴۱۵	۰/۶۲۸	۰/۴۹	۰/۴۹	۰/۰۲۷	۵/۵۱
۹	Carb.Ind	۱/۳۵	۱/۷۱	۱/۵۲	۱/۵۲	۰/۰۵۸	۳/۸۲
۱۰	MSAVI	۰/۰۹۰	۰/۱۹۹	۰/۱۴	۰/۱۴	۰/۰۲۴	۱۷/۱۴۳
۱۱	Clay.Ind	۱/۳۳۰	۱/۵۶	۱/۳۱	۱/۳۱	۰/۰۷۶	۵/۸۰۲
۱۲	B2	۰/۰۷۳	۰/۱۶۲	۰/۱۱۸	۰/۱۲۰	۰/۰۱۷	۱۵/۱۲
۱۳	Temperature	۱۷/۰	۲۰/۰	۱۸/۰	۱۷/۰	۰/۷۹	۴/۳۹۰
۱۴	LS.Factor	۰/۷۳۳	۴۵۲/۴۰۵	۶۰/۳۴	۳۹/۸۴	۶۴/۹۴۱	۱۰۷/۶۲۵
۱۵	TWI	۴/۳۱۵	۱۶/۹۹	۶/۹۲۸	۶/۲۹	۲/۰۷۳	۲۹/۹۲
۱۶	STWI	۴/۸۸۷	۱۰/۹۹	۷/۴۰۰	۷/۲۲	۱/۴۰۲	۱۸/۹۴
۱۷	Pl.Curv	−۰/۰۰۳	۰/۰۰۲۵	۰/۰۰۰۰۲۸	۱۶×۱۰^{-۶}	۰/۰۰۱۰	۳۵۷۱/۴۲۹
۱۸	MrRTF	۰/۰۰۰	۲/۵۹۷	۰/۱۷۸	۸۷×۱۰^{-۵}	۰/۴۳۹	۲۴۵/۸۷
۱۹	MrVBF	۵×۱۰^{-۱۱}	۲/۷۲۸	۰/۲۳۳	۰/۰۰۲	۰/۴۹۵	۲۱۲/۲۲
۲۰	RSP	۰/۳۳۸	۱/۲۳۸	۰/۷۸	۰/۷۷	۰/۱۴۸	۱۹/۱۵
۲۱	V	۰/۲۲	۴۹۲/۵۴	۱۸۴/۲۹	۱۸۵/۹۵	۱۰۶/۸۷	۵۷/۹۹
۲۲	Aspect	۰/۰۲۳	۶/۲۸	۲/۹۶	۳/۲۹	۲/۰۱۱	۶۸/۰۸
۲۴	Max.Curv	−۰/۰۱۲	۰/۰۰۸۲	۰/۰۰۱۹	۰/۰۰۲	۰/۰۰۱۸	۹۴/۲۵

رنگ‌های سبز، زرد و نارنجی در ستون ضریب تغییرات به ترتیب نشان دهندهٔ تغییرپذیری کم، متوسط و زیاد بر اساس طبقه‌بندی وایلدینگ (Wilding., 1985) می‌باشند. مخفف‌های ارائه شده در جدول، در جدول شماره ۲ معرفی شده‌اند.

تفکیک داده‌ها به مجموعه آموزش و آزمون

پیش از اقدام برای مدل‌سازی، داده‌های آزمایشی به دو گروه آموزشی و آزمون تقسیم‌بندی شدند. از داده‌های آموزشی برای آموزش مدل‌ها و از داده‌های آزمون برای ارزیابی مدل‌های ساخته شده استفاده گردید. در این پژوهش، برای تفکیک داده‌ها به مجموعه‌های آموزش و آزمون از روش cLHS استفاده شد (شکل ۵). این روش برای طراحی نسبتاً متوازن مجموعه‌های آموزش و آزمون در مطالعات نقشه‌برداری

رقومی خاک روشی مناسب به‌شمار می‌رود (Minasny and McBratney, 2006). در روش cLHS، انتخاب نمونه‌ها به‌گونه‌ای انجام می‌شود که مجموعه‌های حاصل، تا حد امکان بتوانند دامنه تغییرات موجود در داده‌ها و شرایط محیطی منطقه را بازنمایی کنند. به بیان دیگر، این روش تنها بر پراکنش تصادفی نقاط تکیه ندارد، بلکه تلاش می‌کند نمونه‌های انتخاب‌شده از نظر ویژگی‌های محیطی و نیز موقعیت مکانی، پوشش مناسبی از کل محدوده مطالعه داشته باشند (Roudier et al., 2012). شکل ۵ به‌خوبی نشان می‌دهد که داده‌های مجموعهٔ آموزش و آزمون، از نظر مکانی توزیع نسبتاً متوازن و مناسبی در محدوده مطالعه دارند. همچنین، شکل ۵ نشان می‌دهد که نقاط انتخابی به عنوان مجموعه‌های آموزش و آزمون در تمامی سطوح مقدار کربن آلی خاک، پراکنده شده‌اند. این توزیع نسبتاً متوازن، برای کاهش بیش‌برازش (Overfitting) و ارزیابی بی‌طرفانه‌ی مدل‌های انتخابی ضروری است.



شکل ۵. پراکنش مکانی نمونه‌های آموزش و آزمون انتخاب‌شده با روش cLHS. ۷۰ درصد داده‌ها برای آموزش مدل‌ها و از ۳۰ درصد باقی‌مانده برای ارزیابی و آزمون مدل‌ها استفاده شد.

نتایج بهینه‌سازی مدل‌ها

نتایج بهینه‌سازی مدل‌های مورد استفاده برای پیش‌بینی SOC در منطقه مورد مطالعه در جداول ۴ (RK) و ۵ (RF و XGBoost) نشان داده شده است. بر این اساس در مدل RK بخش رگرسیونی توانسته است **۷۰/۷۷ درصد** از تغییرات SOC را توضیح دهد. تحلیل نیم‌تغییرنمای برازش یافته بر باقی‌مانده‌های حاصل از رابطهٔ رگرسیونی نشان داد که پارامترهای واریوگرام بسته به نوع مدل تفاوت قابل توجهی دارند. به‌گونه‌ای که مدل‌های واریوگرامی Circular، Spherical و Pentaspherical با وابستگی مکانی کمتر از ۲۳ درصد در کلاس وابستگی مکانی ضعیف، و در مدل‌های Exponential و Matérn در کلاس وابستگی مکانی متوسط (۶۶/ ۴۹ درصد) قرار گرفتند. در میان مدل‌های بررسی‌شده مدل Gaussian به‌دلیل توانایی بالا در تبیین ساختار مکانی داده‌های آزمایشی (وابستگی مکانی بسیار قوی، ۹۹/۹۹ درصد) به عنوان مدل واریوگرامی بهینه انتخاب شد. همچنین مقادیر اثر قطعه‌ای ۰/۱۰۹ و دامنه ۷۲۲۲۳۰/۹۱ در این مدل، نشان‌دهنده وجود یک ساختار مکانی منسجم در باقی‌مانده‌ها بود. بنابراین، با توجه به مجموع شاخص‌های ارائه‌شده در جدول ۴ و کارایی بهتر این مدل در تبیین ساختار مکانی داده‌ها، مدل Gaussian به‌عنوان مدل واریوگرامی منتخب برای مدل‌سازی مقادیر باقی‌مانده‌ها با استفاده از روش کریجینگ معمولی انتخاب گردید.

جدول ۴. پارامترهای واریوگرام باقی‌مانده‌های مدل RK

مدل واریوگرامی Model Variogram	اثر قطعه ای Nugget	سقف Total Sill	دامنه Range	وابستگی مکانی Spatial Dependency	کلاس وابستگی مکانی Spatial Dependency Class
Circular	۰/۱۰۴	۰/۱۲۸	۱۶۳۷۹/۶۲	۱۸/۷۹	Weak
Gaussian	۰/۱۰۹	۱۱۷۷/۳۶	۷۲۲۲۳۰/۰۹	۹۹/۹۹	Strong
Exponential	۰/۱۰۰	۰/۰۹۹	۲۷۱۰۷/۹۲	۴۹/۶۶	Moderate
Spherical	۰/۱۰۴	۰/۱۳۴	۲۲۰۳۹/۲۹	۲۲/۱۵	Weak
Matérn	۰/۱۰۰	۰/۱۹۹	۲۷۱۰۷/۹۹	۴۹/۶۶	Moderate
Pentaspherical	۰/۱۰۵	۰/۱۳۴	۲۸۵۶۴/۳۱	۲۱/۹۰	Weak

نتایج بهینه‌سازی مدل جنگل تصادفی نشان داد که مدل بهینه با ۵۰۰ درخت تصمیم (ntree) و تعداد ۱۰ متغیر (mtry) ساخته شده است. بر این اساس، این مدل در مرحله آموزش توانسته است ۴۶/۸۱ درصد از کل تغییرات (واریانس) SOC را در منطقه مطالعاتی تبیین کند. نتایج بهینه‌سازی پارامترهای حاصل از واسنجی مدل XGBoost عبارت بودند از: تابع هدف رگرسیون با خطای مربعات (reg:squarederror) و نرخ یادگیری (eta) برابر با ۰/۰۱، حداکثر عمق درختان (max_depth) برابر با ۵، ضریب نمونه‌گیری از زیرمجموعه‌ها (subsample) برابر با ۰/۹، ضریب نمونه‌گیری از ویژگی‌ها در هر درخت (colsample_bytree) برابر با ۰/۹ و حداقل وزن گرهٔ پایانی (min_child_weight) برابر با واحد (۱). بر این اساس، مدل XGBoost توانسته است ۷۶/۶۲ درصد از تغییرات SOC را در مرحله واسنجی تبیین کند.

جدول ۵. هایپرپارامترهای بهینه مدل‌های جنگل تصادفی (RF) و XGBoost

ردیف	پارامتر	XGBoost	جنگل تصادفی (RF)
۱	تعداد درختان (**ntree/ n_estimators**)	۳۶۱	۵۰۰
۲	تعداد متغیر کاندید در هر گره (**mtry / colsample_bytree**)	۰/۹	۱۰
۳	حداقل وزن گره پایانی (**nodesize / min_child_weight**)	۱	۵
۴	حداکثر عمق درخت (**max_depth**)	۵	–
۵	نرخ یادگیری (**eta**)	۰/۰۱	–
۶	ضریب نمونه‌گیری از زیرمجموعه‌ها (**subsample**)	۰/۹	–
۷	تابع هدف (**objective**)	**reg:squarederror**	–
۸	درصد واریانس توجیه شده توسط مدل	۷۶/۶۲	۴۶/۸۱
۹	روش واسنجی	اعتبارسنجی متقابل ۱۰بخشی با توقف زودهنگام	اعتبارسنجی متقابل۱۰ بخشی با ۳ تکرار

ارزیابی عملکرد مدل‌ها

جدول ۶ عملکرد مدل‌های مورد استفاده در پیش‌بینی محتوای کربن آلی خاک را بر اساس داده‌های آزمون نشان می‌دهد. بر این اساس، مدل **XGBoost** با ضریب تبیین (R^2 =۰/۶۲۸) ، خطای جذر میانگین مربعات (RMSE = ۰/۴۷۳)، خطای مطلق میانگین (MAE = ۰/۳۱۶) و مقدار RPIQ = ۱/۹۸ بهترین عملکرد را در پیش‌بینی کربن آلی خاک نشان داده است. پس از آن، مدل RF با ضریب تبیین (۰/۵۴۵ = R^2)، خطای جذر میانگین مربعات (RMSE = ۰/۵۰۹)، خطای مطلق میانگین (MAE = ۰/۳۷۲) و مقدار RPIQ = ۱/۸۴۱ در رتبه دوم قرار می‌گیرد. همچنین، مدل RK با ضریب تبیین (R^2 = ۰/۴۳۴)، خطای جذر میانگین مربعات (RMSE = ۰/۵۳۴) ، خطای مطلق میانگین (MAE = ۰/۳۹۵) و مقدار RPIQ = ۱/۷۵۴ ضعیف‌ترین عملکرد را از خود نشان داده است.

برتری مدل XGBoost را می‌توان با سازوکار یادگیری آن توضیح داد. در مدل RF، تعداد زیادی درخت تصمیم به‌صورت مستقل و بر پایه نمونه‌گیری تصادفی ساخته می‌شوند و پیش‌بینی نهایی از طریق تجمیع یا میانگین‌گیری نتایج این درخت‌ها به دست می‌آید (Breiman, 2001). این ویژگی اگرچه موجب کاهش واریانس مدل و افزایش پایداری آن می‌شود، اما در برخی موارد می‌تواند به هموار شدن پیش‌بینی‌ها منجر شود، به‌ویژه زمانی که متغیر هدف تحت تأثیر برهم‌کنش‌های پیچیده و غیرخطی بین متغیرهای محیطی قرار داشته باشد. در مقابل، XGBoost از چارچوب gradient boosting استفاده می‌کند؛ به این معنا که درخت‌ها به‌صورت ترتیبی ساخته می‌شوند و هر درخت جدید با هدف کاهش خطاهای باقی‌مانده از مرحله قبل آموزش می‌بیند (Chen and Guestrin., 2016). در نتیجه، این مدل قادر است الگوهای ظریف‌تر، روابط غیرخطی پیچیده‌تر و اثرات تعاملی میان متغیرها را بهتر شناسایی کند.

افزون بر این، وجود مؤلفه‌های منظم‌سازی در تابع هدف XGBoost باعث کنترل پیچیدگی مدل و کاهش خطر بیش‌برازش می‌شود، که این موضوع در داده‌های محیطی با تعداد زیاد متغیرهای کمکی و ساختار پیچیده بسیار حائز اهمیت است (Chen and Guestrin., 2016). همچنین، در مدل XGBoost از یک تابع امتیازدهی شباهت برای هرس گره‌هایی با سود کم در طول فرآیند مدل‌سازی استفاده می‌شود تا از بیش‌برازش جلوگیری شده و پیش‌بینی‌ها از حالت تصادفی دور بمانند. حال‌آن‌که، مدل RF پیش‌بینی کربن آلی خاک را بر اساس خروجی همه درخت‌ها انجام می‌دهد و به همین دلیل شانس بالایی در بروز پیش‌بینی‌های تصادفی وجود دارد. به همین دلیل، در بسیاری از مطالعات نقشه‌برداری رقومی خاک و مدل‌سازی متغیرهای محیطی، روش‌های Boosting در شرایط ناهمگنی بالای محیطی و توپوگرافی پیچیده، عملکرد بهتری نسبت به RF نشان داده‌اند (Hengl et al., 2018). Chen et al., (2019) در مطالعات خود نیز نشان دادند

که مدل XGBoost در پیش‌بینی توزیع مکانی کربن آلی خاک، دقت بالاتری نسبت به مدل RF داشته است. ایشان بیان کردند که این برتری ممکن است به تعداد درخت‌ها، روش‌های هرس درخت و توانایی در مدیریت داده‌های نامتوازن مربوط باشد.

از سوی دیگر، مدل RK ضعیف‌ترین عملکرد را در میان سه مدل بررسی‌شده نشان داد. RK یک روش زمین آماری ترکیبی است که ابتدا رابطه بین متغیر هدف و متغیرهای کمکی را با یک مدل رگرسیونی برآورد می‌کند و سپس باقیمانده‌های حاصل را با استفاده از کریجینگ مدل‌سازی می‌کند (Hengl et al., 2007). مزیت اصلی این روش آن است که می‌تواند هم‌زمان از اطلاعات کمکی و ساختار فضایی باقیمانده‌ها بهره بگیرد. با این حال، کارایی RK به دو شرط اساسی وابسته است: نخست آنکه بخش رگرسیونی بتواند سهم قابل‌توجهی از تغییرات متغیر هدف را توضیح دهد، و دوم آنکه باقیمانده‌ها دارای خودهمبستگی مکانی معنی‌دار و قابل مدل‌سازی باشند (Keskin and Grunwald, 2018). در مطالعه حاضر، به نظر می‌رسد این دو شرط به‌طور کامل برقرار نبوده است. از یک سو، روابط بین SOC و متغیرهای محیطی منتخب، از جمله شاخص‌های توپوگرافی و طیفی، احتمالاً ماهیتی پیچیده، غیرخطی و همراه با برهم‌کنش‌های چندگانه داشته‌اند؛ بنابراین، اگر بخش رگرسیونی RK بر پایه یک فرم خطی یا نسبتاً ساده تعریف شده باشد، توانایی آن در بازنمایی کامل این روابط محدود خواهد بود. از سوی دیگر، بخشی از تغییرات SOC ممکن است ناشی از ناهمگنی‌های موضعی، تغییرات مدیریتی، تفاوت کاربری اراضی یا اثرات ترکیبی متغیرهای محیطی باشد؛ الگوهایی که لزوماً به‌صورت یک ساختار مکانی پایدار و ساده در باقیمانده‌ها ظاهر نمی‌شوند و در نتیجه، قابلیت مدل‌سازی آن‌ها با کریجینگ کاهش می‌یابد (Rossetti et al., 2024; Zhang et al., 2024; De Caires et al., 2025). به همین دلیل، در شرایطی که روابط بین متغیر هدف و عوامل کنترل‌کننده آن پیچیده و غیرخطی باشد، مدل‌های یادگیری ماشین اغلب عملکرد بهتری نسبت به روش‌های کلاسیک یا ترکیبی زمین‌آماری نشان می‌دهند. نتایج مطالعات پیشین نشان می‌دهد که رویکردهای هیبریدی مبتنی بر یادگیری ماشین و کریجینگ باقیمانده‌ها، در مقایسه با مدل‌های یادگیری ماشین منفرد و روش‌های ژئواستاتیستیکی کلاسیک، معمولاً دقت بالاتری در برآورد کربن آلی خاک دارند. این برتری به ترکیب توانایی مدل‌سازی روابط غیرخطی توسط یادگیری ماشین و بازنمایی ساختار فضایی باقیمانده‌ها توسط کریجینگ نسبت داده می‌شود (Zhu et al., 2022; Ho et al., 2024).

نتایج ارزیابی عملکرد نشان داد که مدل **RK**، با وجود انتخاب مدل واریوگرامی مناسب (گوسی) در پیش‌بینی کربن آلی خاک نسبت به مدل‌های RF و XGBoost از دقت پایین‌تری برخوردار بود. این موضوع می‌تواند بیانگر آن باشد که بخش عمده تغییرات کربن آلی خاک در منطقه مورد مطالعه توسط متغیرهای کمکی تبیین شده و باقی‌مانده‌ها از خودهمبستگی فضایی کافی و پایدار برای بهبود محسوس پیش‌بینی از طریق کریجینگ برخوردار نبوده‌اند. در مقابل، مدل‌های RF و XGBoost به دلیل توانایی در شناسایی روابط غیرخطی پیچیده و برهم‌کنش‌های میان متغیرهای محیطی، عملکرد بهتری نشان دادند. این نتیجه با یافته‌های مطالعات اخیر همخوانی دارد که نشان می‌دهند کارایی رویکردهای هیبریدی مانند رگرسیون–کریجینگ به میزان خودهمبستگی فضایی باقی‌مانده‌ها وابسته است و در شرایطی که روابط محیطی پیچیده باشند، مدل‌های یادگیری ماشین اغلب دقت بالاتری ارائه می‌کنند (Hengl et al., 2017; Ho et al., 2024).

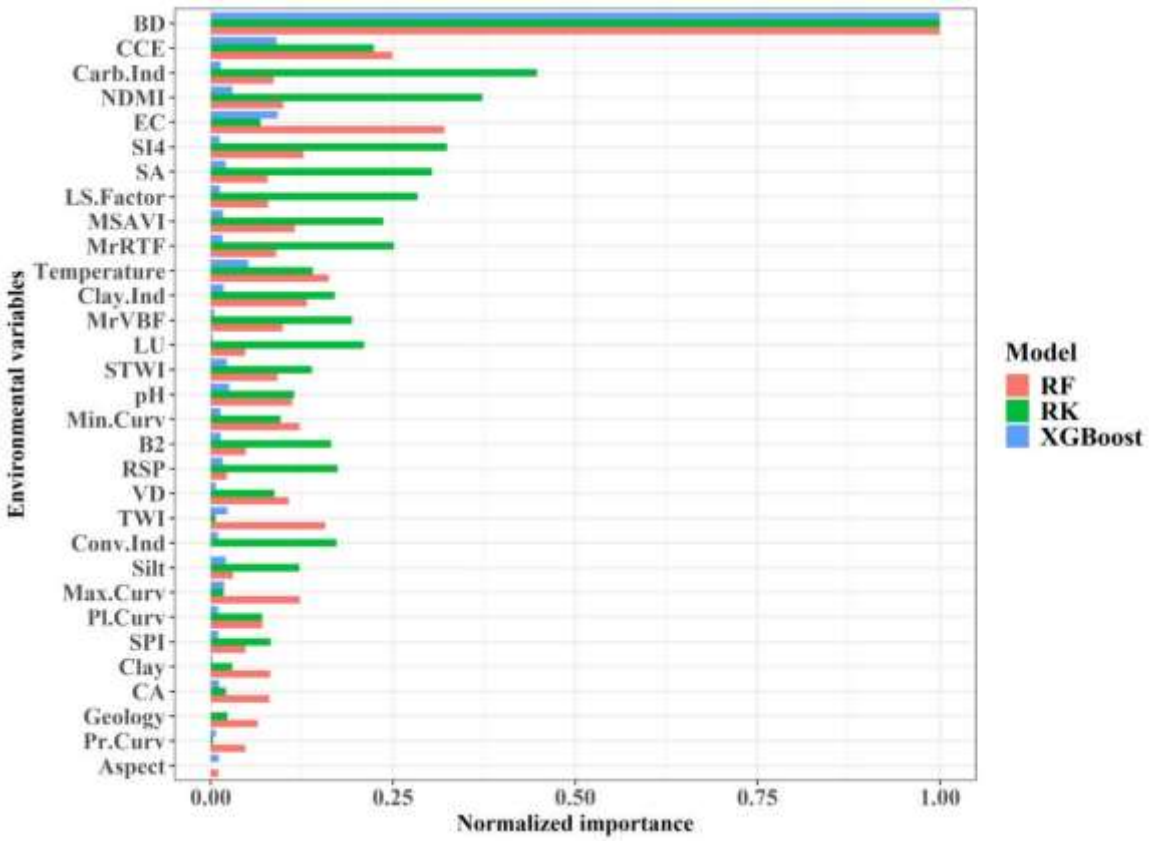
جدول ۶. مقایسه مدل‌های مورد استفاده در مطالعه

Model	R^2	RMSE	MAE	ME	RPIQ
XGBoost	۰/۶۲۸	۰/۴۷۳	۰/۳۱۶	۰/۱۴۶	۱/۹۸۰
Random Forest (RF)	۰/۵۴۵	۰/۵۰۹	۰/۳۷۲	۰/۰۷۴	۱/۸۴۱
Regression Kriging (RK)	۰/۴۳۴	۰/۵۳۴	۰/۳۹۵	۰/۰۸۸	۱/۷۵۴

اهمیت نسبی متغیرهای محیطی در مدل RK و XGBoost و RF

نتایج تحلیل اهمیت متغیرها (شکل ۶) بیانگر آن است که متغیرهای جرم مخصوص ظاهری خاک، محتوای کربنات کلسیم معادل، شاخص کربنات و NDMI ازمهم‌ترین متغیرهای کمکی در برآورد کربن آلی خاک در هر سه مدل مورد بررسی محسوب می‌شوند. در این میان، جرم مخصوص ظاهری خاک در هر سه مدل مورد بررسی، از جمله مدل RK، بالاترین میزان اهمیت را به خود اختصاص داده است. این نتیجه بیانگر آن است که جرم مخصوص ظاهری خاک در منطقه مورد مطالعه، نقش برجسته‌ای در تبیین الگوی پراکنش مکانی کربن آلی خاک ایفا می‌کند. این یافته با مطالعات پیشین همسو است؛ به‌گونه‌ای که مطالعات مختلف، ویژگی‌های فیزیکی خاک و به‌ویژه BD را از متغیرهای مؤثر در مدل‌سازی و پیش‌بینی تغییرات مکانی کربن آلی خاک معرفی کرده‌اند (Hengl et al., 2017; Wadoux et al.,

2020; Ho et al., 2024). پس از BD، Carb.Ind و CCE در رتبه‌های بالای اهمیت قرار گرفتند. این متغیرها بازتابی از شرایط ژئوشیمیایی و منشأ مادری خاک هستند و می‌توانند بر پایداری ماده آلی و سرعت تجزیه آن اثر بگذارند. در محیط‌های آهکی، حضور کربنات‌ها معمولاً با محدودیت‌هایی در انباشت ماده آلی همراه است، اما در عین حال می‌تواند از طریق تغییر در pH، فعالیت زیستی و پیوندهای کاتیونی بر رفتار کربن آلی اثر بگذارد (Cambardella et al., 1994; Wiesmeier et al., 2011). بنابراین، اهمیت بالای این متغیرها نشان می‌دهد که ویژگی‌های شیمیایی خاک در کنار ویژگی‌های فیزیکی، نقش مهمی در توزیع مکانی SOC دارند.



شکل ۶. اهمیت نسبی متغیرهای محیطی در مدل‌های RK و XGBoost و RF

NDMI نیز یکی دیگر از متغیرهای مهم در مدل‌ها بود. این شاخص به‌عنوان نماینده‌ای از وضعیت رطوبتی و در برخی موارد پوشش گیاهی سطح زمین عمل می‌کند. رطوبت بیشتر معمولاً شرایط را برای تجمع ماده آلی و کاهش سرعت تجزیه آن مساعدتر می‌کند، در حالی‌که مناطق خشک‌تر با تجزیه پذیری و محدودیت در تولید زیست‌توده متفاوت عمل می‌کنند. اهمیت NDMI با این دیدگاه سازگار است که شاخص‌های طیفی رطوبت می‌توانند به‌طور غیرمستقیم الگوهای مکانی کربن را آشکار کنند (Hengl et al., 2017; Batjes, 2014).

نکته قابل توجه، انتخاب متغیر EC به‌عنوان یکی از متغیرهای تاثیرگذار در مدل RF برای پیش‌بینی محتوای SOC بوده است. این در حالی است که دو مدل دیگر (RK و XGBoost) در پیش‌بینی مقادیر SOC وزن کمتری را برای این متغیر در نظرگرفته‌اند. بر این اساس، می‌توان گفت که اگرچه EC معیاری از وضعیت شوری و تجمع املاح در خاک است و می‌تواند رشد گیاه، تراکم پوشش گیاهی و در نتیجه ورودی بقایای آلی به خاک را تحت تأثیر قرار دهد، با این حال، با توجه به پایین بودن مقادیر EC و غیرشور بودن خاک‌های منطقه مطالعاتی، انتخاب آن به عنوان یک متغیر تاثیرگذار توسط مدل RF را می‌توان از دلایل عدم موفقیت نسبی این مدل نسبت به دو مدل دیگر برای پیش‌بینی مقادیر SOC نظر گرفت.

در کنار متغیرهای BD، Carb.Ind، NDMI برخی متغیرهای توپوگرافیِ باقی‌مانده پس از غربالگری همخطی، از جمله LS.Factor، TWI، Aspect و انواع انحنا (Profile curvature، Plan curvature، General curvature) نیز در مدل‌ها اهمیت قابل‌توجهی نشان

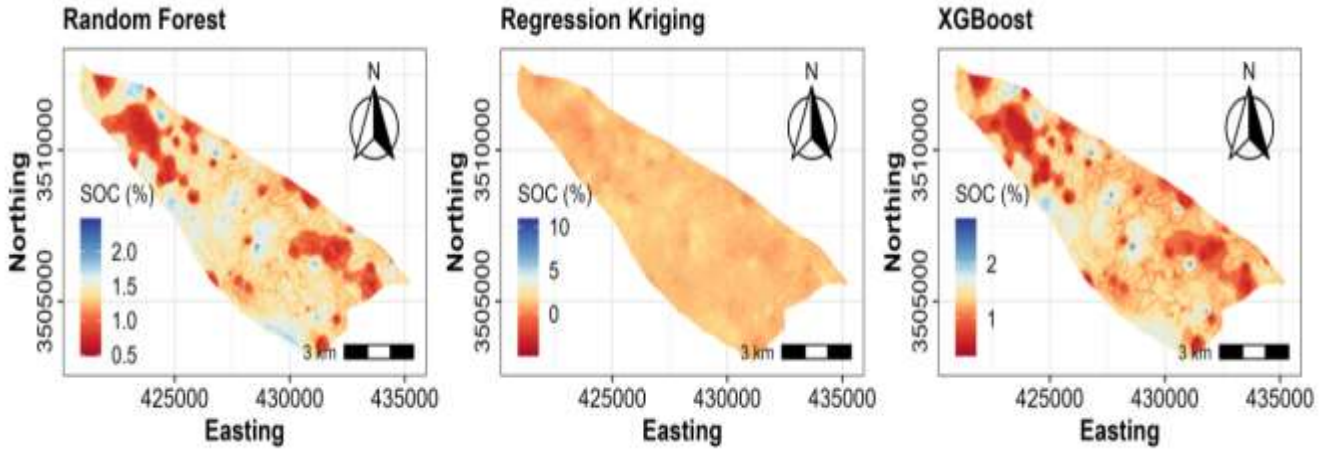
دادند. این متغیرها به‌طور غیرمستقیم از طریق کنترل فرآیندهای رواناب، تجمع رطوبت، فرسایش، رسوب‌گذاری و بازتوزیع مواد آلی بر الگوی مکانی کربن آلی خاک اثر می‌گذارند. در این میان، اگرچه شیب به دلیل همخطی از مدل نهایی حذف شد، اما بخشی از اثر آن می‌تواند در شاخص LS.Factor بازتاب یافته‌باشد. همچنین، نواحی پست‌تر و همگرا با شاخص رطوبتی بالاتر معمولاً شرایط مساعدتری برای تجمع مواد آلی دارند، در حالی‌که نواحی محدب‌تر ممکن است با خروج رواناب و کاهش ذخیره SOC مواجه شوند. این الگو با روندهای گزارش‌شده در مطالعات جدید نقشه‌برداری رقومی کربن آلی خاک نیز همسو است که بر نقش ترکیبی متغیرهای زمین‌سنجی، سنجش‌ازدور و اعتبارسنجی فضایی تأکید دارند (Ho et al., 2024).

هم‌سویی نتایج RF و XGBoost در شناسایی متغیرهای کلیدی، به‌ویژه BD، EC و CCE، نشان می‌دهد که این عوامل کنترل‌کننده‌های اصلی پراکنش مکانی کربن آلی خاک در منطقه مورد مطالعه هستند. در هر دو مدل، BD در صدر اهمیت قرار گرفته است که می‌تواند بیانگر نقش بنیادی ویژگی‌های فیزیکی خاک در توزیع مکانی کربن آلی در منطقه مطالعاتی باشد. این هم‌خوانی بین دو روش یادگیری ماشین، اعتماد به نتایج را افزایش می‌دهد، زیرا نشان می‌دهد که اهمیت این متغیرها وابسته به یک مدل خاص نیست و در دو چارچوب تحلیلی مستقل نیز تکرار شده است. این الگو با نتایج مطالعات پیشین نیز هم‌خوانی دارد؛ Wiesmeier et al., (2011) در پژوهشی که با هدف نقشه‌برداری رقومی کربن آلی خاک در یک اکوسیستم نیمه‌خشک انجام شد، گزارش کردند که جرم مخصوص ظاهری خاک و برخی متغیرهای محیطی از مهم‌ترین عوامل مؤثر در پیش‌بینی محتوای کربن آلی خاک هستند. همچنین، Hengl et al., (2017) نیز نشان دادند که استفاده هم‌زمان از متغیرهای محیطی، توپوگرافی و داده‌های کمکی می‌تواند دقت پیش‌بینی الگوهای مکانی خصوصیات خاک، از جمله کربن آلی، را به‌طور قابل‌توجهی بهبود دهد.

نقشه‌های پیش‌بینی مکانی

نقشه‌های پیش‌بینی حاصل از مدل‌های مختلف در شکل ۷ ارائه شده‌اند. بررسی این نقشه‌ها گویای الگوی نسبتاً مشابه در توزیع مکانی کربن آلی خاک است، هرچند تفاوت‌هایی در جزئیات برآوردها وجود دارد. بیشترین مقادیر کربن آلی در بخش‌های شمالی و شمال‌غربی منطقه مشاهده می‌شود. نتایج مدل‌های RF و XGBoost با تأکید بر متغیرهای محیطی، نشان می‌دهند که در این نواحی، شرایط برای تجمع کربن مهیاتر است. در مقابل، بخش‌های مرکزی و جنوب‌شرقی دارای مقادیر کمتری از کربن آلی هستند.

تحلیل اهمیت متغیرها (شکل ۴) نشان داد که تفاوت در میزان کربن آلی خاک‌ها در این مناطق، بیش از آنکه صرفاً ناشی از یک عامل باشد، تابع ترکیب پیچیده‌ای از ویژگی‌های فیزیکی و شیمیایی خاک است. همان‌طور که پیش‌تر ذکر شد، متغیرهای BD، EC و CCE به‌عنوان پیش‌ران‌های اصلی شناسایی شدند. اگرچه دامنه تغییرات EC در منطقه گسترده است، اما تحلیل‌های همبستگی (شکل ۴) تأیید می‌کند که این متغیر در کنار سایر فاکتورهای خاکی، نقش تعدیل‌کننده در توزیع SOC دارد. این یافته با مطالعات اخیر Ho et al., (2024) هم‌سو است که بیان می‌دارند در اراضی با ویژگی‌های مشابه، تعاملات متغیرهای فیزیکی–شیمیایی، عاملی تعیین‌کننده‌تر از صرفاً یک کاربری یا یک شاخص خاص است.



شکل ۷. نقشه‌های پیش‌بینی مکانی SOC با استفاده از مدل‌های RF، RK و XGBoost

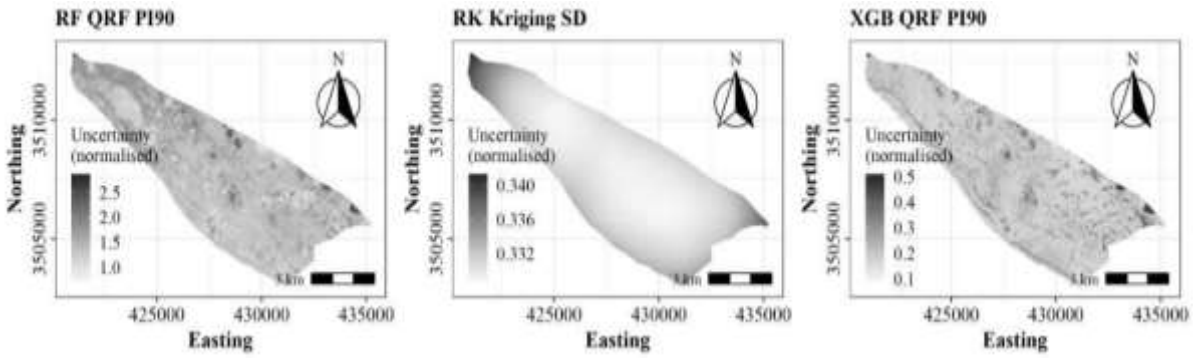
مدل RK برخلاف مدل‌های یادگیری ماشین، پهنه‌بندی یکنواخت‌تری را ارائه داده است. نکته قابل تامل در نقشه RK، بیش‌برآوردی (Overestimation) احتمالی در برخی پیکسل‌های پراکنده است که منجر به تفاوت در اعداد ارائه شده برای راهنمای نقشه RK نسبت به دو مدل دیگر شده است؛ در مقابل، مدل‌های XGBoost و RF تفکیک فضایی بهتری از الگوهای پیچیده ارائه داده‌اند. مدل XGBoost به ویژه در شناسایی نقاط با محتوای بیشتر کربن آلی در شمال منطقه، مرزهای دقیق‌تری را ترسیم کرده است. این موضوع احتمالاً به دلیل توانایی این مدل در شناسایی اثرات آستانه‌ای متغیرهایی نظیر کربنات کلسیم معادل و شاخص‌های طیفی است که در پژوهش انجام شده توسط Hengl et al., (2017) نیز به عنوان مزیت اصلی مدل‌های Boosting در مناطق ناهمگن ذکر شده است.

تفاوت بافت نقشه‌ها نشان می‌دهد که مدل‌های مبتنی بر درخت تصمیم (RF و XGBoost) به دلیل میانگین‌گیری از چندین ساختار تصمیم‌گیری، نویزهای محیطی را بهتر مدیریت کرده و پیوستگی معناداری بین اراضی با کاربری‌های مختلف و توپوگرافی منطقه ایجاد کرده‌اند. این در حالی است که در مدل RK، نوسانات موضعی ناشی از واریانس باقیمانده‌ها، وضوح روابط محیطی را تحت‌تأثیر قرار داده است.

تحلیل عدم قطعیت

نقشه‌های عدم قطعیت پیش‌بینی محتوای کربن آلی خاک که از مدل‌های مختلف به‌دست آمده‌اند، در شکل ۸ ارائه شده‌اند. بر این اساس، توزیع عدم‌قطعیت در سطح منطقه در هر سه مدل، یکنواخت نیست و به‌طور مشخص تحت تأثیر تراکم نمونه‌برداری و نیز پیچیدگی شرایط محیطی قرار دارد. در هر سه مدل، بخش‌های مرکزی منطقه عموماً عدم‌قطعیت کمتری نشان می‌دهند، در حالی‌که نواحی پیرامونی، به‌ویژه لبه‌های شمال‌غربی و بخش‌هایی از حاشیه شرقی و جنوبی، از عدم‌قطعیت بیشتری برخوردارند. در میان مدل‌های مورد بررسی، مدل RK نقشه‌ای هموارتر و پیوسته‌تر از عدم‌قطعیت ارائه می‌دهد. این الگو با ماهیت کریجینگ سازگار است؛ زیرا در این روش، عدم‌قطعیت مدل عمدتاً بر اساس ساختار مکانی باقیمانده‌ها (واریوگرام) و فاصله و آرایش نقاط نمونه‌برداری برآورد می‌شود. از این‌رو، انتظار می‌رود تغییرات مکانی عدم‌قطعیت در RK بیشتر تدریجی باشد و از الگوهای لکه‌ای و نقطه‌ای کمتر تبعیت کند (Hateffard et al., 2024.; Rossetti et al., 2022).

در مقابل، مدل‌های RF و XGBoost الگوهای ناهمگن‌تر و موضعی‌تری از عدم‌قطعیت نشان می‌دهند. این موضوع بیانگر آن است که عدم‌قطعیت در این دو مدل صرفاً تابع فاصله مکانی از نقاط نمونه‌برداری نیست، بلکه به کیفیت پوشش‌دهی داده‌های آموزشی در فضای متغیرهای کمکی نیز وابسته است؛ به عبارت دیگر، در نواحی که شرایط محیطی مشابه آن‌ها در داده‌های آموزشی کمتر وجود داشته است، پیش‌بینی مدل با اطمینان کمتری همراه است؛ بنابراین مقدار عدم‌قطعیت افزایش می‌یابد. افزون بر این، غیرخطی بودن روابط و پیچیدگی پاسخ مدل می‌تواند باعث شود تغییرات عدم‌قطعیت در مدل‌های درختی به‌صورت لکه‌ای‌تر و با جزئیات مکانی بیشتر ظاهر شود (Wadoux et al., 2020; Hateffard et al., 2024).



شکل ۸. نقشه‌های عدم قطعیت پیش‌بینی SOC با استفاده از مدل‌های RF، RK و XGB

به بیان دیگر، جایی که ترکیب ویژگی‌های محیطی از الگوهای آموخته‌شده توسط مدل فاصله بیشتری دارد، بازه پیش‌بینی گسترده‌تر و عدم‌قطعیت افزایش می‌یابد. این یافته با مطالعات پیشین نیز هم‌خوانی دارد. Lilburne et al., (2024) در مطالعه‌ای به بررسی و ارزیابی عدم‌قطعیت پیش‌بینی‌های حاصل از نقشه‌برداری رقومی داده‌های بافت خاک از پایگاه جهانی SoilGrids 2.0 پرداخته و گزارش نمودند که

پیش‌بینی‌های ارائه شده برای کشور هلند به دلیل وجود داده‌های آموزشی بیشتر از عدم‌قطعیت کمتری نسبت به نیوزیلند برخوردار هستند. همچنین Nikou and Tziachris (2022) در پژوهشی به ارزیابی توانایی مدل‌های جنگل تصادفی چندکی در پیش‌بینی و برآورد عدم‌قطعیت توزیع مکانی ماده آلی خاک پرداخته و گزارش نمودند که بسیاری از روش‌های یادگیری ماشین فاقد قابلیت ذاتی برای برآورد عدم‌قطعیت هستند. حال آن‌که، مدل‌های مبتنی بر جنگل تصادفی مانند مدل جنگل تصادفی چندکی عملکرد بسیار خوبی در پیش‌بینی ماده آلی خاک داشته و ضریب تبیین آن به‌مراتب بالاتر از روش‌های زمین‌آماری سنتی است. در همین ارتباط، Hengl et al., (2018) نیز گزارش نمودند که در مطالعات نقشه‌برداری رقومی خاک، نمایش عدم‌قطعیت باید هم‌زمان با نقشه‌های پیش‌بینی ارائه شود تا کاربر بتواند علاوه بر مقدار برآوردشده، میزان اعتماد به آن را نیز ارزیابی کند.

بر این اساس می‌توان گفت که نقشه‌های عدم‌قطعیت، علاوه بر تکمیل نقشه‌های پیش‌بینی SOC، در تصمیم‌گیری‌های مدیریتی و طراحی مطالعات آینده مفید هستند. این نقشه‌ها امکان شناسایی نواحی با قابلیت اعتماد کمتر را فراهم می‌کنند؛ بنابراین، می‌توان از آن‌ها برای اولویت‌بندی نمونه‌برداری‌های تکمیلی، به‌ویژه در حاشیه‌های شمال‌غربی، شرقی و جنوبی منطقه که عدم‌قطعیت بالاتری نشان دادند، استفاده کرد. تمرکز نمونه‌برداری‌های آینده بر این نواحی می‌تواند پوشش فضای محیطی داده‌های آموزشی را بهبود داده و در نتیجه به افزایش دقت نقشه‌هایSOC منجر شود.

نتیجه‌گیری کلی

پژوهش حاضر با هدف بهبود دقت نقشه‌برداری رقومی SOC در حوضه آبخیز کارون ۳ انجام شد. در پاسخ به هدف اول تحقیق مبنی بر مقایسه عملکرد مدل‌ها، نتایج نشان داد که مدل‌های یادگیری ماشین و در رأس آن‌ها مدل XGBoost، در مقایسه با روش زمین‌آماری رگرسیون کریجینگ (RK) و مدل جنگل تصادفی (RF)، از کارایی بالاتری در برآورد مکانی کربن آلی برخوردارند. این یافته، فرض اصلی پژوهش را تأیید می‌کند که مدل‌های تقویت‌شونده به دلیل توانایی در ثبت تعاملات پیچیده و غیرخطی میان ویژگی‌های خاکی و متغیرهای محیطی، ابزار کارآمدتری برای مدل‌سازی در مناطق دارای ناهمگنی شدید توپوگرافیکی هستند.

در راستای هدف دوم و تحلیل اهمیت متغیرهای محیطی، نتایج نشان داد که ترکیبی از ویژگی‌های ذاتی خاک (نظیر جرم مخصوص ظاهری و کربنات کلسیم معادل) و متغیرهای استخراج شده از داده‌های سنجش از دوری و مدل رقومی ارتفاع (مانند NDMI، شاخص شوری SI4 و فاکتور LS) نقش تعیین‌کننده‌ای در تبیین تغییرپذیری کربن آلی دارند. برجستگی این متغیرها نشان‌دهنده کنترل شدید فرآیندهای هیدرولوژیکی بر پویایی کربن در محیط‌های نیمه‌خشک است.

در نهایت، با تحقق هدف سوم، نقشه رقومی SOC و نقشه عدم‌قطعیت متناظر با آن تولید گردید. توزیع مکانی کربن آلی الگوی مشخصی را نشان داد که در آن نواحی شمالی و شمال‌غربی به دلیل شرایط اقلیمی و پوشش گیاهی، محتوای کربن آلی بیشتری را نسبت به بخش‌های مرکزی و جنوبی دارا هستند. تحلیل نقشه‌های عدم‌قطعیت نشان داد که قابلیت اطمینان مدل در نواحی مرکزی (با تراکم نمونه‌برداری بالاتر) حداکثر است، در حالی که در حاشیه‌های حوضه، عدم‌قطعیت افزایش می‌یابد.

به‌طور کلی، با توجه به نتایج این پژوهش پیشنهاد می‌شود در مطالعات آتی، پویایی زمانی کربن آلی با استفاده از داده‌های سری‌های زمانی سنجش از دور و داده‌های عمقی خاک برای کاهش ابهامات مدل‌سازی مورد توجه قرار گیرد.

ملاحظات اخلاقی

حامی مالی

هزینه‌های اجرای این پژوهش توسط معاونت پژوهشی، فناوری و ارتباط با جامعه دانشگاه شهید چمران اهواز (**SCU.AS1402.365**) پرداخت شده است.

مشارکت نویسندگان

مشارکت نویسندگان در مقاله مستخرج از پایان‌نامه تقریباً به شکل زیر باشد:

نویسنده اول: دانشجو: تهیه و آماده‌سازی نمونه‌ها، انجام آزمایش و گردآوری داده‌ها، انجام محاسبات، تجزیه و تحلیل آماری داده‌ها، تحلیل و تفسیر اطلاعات و نتایج، تهیه پیش‌نویس مقاله

نویسنده دوم: استاد راهنمای پایان‌نامه، طراحی پژوهش، نظارت بر مراحل انجام پژوهش، بررسی و کنترل نتایج، اصلاح، بازبینی و نهایی‌سازی مقاله

نویسنده سوم: استاد مشاور پایان‌نامه، مشارکت در طراحی پژوهش، نظارت بر پژوهش، مطالعه و بازبینی مقاله

اعلامیه هوش مصنوعی مولد و فناوری‌های مبتنی بر هوش مصنوعی در فرایند نگارش

در طول آماده‌سازی این اثر، نویسنده(گان) از ابزار هوش مصنوعی **editGPT** برای ویرایش گرامری و غلط‌های املایی احتمالی در متن انگلیسی مقاله استفاده کردند. پس از استفاده از این ابزارنویسنده(گان) مطالب را در صورت نیاز بررسی و ویرایش کرده و مسئولیت کامل محتوای نشریه را بر عهده می‌گیرند.

بیانیه دسترسی به داده‌ها

داده‌هایی پژوهش حاضر از طریق درخواست از نویسندگان قابل دسترسی است.

سپاسگزاری

از معاونت محترم پژوهشی دانشگاه دانشگاه شهید چمران اهواز به خاطر حمایت مالی در اجرای پژوهش حاضر سپاسگزاری می‌شود.

از داوران محترم به خاطر ارائه نظرهای ساختاری و علمی سپاسگزاری می‌شود.

پیروی از اصول اخلاق پژوهش

نویسندگان اصول اخلاقی را در انجام و انتشار این پژوهش علمی رعایت نموده‌اند و این موضوع مورد تأیید همه آنهاست.

تعارض منافع

بنا بر اظهار نویسندگان این مقاله تعارض منافع ندارد.

REFERENCES

Banaei, M. (1998). Map of soil moisture and temperature regimes in Iranian soils. Soil and Water Research Institute, Tehran, Iran.

Batjes, N. H. (2014). A standardized input data set for global soil carbon model evaluation. *Geoderma, 214-215*, 119–126.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test, 25*(2), 197–227.

Borůvka, L., Donátová, H., and Němeček, J. (2005). Effect of land use and soil properties on soil acidification in the Czech Republic. *Soil and Tillage Research, 78*(1), 75–85.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Cambardella, C. A., Moorman, T. B., Novak, J. M., Parkin, T. B., Karlen, D. L., Turco, R. F., and Konopka, A. E. (1994). Field-scale variability of soil properties in central Iowa soils. *Soil Science Society of America Journal, 58*(5), 1501–1511.

Carré, F., McBratney, A. B., Mayr, T., and Montanarella, L. (2007). Digital soil assessments: Beyond DSM. *Geoderma, 142*(1-2), 69–79.

Catani, F., Lagomarsini, D., Segoni, S., and Tofani, V. (2013). Landslide susceptibility estimation by random forests technique: Sensitivity and scaling issues. *Natural Hazards and Earth System Sciences, 13*(11), 2815–2831.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.

Chen, W., Li, Y., Wang, J., Chen, G., and Chen, G. (2019). Spatial prediction of soil organic matter content using machine learning algorithms. *Catena, 174*, 170–185.

Corwin, D. L., and Lesch, S. M. (2005). Apparent soil electrical conductivity measurements in agriculture. *Computers and Electronics in Agriculture, 46*(1–3), 11–43.

Dadgar, M., & Faramarzi, S.E. (2024). Assessing the performance of machine learning models for predicting soil organic carbon variability across diverse landforms. *Environmental Earth Sciences*, 83, 657. https://doi.org/10.1007/s12665-024-11960-0

De Caires, S.A., St Martin, C., Atwell, M.A., Kaya, F., Wuddivira, G., & Wuddivira, M.N. (2025). Advancing soil mapping and management using geostatistics and integrated machine learning and remote sensing techniques: A synoptic review. Discover Soil, 2: 53. https://doi.org/10.1007/s44378-025-00082-z

Fox, J., and Weisberg, S. (2019). An R companion to applied regression (3rd ed.). Sage. https://www.john-

fox.ca/Companion/
Hateffard, F., Steinbuch, L., & Heuvelink, G. B. M. (2024). Evaluating the extrapolation potential of random forest digital soil mapping. *Geoderma*, 441, 116740. https://doi.org/10.1016/j.geoderma.2023.116740
Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., … & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE, 12*(2), e0169748.
Hengl, T., Heuvelink, G. B. M., and Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. *Computers & Geosciences, 33*(10), 1301–1315.
Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., and Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ, 6*, e5518.
Hijmans, R.J. (2023). *raster: Geographic Data Analysis and Modeling*. R package version 3.6-26. https://CRAN.R-project.org/package=raster
Hocking, R. R. (2013). Methods and applications of linear models: Regression and the analysis of variance (3rd ed.). Wiley.
Ho, V. H., Nguyen, T. T., Nguyen, T. A., Pham, T. D., Bui, D. T., & Nguyen, Q. H. (2024). Random forest regression kriging modeling for soil organic carbon density estimation using multi-source environmental data in central Vietnamese forests. Modeling Earth Systems and Environment.
Hojati, S. (2019). Use of spatial statistics to identify hotspots of lead and copper in selected soils from north of Khuzestan Province, southwestern Iran, *Archives of Agronomy and Soil Science*, 65(5), 654-669. https://doi.org/10.1080/03650340.2018.1520977
Jena, R. K., Moharana, P. C., Dharumarajan, S., Sharma, G. K., Ray, P., Deb Roy, P., and Gaber, A. (2023). Spatial prediction of soil particle-size fractions using digital soil mapping in the North Eastern region of India. *Remote Sensing, 15*(5), 1295.
Kaya, F., Keshavarzi, A., Gümüş, İ., Erşahin, S., and Başayiğit, L. (2022). Evaluation of machine learning algorithms for predicting soil organic carbon stock in a semi-arid environment. *Geoderma Regional, 28*, e00466.
Keskin, H., and Grunwald, S. (2018). Regression kriging as a workhorse in the digital soil mapper's toolbox. *Geoderma*, *326*, 22–41.
Kasraei, B., Schmidt, M.G., Zhang, J., Bulmer, C.E., Filatow, D.S., Arbor, A., Pennell, T., & Heung, B. (2024). A framework for optimizing environmental covariates to improve model parsimony in digital soil mapping. *Geoderma*, 446, 116890. https://doi.org/10.1016/j.geoderma.2024.116873
Khaledian, Y., and Miller, B. A. (2020). Selecting appropriate machine learning methods for digital soil mapping. *Applied Mathematical Modelling, 81*, 401–418.
Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5), 1–26. https://doi.org/10.18637/jss.v028.i05
Lilburne, L., Helfenstein, A., Heuvelink, G.B.M., & Eger, A. (2024). Interpreting and evaluating digital soil mapping prediction uncertainty: A case study using texture from SoilGrids. *Geoderma*, 450, 117052. https://doi.org/10.1016/j.geoderma.2024.117052.
Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3), 18-22.
Matinfar, H.R., Jalali, M., & Dibaei, Z. (2021). Estimation of the amount of soil organic carbon using spectral data in the VIS-NIR-SWIR-TIR spectral range. *Journal of Water and Soil*, 35(4), 567-581.
McBratney, A. B., Santos, M. M., and Minasny, B. (2003). On digital soil mapping. *Geoderma, 117*(1-2), 3–52.
Minasny, B., and McBratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences, 32*(9), 1378–1388.
Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7, 983-999. http://jmlr.csail.mit.edu/papers/v7/
Minasny, B., and McBratney, A. B. (2016). Digital soil mapping: A brief history and some directions. *Geoderma*, *264*, 301–311.
Moharana, P. C., Meena, R. L., Nogiya, M., Jena, R. K., Sharma, G. K., Sahoo, S., and Vara Prasad, P. V. (2022). Impacts of land use on pools and indices of soil organic carbon and nitrogen in the Ghaggar flood plains of arid India. *Land, 11*(8), 1180.
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.
Moradinasab, V., Hojati, S., Landi, A., Faz Cano, A. (2021). Comparing soil development in two topo-sequences with different parent materials in part of Karoon 3 Basin, east of Khuzestan Province. *Iranian*

Journal of Soil and Water Research, 52(1), 143-159.
Mousavi, A., Shahbazi, F., Oustan, S., Jafarzadeh, A.A., & Minasny, B. (2020). Application of two data mining techniques for mapping the spatial distribution of soil organic carbon (Case study: East shore of Urmia Lake). *Journal of Water and Soil*, 34(3), 689-705.
Naimi, B. (2023). *usdm: Uncertainty analysis for species distribution models* (R package version 2.1-7) [Computer software]. https://CRAN.R-project.org/package=usdm
Nikou, M., & Tziachris, P. (2022). Prediction and uncertainty capabilities of quantile regression forests in estimating spatial distribution of soil organic matter. *ISPRS International Journal of Geo-Information*, 11(2), 130. https://doi.org/10.3390/ijgi11020130
Pebesma, E.J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, 30(7), 683–691. https://doi.org/10.1016/j.cageo.2004.03.012
Pebesma, E., & Bivand, R. (2005). Classes and methods for spatial data in R. *R News*, *5*(2), 9–13. https://CRAN.R-project.org/doc/Rnews/
Rahmati, F., Hojati, S., Rangzan, K., & Landi, A. (2024). Modelling and digitally mapping of soil organic carbon in Semiron County employing several machine learning algorithms. *Agricultural Engineering*, 47(4), 585-605.
R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
Rohmer, J., Belbeze, G., & Guyonnet, D. (2024). Insights into the prediction uncertainty of machine-learning-based digital soil mapping through a local attribution approach. *SOIL*, 10(2), 679–697. https://doi.org/10.5194/soil-10-679-2024.
Rossetti, C., Piccini, C., & Conforti, M. (2024). Spatial prediction of soil organic carbon: Combining machine learning with residual kriging in an agricultural lowland area (Lombardy region, Italy). *Geoderma*, *448*, 116953. https://doi.org/10.1016/j.geoderma.2024.116953.
Stavi, I., and Lal, R. (2015). Achieving zero net land degradation: Challenges and opportunities. *Journal of Arid Environments, 112*, 44-51.
Stoorvogel, J. J., Kempen, B., Heuvelink, G. B. M., & De Bruin, S. (2009). Implementation and evaluation of existing knowledge for digital soil mapping in Senegal. *Geoderma*, *149*(3-4), 161–170.
Talebian, M., Raisossadat, S.N., and Rahimzadeh, F. (1999). Dehdez geological quadrangle map of Iran, no. 6052. Scale 1:100000. Ministry of Industry and Mines and Geological Survey f Iran.
Tassinari, D., Richer-de-Forges, A. C., and Brus, D. J. (2016). Accounting for geologic and geomorphic processes in modeling soil depth and soil classes in Barro Colorado Nature Monument, Panama. *Geoderma*, *277*, 21–38.
Tziolas, N., Tsakiridis, N., Heiden, U., & van Wesemael, B. (2024). Soil organic carbon mapping utilizing conventional neural networks and Earth observation data, a case study in Bavaria state Germany. *Geoderma*, 444, 116867. https://doi.org/10.1016/j.geoderma.2024.116867
Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.
Wadoux, A. M. J.-C., Minasny, B., and McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews, 210, 103359.
Wang, X., Li, Z., and Li, P. (2009). Topography and spatial variability of soil physical properties. *Scientia Agricola*, *66*(3), 338–352.
Wang, H., Zhang, X., Wu, W., and Liu, H. (2021). Prediction of Soil Organic Carbon under Different Land Use Types Using Sentinel-1/-2 Data in a Small Watershed. *Remote Sensing*, *13*(7), 1229.
Wei, T., & Simko, V. (2024). *R package 'corrplot': Visualization of a correlation matrix* (Version 0.95) [Computer software]. https://github.com/taiyun/corrplot
Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, 260p. https://doi.org/10.1007/978-3-319-24277-4
Wickham, H., & Bryan, J. (2026). *readxl: Read Excel files* (R package version 1.5.0) [Computer software]. https://readxl.tidyverse.org
Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2026). *dplyr: A grammar of data manipulation* (R package version 1.2.1) [Computer software]. https://dplyr.tidyverse.org
Wiesmeier, M., Barthold, F., Blank, B., and Kögel-Knabner, I. (2011). Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. *Plant and Soil*, *340*(1), 7–24.
Wiesmeier, M., Urbanski, L., Hobley, E., Lang, B., von Lützow, M., Marin-Spiotta, E., van Wesemael, B., Rabot, E., Ließ, M., Garcia-Franco, N., Wollschläger, U., Vogel, H.-J., & Kögel-Knabner, I. (2019). Soil

organic carbon storage as a key function of soils - A review of key indicators and biological pathways. Geoderma, 333, 149-162.

Wilding, L. P. (1985). Spatial variability: Its documentation, accommodation and implication to soil surveys. In *Proceedings of the Soil Spatial Variability Workshop* (pp. 166–194). Las Vegas, NV.

Zhang, Y., Wang, H., & Li, X. (2022). Peatland restoration reduces carbon emissions by altering pore characteristics and water content. *Land*, *11*(9), 1646.

Zhang, C., Chen, Y., Wei, Y., Yu, P., Hong, Y., Hu, Y., Wang, J., & Shi, Z. (2024). Unraveling the threshold and interaction effects of environmental variables on soil organic carbon mapping in plateau watershed. *Geoderma*, 450, 117032. https://doi.org/10.1016/j.geoderma.2024.117032