Research Article

Agricultural Engineering., 47(4) (2024) 585-605 ISSN (P): 2588-526X

DOI: 10.22055/agen.2024.47815.1740 ISSN (E): 2588-5944

Modelling and Digitally Mapping of Surface Soil Organic Carbon in Semirom County Employing Several Machine Learning Algorithms

F. Rahmati[1], S. Hojati[2], K. Rngzan[3] and A. Landi[2]

1. Former PhD Student, Department of Soil Science, Faculty of Agriculture, Shahid Chamran University of Ahvaz, Ahvaz, Iran
2. Professor,Department of Soil Science, Faculty of Agriculture, Shahid Chamran University of Ahvaz, Ahvaz,, Iran
3. Professor,Department of Remote Sensing and GIS, Faculty of Earth Sciences, Shahid Chamran University of Ahvaz,, Ahvaz, Iran

Received: 22 August 2024 Accepted: 16 December 2024 *Corresponding Author: s.hojati@scu.ac.ir

Abstract

Introduction: Knowledge of the spatial distribution of soil organic carbon (SOC) is a practical tool for determining sustainable land management strategies. Estimating carbon content and stocks is essential for carbon sequestration, greenhouse gas emissions and national carbon balance inventories. Accurate mapping of SOC's spatial distribution is a key assumption for soil resource management and land use planning. Over the last two decades, the utilization of data mining approaches in the spatial modeling of SOC using machine learning algorithms has gained considerable attention. The digital environment requires continuous soil maps at local and regional scales; however, such information is not always available at the necessary scale. Therefore, the digital soil mapping (DSM) approach is a crucial solution for quantifying and assessing variations in soil properties such as SOC, using remotely sensed indices and digital elevation model (DEM) as the most commonly used ancillary data for SOC prediction. In this context, the data mining techniques serves as the pathway to create digital soil maps. This study aims to compare two common machine learning algorithms, random forest and multiple linear regression, in the digital mapping of surface SOC in Semirom County, Isfahan Province. Digital maps of SOC using these two algorithms were created, and the most important variables affecting the spatial distribution of SOC in the study area were reported.

Materials and Methods: A total of 200 surface soil samples (0-10 cm) were collected from the Semirom area (51º 17' - 52º 3' E; 30º 42' - 31º 51' N) in Isfahan, Iran. According to the reports from synoptic meteorological station, the annual average temperature ranged from 7.5 to 12.5 ◦C, and the annual precipitation varied between 350 and 450 mm. The soil moisture and temperature regimes are classified as Xeric and Mesic, respectively. Soil samping was conducted using the Global Positioning System (GPS) at the surface layer (0-10 cm). The preparation of the soil samples involved air drying, pounding and softening, followed by seiving through a 2 mm mesh. The amount of organic carbon in the samples was determined using the Walkley-Black method. Additionally, in order to evaluate the effect of other soil properties on the organic content of the soils, laboratory analyses were conducted, including measurements of saturated soil moisture content, soil texture, soil pH in saturated pastes, electrical conductivity of the soil saturation extracts and the calcium carbonate equivalent of the soils, following standard laboratory protocols.

In this research, auxiliary variables including terrain parameters and vegetation indices were derived from DEM and the Landsat 8 OLI satellite images using ArcMap version 10.4.10 and SAGAGIS version 6.0.4. All auxiliary layers were then converted to raster format using the "raster" package and

merged using the “Covstack” function. The values of all environmental covariates at each sampling point were extracted into a single file using the “extract” function from the “sp” package in the RStudio environment. Subsequently, using SPSS software version 19 and the principal component analysis (PCA) method, the most important auxiliary covariates among the 29 variables used in this research were identified for the modeling process. The dataset was then split into two groups; calibration (80%) and validation (20%) subsets. Finally, SOC contents were predicted and mapped using multiple linear regression (MLR) and random forest (RF) algorithms in the RStudio environment. The MLR and RF algorithms were executed using the “lm” and “randomForest” packages, respectively. Five different statistics were utilized to evaluate the performance of each model, including the coefficient of determination (R^2), bias, root mean square error (RMSE), normalized RMSE (nRMSE), and mean bias error (MBE).

Results and Discussion: Based on the descriptive analysis of the soil samples, the soils in the study area were characterized as non-saline, alkaline, and calcareous. The SOC contents of the soils ranged from 0.3 % to 2.2% with a mean value of 0.89%. The coefficient of variation for the SOC contents was 21.7%, which classifies the soils of the study area as having moderate variability, according to the values proposed by Wilding (1985). The results of the PCA indicated that the most important auxiliary variables for the modeling process included slope aspect, channel network base level, catchment slope, total curvature, height, longitudinal curvature, mass balance index, modified catchment area, slope degree, slope length, topographic position index, vertical distance to the channel network, soil adjusted vegetation index, transformed vegetation index, difference vegetation index, ratio vegetation index, and general curvature.These variables explained 80% of the total variance in the study area. A comparison of two different SOC prediction models, demonstrated that the RF model (ntree = 1000 and mtry = 10) outperformed the MLR model, with R^2, RMSE, nRMSE, and bias values of 0.79, 0.12, 0.13, and 0.002 respectively. The five most important variables identified by the RF algorithm for predicting SOC contents in the study area were the transformed vegetation index, ratio vegetation index, soil adjusted vegetation index, and slope degree. The final map of surface SOC distribution reveals that, although the RF algorithm provided better predictions than the MLR model, it also resulted in overestimation and/or underestimation of the minimum and maximum values of surface SOC contents, respectively.

Conclusion: The results of this study demonstrated that the RF regression algorithms outperformed the MLR method, thanks to its ability to account for the nonlinear and complex relationships between SOC content and environmental covariates.

Key Words: Environmental covariates, machine learning, performance, spatial distribution

مهندسی زراعی (مجله علمی کشاورزی) جلد ۴۷ شماره۴، زمستان ۱۴۰۳

مقاله پژوهشی

مدل‌سازی و نقشه‌برداری رقومی کربن آلی در خاک سطحی اراضی شهرستان سمیرم با استفاده از برخی روش‌های یادگیری ماشین

فاطمه رحمتی[۱]، سعید حجتی[۲*]، کاظم رنگزن[۳] و احمد لندی[۴]

۱- دانش آموخته دکتری گروه مهندسی علوم خاک، دانشکده کشاورزی، دانشگاه شهید چمران اهواز، اهواز، ایران
۲- استاد گروه مهندسی علوم خاک، دانشکده کشاورزی ، دانشگاه شهید چمران اهواز، اهواز، ایران
۳- استاد گروه سنجش ازدور و GIS، دانشکده علوم زمین، دانشگاه شهیدچمران اهواز، اهواز، ایران

تاریخچه مقاله

دریافت: ۱۴۰۳/۰۶/۰۱

پذیرش نهایی: ۱۴۰۳/۰۹/۲۶

کلمات کلیدی:
یادگیری ماشین،
متغیرهای کمکی،
عملکرد،
پراکنش مکانی

* عهده دار مکاتبات

Email: s.hojati@scu.ac.ir

چکیده

آگاهی از توزیع مکانی کربن آلی خاک گامی موثر در دستیابی به استفاده پایدار از اراضی و تعیین استرازی‌های مدیریتی مربوط به آن است. از این رو، این مطالعه با هدف مدل‌سازی و نقشه‌برداری رقومی کربن آلی خاک سطحی (۱۰-۰ سانتی‌متری) شهرستان سمیرم با استفاده از روش‌های رگرسیون جنگل تصادفی و رگرسیون خطی چند متغیره انجام شد. به این منظور ۲۰۰ نمونه خاک سطحی به صورت منظم و با فواصل نمونه‌برداری ۵ کیلومتر × ۵ کیلومتر از سطح منطقه برداشت گردید و سپس کربن آلی نمونه‌ها با استفاده از روش واکلی- بلک اندازه‌گیری شد. در پایان، نقشه رقومی کربن آلی در خاک سطحی منطقه با روش‌های مزبور و به کمک متغیرهای کمکی استخراج شده از مدل رقومی ارتفاع و تصاویر ماهوارهٔ لندست ۸ در محیط نرم‌افزار RStudio تهیه شد. یافته‌های این مطالعه حاکی از آن است که الگوریتم جنگل تصادفی برای برآورد میزان کربن آلی خاک به ترتیب با مقادیر RMSE و R^2 معادل ۰/۱۲ و ۰/۷۹ نسبت به روش رگرسیون خطی چندمتغیره با RMSE و R^2 معادل ۰/۱۹۲ و ۰/۵۷ پیش‌بینی‌های بهتری ارائه داده است. نتایج نشان داد که مهم‌ترین متغیرهای محیطی مؤثر بر توزیع کربن آلی خاک در منطقه مطالعاتی در مدل‌های مورد استفاده یکسان نیستند. به‌گونه‌ای که در مدل جنگل تصادفی شاخص‌های مستخرج از پوشش گیاهی و در رگرسیون خطی چندمتغیره شاخص‌های توپوگرافی نقش بیشتری در توزیع کربن آلی داشته است. بررسی نقشه نهایی پراکنش کربن آلی خاک در منطقه مطالعاتی نشان داد که تخمین‌های انجام شده با روش جنگل تصادفی اگرچه در مقایسه با روش رگرسیون خطی چندمتغیره تخمین‌های بهتری را ارائه داده اما در تخمین مقادیر کمینه و بیشینه مقادیر کربن آلی سطحی خاک‌ها موفق نبوده است.

مقدمه

مواد آلی بر بسیاری از ویژگی‌های خاک‌ها از جمله حاصلخیزی، ساختمان، نفوذ آب در خاک، ظرفیت نگهداشت آب، تراکم و فعالیت میکروبی خاک تأثیر دارد (۱۵ و ۲۷). به علاوه، کربن آلی نقشی حیاتی در پایداری محیط زیست دارد و از شاخص‌های مهم در ارزیابی کیفیت و سلامت خاک است؛ از این رو، مدل‌سازی و نمایش تغییرات توزیع مکانی کربن آلی به‌ویژه در خاک‌های مناطق خشک و نیمه خشک می‌تواند در مدیریت پایدار اراضی بسیار حائز اهمیت تلقی گردد. نقشه‌برداری ویژگی‌های خاک به روش سنتی زمان‌بر و پرهزینه است. این در حالی است که در سال‌های اخیر استفاده از تکنیک‌های نقشه‌برداری رقومی خاک با صرف حداقل هزینه و وقت برای پیش‌بینی مکانی خصوصیات خاک مورد توجه بسیاری از محققان قرار گرفته است. معادله ینی (۱۱) پایه و اساس نقشه‌برداری رقومی خاک است که در آن تشکیل خاک وابسته به پنج فاکتور اقلیم، موجودات زنده، پستی و بلندی، مواد مادری و زمان می‌باشد. مک براتنی و همکاران[1] (۱۵) معادله ینی را بازبینی کردند و مدل اسکورپن ($S_{c,a}=f(s,c,o,r,p,a,n)$) را ارایه دادند که برای توصیف کمی روابط بین خاک و عوامل مکانی دیگر به کار می‌رود و شامل هفت عامل سایر خصوصیات خاک دریک نقطه (s) خصوصیات اقلیمی(c)، پوشش گیاهی، جانوری یا فعالیت انسانی (o)، توپوگرافی (r)، مواد مادری (p)، سن (a)، و موقعیت جغرافیایی (n) است. در مدل اسکورپن $S_{c,a}$ کلاس‌های خاک یا خصوصیات خاک است.

سه جزء اصلی معادله اسکورپن ویژگی‌های خاک، متغیرهای محیطی و مدل‌های (توابع) ارتباط دهنده بین ویژگی-های خاک و متغیرهای محیطی می‌باشند. متغیرهای کمکی محیطی در واقع نماینده عوامل خاک‌سازی هستند. اکثر متغیرهای محیطی با استفاده از مدل‌های رقومی ارتفاع، داده‌های طیفی سنجش از دور و نقشه‌های حاصل از مطالعات گذشته به دست می‌آید (۱). استفاده از داده‌های سنجش از دور و مدل‌های رقومی ارتفاع این امکان را فراهم می‌سازد تا تأثیر ژئومورفولوژی و موقعیت خاک در زمین نما، تکامل خاک، پراکنش مکانی رطوبت بر اساس انحناءهای اراضی، پوشش گیاهی، توپوگرافی و عمق رسوبات زمین‌شناسی و انسان در محاسبات در نظر گرفته شود (۱ و ۲۱).

روش های مدل سازی متعددی برای نشان دادن ارتباط متغیرهای محیطی و ویژگی‌های خاک در نقشه‌برداری رقومی خاک وجود دارد. از بین مدل‌ها می‌توان به رگرسیون خطی چند متغیره و جنگل تصادفی اشاره کرد. مدل‌های رگرسیون خطی یک مدل آماری است که برای تخمین ویژگی های خاک از همبستگی یا ارتباط خطی بین ویژگی خاک و متغیر های کمکی استفاده می‌کند. در روش رگرسیون خطی چند متغیره، می‌توان رابطه بین متغیر هدف و بیش از یک متغیر کمکی را به‌صورت همزمان ارزیابی نمود. این روش در مقابل اطلاعات نادرست، حساسیت بالایی دارد و ورود چنین داده‌هایی ممکن است منجر به خطاهای بزرگی در نتایج به دست آمده شود. علاوه بر این، برای استفاده از این روش متغیر ها باید توزیع نرمال داشته باشند و تغییر آنها از یک رابطه خطی پیروی کند (۳۶).

الگوریتم جنگل تصادفی یک روش قدرتمند در فرآیند مدل سازی است که پیشرفت قابل توجهی در روش های داده‌کاوی ارایه داده است (۲۱). الگوریتم جنگل تصادفی مدل توسعه یافته‌ای از مدل طبقه‌بندی و رگرسیون درختی است که داده‌های مشاهداتی و متغیرهای کمکی را به دفعات برای به دست آوردن ارتباط بهینه بین متغیر پاسخ و متغیرهای مستقل و انجام تخمین‌های بعدی تفیک و دسته‌بندی می‌کند. در این روش، مجموعه‌ای از شرط‌ها به صورت یک الگوریتم درختی با ساختار منطقی (اگر-آنگاه[2]) برای طبقه‌بندی یا پیش‌بینی کمی یک متغیر به کار می‌رود (۱).

روش جنگل تصادفی به‌واسطه مزایای فراوانی مانند کارایی بالا در پیش‌بینی، قابل اجرا بودن برای حجم فراوانی از داده ها، برآورد داده های گم شده، مقاوم بودن در برابر نویز و عدم تأثیرپذیری از واریانس متغیرهای محیطی و خاک در سال های اخیر مورد توجه بسیاری از محققان قرار گرفته است (۱ و ۲۱) و نتایج بسیاری از تحقیقات نشان داده است که مدل جنگل

1- McBratney *et al.*

2- if-then

تصادفی روش مناسبی برای برآورد ویژگی‌های خاک می‌باشد. هنگل و همکاران[1] (۶) نقشه رقومی خصوصیات خاک را در آفریقا با دو مدل رگرسیون چندگانه خطی و جنگل تصادفی تهیه کردند و نتیجه گرفتند که مدل جنگل تصادفی در مقایسه با مدل رگرسیون خطی چندگانه دقت بیشتری داشته است. رادیانتو و همکاران[2] (۲۵) از سه مدل شبکه عصبی مصنوعی، جنگل تصادفی و رگرسیون درختی برای تهیه نقشه رقومی کربن آلی خاک در اندونزی استفاده کردند و بهترین ضرایب تبیین را برای این مدل‌ها به ترتیب ۰/۸۶- ۰/۵۹، ۰/۹۸- ۰/۹ و ۰/۹۹- ۰/۹۵ گزارش کردند. هیونگ و همکاران[3] (۷) نیز در پژوهشی به این نکته اشاره کردند که مدل های جنگل تصادفی و رگرسیون درختی به دلیل سرعت پردازش پارامترها و تفسیر آسان داده های خروجی، مدل‌های مناسبی به منظور استفاده در اهداف طبقه‌بندی هستند

بررسی منابع انجام شده نشان می‌دهد تا کنون مطالعه‌ای در مورد مدل‌سازی تغییرات مکانی کربن آلی و شناسایی متغیرهای زودیافت تاثیرگذار بر پراکنش مکانی کربن آلی خاک در منطقه سمیرم استان اصفهان با استفاده از الگوریتم‌های مرسوم داده-کاوی انجام نشده است. از این رو، این پژوهش با اهداف مقایسه دو الگوریتم جنگل تصادفی و رگرسیون خطی چند متغیره، شناسایی مهترین عوامل محیطی کنترل کننده تغییرات کربن آلی و تهیه نقشه روند تغییرات کربن آلی در این منطقه انجام شد.

مواد و روش‌ها

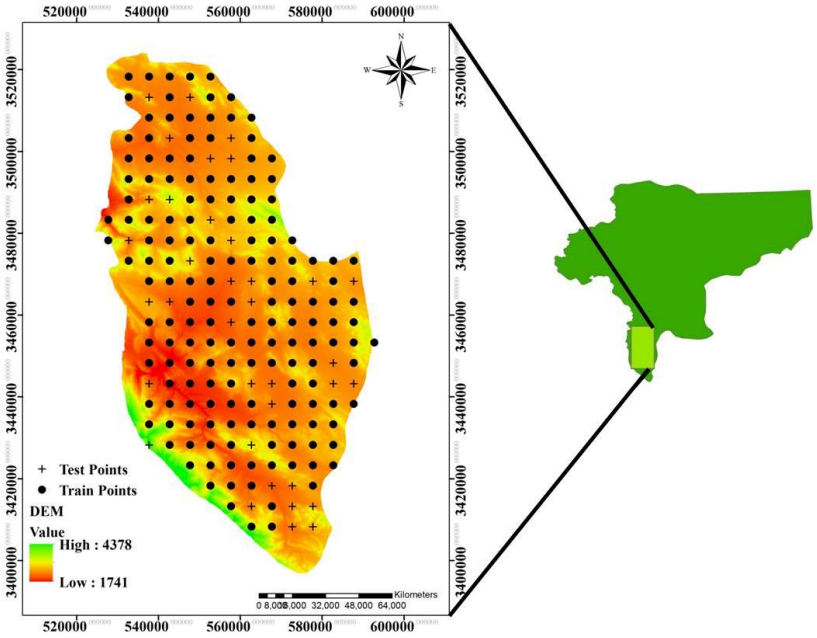
موقعیت منطقه مورد مطالعه

شکل ۱ موقعیت منطقه مورد مطالعه در شهرستان سمیرم استان اصفهان را نشان می‌دهد. شهرستان سمیرم در جنوب غربی استان اصفهان با مساحت ۵۲۲۴ کیلومترمربع بین طول‌های جغرافیایی ´ ۱۷ °۵۱ تا´ ۳ °۵۲ شرقی و عرض´۴۲ °۳۰ تا ´۵۱ °۳۱ شمالی واقع شده است.

میانگین دمای شهرستان ۷/۵ تا ۱۲/۵ درجه است و بارندگی سالانه ۳۵۰ تا ۴۵۰ میلی‌متر می‌باشد. ارتفاع منطقه ۲۰۰۰ تا ۳۰۰۰ متر بالاتر از سطح دریا قرار دارد و رژیم رطوبتی خاک، زریک[4] و رژیم حرارتی آن، مزیک[5] می‌باشد. در این منطقه تشکیلات دوره دوم زمین‌شناسی تا عهد حاضر مشاهده می‌شود. از نظر چینه‌شناسی، سنگ‌های مربوط به دوره‌های مختلف در کل منطقه قابل مشاهده است. نهشته‌های کواترنری شامل تراس‌های جدید و قدیم رودخانه‌های فصلی و دائمی است که عمدتاً اراضی زراعی را شامل می‌باشد. در مجموع ۳۷/۶ درصد کل اراضی منطقه را اراضی مواج یا همان فلات‌ها یا دشت‌های بریده می‌پوشانند. خاک‌های منطقه اغلب متعلق به رده‌های انتی‌سولز، آلفی‌سولز و اینسپتی‌سولز می‌باشند (۱۰).

شهرستان سمیرم از نظر موقعیت قرارگیری در حاشیه جنوب شرقی رشته کوه‌های زاگرس قرار گرفته به طوری که کوه‌های دنا به صورت حصاری در غرب و جنوب غربی آن کشیده شده است. بخش مرکزی قسمت‌های مرکزی و شمالی شهرستان را شامل می‌شود و بخش پادنا قسمت‌های جنوبی شهرستان را در برمی‌گیرد. شهر سمیرم واقع در مرکز شهرستان با ارتفاع ۲/۴۶۰ از سطح دریا به صورت دشت نیمه هموار با ارتفاع زیاد و شیب فراوان توسط کوه‌ها محصور شده است. با توجه به نقشه کاربری اراضی (شکل ۲) بیشترین مساحت منطقه مراتع فقیر با پوشش گیاهی کم می‌باشد که با رنگ بنفش نشان داده شده است. بر این اساس، ۷۲ درصد مساحت منطقه مراتع کوهستان‌های نسبتاً سرد است و دارای گیاهان یکساله می‌باشد. نقشه ژئومرفولوژی منطقه نشان می‌دهد در این منطقه نهشته‌های کواترنری وجود دارند که کاربری کشاورزی دارند (۱۰). در بخش مرکزی در مناطق با ارتفاع زیاد و در منطقه جنوبی در مناطق با ارتفاع کم سنگ‌های کربناته وجود دارد. این سنگ‌ها در منطقه مرکزی، با چشم‌انداز کوه و تپه توأم می‌باشند.. در حالی که درمنطقه جنوبی این سازند با اراضی مسطح و کم ارتفاع توأم شده است. آب و هوای این شهرستان بر اساس تقسیم‌بندی کوپن جزء مناطق معتدل سرد با تابستان‌های گرم و خشک و در ارتفاعات جنوبی جزء مناطق معتدل سرد با تابستان‌های خنک وخشک می‌باشد (۱۰).

1- Hengl *et al.*
2- Rudiyanto *et al.*
3- Heung *et al.*
4- Xeric
5- Mesic



شکل(۱) منطقه مورد مطالعه که نقاط آموزش (علامت دایره) و آزمون (به صورت +) را نشان می‌دهد

Figure (1) The study area showing training (Circle symbol) and testing (+) points

نمونه‌برداری و تجزیه و تحلیل آزمایشگاهی

برای نمونه‌برداری از منطقه مورد مطالعه بعد از جمع‌آوری و بررسی اطلاعات اولیه و نقشه‌ها با استفاده از نرم افزار گوگل ارث موقعیت جغرافیایی تعداد ۲۰۰ نقطه با فواصل نمونه‌برداری ۵ کیلومتر در ۵ کیلومتر به صورت یک شبکه منظم تعیین گردید. سپس با استفاده از سیستم موقعیت یاب جهانی[1] (GPS) از لایه سطحی خاک (۱۰-۰ سانتی‌متری) نمونه‌برداری انجام شد. آماده‌سازی نمونه‌های خاک شامل هوا خشک کردن، کوبیدن و نرم کردن نمونه‌های برداشت شده انجام و سپس نمونه‌ها از الک دو میلی‌متری عبور داده شدند. آنگاه میزان کربن آلی نمونه‌ها به روش واکلی- بلک (۳۵) تعیین گردید. همچنین به منظور ارزیابی تأثیر ویژگی‌های خاک بر میزان کربن آلی خاک، تجزیه‌های آزمایشگاهی شامل تعیین درصد رطوبت اشباع خاک به روش وزنی (۸) بافت خاک به روش هیدرومتری (۲)، واکنش خاک (pH) در گل اشباع با استفاده از دستگاه pH متر (۱۶)، قابلیت هدایت الکتریکی عصاره اشباع خاک با استفاده از دستگاه هدایت‌سنج الکتریکی (۲۷) و مقدار آهک به روش تیترا سیون برگشتی (۲۰) انجام گردید.

انتخاب متغیرهای محیطی و مدل‌سازی مکانی

در این پژوهش، تعداد ۲۸ متغیر کمکی شامل شیب[2]، جهت شیب[3]، سطح مبنای شبکه زهکشی[4]، فاصله عمودی تا شبکه زهکشی[5]، شیب حوضه[6]، انحنای کلی[7]، انحنای طولی[8]، انحنای حداکثر[9]، ارتفاع از سطح دریا[10]، شاخص توازن جرم[11]، مساحت اصلاح شده حوضه[12]، طول شیب[13]، شاخص موقعیت توپوگرافی[14]، شاخص

1- Global Positioning System

2- Slope
3- Aspect
4- Channel Network Base Level
5- Vertical distance to channel network
6- Catchment Slope
7- Total Curvature
8- Longitudinal Curvature
9 - Maximum Curvature
10- Height
11- Mass Balance Index
12- Modified catchment area
13- Slope length
14- Topographic position Index

پوشش گیاهی نرمال شده[1]، شاخص پوشش پوشش گیاهی تعدیل شده برای خاک[2]، شاخص پوشش گیاهی تغییر یافته[3]، شاخص پوشش گیاهی تغییریافته تصحیح شده[4]، شاخص تفاضل پوشش گیاهی[5]، شاخص پوشش گیاهی نسبی[6]، انحنای عمومی[7]، مساحت ویژه[8]، شاخص خیسی توپوگرافی[9]، عمق دره[10]، شاخص ناهمواری توپوگرافی[11]، بافت سطح زمین[12]، شاخص همواری دره با درجه تفکیک بالا[13] و شاخص انباشت جریان[14] با استفاده از مدل رقومی ارتفاع با قدرت تفکیک مکانی ۹۰ متر و باندهای چندگانه تصاویر ماهواره لندست ۸ با استفاده از نرم‌افزار های ArcMap نسخه ۱۰/۴/۱ و SAGAGIS نسخه ۶/۴/۰ (۱۳) تهیه شدند (۵). شکل ۲ بعضی از متغیرهای محیطی تهیه شده در این پژوهش را نمایش می‌دهد.

سپس با استفاده از نرم‌افزار SPSS نسخه شماره ۱۹ و روش تجزیه به مؤلفه‌های اصلی (PCA)، از بین ۲۸ متغیر کمکی مورد استفاده در این پژوهش، مهم‌ترین متغیرهای کمکی برای فرآیند مدل سازی انتخاب گردیدند (۹). برای پیش‌بینی پراکنش ویژگی‌های خاک‌ها در منطقه مورد مطالعه، مقادیر متغیرهای محیطی و داده‌های مربوط به ویژگی‌های خاک در هر نقطه نمونه‌برداری شده به عنوان ورودی وارد محیط نرم‌افزار RStudio شدند و مقادیر متناظر آن‌ها در نقاط نمونه‌برداری با استفاده از تابع extract در بسته نرم‌افزاری sp استخراج گردید. چگونگی استخراج این پارامترها در روش ارائه شده توسط هنگل و روزیتر (۴ و ۵) تشریح گردیده است. لازم به ذکر است که پیش از استفاده از تابع extract برای استخراج مقادیر عددی متغیرهای محیطی در نقاط نمونه‌برداری شده، ابتدا تمام لایه‌های اطلاعاتی کمکی با استفاده از بسته raster به فرمت رستری تبدیل شده و سپس با استفاده از تابع Covstack با یکدیگر ادغام شدند.

روش رگرسیون خطی چند متغیره

در این روش، بر اساس رگرسیون، رابطه بین متغیر هدف (کربن آلی خاک) و بیش از یک متغیر کمکی ارزیابی می‌شود (معادله ۱).

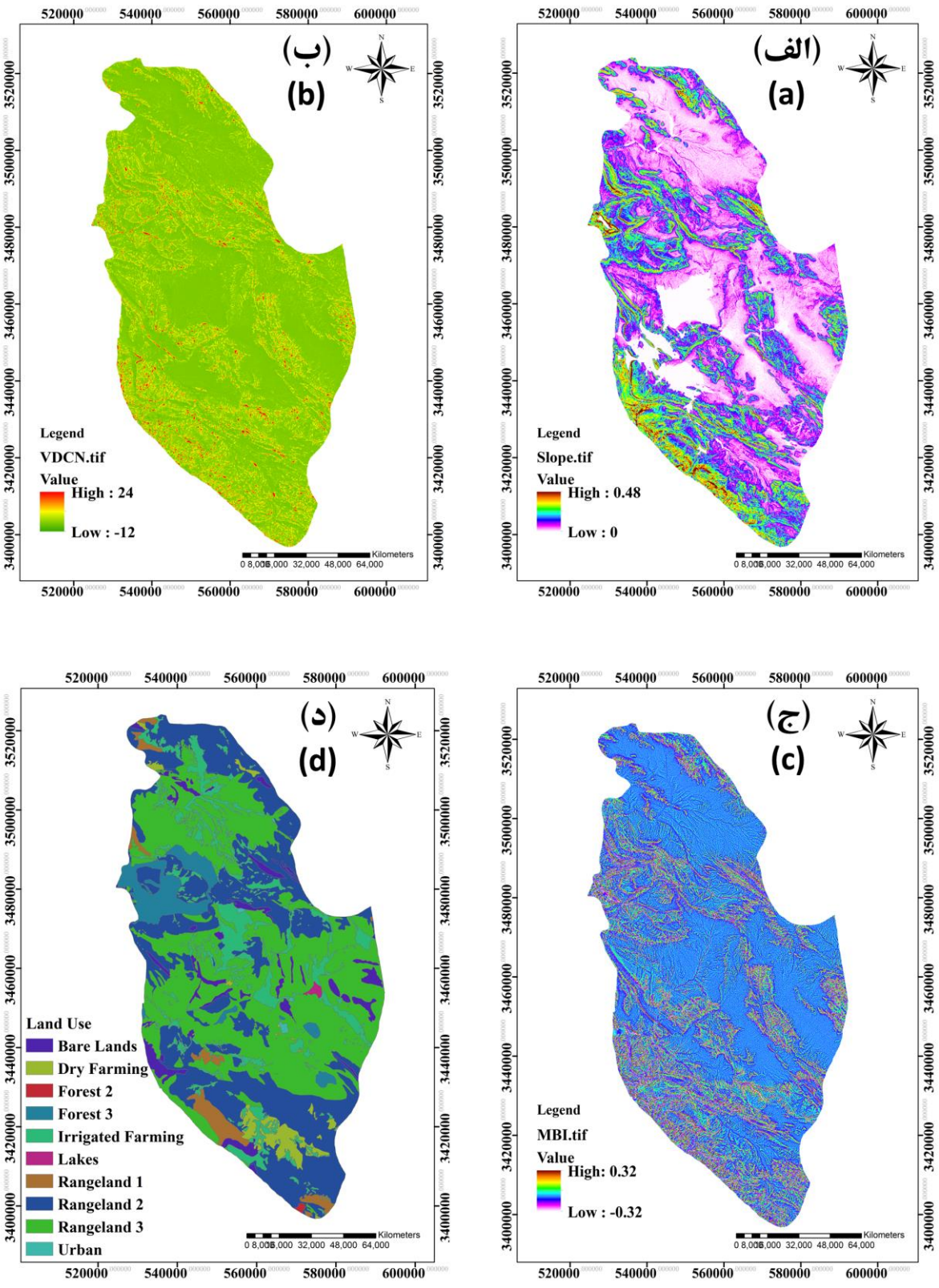
$$y_i = \beta_0 + \sum_{k=1}^{P} \beta_k x_{ik} \qquad (۱)$$

در معادله ۱، y_i مقدار متغیر هدف در یک مشاهده معین β_0 مقدار عرض از مبدا، β_k ضریب رگرسیون، x_{ik} داده‌های هر متغیر کمکی مورد نظر مربوط به نقطه مشاهداتی و P تعداد متغیرهای کمکی است. لازم به توضیح است که در روش رگرسیون خطی چند متغیره فرض بر خطی بودن روابط بین متغیرهای کمکی و متغیر هدف است. در این پژوهش برای اجرای الگوریتم رگرسیون خطی چند متغیره از بسته "lm" در محیط RStudio استفاده شد (۳). اما از آن‌جایی که ممکن است همیشه روابط بین متغیرهای کمکی و متغیر هدف خطی نباشد. لذا در این پژوهش از رگرسیون جنگل تصادفی نیز استفاده شد.

رگرسیون جنگل تصادفی

مدل جنگل تصادفی یکی از الگوریتم‌های پرکاربرد یادگیری ماشین محسوب می‌شود. رویکرد جنگل تصادفی مبتنی بر روش‌های جدید ترکیب اطلاعات است. این مدل دقیق و سریع است و به جای رشد دادن یک درخت، در این الگوریتم تعداد زیادی درخت مستقل تصمیم‌گیری (صدها یا هزارها) برای برقراری ارتباط بین متغیرهای کمکی و خصوصیات خاک رشد داده می‌شود (۱).

1 - Normalized Difference Vegetation Index
2- Soil Adjusted Vegetation Index (SAVI)
3- Transformed Vegetation Index (TVI)
4- Corrected Transformed Vegetation Index (CTVI)
5- Difference Vegetation Index (DVI)
6- Ratio Vegetation Index (RVI)
7- General Curvature
8- Surface Area
9- Topogrphic Wetness Index
10- Valley Depth
11- Topographic Ruggedness Index
12- Terrain Surface Texture
13- Multi-resolution Valley Bottom Flatness
14- Flow Accumilation



شکل(۲) نقشه‌های (الف) شیب بر حسب درجه (ب) نقشه فاصله عمودی تا شبکه زهکشی، (ج) نقشه شاخص توازن جرم و (د) کاربری اراضی در منطقه مطالعاتی

Figure (4) (a) The slope (%), (b) Vertical distance to channel network, (c) Mass Balance Index, and (d) Land Use maps of the study area

ایجاد هر درخت تصمیم شامل دو مرحله است. مرحله اول، شامل ایجاد گره ها و انشعابات می باشد؛ گره در حقیقت اختصاص دادن متغیر های مستقل به گروه های تعریف شده، براساس متغیر وابسته می‌باشد. انشعاب نیز عبارت است از انتخاب نقطه شکست از طریق انتخاب متغیری که بتواند بهترین شاخه‌های جدید را ایجاد کند. مرحله دوم، مرحله توقف و هرس است. در واقع مرحله توقف، شامل قواعدی است که میزان توسعه انشعاب در گره ها را محاسبه می‌کند. هرس نیز شامل حذف شاخه‌هایی است که اثر ناچیزی در مقادیر تخمینی حاصل از مدل نهایی دارند.

هدف از این مرحله به حداقل رساندن خطای پیش‌بینی است (۱). در این روش، یک نمونه به طور تصادفی از کل داده‌ها انتخاب شده و یک درخت روی این نمونه ساخته می شود. سپس در هر گره درخت، گروهی تخمین‌گر از کل تخمین‌گرها انتخاب می‌شوند و بهترین انشعاب با استفاده از این تخمین‌گر ها تعیین می‌گردد و به این شکل خطای کلی مدل کاهش می‌یابد. سپس مدل سازی با روش جنگل تصادفی، ابتدا روی داده های آموزشی سپس روی داده های اعتبار سنجی اجرا می شود و در نهایت مدلی انتخاب می‌شود که دارای کمترین خطا است. در این مدل، تقریباً یک سوم داده‌های اصلی در ایجاد هر درخت استفاده نمی‌شوند. در این مدل، تقریباً یک سوم داده‌های اصلی در ایجاد هر درخت استفاده نمی شوند. چون این نمونه‌ها در آموزش استفاده نشده‌اند، و در مرحله آزمون مدل مورد استفاده قرار می‌گیرند (۱). برای اینکه مدل جنگل تصادفی یک مدل پیش‌بینی تولید کند، باید دو پارامتر مشخص شود. تعداد درختان رگرسیون در جنگل (ntree) و تعداد ویژگی های تصادفی انتخاب شده در هر گره (mtry) از جمله این عوامل هستند.

الگوریتم جنگل تصادفی برای تعیین اهمیت متغیر های کمکی به طور پیش فرض از دو روش کاهش ضریب جینی[1] یا درصد افزایش در مقدار میانگین مربعات خطا (%IncMSE) استفاده می‌کند که در این مطالعه از روش دوم یعنی در صد افزایش میانگین مربعات خطا برای تعیین اهمیت متغیر های کمکی استفاده گردید. در این روش ابتدا برای هر درخت، خطای پیش بینی برای نمونه‌های خارج از کیسه (OOB) محاسبه می‌گردد. سپس این کار با جابجایی تمامی متغیر های پیش‌بینی کننده تکرار می‌گردد و در پایان، میانگین اختلاف بین این دو مرحله در رابطه با تمامی درختان رشد یافته در جنگل محاسبه می‌شود و با انحراف معیار تفاوت‌ها نرمال می‌شود. (۳۰). این کار برای تمامی متغیرهای مورد استفاده در مدل سازی متغیر هدف انجام و در نهایت نمودار اهمیت متغیرهای به کاربرده شده تهیه شد.

در این پژوهش برای اجرای الگوریتم جنگل تصادفی از بستهٔ نرم‌افزاری randomForest در محیط RStudio استفاده شد. برخی دیگر از بسته‌های استفاده شده در این مطالعه شامل readxl به منظور فراخوانی فایل‌های اکسل، GGally به منظور بررسی همزمان همبستگی بین متغیر ها و توزیع داده ها و MASS به منظور شناسایی داده‌های گم شده است. همچنین از تابع importance در بسته randomForest برای تعیین اهمیت متغیرهای کمکی به‌کاررفته در فرآیند مدل‌سازی استفاده شد. از روش نمونه‌گیری بوت استرپ (نمونه‌گیری با جایگزینی به‌گونه‌ای که یک نمونه می‌تواند بیش از یک بار در فرآیند نمونه‌گیری انتخاب شود) و با استفاده از بسته caret به منظور تفکیک داده های آموزشی (حدوداً ۸۰ درصد) و آزمون (حدوداً ۲۰ درصد) و اعتبارسنجی نتایج استفاده گردید.(۱).

ارزیابی نتایج مدل‌ها

در این پژوهش برای ارزیابی نتایج مدل سازی کربن آلی خاک و بررسی دقت مدل از ضریب تبیین (معادله ۳)، شاخص میانگین انحراف خطا (معادله ۴)، میانگین ریشه دوم مربعات خطا (معادله ۵)، میانگین ریشه مربعات خطای نرمال شده (معادله ۶) و بایاس (معادله ۷) استفاده شد. در آمار تابع اریبی یا بایاس یک برآوردگر، اختلاف بین امید ریاضی مقادیر واقعی و تخمین زده شده را نشان می‌دهد.

1- Gini

$$R^2 = \left[\frac{\sum_{i=1}^{n}(Z(xi) - Z(ave))(Z^*(xi) - Z^*(ave))}{\sqrt{\sum_{i=1}^{n}(Z(xi) - Z(ave))^2(Z^*(xi) - Z^*(ave))^2}}\right]^2 \quad (۳)$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n} Z(xi) - Z^*(xi) \quad (۴)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\left(Z^*(x_i) - Z(x_i)\right)^2\right]} \quad (۵)$$

$$nRMSE = RMSE/Mean \quad (۶)$$

$$Bias = Z(ave) - Z^*(ave) \quad (۷)$$

در معادله‌های ۳ تا ۷، n تعداد نمونه‌ها، $Z^*(xi)$ مقدار پیش‌بینی شده، $Z^*(ave)$ میانگین مقادیر پیش‌بینی شده، $Z(xi)$ مقدار اندازه‌گیری شده و $Z(ave)$ میانگین مقادیر اندازه‌گیری شده است.

تهیه نقشه پراکنش مکانی کربن آلی

نقشه پراکنش مکانی کربن آلی در خاک های سطحی منطقه مطالعاتی با استفاده از مقادیر پیش‌بینی شده از الگوریتم‌های جنگل تصادفی و رگرسیون خطی چندمتغیره استفاده گردید. برای درون‌یابی در نقشه‌های تولیدی نیز از روش وزن‌دهی معکوس فاصله استفاده شد که تمامی این فرآیندها در محیط نرم افزار ArcMap انجام شد.

نتایج و بحث

در جدول (۱) خلاصه آماری داده‌های خاکی ویژگی‌های خاک مورد بررسی ارائه شده است. بر این اساس، میانگین کربن آلی خاک در منطقه مطالعاتی برابر با ۰/۸۹ درصد به‌دست آمده است که بر اساس طبقه‌بندی سایس (۳۱) برای خاک‌های مناطق نیمه خشک در کلاس با مقادیر زیاد کربن آلی خاک طبقه‌بندی می‌شود. بر این اساس، مقدار حداقل و حداکثر کربن آلی در منطقه مطالعاتی به ترتیب ۰/۳ و ۲/۱۷ درصد می‌باشد. براساس طبقه‌بندی وایلدینگ (۳۹) ضریب تغییرات کمتر از ۱۵ درصد نشان دهنده تغییرپذیری کم، ضریب تغییرات ۱۵ تا ۳۵ درصد نشان دهنده تغییرپذیری متوسط و ضریب تغییرات بالاتر از ۳۵ درصد نشان دهنده تغییرات زیاد ویژگی‌های خاک در منطقه می‌باشند. بر این اساس، تغییرات کربن آلی، رس، سیلت و EC در منطقه در کلاس متوسط قرار دارد. تغییرات مقادیر کربنات کلسیم معادل و pH در منطقه کم و تغییرپذیری مقدار شن در منطقه زیاد است.

شکل ۳ نتایج همبستگی بین ویژگی‌های خاک را نشان می‌دهد. بر این اساس، درصد کربن آلی با واکنش خاک (pH) همبستگی منفی معنی‌دار (۰/۲۲۳-) در سطح ۰/۰۰۱ دارد. در این رابطه به نظر می‌رسد که با افزایش pH شرایط برای تجزیه موا آلی خاک فراهم تر می‌شود و لذا محتوای کربن آلی خاک را حت تاثیر قرار میدهد. نتایج اسکالبرگ (۲۹) پیرامون تغییرات pH در لایه‌های مختلف خاک نیز حاکی از آن است که محتوای pH خاک‌ها می‌تواند ارتباط معنی‌داری با کربن آلی خاک داشته باشد. همچنین یافته‌های این مطالعه حاکی از آن است که بین محتوای سیلت با درصد کربنات کلسیم معادل (۰/۸۵۴) و هدایت الکتریکی (۰/۶۲۸) همبستگی مثبت معنی‌داری در سطح ۰/۰۰۱ و با شن همبستگی منفی (۰/۷۵۹) معنی‌دار در سطح ۰/۰۰۱ وجود دارد. بر این اساس، می‌توان چنین نتیجه‌گیری نمود که عمده کربنات کلسیم در خاک‌های مورد مطالعه عمدتاً در اندازه سیلت قرار دارند و با افزایش مقدار سیلت خاک‌ها بر فراوانی آهک در خاک‌های مورد مطالعه افزوده می‌شود.

جدول(۱) توصیف آماری ویژگی‌های خاک در منطقه مطالعاتی

Table (1) Descriptive statistics of soil properties in the study area

ویژگی آماری Statistic	سیلت (٪) Silt (%)	شن (٪) Sand (%)	رس(٪) Clay (%)	کربن آلی (٪) Organic Carbon (%)	کربنات کلسیم معادل Calcium (٪) Carbonate Equivalent (%)	واکنش خاک pH	محتوای رطوبت وزنی خاک (٪) Soil moisture content (%)	رسانایی الکتریکی خاک (dS/m) Electrical Conductivity (dS/m)
میانگین (Mean)	40	31	28	0.9	31.6	7.6	36.0	0.6
حداکثر (Max)	56	70	15	2.2	70.0	8.2	46.0	1.5
حداقل (Min)	10	5	40	0.3	5.0	7.0	26.0	0.3
دامنه (Range)	46	65	25	1.9	65.0	1.2	2.0	1.2
انحراف معیار (Standard Deviation)	10	16	5	0.2	3.9	0.1	3.1	0.2
ضریب تغییرات (٪) (Coefficient of Variation (%))	26	50	19.0	21.7	12.4	1.3	8.6	28.7

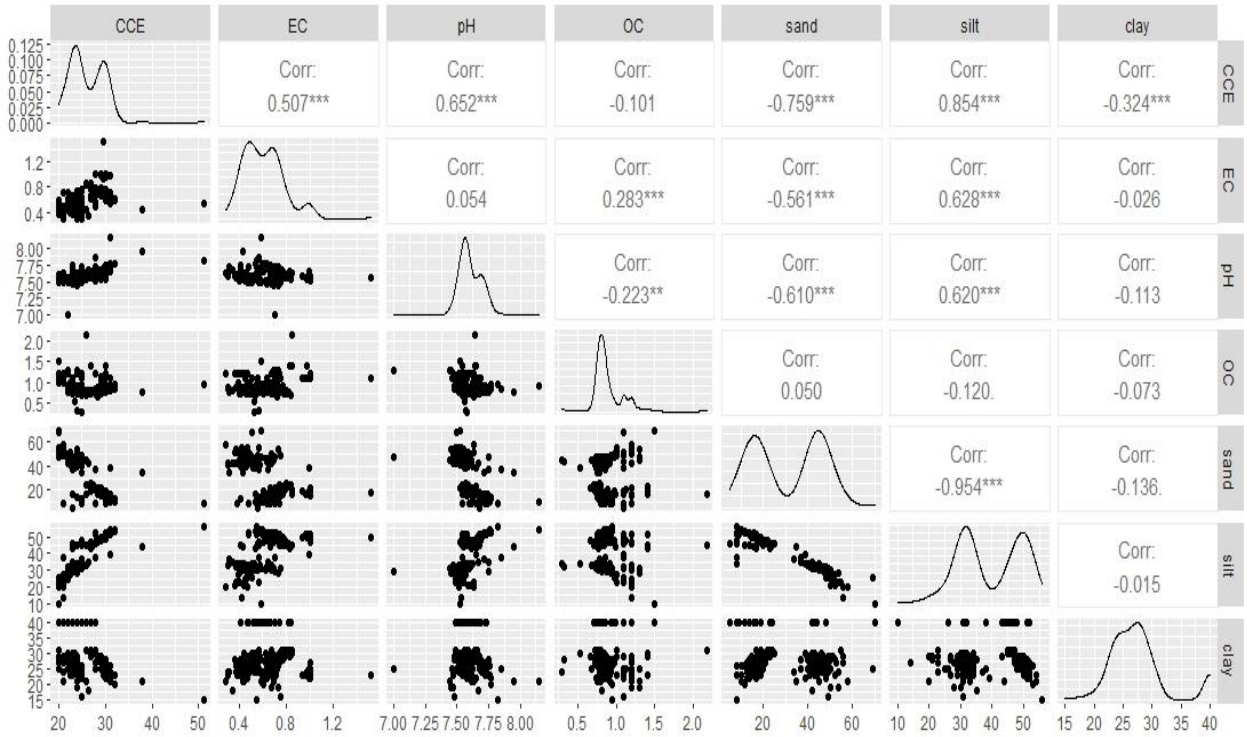
انتخاب متغیرهای کمکی برای فرآیند مدل‌سازی

جدول ۲ نتایج آزمون کرویت KMO و بارتلت را نشان می‌دهد. مقدار این شاخص همواره در محدوده صفر و ۱ قرار دارد. در صورتی که مقدار این شاخص بیشتر از ۰/۷ و نزدیک به ۱ باشد داده های مورد نظر برای تحلیل عاملی مناسب هستند. در مقابل، اگر مقدار این شاخص کمتر از ۰/۵ باشد برای تحلیل عاملی مناسب نخواهد بود. حال آن‌که، اگر مقدار این شاخص بین ۰/۵ تا ۰/۷ باشد، می‌توان با احتیاط به تحلیل عاملی پرداخت (۹). بر این اساس، در پژوهش حاضر داده‌های مورد نظر برای تحلیل عاملی مناسب هستند.

شکل ۴ نمودار اسکری (بازویی) و جدول ۳ نتایج تجزیه مؤلفه‌های اصلی را نشان می‌دهد. بر این اساس و با در نظر گرفتن مقادیر ویژه[1] بزرگتر از ۱، تعداد ۷ مولفه اصلی برای ادامه مسیر و انتخاب متغیرهای موثر و کارآمد در فرآیند مدل‌سازی انتخاب شد که این مؤلفه‌ها در مجموع بیش از ۸۰ درصد از تغییرات کربن آلی را در منطقه مطالعاتی توجیه می‌نمایند. درون هر مؤلفه اصلی متغیرهای کمکی که دارای بیشترین مقادیر همبستگی بودند، پس از انطباق با ضریب همبستگی بین آنها، در صورت وجود همبستگی بالا (بیشتر از ۰/۶)، متغیر با بالاترین مقدار همبستگی نگه داشته شد و بقیه حذف گردیدند (۴۰).

بر این اساس، مهمترین متغیرهای کمکی مورد استفاده در فرآیند مدل‌سازی شامل جهت شیب، سطح مبنای شبکه زهکشی، شیب حوضه، انحنای کلی، ارتفاع، انحنای طولی، شاخص توازن جرم، مساحت اصلاح شده حوضه، شیب، طول شیب، شاخص موقعیت توپوگرافی، فاصله عمودی تا شبکه زهکشی، شاخص پوشش گیاهی تعدیل شده برای خاک، شاخص پوشش گیاهی تغییر یافته، شاخص تفاضل پوشش گیاهی، شاخص پوشش گیاهی نسبی، و انحنای عمومی بودند.

1- Eigenvalue



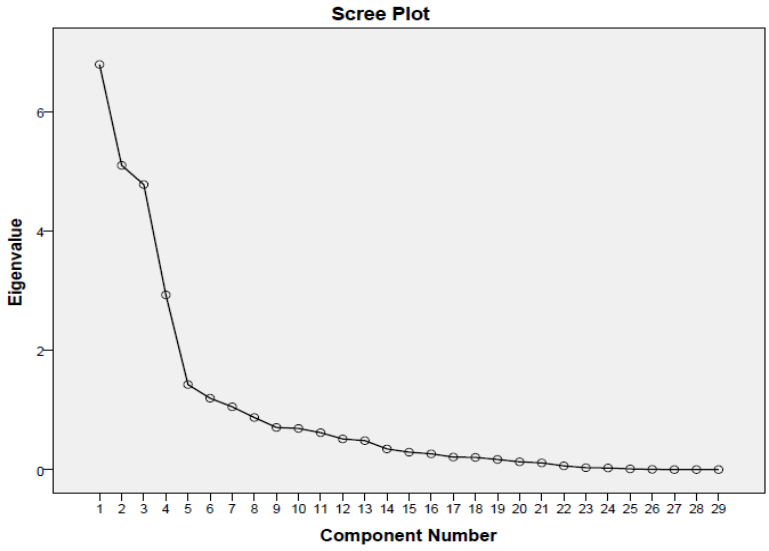
شکل(۳) ضریب همبستگی پیرسون بین ویژگی‌های خاک

Figure (3) Correlation coefficients between soil properties

جدول (۲) نتایج آزمون KMO و بارتلت

Table (2) Results of KMO and Bartlett`s Test

پارامتر		مقدار عددی
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.788
Bartlett`s Test of Sphericity	Approx. Chi-Square	3182.879
	Df	120
	Sig	0.000



شکل(٤) نمودار اسکری (بازویی) برای تعیین تعداد مولفه‌های مناسب

Figure (4) The scree plot to determine suitable number of components

جدول(۳) واریانس تجزیه مؤلفه های اصلی متغیرهای محیطی
Table (3)The PCA variance of environmental variable

مؤلفه Component	مقادیر ویژه اولیه Initially Eigenvalues			مجموع مربعات مؤلفه‌های استخراج شده Extraction Sums of Squared Loadings			مجموع مربعات مؤلفه ها پس از چرخش Rotation Sums of Squared Loadings		
	کل Total	واریانس (٪) variance (%)	درصد تجمعی Cumulative Percent	کل Total	واریانس (٪) variance (%)	درصد تجمعی Cumulative Percent	کل Total	واریانس (٪) Variance (%)	درصد تجمعی Cumulative Percent
PC1	6.7	23.4	23.4	6.7	23.4	23.4	6.5	22.5	22.5
PC2	5.09	17.5	40.9	5.09	17.6	40.9	3.7	13.1	35.6
PC3	4.7	16.4	57.4	4.7	16.4	57.4	3.7	12.9	48.6
PC4	2.9	10.1	67.5	2.9	10.1	67.5	3.2	11.3	59.9
PC5	1.4	4.9	72.4	1.4	4.9	72.4	3.1	10.7	70.6
PC6	1.19	4.1	76.5	1.19	4.1	76.5	1.6	5.5	76.2
PC7	1.05	3.6	80.2	1.05	3.6	80.2	1.1	3.9	80.2

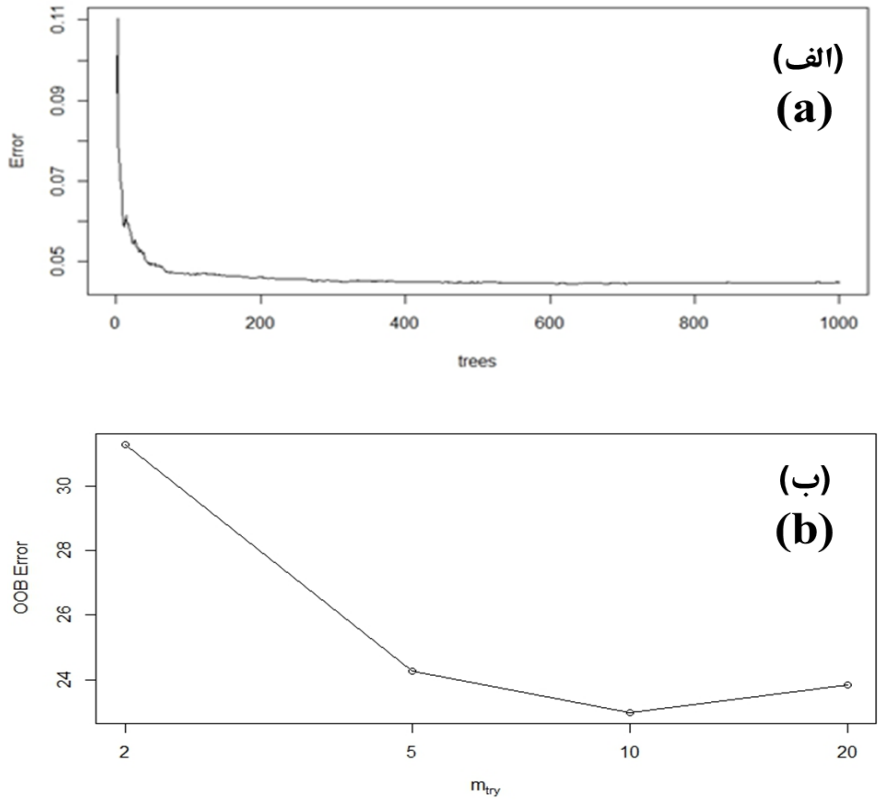
تعیین پارامترهای الگوریتم جنگل تصادفی

در الگوریتم جنگل تصادفی تعیین دو پارامتر ntree و mtry اهمیت فراوانی دارد. بر این اساس، با افزایش تعداد مؤلفه‌های کمکی به عنوان ورودی مدل، مقدار خطای OOB کاهش می‌یابد. شکل ۵ الف نشان دهنده رابطه تعداد درختان و مقدار خطا در الگوریتم جنگل تصادفی است. بر این اساس، همانگونه که مشاهده می‌شود با افزایش تعداد درختان، مقدار خطا کاهش می‌یابد، به گونه‌ای که با افزایش تعداد درختان در جنگل تصادفی ساخته شده به ۱۰۰۰ درخت کمترین مقدار خطای تخمین به دست آمد. به همین دلیل برای افزایش دقت مدل، تعداد ntree (۱۰۰۰) به عنوان مقدار بهینه انتخاب گردید هر گره حاوی اطلاعاتی در مورد تعداد در آن گره و توزیع مقادیر متغیر وابسته است. نمونه‌ها موارد موجود در گره ریشه همه مشاهدات موجود در مجموعه آموزش است. در این پژوهش تعداد گره، ۲۴ در نظر گرفته شد. شکل ۵ ب ارتباط بین مقدار خطای تخمین خارج از کیسه (OOB) و تعداد متغیرهایی که به صورت تصادفی در هر بار نمونه‌گیری انتخاب می‌شوند (mtry) را نشان می‌دهد. با افزایش تعداد متغیرهای کمکی به ۱۰ متغیر به عنوان ورودی مدل، مقدار خطای OOB کاهش می‌یابد. بر اساس شکل ۵ب، همچنین با افزایش تعداد متغیرهای کمکی به بیش از ۱۰ متغیر مقدار خطای پیش‌بینی را به طور فزاینده‌ای بالا برده است. از این رو، در نقطه‌ای که مقدار خطای OOB کمینه مقدار خود را دارد، بهترین مقدار پارامتر mtry به دست می‌آید که از آن برای تخمین ویژگی‌های خاک استفاده می‌گردد. بنابراین مقدار mtry بهینه در این مطالعه ۱۰ انتخاب گردید.

تعیین اهمیت متغیرهای کمکی برای پیش‌بینی کربن آلی در مدل‌های مورد استفاده

شکل ۶ اهمیت متغیر های کمکی مورد استفاده در مدل‌سازی کربن آلی خاک را با الگوریتم‌های جنگل تصادفی و رگرسیون خطی چندمتغیره نشان می‌دهد. همانگونه که مشاهده می گردد متغیر های تاثیر گذار تشخیص داده شده در الگوریتم‌های مورد بررسی با یکدیگر متفاوت هستند. بر این اساس، در الگوریتم جنگل تصادفی شاخص های پوشش گیاهی تغییر یافته، پوشش گیاهی نسبی، شاخص پوشش گیاهی تعدیل کننده اثر خاک و شیب به عنوان پراهمیت‌ترین متغیر ها در پیش‌بینی کربن آلی خاک شناسایی شده‌اند. تقی‌زاده مهرجردی و همکاران[1] (۳۲) برای تهیه نقشه رقومی کلاس خاک و کربن آلی خاک در شهربانه در استان کردستان، از متغیر های کمکی داده های سنجش از دور و پارامتر های سرزمینی استفاده کردند و به این نتیجه رسیدند که مناسبترین متغیرهای کمکی جهت پهنه‌بندی کربن آلی خاک، شاخص پوشش گیاهی، شاخص رس، شیب، جهت شیب، انحنای سطح، شاخص همواری دره با در جه تفکیک بالا و شاخص خیسی بوده است.

1- Taghizadeh-Mehrjardi *et al*



شکل (۵) (الف) میزان خطا برای تعیین مقدار بهینه ntree (ب) میزان خطای OOB برای تعیین مقدار mtry
Figure (5) (a) The error to determining optimal ntree and (b) the OOB error to determining optimal mtry

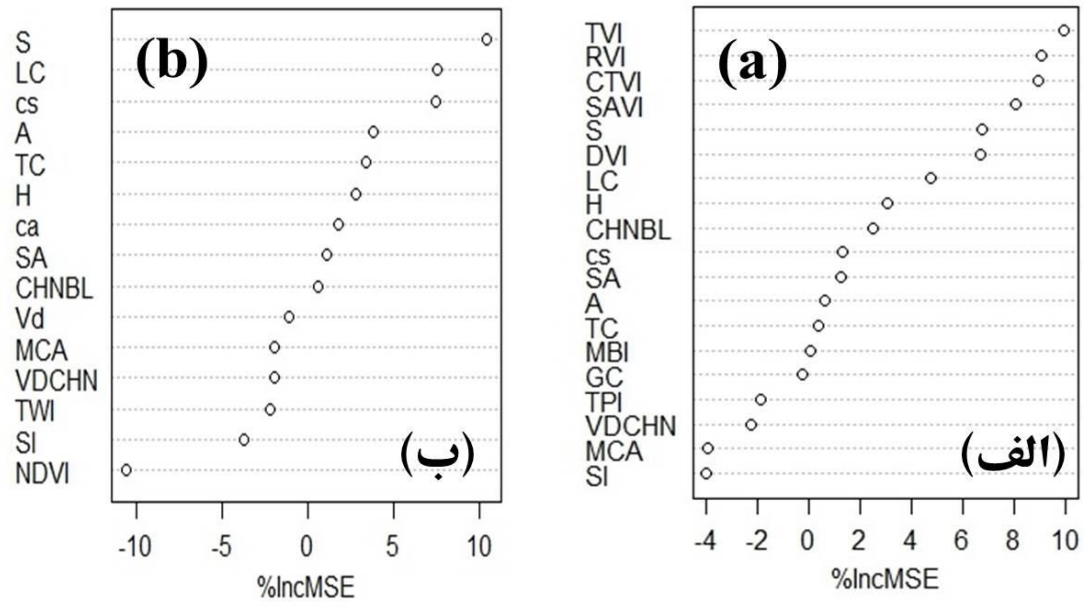
همانطور که در شکل ۶ مشاهده می‌شود، در الگوریتم جنگل تصادفی مهم‌ترین متغیرهای محیطی مؤثر بر توزیع کربن آلی خاک به‌ترتیب شامل شاخص‌های پوشش گیاهی (به دلیل برگشت بقایای گیاهی به خاک) و توپوگرافی می‌باشند؛ در صورتی که در روش رگرسیون خطی چندمتغیره، متغیرهای کمکی مرتبط با توپوگرافی اعم از درجه شیب، انحنای طولی، جهت شیب و انحنای کلی در مقایسه با شاخص‌های پوشش گیاهی به عنوان تاثیرگذارترین عوامل کنترل‌کنندهٔ پراکنش کربن آلی خاک در منطقه مطالعاتی تشخیص داده شده است. به اعتقاد نویسندگان تفاوت مشاهده شده بین دو روش در انتخاب متغیرهای تاثیرگذار در پیش‌بینی محتوای کربن آلی خاک تا حدودی به ماهیت عملکردی آنها برمی‌گردد؛ به گونه‌ای که در الگوریتم جنگل تصادفی علاوه بر روابط خطی از روابط غیرخطی نیز در شناسایی مولفه‌های تاثیرگذار در برآورد کربن آلی استفاده کرده است، حال آن‌که روش رگرسیون خطی چند متغیره تنها پارامترهایی را که با محتوای کربن آلی خاک رابطه مستقیم خطی دارند به عنوان متغیرهای تاثیرگذار معرفی نموده است. مصلح و همکاران[1] (۱۸) اظهار کردند که پارامترهای مستخرج از مدل رقومی ارتفاع، حتی در مناطق دارای شدت پستی وبلندی کم، به عنوان متغیرهای محیطی مناسب در مدل‌سازی کلاس‌ها و خصوصیات خاک‌ها محسوب می‌شوند. وانگ و همکاران[2] (۳۸) در بررسی تغییرات مکانی کربن آلی خاک دریافتند که شاخص پوشش گیاهی و جهت شیب از متغیرهای مهم در توزیع مکانی کربن آلی خاک هستند. ژو و همکاران[3] (۴۳ و ۴۴) نیز در مطالعاتی گزارش کردند که متغیرهای توپوگرافی دارای

1- Mosleh *et al*
2- Wang *et al*
3- Zhou *et al*

همبستگی بیشتری نسبت به سایر متغیرهای کمکی مورد استفاده با محتوای کربن آلی خاک و نیتروژن خاک بوده است. در واقع، متغیرهای توپوگرافی با کنترل جریان آب و رسوبگذاری بر توسعه خاک و توزیع مکانی ویژگی‌های خاک تأثیر می‌گذارند. میناسنی و مک برتنی (۱۹) در دو منطقه اندونزی نقشه رقومی ذخیره کربن آلی خاک را با استفاده از متغیرهای توپوگرافی شامل شیب، جهت شیب و شاخص رطوبت و به کارگیری مدل رگرسیون خطی، جنگل تصادفی و شبکه عصبی مصنوعی تهیه کردند و ضرایب تبیین را به ترتیب ۰/۹۹، ۰/۹۵ و ۰/۹۸ گزارش کردند.

ارزیابی نتایج مدل‌سازی پراکنش کربن آلی در منطقه

نتایج اعتبارسنجی و ارزیابی مدل‌های جنگل تصادفی و رگرسیون خطی چند متغیره در جدول ۴ نشان می‌دهد که که مدل جنگل تصادفی با مقادیر RMSE، R^2، nRMSE، MBE و بایاس ۰/۱۲، ۰/۷۹، ۰/۱۳، ۰/۰۱۳ و ۰/۰۰۲- نسبت به مدل رگرسیون خطی چند متغیره به ترتیب با مقادیر RMSE، R^2، nRMSE، MBE و بایاس ۰/۱۹، ۰/۵۷، ۰/۲۲، ۰/۰۳۷ و ۰/۰۰۵ برای برآورد میزان کربن آلی خاک مناسب‌تر است. به طور کلی، هرچه میزان MBE، RMSE و nRMSE مدل به صفر و هر چه میزان R^2 به یک نزدیک‌تر باشد، مدل دارای دقت بالاتری می‌باشد. با توجه به این که مقدار RMSE متناسب با واحد اندازه‌گیری متغیر است مقایسه مقدار آن بین مدل‌های ساخته شده با دو متغیر با واحدهای متفاوت درست نخواهد بود لذا مقدار RMSE را به دامنه داده‌های متغیر وابسته تقسیم کرده و آن را nRMSE یا RMSE نرمال شده می‌نامند. لازم به ذکر است nRMSE زیر ۱۰درصد نشان دهنده دقیق بودن مدل، ۲۰-۱۰ درصد مناسب بودن مدل، ۳۰-۲۰ درصد دقت متوسط و بیش از ۳۰ درصد نشانه ضعیف بودن مدل است.



شکل (٦) اهمیت متغیرهای محیطی مورد استفاده در مدل‌سازی با الف الگوریتم جنگل تصادفی و (ب) رگرسیون خطی چند متغیره

Figure (6) The importance of environmental variable in (a) Random Forest and (b) Multiple Linear Regression modeling

یافته‌های این مطالعه همچنین نشان داد که الگوریتم جنگل تصادفی به کار رفته در این پژوهش برای پیش‌بینی کربن آلی خاک منطقه مطالعاتی قادر به توجیه ۷۹ درصد از تغییرات مکانی کربن آلی است و متغیرهای به کاررفته در فرآیند مدل‌سازی قادر به پیش‌بینی حدود ۲۱ درصد از تغییرپذیری کربن آلی خاک در این منطقه نیست که به نظر می‌رسد تا اندازه زیادی به دلیل در نظر نگرفتن پارامترهای مدیریتی می‌باشد.

پایین بودن ضریب تبیین مدل رگرسیون چند متغیره احتمالاً به عدم استفاده از رابطه‌های غیرخطی و پیچیده بین مقدار کربن آلی خاک و پارامترهای کمکی مربوط می‌باشد. در الگوریتم جنگل تصادفی، مقادیر ویژگی های خاک با استفاده از یک فرمول ریاضی به دست نمی‌آید بلکه روابط بین متغیرهای محیطی و ویژگی خاک با استفاده از داده های آموزش کشف می شود (۲۶). ژانگ و همکاران[1] (۴۱) نیز برتری روش جنگل تصادفی را نسبت به روش رگرسیون خطی در خصوص پیش‌بینی کربن آلی خاک نشان دادند. جیونگ و همکاران[2] (۱۲) نیز برای پیش‌بینی توزیع کربن آلی خاک، مدل‌های جنگل تصادفی و رگرسیون خطی چندگانه را به‌کار برده و نشان دادند که روش جنگل تصادفی با مقادیر R^2=۰/۹۷، RMSE=۰/۶۶۲، MSE=۰/۴۳۸ و ME=-۰/۰۲۴۸ نتایج قابل قبولی در پیش‌بینی تغییرات کربن آلی خاک داشته است. سرینیواس و همکاران[3] (۳۰) نیز نقشه رقومی کربن آلی و غیرآلی 1198 نمونه خاک را در هندوستان با استفاده از مدل جنگل تصادفی تهیه و گزارش کردند که مدل جنگل تصادفی مدل مناسبی بوده است.

نقشه نهایی پیش‌بینی کربن آلی خاک

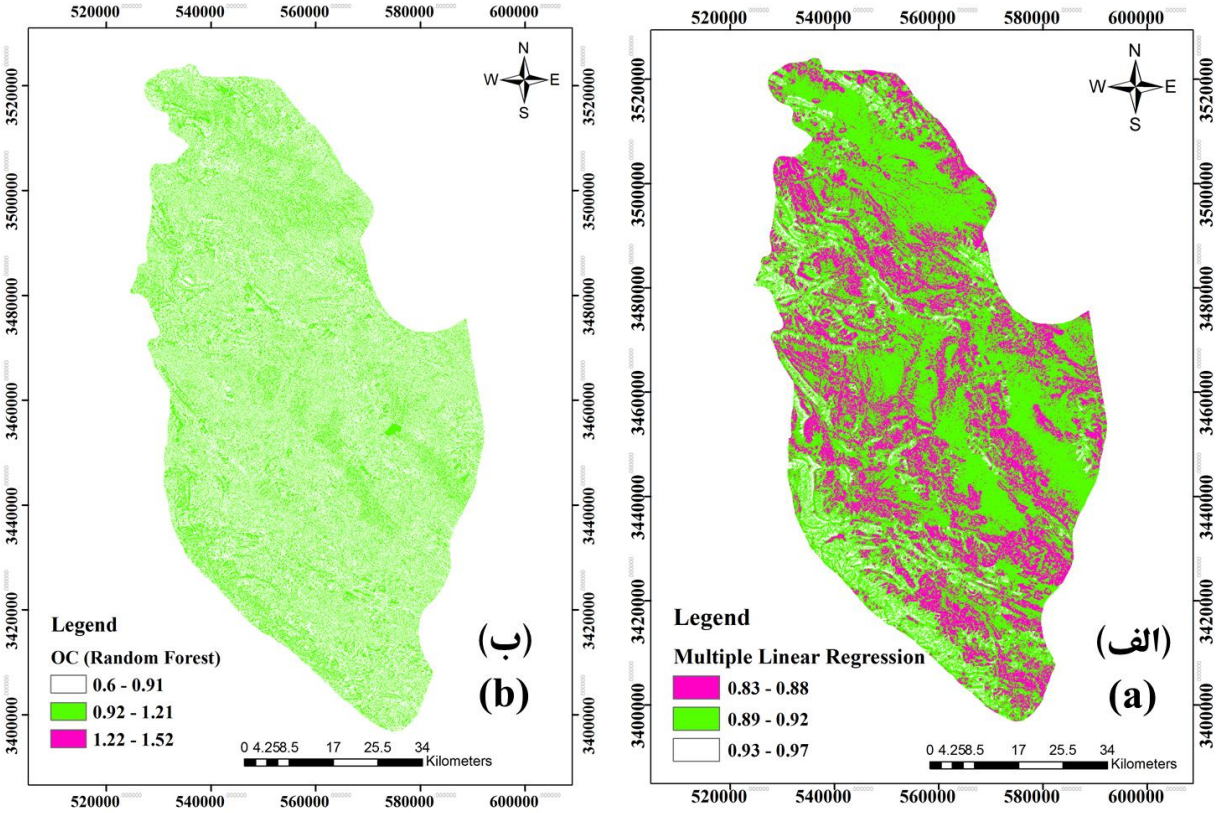
شکل ۷ نقشه نهایی پیش‌بینی کربن آلی خاک را به روش‌های جنگل تصادفی و رگرسیون خطی چند متغیره نشان می‌دهد. لازم به توضیح است که در این مطالعه، محتوای کربن آلی خاک‌ها در سه کلاس کم (کمتر از ۰/۴ درصد)، متوسط (۰/۴ تا ۰/۸در صد) و زیاد (بیش از ۰/۸ درصد) دسته‌بندی شده است (۳۰ و ۳۱). همان‌گونه که ملاحظه می‌گردد در نقشه به دست آمده به روش جنگل تصادفی، محتوای کربن آلی خاک سطحی منطقه از ۰/۵۹ تا ۱/۵۲ درصد تخمین زده شده است. در حالی که، محتوای کربن آلی تخمین زده شده با روش رگرسیون خطی چند متغیره در محدوده ۰/۸۳ تا ۰/۹۷ درصد می‌باشد که اختلاف نسبتاً زیادی را با مقادیر واقعی (اندازه‌گیری شده) کربن آلی در خاک سطحی منطقه نشان می‌دهد (جدول ۱). لذا به نظر می‌رسد روش جنگل تصادفی محتوای کربن آلی را در خاک‌های سطحی منطقه به نسبت بهتر پیش‌بینی کرده است، اگرچه به نوعی در پیش‌بینی مقادیر زیاد کربن آلی دچار کم تخمینی شده است. لازم به توضیح است که در روش جنگل تصادفی، ۹۹ درصد سطح منطقه از لحاظ محتوای کربن آلی در کلاس کربن آلی زیاد و تنها ۱ در صد سطح منطقه در کلاس کربن آلی متوسط قرار داشته است.

1- Zhou *et al*
2- Jeong *et al*
3- Sreenivas *et al*

جدول (٤) نتایج ارزیابی مدل‌های جنگل تصادفی و رگرسیون خطی چندمتغیره در پیش بینی کربن آلی خاک منطقه مطالعاتی

Table (4) Evaluation of Random Forest and Multiple Linear Regression models in prediction of Soil Organic Carbon content of the study area

Bias	R^2	nRMSE	RMSE	MBE	مدل Model
-0.0023	0.79	0.13	0.12	0.013	رگرسیون جنگل تصادفی Random Forest Regression
0.0055	0.57	0.22	0.19	0.037	رگرسیون چند متغیره خطی Multiple Linear Regression



شکل (۷) نقشه توزیع کربن آلی در منطقه به روش (الف) رگرسیون خطی چند متغیره و (ب) جنگل تصادفی

Figure (7) Map of the soil organic carbon produced by (a) Multiple Linear Regression, and (b) Random Forest algorithms

نتیجه‌گیری

این مطالعه با هدف مدل‌سازی و تهیه نقشه رقومی کربن آلی خاک سطحی منطقه سمیرم استان اصفهان با استفاده از مدل‌های جنگل تصادفی و رگرسیون خطی چندمتغیره انجام شد. نتایج حاصل از مدل‌سازی‌های انجام گرفته بیانگر عملکرد بهتر روش رگرسیون جنگل تصادفی به دلیل توانایی آن در لحاظ نمودن روابط غیرخطی و پیچیده بین کربن آلی خاک با عوامل محیطی نسبت به روش رگرسیون خطی چندمتغیره در تخمین کربن آلی سطحی خاک بود. بر این اساس، در الگوریتم جنگل تصادفی

ساخته شده با ۱۰۰۰ درخت تصمیم (ntree) و mtry معادل ۱۰، مهمترین متغیرهای موثر بر تخمین کربن آلی در خاک سطحی منطقه مطالعاتی، به ترتیب شامل شاخص‌های مختلف پوشش گیاهی و درجه شیب می‌باشند. با این حال بررسی نقشه نهایی پراکنش کربن آلی در منطقه مطالعاتی نشان می‌دهد تخمین‌های انجام شده با روش جنگل تصادفی اگرچه در مقایسه با روش رگرسیون خطی چندمتغیره تخمین‌های بهتری را ارائه داده است اما در پیش‌بینی کمینه و بیشینه مقادیر کربن آلی سطحی خاک‌ها دچار کم تخمینی و/یا بیش‌تخمینی شده است. لذا پیشنهاد می‌شود در تهیه نقشه پراکنش کربن آلی خاک در منطقه مطالعاتی سایر روش‌های مدل‌سازی مانند ماشین‌های بردار پشیبان، جنگل تصادفی چندکی، الگوریتم ژنتیک، یادگیری عمیق نیز بررسی و نتایج این روش‌ها با نتایج مطالعه حاضر مقایسه و بهترین روش برای نقشه‌سازی ویژگی‌های خاک منطقه اعم از کربن آلی مورد استفاده قرار بگیرد.

سپاس گزاری

هزینه‌های اجرای این پژوهش توسط دانشگاه شهید چمران اهواز (پژوهانه شماره SCU.AS99.365) پرداخت شده است.

References

1. Breiman, L., 2001. Random forests. Journal of Machine learning, 45(1): 5-32. DOI: 10.1023/A: 1010950718922.

2. Bouyoucos, G. J. 1951. A recalibration of hydrometer method for making mechanical analysis of soil. Agronomy, 43: 434-438. DOI: 10.2134/agronj1951.00021962004300090005x

3. Cutler, D.R., Edwards, J.T.C Beard, A. Cutler. K. H., and Hess, K.T. 2007. Random forests for classification in ecology. Journal of Ecology, 88(11): 2783-2792. DOI: 10.1890/07-0539.1

4. Hengl T., Rossiter D. G., and Stein A. 2003. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Geoderma, 120: 75–93. DOI: 10.1071/SR03005.

5. Hengl, T., Heuvelink, G.B.M., and Stein, A. 2004. A generic framework for spatial prediction of soil variables based on regression kriging. Geoderma, 120: 75–93. DOI: 10.1016 / j.geoderma. 2003.08.018.

6. Hengl, T., Heuvelink, B. M., Kempen, B., Leenaars, J.G. B., Walsh., M. G., Shepherd, K.D., Sila, A., MacMillan, R.A., Jesus, J. M., Tamene, L., and Tondoh, J.E. 2015. Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions, PLOS ONE 10 (6): e0125814. DOI: 10.1371/journal.pone.0125814.

7. Heung, B., H.C. Ho., J. Zhang., A. Knudby., C.E. Bulmer., and M.G. Schmidt. 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma, 265: 62-77. DOI: 10.1016 / j.geoderma.2015.11.014.

8. Fathololoumi, S., Vaezi, A. R., Alavipanah, S. K., Ghorbani, A., Saurette and D., Biswas, A. (2021). Effect of multi-temporal satellite images on soil moisture prediction using a digital soil mapping approach. Geoderma, 385, 114901. DOI: 10.1016/j.geoderma.2020.114901.

9. IBM Corp (2010). IBM SPSS Statistics for Windows, version 19, SPSS Inc., Chicago, Ill., USA.

10. Jalalian A. 1997. The studies of land resources and capability determination in Semirom area. The Ministry of Jahad Sazandegi, Isfahan Province. (in Persian)

11. Jenny, H. 1941. Factors of soil formation: A system of quantitative pedology. McGraw-Hill, New York.

12. Jeong, G.H., Oeverdieck, S.J., Park, Huwe, B. and Ließ, M. 2017. Spatial soil nutrients prediction using three supervised learning methods for assessment of land potentials in complex terrain. Catena, 154: 73-84. DOI: 10.1016/j.catena.2017.02.006.

13. Lanyon, L. E. and Heald, W. R. 1982.Magnesium, calcium, strontium and barium. In: Page, A.L., et al. (Eds.), Methods of Soil Analysis. Part II, Agronomy. Monograph, American Society of Agronomy and Soil Science Society of America, Madison, Wisconsin, pp. 247-260.

14. Mahmoudzadeh, H., Matinfar, H. R., Taghizadeh-Mehrjardi, R., and Kerry, R. (2020). Spatial prediction of soil organic carbon using machine learning techniques in western Iran. Geoderma Regional, 21, e00260. DOI: 10.1016/j.geodrs.2020.e00260.

15. McBratney, A.B., Mendonça Santos, and M.L., Minasny, B. 2003. On digital soil mapping. Geoderma, 117: 3-52. DOI: 10.1016/S0016-7061(03)00223-4.

16. McBratney, A.B., Stockmann, U., Angers, D., Minasny, B., and Field, D. 2014. Challenges for soil organic carbon research. In: Hartemink, A.E., McSweeney, K., (Eds.), Soil Carbon, Cham: Springer, pp. 3-16. DOI: 10.1007/978-3-319-04084-4-1.

17. Mondal, A., Khare, D., Kundu, S., Mondal, S., Mukherjee, S. and A. Mukhopadhya. 2016. Spatial soil organic carbon (SOC) prediction by regression kriging using remote sensing data. Egypt. Journal of Remote Sensing and Space Science, 20 (1): 61-70. DOI: 10.1016/j.ejrs.2016.06.004.

18. Mosleh, Z., Salehi, M.H., Jafari, A., Borujeni, I.E., and Mehnatkesh, A. 2016. The effectiveness of digital soil mapping to predict soil properties over low-relief areas. Environmental Monitoring and Assessment, 188 (3): 195. DOI: 10.1007/s10661-016-5204-8.

19. Minasny, B., and McBratney, A.B. (2016). Digital soil mapping: A brief history and some lessons. Geoderma, 264: 301-311. DOI: 10.1016/j.geoderma.2015.07.017

20. Nelson, R. E. 1982. Carbonate and gypsum. In: Page, A.L., et al. (Eds.), Methods of Soil Analysis. Part 2: Chemical Methods, 2nd Ed., Agronomy Monograph, No. 9, American Society of Agronomy and Soil Science Society of America, Madison, WI. pp. 181-196.

21. Olaya, V. 2004. A gentle introduction to SAGA GIS. The SAGA User Group eV, Gottingen, Germany. 208 pp.

22. Osat, M., Heidari, A., Karimian Eghbal, M., and Mahmoodi, Sh. (2016). Spatial variability of soil development indices and their compatibility with soil taxonomic classes in a hilly landscape: a case study at Bandar village, Northern Iran. Journal of Mountain Science, 13(10): 1746-1759. DOI: 10.1007/s11629-016-3952-0.

23. R Development Core Team. 2015. R: a language and environment for statistical computing. R. Foundation for Statistical Computing, Vienna, Austria. http://www.Rproject.org.

24. Rossel, R.A.V., and McBratney, A.B. 2009. Diffuse reflectance spectroscopy as a tool for digital soil mapping. In: Hartemink, A.E., et al., (Eds.), Digital Soil Mapping with Limited Data. Springer, Dordrecht, pp. 165-172. DOI: 10.1007/978-1-4020-8592-5.

25. Rudiyanto, R., Minasny, B., Setiawan, B.I., Arif, C., Saptomo, S.K., and Chadirin, Y. (2016). Digital mapping for cost-effective andaccurate prediction of the depth and carbon stocks in Indonesian peatlands. Geoderma, 272: 20–31. DOI: 10.1016/j.geoderma.2017.10.018.

26. Richards, L.A. 1954. Diagnosis and Improvement of Saline-Alkali Soils. USDA Hand book, No. 60. Washington, D.C., U.S.A.

27. Rhoades, J.D. 1996. Salinity: electrical conductivity and total dissolved soils. In: Sparks, D.L. (Ed.), Methods of Soils Analysis, Part 3: Chemical Methods. Soil Science Society of America Book series Number 5, Soil Science Society of America, Madison, Wisconsin, pp. 417-435.

28. Tarkalson, D.D., Brown, B., Kok, H., and Bjorneberg, D.L. 2009. Irrigated small-grain residue management effects on soil chemical and physical properties and nutrient cycling. Soil Science, 174:303-311. DOI: 10.1097/SS.0b013e3181a82a5f

29. Skullberg, U. 1991. Seasonal Variation of pH H_2O and pH $CaCl_2$ in centimeter- layers of mor Humus in a *Picea Abies* (L.) Karst stand. Sweden University of Agri Science, Department of Forest Site Research.

30. Sreenivas, K., Dadhwal, V.K., Kumar, S., Harsha, G.S., Mitran, T., Sujatha., G., Janaki Rama, S., Fyzee, M.A, and Ravisankar, T. 2016. Digital mapping of soil organic and inorganic carbon status in India. Geoderma, 269: 160-173. DOI: 10.1016/j.geoderma.2016.02.002

31. Sys, C., Van Ranst, E. and Debaveye, J. 1991. Land Evaluation. Agricultural Publication No. 7, General Administration for Development Cooperation, Brussels.

32. Taghizadeh-Mehrjardi, R., Nabiollahi, K. and Kerry, R. 2016. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. Geoderma, 266: 98–110. DOI: 10.1016/j.geoderma.2015.12.003

33. Vaysse, K. and Lagacherie, K. 2015. Evaluating digital soil mapping approaches for mapping Global Soil Map soil properties from legacy data in Languedoc-Roussillon (France). Geoderma Regional, 4: 20-30. DOI: 10.1016/j.geodrs.2014.11.003

34. Venables, W.N., and B.D. Ripley. 2013. Modern applied statistics with S-PLUS. Springer, Dordrecht, 498 p.

35. Walkley A. and Black, I.A. 1934. An examination of the Degtjareff method for determining organic carbon in soils: effect of variations in digestion conditions and of inorganic soil constituents. Soil Science, 63: 251-263. DOI: 10.1097/00010694-194704000-00001

36. Wallach, D., Makowski, D., Jones, J.W., and Brun, F. 2006. Working with dynamic crop models: Evaluation, analysis, parameterization, and applications. Elsevier.

37. Wang Sh., Jin X., Adhikari, K., Li, W., Yu, M., Bian, Zh. and Wang, Q. 2018. Mapping total soil nitrogen from a site in northeastern China. Catena, 166: 134-146. DOI: 10.1016/j.catena.2018.03.023.

38. Wang, S., Wang, Q., Adhikari, K., Jia, S., Jin, X. and Liu, H. 2016. Spatial-temporal changes of soil organic carbon content in Wafangdian, China. Sustainability, 8: 1154. DOI: 10.3390/su8111154.

39. Wilding, L. 1985. Soil spatial variability: Its documentation, accommodation, and implication to soil surveys. In: Nielson, D.R., Bouma, J. (Eds.), Wagenigen, Netherland, pp. 166-194.

40. Yarali, J., Esmaili, A. and Esmaili, G.H. 2013. Statistical Analyze with SPSS 20. Kankash Publication, pp. 220-234.

41. Zhang, H., Wu, P., Yin, A., Yang, X., Zhang, M. and Gao, C. 2017. Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: A comparison of multiple linear regressions and the random forest model. Science of the Total Environment, 592: 704-713. DOI: 10.1016/j.scitotenv.2017.02.146

42. Zhao, Z., Yang, Q., Benoy, G., Chow, T.L., Xing, Z., Rees, H.W., and F.R. Meng. 2010. Using artificial neural network models to produce soil organic carbon content distribution maps across landscapes. Canadian Journal of Soil Science, 90 (1): 75–87. DOI: 10.4141/CJSS08057

43. Zhou, T., Geng, Y., Chen, J., Pan, J., Haase, D. and Lausch, A. 2020. High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms. Science of the Total Environment, 729: 138244. DOI: 10.1016/j.scitotenv.2020.138244

44. Zhou, Y., Hartemink, A.E., Shi, Z., Liang, Z., and Lu, Y. 2019. Land use and climate change effects on soil organic carbon in North and Northeast China. Science of the Total Environment, 647: 1230-1238. DOI: 10.1016/j.scitotenv.2018.08.016.