Research Article

Agricultural Engineering., 46(3) (2023) 289-308 ISSN (P): 2588-526X

DOI: 10.22055/AGEN.2023.44688.1680 ISSN (E): 2588-5944

Using Ensemble Model Approach for Spatial Modeling of Soil Imbalanced Classes

M. Rahimi Mashkaleh[1], M. Amir Delavar[2,*] and M. Jamshidi[3]

1. Ph.D. Student of Department of Soil Science, Faculty of Agriculture, University of Zanjan, Zanjan, Iran
2. Associate Professor, Department of Soil Science, Faculty of Agriculture, University of Zanjan, Zanjan Iran
3. Assistant Professor, Soil and Water Research Institute, Agricultural Research, Education and Extension Organization, Karaj, Iran

Received: 1 September 2023 Accepted: 20 November 2023 *Corresponding Author: amir-delavar@znu.ac.ir

Abstract

Introduction: Imbalanced data remains a widespread and significant challenge, particularly impacting machine learning algorithms. Therefore, addressing imbalanced data classification has emerged as a crucial research area within the field of data mining. This issue, often characterized by a limited number of instances in one class and a substantial number in other classes, poses substantial hurdles for machine learning algorithms. Consequently, data mining experts and machine learning professionals are actively working on refining methods and models for classifying imbalanced data with the aim of improving the accuracy of such classifications. The principal objective of this study is to precisely detect and categorize samples from the minority class, ultimately enhancing the precision of soil class classification. This research is conducted in a specific region, encompassing the southwestern territories of Zanjan province.

Materials and Methods: To achieve this objective, a total of 148 soil profiles were excavated using a regular grid pattern with an average spacing of 500 meters (and in some locations, up to 700 meters based on expert recommendations). After the samples were air-dried, they were transported to the laboratory. Physical and chemical analyses were conducted on all collected samples, including assessments of soil texture, soil pH, calcium carbonate equivalent, cation exchange capacity, electrical conductivity, organic carbon content, and gypsum content. Subsequently, the soil samples were meticulously classified and described up to the family level, following the comprehensive standards of the soil classification system. The most appropriate covariates were selected among 57 covariates including geomorphological and geological maps, digital elevation model (DEM), and data from Landsat 8 satellite images, using principal component analysis (PCA) and expert knowledge approaches for predicting soil classes selected. Saga-GIS and ENVI software were used to extract environmental covariates. Modeling of the soil-landscape relationship was performed using three algorithms, namely multinomial logistic regression (MNLR), random forest (RF), boosted regression tree (BRT) and ensemble model (after data balancing) in "R studio" software. To check the accuracy of the used model, the data was randomly divided into training and validation data. 80% of the data (118 profiles) were used for model training and 20% (30 profiles) were used as validation data for evaluation.

Results and Discussion: The results of the selection of covariates showed that 10 information covariates of geomorphological maps, geological information and features extracted from the

digital elevation model (DEM), including Analytical hill shading (AHS), sunrise, valley depth (VD), LS Factor, Channel network distance (CND), Topographic wetness index (TWI) and Multi-resolution ridge top flatness (MRRTF) were selected as input variables. Based on the results of profile analysis, the soils of the region at the subgroup level were categorized into five classes, with imbalanced distribution, including Typic Calcixerepts, Typic Haploxerepts, Gypsic Haploxerepts, Typic Xerorthents, and Lithic Xerorthents. The results of evaluation metrics such as overall accuracy and Kappa index were 65% and 0.32 for the RF algorithm, %60 and 0.35 for the boosted regression tree algorithm, 65% and 0.41 for the MNLR algorithm and after balancing the data with the ensemble model approach, it was 70% and 0.62 respectively. The results of two statistics of user's accuracy and producer's accuracy showed that among individual models, the multinomial logistic regression model has higher accuracy in predicting soil classes. Although the ensemble model has succeeded in predicting the soil minority classes well, due to the fact that the two weaker models of the RF and BRT are involved in the modeling, It showed lower values compared to the individual multinomial logistic regression model, in predicting some classes of the majority of soil, especially the two classes of Typic Haploxerepts and Typic Xerorthents.

Conclusions: Conclusions: In summary, the results have demonstrated that when learning algorithms are individually applied, they do not exhibit high accuracy in spatially predicting soil classes. However, when these algorithms are amalgamated into an ensemble model, they exhibit remarkable accuracy in spatial soil class prediction, outperforming individual models in terms of performance and accuracy. Moreover, the ensemble model substantially enhances prediction accuracy and reduces the occurrence of misclassifications, especially at the subgroup level. While each specific model excels in predicting a particular soil classification, the cumulative ensemble models consistently outperform individual models in terms of overall performance and accuracy, underscoring the effectiveness of ensemble modeling in improving spatial soil classification.

Keywords: Boosted regression trees, data balancing, imbalanced dataset, minority

مهندسی زراعی (مجله علمی کشاورزی) جلد ۴۶ شماره۳، پاییز ۱۴۰۲

مقاله پژوهشی

استفاده از رویکرد مدل تجمعی برای مدل‌سازی مکانی کلاس‌های نامتعادل خاک

مستانه رحیمی مشکله[۱]، محمد امیر دلاور[۲*] و محمد جمشیدی[۳]

۱- دانشجوی دکتری گروه علوم خاک دانشکده کشاورزی، دانشگاه زنجان، زنجان، ایران

۲- دانشیار گروه علوم خاک دانشکده کشاورزی، دانشگاه زنجان، زنجان، ایران

۳- استادیار موسسه تحقیقات خاک و آب، سازمان تحقیقات، آموزش و ترویج کشاورزی، کرج، ایران

تاریخچه مقاله

دریافت: ۱۴۰۲/۰۶/۱۰

پذیرش نهایی: ۱۴۰۲/۰۸/۲۹

کلمات کلیدی:

رگرسیون لجستیک چندجمله‌ای،

داده‌های نامتعادل،

متعادل‌سازی داده،

کلاس اقلیت

* عهده‌دار مکاتبات

Email: amir-delavar@znu.ac.ir

چکیده

عدم تعادل داده‌ها یکی از رایج‌ترین مشکلاتی است که بر الگوریتم‌های یادگیری ماشینی تأثیر می‌گذارند. ازاین‌رو طبقه‌بندی داده‌های نامتعادل به یک موضوع تحقیقاتی مهم در زمینه داده‌کاوی تبدیل شده است. هدف از انجام این پژوهش شناسایی صحیح نمونه‌های کلاس اقلیت و افزایش دقت طبقه‌بندی کلاس‌های خاک با استفاده از رویکرد مدل تجمعی در بخشی از اراضی جنوب غربی استان زنجان است. برای دستیابی به این هدف تعداد ۱۴۸ خاک‌رخ با روش الگوی شبکه‌بندی منظم و میانگین فاصله ۵۰۰ متر حفر، تشریح و با تجزیه‌وتحلیل آزمایشگاهی تا سطح فامیل رده‌بندی گردید. مناسب‌ترین متغیرهای محیطی بر اساس نظر کارشناسی و رویکرد تحلیل مؤلفه اصلی از میان ۵۷ متغیر شامل اطلاعات نقشه‌های ژئومورفولوژی و زمین‌شناسی، مدل رقومی ارتفاع و داده‌های حاصل از تصاویر ماهواره‌ای لندست ۸ برای پیش‌بینی کلاس‌های خاک انتخاب شد. مدل‌سازی رابطه خاک - زمین‌نما با استفاده از الگوریتم‌های یادگیرنده جنگل تصادفی، درخت تصمیم توسعه‌یافته و رگرسیون لجستیک چندجمله‌ای و مدل تجمعی (بعد از متعادل‌سازی داده‌ها) در محیط نرم‌افزار "R studio" انجام شد. نتایج انتخاب متغیرهای محیطی نشان داد که ۱۰ متغیر اطلاعات نقشه‌های ژئومورفولوژی، اطلاعات زمین‌شناسی و ویژگی‌های مستخرج از مدل رقومی ارتفاع شامل تجزیه‌وتحلیل سایه‌اندازی تپه‌ها، طلوع خورشید، عمق دره، شاخص طول در جهت شیب، فاصله تا شبکه آبراهه، شاخص رطوبتی توپوگرافی و شاخص همواری بالای پشته با درجه تفکیک بالا به‌عنوان متغیر ورودی انتخاب شدند. خاک‌های منطقه در سطح زیرگروه در پنج کلاس با توزیع نامتعادل شامل تیپیک کلسی-زرپتز، تیپیک هاپلوزرپتز، جیپسیک هاپلوزرپتز، تیپیک زراورتنتز و لیتیک زراورتنتز طبقه‌بندی شدند. صحت کلی و ضریب کاپا برای ارزیابی

کلاس‌های خاک در سطح زیرگروه به ترتیب در مدل‌های فردی رگرسیون لجستیک چندجمله‌ای ٦٥ درصد و ۰/٤١، جنگل تصادفی ٦٥ درصد و ۰/۳۲، درخت تصمیم توسعه‌یافته ٦۰ درصد و ۰/۳٥ و در مدل تجمعی ۷۰ درصد و ۰/٦۲ به دست آمد. نتایج صحت کاربر و صحت تولیدکننده نشان داد در میان مدل‌های فردی، مدل رگرسیون لجستیک چندجمله‌ای دقت بالاتری در پیش‌بینی کلاس‌های خاک دارد. مدل تجمعی با اینکه موفق به پیش‌بینی خوب کلاس‌های اقلیت خاک شده است اما به دلیل اینکه دو مدل ضعیف‌تر جنگل تصادفی و درخت تصمیم توسعه‌یافته در مدل‌سازی دخیل هستند نسبت به مدل فردی رگرسیون لجستیک چندجمله‌ای مقادیر کمتری در پیش‌بینی برخی کلاس‌های اکثریت خاک به‌ویژه دو کلاس تیپیک هاپلوزرپتز و تیپیک زراورتنتز نشان داد. به‌طورکلی نتایج نشان داد که الگوریتم‌های یادگیرنده زمانی که به‌تنهایی و به‌صورت فردی مدل‌سازی می‌شوند از دقت بالایی در پیش‌بینی مکانی کلاس‌های خاک برخوردار نیستند اما مجموعه این الگوریتم‌ها در قالب مدل تجمعی دقت بالایی در پیش‌بینی مکانی کلاس‌های خاک از خود نشان می‌دهند و مدل‌های تجمعی عملکرد و دقت بالاتری در مقایسه با مدل‌های فردی دارند.

مقدمه

تهیه نقشه‌های خاک که به‌عنوان نقشه‌های پایه و مرجع در مدیریت خاک‌های کشاورزی محسوب می‌شوند، از ضروریات مطالعات خاک‌شناسی است (۴۸). نقشه‌های خاک می‌توانند مقدمات پیش‌بینی رفتار خاک‌ها در مقابل کاربری‌های مختلف را فراهم آوردند و تفسیر این نوع مطالعات، قابلیت‌های کاربردی خاک مانند توانایی نگهداری آب، مواد مغذی، وضعیت کربن آلی و غیره را در اختیار کاربران قرار می‌دهند (۲). روش‌های نقشه‌برداری رقومی خاک ضمن استفاده از الگوریتم‌های پیش‌بینی در یادگیری ماشین و داده‌کاوی[1]، علاوه بر تهیه نقشه-های جدید (۳) می‌توانند جایگزین کارآمدی برای مطالعات خاک‌شناسی با روش‌های مرسوم برای به‌روزرسانی نقشه کلاس‌های خاک باشند (۲۷)؛ اما علی‌رغم افزایش دقت نقشه‌برداری رقومی خاک در سال‌های اخیر، تولید نقشه‌های خاک در مقیاس منطقه‌ای با دقت بالا همچنان یک کار چالش‌برانگیز است (۳۵). برای حل چالش مدیریت کاربری زمین نیاز به پیش‌بینی ویژگی‌های خاک با بالاترین دقت ممکن است. ازاین‌رو، بررسی کاربرد روش‌هایی برای ترکیب پیش‌بینی‌های موجود و بهبود دقت پیش‌بینی آن‌ها لازم است (۱۶ و ۲۰).

با توجه به ماهیت توزیع خاک‌ها، یک مسئله مهم در مدل‌سازی و در فرآیند نقشه‌برداری کلاس‌های خاک، عدم تعادل کلاس‌های مشاهده شده است (۴۲). داده‌های نامتعادل به این واقعیت اشاره دارد که تعداد مشاهده‌ها با فراوانی بیشتر (کلاس اکثریت)، بسیار بیشتر از تعداد مشاهده‌ها با فراوانی کمتر (کلاس اقلیت) است (۷ و ۵۸). به‌طورمعمول کلاس‌های خاک توسط انواع مختلفی از روش‌های آماری و داده‌کاوی از طریق نقشه‌برداری رقومی خاک پیش‌بینی می‌شوند. از جمله این

1- Data mining

روش‌ها رگرسیون لجستیک چندجمله‌ای[1]، جنگل‌های تصادفی[2] و درخت تصمیم توسعه‌یافته[3] است. در روش‌های معمول پیش‌بینی مکانی، چندین مدل ارزیابی و بهترین مدل اجرایی انتخاب می‌گردد (۱۰ و ۵۱). عدم تعادل داده‌ها بر الگوریتم‌های یادگیری معمول تأثیر منفی دارد و این اثر منفی در ترکیب با عواملی همچون وجود نقاط پرت و تعداد ناکافی مشاهده‌های آموزشی، شدیدتر است و می‌تواند منجر به برازش بیش‌ازحد مدل شود (۲۹ و ۵۹). همچنین انتخاب یک الگوریتم طبقه‌بندی کننده مناسب به‌طور بالقوه مشکل‌ساز است (۲۱)؛ زیرا با توجه به ساختار مدل، یک مدل واحد دارای مزایا و معایبی است، بنابراین یک الگوریتم طبقه‌بندی‌کننده ممکن است در معین، دقت بیشتری نسبت به سایر الگوریتم‌ها داشته باشد و بالعکس (۲۱ و ۴۹). برای غلبه بر این مشکل، مجموعه چندین مدل آموزش‌دیده (مدل تجمعی)، جایگزینی مناسبی است که با کمک به اطلاعات به‌دست‌آمده از مدل‌های طبقه‌بندی منجر به‌دقت بیشتر طبقه‌بندی شود (۴۹). مزیت الگوریتم‌های تجمعی به‌عنوان تکنیک‌های قدرتمند این است که نسبت به روش‌ها و الگوریتم‌های منفرد یا تک‌پایه با حذف نقاط ضعف این الگوریتم‌ها، دقت خوبی در برازش و پیش‌بینی بالاتر دارند (۵۴).

به‌طورکلی نقشه‌برداری رقومی خاک پتانسیل زیادی برای استفاده از همه مدل‌های یادگیری ماشین از طریق مدل‌سازی تجمعی دارد (۱۶). مدل‌های یادگیری تجمعی یکی از محبوب‌ترین رویکردها برای کنترل عدم تعادل کلاس است (۵، ۱۹ و ۳۰). این رویکرد معمولاً با برازش یک مدل جدید با استفاده از پیش‌بینی‌های چندین مدل پیش‌بینی‌شده و سپس خروجی‌سازی پیش‌بینی مدل جدید اجرا می‌شود (۱۳). علاوه بر این، مدل‌های تجمعی به‌طور بالقوه منجر به پیش‌بینی‌های بهتر و پایدارتر و همچنین کاهش ریسک انتخاب طبقه‌بندی اشتباه می‌شوند (۲۱). در رویکرد تجمعی فرض بر این است که مدل جدید، به‌خوبی هر یک از مدل‌های فردی خواهد بود و از تمام اطلاعات موجود به‌طور مؤثر استفاده می‌کند (۱۵). مطالعات محققینی همچون لودویگ و همکاران[4] (۳۳)، چن و همکاران[5] (۱۲)، ژانگ و هارتمینک[6] (۶۰)، ووهلند و همکاران[7] (۵۶) و گروشچینسکی و گروشچینسکی[8] (۲۲) نشان داد که روش‌های تجمعی دقت پیش‌بینی را بهبود می‌بخشد و عملکرد بهتری نسبت به رویکردهای فردی دارند.

این پژوهش با هدف تولید نقشه‌های دقیق و کارآمد و مدیریت بهتر اراضی کشاورزی، به مقایسه عملکرد مدل‌های مختلف یادگیری ماشین از جمله جنگل تصادفی، درخت تصمیم توسعه‌یافته و رگرسیون لجستیک چندجمله‌ای برای پیش‌بینی مکانی کلاس‌های خاک در سطح زیرگروه و ارزیابی توانایی مدل تجمعی در بخشی از اراضی استان زنجان می‌پردازد.

مواد و روش‌ها

منطقه موردمطالعه

پژوهش حاضر در بخشی از اراضی جنوب غربی استان زنجان با مختصات جغرافیائی ۴۷ درجه و ۹۱ دقیقه تا ۴۸ درجه و ۱۱ دقیقه طول شرقی و ۳۶ درجه و ۳۷ دقیقه تا ۳۶ درجه و ۳۱ دقیقه عرض شمالی و به مساحت ۱۳۸۲۳ هکتار اجرا گردید (شکل ۱). متوسط ارتفاع منطقه ۱۴۸۲ متر از سطح دریای آزاد بوده و دارای متوسط بارندگی سالیانه منطقه ۳۴۰ میلی‌متر و متوسط دمای سالیانه ۱۴ درجه سلسیوس است (۴۶). بر اساس اطلاعات به‌دست‌آمده رژیم حرارتی منطقه مزیک[9] و رژیم رطوبتی آن زریک[10] است (۴۳). سازندهای زمین‌شناسی عمده منطقه شامل لایه‌های کربناته، سنگ‌آهک، کنگلومرا و مواد آتش‌فشانی و فیزیوگرافی منطقه شامل دو واحد اراضی تپه‌ماهور[11] و دشت‌های دامنه‌ای[12] است. اراضی تپه‌ماهوری در سطح پستی‌وبلندی به تپه‌های با ارتفاع متوسط[13] و تپه‌های با

1- Random forests, RF
2- Boosted regression trees, BRT
3- Multinomial logistic regression, MNLR

4- Ludwig *et al.*
5- Chen *et al.*
6- Zhang *et al.*
7- Vohland *et al.*
8- Gruszczynski and Gruszczynski
9- Mesic
10- Xeric
11- Hill lands
12- Piedmont plains
13- Medium hill

ارتفاع کم[1] تفکیک شده‌اند. با توجه به تقسیم‌بندی زمین‌شناسی، تپه‌های با ارتفاع متوسط غالباً رسوبات واریزه‌ای - بادرفتی آهکی همراه با مارن[2] و پستی‌وبلندی تپه‌های با ارتفاع کم غالباً شامل مخلوط رسوبات واریزه‌ای - بادرفتی و مارن[3] هستند. بر اساس سامانه آمریکایی رده‌بندی خاک‌ها، خاک‌های منطقه در دو رده انتی‌سولز[4] و اینسپتی‌سولز[5] طبقه‌بندی شده‌اند.

مطالعات صحرایی و آزمایشگاهی

پس از بررسی نقشه‌های توپوگرافی و استفاده از اطلاعات موروثی در محیط سامانه‌های اطلاعات جغرافیایی، تصاویر برگرفته از گوگل ارث و انجام بازدیدهای صحرایی، ۱۴۸ خاک‌رخ بر اساس الگوی طبقه‌بندی تصادفی با میانگین فاصله ۵۰۰ متر مطابق روش‌های استاندارد مطالعه (۴۴) و نمونه‌برداری از افق‌های مشخصه سطحی و زیرسطحی هر خاک‌رخ انجام شد. نمونه‌ها پس از هوا خشک شدن از الک ۲ میلی‌متری عبور داده شدند و تجزیه‌های فیزیکی و شیمیایی شامل بافت (۶)، واکنش (۳۷)، کربنات کلسیم معادل (۳۲)، ظرفیت تبادل کاتیونی (۴۷)، قابلیت هدایت الکتریکی (۳۹)، کربن آلی (۵۷) بر روی تمام نمونه‌ها و در صورت مشاهده صحرایی گچ بر روی برخی نمونه‌ها اندازه‌گیری گچ (۴) انجام شد. خاک‌رخ‌ها بر اساس نتایج تشریح صحرایی و اطلاعات تجزیه‌های فیزیکی و شیمیایی نمونه‌های خاک در سیستم جامع رده‌بندی خاک به روش آمریکایی (۴۵) تا سطح فامیل طبقه‌بندی شدند.

استخراج متغیرهای محیطی

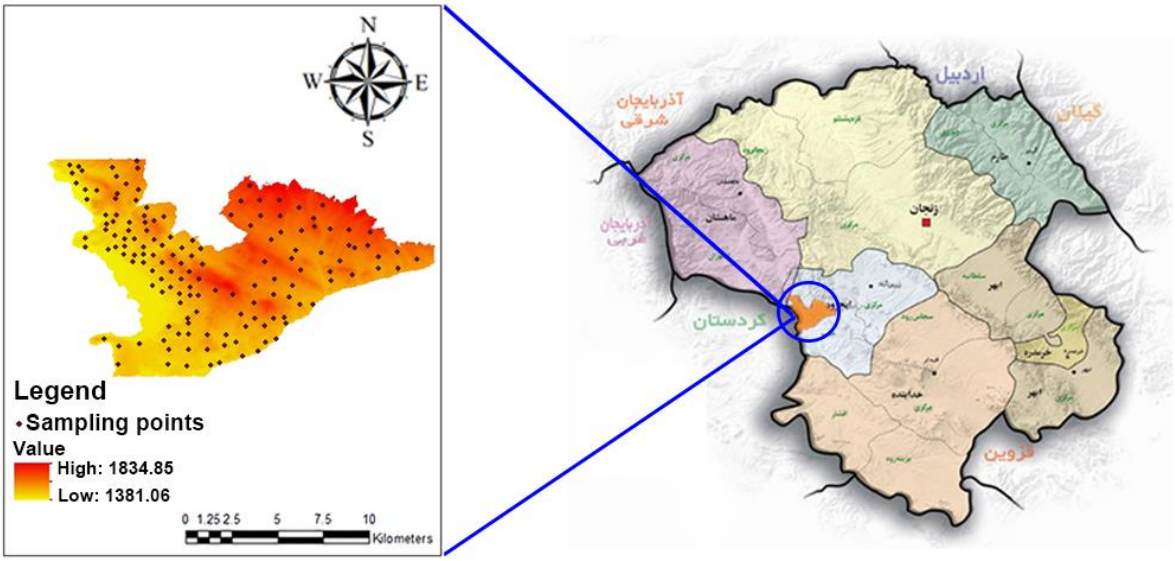
متغیرهای محیطی مورداستفاده در این پژوهش شامل اطلاعات نقشه‌های ژئومورفولوژی، نقشه زمین‌شناسی، داده‌های سنجش‌ازدور و توپوگرافی هستند. برای این منظور نقشه زمین-شناسی با مقیاس ۱:۲۵۰۰۰۰ منطقه تهیه‌شده توسط سازمان زمین-شناسی کشور در محیط Arc-GIS (نسخه 10.7) زمین مرجع و رقومی شد. از مدل رقومی ارتفاع با قدرت تفکیک مکانی ۳۰×۳۰ متر سنجنده استر، ۱۸ شاخص پستی‌وبلندی در محیط نرم افزار SAGA GIS (نسخه 7.9) استخراج شد. ۳۶ شاخص سنجش‌ازدور از تصاویر سنجنده (OLI/TIRS) ماهواره لندست ۸ با قدرت تفکیک مکانی ۳۰×۳۰ متر (USGS 2014) پس از اعمال تصحیحات رادیومتریکی و اتمسفری در محیط نرم‌افزار ENVI (نسخه 5.3) تهیه و استخراج شد (شکل ۲- الف). نقشه ژئومورفولوژی منطقه بر اساس تلفیق لایه‌های اطلاعاتی شامل واحدهای لندفرم و مواد مادری به همراه تفسیر تصاویر ماهواره‌ای بر اساس رویکرد سلسله مراتبی ارائه‌شده توسط زینک (۶۱) تهیه گردید (شکل ۲- ب و جدول ۱). از میان متغیرهای محیطی تهیه‌شده، تعدادی متغیر محیطی بر اساس رویکرد تحلیل مؤلفه اصلی[6] در نرم‌افزار SPSS (نسخه 26.0) و رتبه‌بندی اهمیت نسبی مدل یادگیری ماشین به همراه نظارت کارشناس برای ورود به مدل انتخاب گردید (۳۱).

مدل‌سازی خاک_ زمین‌نما

مدل رگرسیون لجستیک چندجمله‌ای

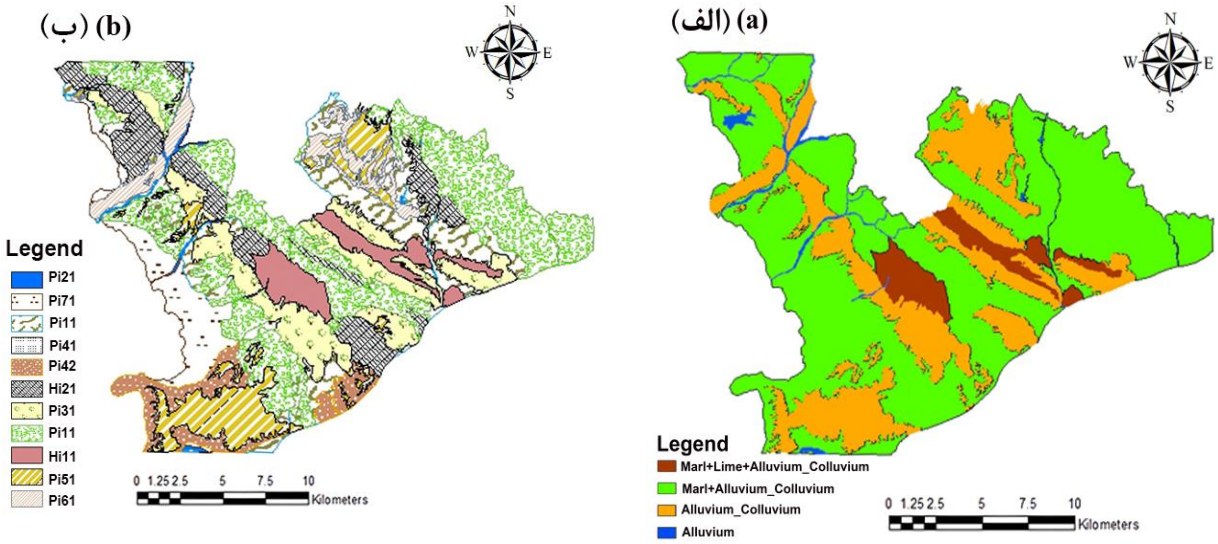
مدل رگرسیون لجستیک یک نوع مدل خطی تعمیم‌یافته است و برای مجموعه داده‌هایی مناسب است که متغیر وابسته به‌صورت کیفی است. این مدل‌ها قادر به توصیف روابط بین مجموعه‌ای از متغیرهای پیش‌بینی کننده و یک متغیر وابسته دو بخشی است که دارای مقادیر صفر یا چهار است (۲۸). در هر دو حالت مدل رگرسیون لجستیک چندجمله‌ای برای هر کلاس خاک در منطقه موردمطالعه توسعه و روابط توپوگرافی و واحدهای طبقه-بندی خاک از داده‌ها تعیین شدند. مدل‌سازی رگرسیون لجستیک در نرم‌افزار R_{studio} بر اساس بسته "Caret" انجام شد.

1- Low hill
2- Marl+lime+alluvio-colluvium
3- Marl+Alluvio-colluvium
4- Entisols
5- Inceptisols
6- Principal component analysis, PCA



شکل (۱) موقعیت منطقه و نقاط نمونه‌برداری

Figure (1) Location of the study area and sampling points



شکل (۲) نقشه الف) زمین‌شناسی و ب) ژئومورفولوژی منطقه موردمطالعه

Figure (2) Map of a geology and b) Geomorphologi in study area

جدول (۱) واحدهای تفکیک‌شده در سطح لندفرم بر اساس اطلاعات ژئومورفولوژی در منطقه موردمطالعه (۴۳)

Table (1) Separated units at the landform level based on geomorphological information in the studied area (43)

واحد نقشه Map unit	اجزای لندفرم Landform components	شکل زمین Landform	سنگ‌شناسی Lithology	توپوگرافی Topography	زمین‌نما Landscape
Consociation	Hi111	Slope facet complex	Marl + Lime + Alluvium - Colluvium	Medium hill	تپه‌ماهور Hill lands
Complex	Hi211	Summit	Marl + Alluvium - Colluvium	Low hill	
	Hi212	Shoulder			
	Hi213	Backslope			
	Hi214	Footslope			
	Hi215	Toeslope			
Association	Pi111	High glacic	Alluvium - Colluvium	Glacic, Dissected	دشت دامنه‌ای Piedmont plains
Complex	Pi211	Middle glacic	Marl + Alluvium - Colluvium	Glacic, Moderately dissected	
Association	Pi311	Low glacic	Marl + Alluvium - Colluvium	Glacic, Low dissected	
Association	Pi411	Side slope	Alluvium	Glacic terrace, Dissected	
Consociation	Pi421	Tread	Alluvium - Colluvium		
Association	Pi511	Side slope	Marl + Alluvium - Colluvium	Glacic terrace, Slightly eroded	
Association	Pi611	Side slope	Alluvium - Colluvium	Coalescing fan	
Association	Pi711	Upper part	Marl + Alluvium - Colluvium	Channeled recent alluvial deposits	
	Pi712	Middle part			
	Pi713	Lower part			

مدل جنگل تصادفی

مدل جنگل تصادفی یک تکنیک یادگیرنده فعال و توسعه‌یافته از مدل طبقه‌بندی و رگرسیون درختی است. در این روش داده‌ها به‌طور تکراری برای به دست آوردن ارتباط بین متغیر پاسخ و متغیرهای مستقل و انجام تخمین جداسازی می‌شوند. در روش جنگل تصادفی برخلاف سایر روش‌های درختی که تعداد محدودی درخت ترسیم می‌کنند، صدها یا هزاران درخت طبقه‌بندی تولید می‌شود (۹). این روش یک روش یادگیری گروهی است و برای طبقه‌بندی با ساختن تعداد درختان زیاد عمل می‌نماید (۸). کلیه مراحل مدل‌سازی با استفاده از روش یادگیری جنگل تصادفی با استفاده از بسته Random Forest به همراه کدنویسی در محیط نرم‌افزار R_{Studio} انجام شد.

مدل درخت تصمیم توسعه‌یافته

رگرسیون درختی توسعه‌یافته به‌عنوان یکی از الگوریتم-های یادگیری ماشین ترکیبی از دو تکنیک آماری بوستینگ[1] و رگرسیون درختی است (۱). بوستینگ یک روش مرحله‌ای روبه‌جلو است که در آن مدل‌های درختی به‌صورت تکرارپذیر با زیرمجموعه‌ای از داده‌های آموزشی برازش داده می‌شوند. در برازش رگرسیون درختی توسعه‌یافته باید دو پارامتر نرخ یادگیری[2] و پیچیدگی درخت[3] مشخص گردند. نرخ یا مقدار یادگیری سهم هر درخت متوالی را در مدل نهایی تعیین می‌کند. پیچیدگی درخت اثرات اصلی یا اثرات متقابل بین متغیرها را نشان می‌دهد (۱۷). مدل‌سازی رگرسیون درختی توسعه‌یافته در نرم‌افزار R_{studio} و با بسته C5.0 انجام شد.

1- Boosting
2- Learning rate
3- Tree complexity

مدل تجمعی

در تجزیه‌وتحلیل داده‌ها و مدل‌سازی پیش‌بینی، یک مدل فردی مبتنی بر یک نمونه داده می‌تواند دارای تنوع زیاد، تعداد زیادی نادرستی یا سوگیری‌های گسترده باشد که بر قابلیت اطمینان نتایج تأثیر می‌گذارد (۴۰). اثرات این محدودیت‌ها را می‌توان با تجزیه‌وتحلیل نمونه‌های متعدد یا ترکیب مدل‌های مختلف کاهش داد که می‌تواند به ارائه اطلاعات بهتر به تصمیم‌گیرندگان و بهبود الگوریتم‌های مدل‌سازی کمک کند (۱۱ و ۴۰). در روش تجمعی از چندین الگوریتم یادگیری اصطلاحاً ضعیف[1] که مدل‌های پایه هستند، استفاده و با ترکیب چندین مدل ضعیف، مدل پیچیده‌ای ایجاد می‌شود در این مطالعه، مدل‌های با پیش‌بینی مکانی ضعیف‌تر تحت یک سناریو با استفاده از تابع combine (موجود در بسته EnsembleCV) در محیط نرم‌افزار R_{studio} ادغام و مدلی جدید با عملکرد بهتر ایجاد شد.

ارزیابی دقت مدل‌های پیش‌بینی کننده

برای بررسی صحت مدل مورداستفاده، داده‌ها به‌طور تصادفی به داده‌های آموزشی و اعتبارسنجی تقسیم شدند. ۸۰ درصد داده‌ها (۱۱۸ پروفیل) برای آموزش مدل و ۲۰ درصد (۳۰ پروفیل) دیگر به‌عنوان داده‌های اعتبارسنجی برای ارزیابی مورداستفاده قرار گرفتند. هر مدل با داده‌های آموزشی برازش داده شد و سپس پیش‌بینی برای داده‌های اعتبارسنجی انجام شد. کلاس‌های پیش‌بینی‌شده با استفاده از ماتریس خطا برحسب درصد بیان شد. پارامترهای استخراج‌شده از ماتریس خطا شامل صحت کلی نقشه[2]، صحت تولیدکننده[3]، صحت کاربر[4] و ضریب کاپا[5] برای اعتبارسنجی مورداستفاده قرار گرفت (۲۵).

صحت کلی

رابطه (۱) صحت کلی نقشه را نشان می‌دهد که از تقسیم کل تعداد کلاس به‌درستی پیش‌بینی‌شده بر تعداد کل پیکسل‌های ماتریس خطا (N) به دست می‌آید. در این معادله Xiiتعداد مشاهده‌ها در ردیف i و ستون i است، k تعداد سطرها در ماتریس خطا است. صحت کلی طبقه‌بندی ارتباط بین همه داده‌های مورداستفاده و داده‌های طبقه‌بندی‌شده را نشان می‌دهد و از جمله پارامترهای اندازه‌گیری است که فقط دقت کلی را گزارش می‌کند و در مورد هر کدام از طبقات به‌طور مجزا اطلاعاتی ارائه نمی‌کند.

$$OA = \sum_{i=1}^{k} X_{ii}/N$$ (رابطه ۱):

شاخص کاپا

آماره کاپا یک شاخص قوی است که نسبت احتمال حضور یا عدم حضور کلاس‌هایی که به‌درستی به‌وسیله مدل پیش‌بینی شدند را محاسبه می‌کند. شاخص کاپا معیاری برای مقایسه طبقه‌بندی مدل خودکار با طبقه‌بندی تصادفی است (رابطه ۲). دامنه تغییرات آماره کاپا بین صفر تا یک است. اگر کاپا برابر با صفر باشد نشان‌دهنده طبقه‌بندی کاملاً تصادفی و مقدار منفی نشان‌دهنده خطا در طبقه‌بندی و اگر این مقدار برابر با یک باشد نشان‌دهنده طبقه‌بندی کاملاً صحیح است.

$$Kappa = N\sum_{i=1}^{k} X_{ii} - \sum_{i=1}^{k}(X_{i+} \times X_{+j})/N^2 - \sum_{i=1}^{k}(X_{i+} \times X_{+j})$$ (رابطه ۲):

که در آن Xi+ و X+j به ترتیب مجموع حاشیه‌ای برای ردیف i و ستون j هستند. مقادیر کاپا بیشتر از ۰/۸ نشان‌دهنده توافق یا دقت قوی بین نقشه طبقه‌بندی و اطلاعات مرجع زمینی است. مقادیر بین ۰/۴ و ۰/۸ نشان‌دهنده توافق متوسط و مقادیر کمتر از ۰/۴ نشان‌دهنده توافق ضعیف است (۱۴).

صحت تولیدکننده

قابلیت اطمینان تولیدکننده ارتباط بین همه کلاس‌های صحیح پیش‌بینی‌شده و مجموع کلاس‌های پیش‌بینی‌شده (کلاس‌های حضور مشاهده‌شده که به اشتباه جزء کلاس‌های

1- Weak learn
2- Overall accuracy, OA
3- Producer accuracy, PA
4- Users accuracy, UA
5- Kappa index

عدم حضور پیش‌بینی شده‌اند) است. با توجه به رابطه (۳) تعداد کل پیکسل‌های صحیح در یک کلاس تقسیم بر تعداد کل پیکسل‌های آن کلاس که از داده‌های مرجع زمین (کل ستون) تعیین می‌شود، صحت تولیدکننده نامیده می‌شود.

$$PA = \frac{X_{jj}}{X_{+j}}$$

(رابطه ۳):

صحت کاربر

اگر تعداد کل پیکسل‌های صحیح در یک کلاس بر تعداد کل پیکسل‌هایی که واقعاً در آن دسته طبقه‌بندی شده‌اند (کل ردیف) تقسیم شود صحت کاربر نامیده می‌شود و از رابطه (۴) به دست می‌آید. دامنه تغییرات صحت تولیدکننده و صحت کاربر حد واسط صفر و یک است که در نتیجه مقادیر بالاتر نشان‌دهنده عملکرد مناسب مدل است.

$$UA = \frac{X_{ii}}{X_{+i}}$$

(رابطه ۴):

نتایج و بحث

متغیرهای محیطی منتخب برای پیش‌بینی کلاس‌های خاک

بر اساس اهداف تحقیق و روش‌های انتخاب متغیر محیطی از میان ۵۷ متغیر محیطی تولیدشده، ۱۰ متغیر محیطی شامل اطلاعات نقشه‌های ژئومورفولوژی، اطلاعات زمین‌شناسی و ویژگی‌های مستخرج از مدل رقومی ارتفاع شامل تجزیه‌وتحلیل سایه‌اندازی تپه‌ها[1]، طلوع خورشید[2]، عمق دره[3]، شاخص طول در جهت شیب[4]، فاصله تا شبکه آبراهه[5]، شاخص رطوبتی توپوگرافی[6] و شاخص همواری بالای پشته با درجه تفکیک بالا[7] به‌عنوان مؤثرترین متغیرهای محیطی برای پیش‌بینی کلاس‌های خاک شناخته‌شده و به‌عنوان ورودی مدل انتخاب شدند (جدول ۲). اثر فرآیندهای ژئومورفولوژی در توسعه خاک توسط تعدادی از محققین مانند اسکول و همکاران[8] (۴۱)، جعفری و همکاران[9] (۲۳)، تقی‌زاده و همکاران[10] (۵۱) بررسی شده است. آن‌ها تأکید کردند اطلاعات ژئومورفولوژی مهم‌ترین متغیر کمکی برای تخمین نقشه کلاس‌های خاک در مناطق خشک و نیمه‌خشک است. ادهیکاری و همکاران[11] (۲) در مطالعه خود بیان کردند نقشه زمین‌شناسی در کنار مقدار رس خاک سطحی و ارتفاع، مهم‌ترین پارامترهای محیطی برای بیان تغییرات خاک با استفاده از سامانه رده‌بندی جهانی هستند. جعفری و همکاران (۲۴) و ویس و لاگچری[12] (۵۵) در مطالعه خود گزارش کردند که ویژگی‌های توپوگرافی و فرآیندهای ژئومورفیک در پیش‌بینی کلاس‌های خاک از عوامل مؤثر بر تشکیل خاک در مناطق موردمطالعه هستند.

نتایج آماری

نتایج طبقه‌بندی خاک‌رخ‌ها نشان داد که خاک‌های منطقه در دو رده انتی‌سولز و اینسپتی‌سولز قرار می‌گیرند. این خاک‌ها در سطح زیرگروه در پنج کلاس تیپیک کلسی-زرپتز[13]، تیپیک هاپلوزرپتز[14]، جیپسیک هاپلوزرپتز[15]، تیپیک زراورتنتز[16] و لیتیک زراورتنتز[17] طبقه‌بندی شدند. با توجه به درصد فراوانی مشاهده‌های کلاس‌های خاک در منطقه، زیرگروه‌های خاک تیپیک کلسی‌زرپتز، تیپیک هاپلوزرپتز و تیپیک زراورتنتز به ترتیب با ۳۲/۴۳ درصد، ۱۷/۵۶ درصد و ۲۰/۹۴ درصد در کلاس اکثریت (با فراوانی داده بیشتر) و زیرگروه‌های خاک جیپسیک هاپلوزرپتز و لیتیک زراورتنتز با ترتیب با ۸/۱ درصد و ۷/۳۴ درصد در کلاس اقلیت (با فراوانی داده کمتر) قرار می‌گیرند (جدول ۳).

1- Analytical hillshading
2- Sunrise
3- Valley depth
4- LS_Factor
5- Channel network distance, CND
6- Topographic wetness index, TWI
7- Multi-Resolution ridge top flatness index, MRRTF

8- Scull *et al.*
9- Jafari *et al.*
10 Taghizadeh-Mehrjardi *et al.*
11- Adhikari *et al.*
12- Vaysse and Lagacherie
13- Typic Calcixerepts
14- Typic Haploxerepts
15- Gypsic Haploxerepts
16- Typic Xerorthents
17- Lithic Xerorthents

نتایج مقادیر صحت‌سنجی پیش‌بینی مکانی هر یک از کلاس‌های خاک بر اساس چهار شاخص صحت کلی، شاخص کاپا، صحت کاربر و صحت تولیدکننده توسط الگوریتم‌های یادگیرنده رگرسیون لجستیک چندجمله‌ای، جنگل تصادفی و درخت تصمیم توسعه‌یافته در جداول ۴ و ۵ ارائه شده است. مطابق با نتایج ارائه‌شده، شاخص کاپا در مدل رگرسیون لجستیک چندجمله‌ای ۰/۴۱، در مدل جنگل تصادفی ۰/۳۲ و در مدل درخت تصمیم توسعه‌یافته ۰/۳۵ است که مؤید آن است که این شاخص در مدل رگرسیون لجستیک چندجمله‌ای دارای توافق متوسط و در دو مدل دیگر دارای توافق ضعیف است.

جدول (۲) متغیرهای محیطی منتخب برای مدل‌سازی جهت پیش‌بینی کلاس‌های خاک
Table (2) Selected environmental covariates for modeling to predict soil classes

متغیر محیطی مورداستفاده Environmental Covariate used	نوع متغیر Covariate type	مقیاس Scale	نام متغیر Covariate name	منبع Source
نقشه ژئومورفولوژی Geomorphology map	وکتوری Vector	۱:۵۰۰۰۰ 1:50000	نقشه ژئومورفولوژی Geomorphology map	Zinck et al (2016)
نقشه زمین‌شناسی Geology map	وکتوری Vector	۱:۲۵۰۰۰۰ 1:250000	نقشه زمین‌شناسی Geology map	Iran National Cartographic Center
مدل رقومی ارتفاع Digital elevation model	رستری Raster	۳۰ متر 30 meters	مدل رقومی ارتفاع DEM	ALOS PLASAR (2011)
			تجزیه‌وتحلیل سایه‌اندازی تپه‌ها Analytical Hill shading	Olaya (2004)
			زمان طلوع خورشید Sunrise	Olaya (2004)
			عمق دره Valley Depth	Olaya (2004)
			شاخص طول در جهت شیب LS Factor	Olaya (2004)
			فاصله تا شبکه آبراهه Channel Network Distance	Olaya (2004)
			شاخص رطوبتی Topographic Wetness Index (TWI)	Olaya (2004)
			شاخص همواری بالای پشته با درجه تفکیک بالا Multi-resolution Ridge Top Flatness (MrRTF)	Olaya (2004)

جدول(۳) تعداد و درصد فراوانی کلاس‌های خاک در سطح زیرگروه

Table(3) The number and percentage of abundance of soil classes at the subgroup level

تحت گروه‌های خاک Soil subgroups	تعداد مشاهده‌ها Number of observations	درصد مشاهده‌ها Percentage of observations
تیپیک کلسی‌زرپتز Typic Calcixerepts	68	32.43
تیپیک هاپلوزرپتز Typic Haploxerepts	26	17.56
جیپسیک هاپلوزرپتز Gypsic Haploxerepts	12	8.1
تیپیک زراورتنتز Typic Xerorthents	31	20.94
لیتیک زراورتنتز Lithic Xerorthents	11	7.34

مقادیر شاخص صحت کلی برای مدل جنگل تصادفی، مدل درخت تصمیم توسعه‌یافته و رگرسیون لجستیک چندجمله‌ای به ترتیب برابر ۶۵، ۶۰ و ۶۶ درصد پیش‌بینی‌شده و مدل رگرسیون لجستیک چندجمله‌ای بهترین عملکرد را در میان مدل‌های فردی برای پیش‌بینی زیرگروه‌های خاک را نشان داد. در پژوهشی مشابه توسط فاتحی و همکاران (۱۸) در بخشی از اراضی استان کرمانشاه نشان داده شد که صحت کلی نقشه پیش‌بینی زیرگروه‌های خاک با روش‌های رگرسیون لجستیک چندجمله‌ای و درختان طبقه‌بندی به ترتیب ۵۰ و ۴۷ درصد است که مؤید آن است که مدل رگرسیون لجستیک توانایی بیشتری برای پیش‌بینی زیرگروه‌های خاک در منطقه موردمطالعه داشته است. جین و همکاران[1] (۲۶) در مطالعه نقشه‌برداری رقومی کلاس‌های خاک در غرب هائیتی با مدل رگرسیون لجستیک چندجمله‌ای و مدل درختان تصادفی به این نتیجه رسیدند که رگرسیون لجستیک چندجمله‌ای با مقدار شاخص کاپا ۰/۴۵ کارایی بهتری نسبت به مدل درخت تصادفی داشته است.

نتایج برازش مدل‌های یادگیری ماشین بر روی کلاس‌های خاک نشان داد از میان سه الگوریتم جنگل تصادفی، درخت تصمیم توسعه‌یافته و رگرسیون لجستیک چندجمله‌ای، دو الگوریتم جنگل تصادفی و درخت تصمیم توسعه‌یافته به‌عنوان مدل ضعیف‌تر شناخته شدند؛ بنابراین این دو الگوریتم ضعیف تحت یک سناریو با هم ترکیب‌شده و مدل‌سازی با استفاده از آن‌ها انجام شد. مدل تجمعی جنگل تصادفی و درخت تصمیم توسعه‌یافته تحت عنوان مدل RF-BRT[2] نامیده شدند. مطابق با نتایج ارائه‌شده در جدول (۴)، پس از متعادل‌سازی داده‌ها با رویکرد مدل تجمعی مقادیر شاخص کاپا و صحت کلی به ترتیب برابر با ۰/۶۲ و ۷۰ درصد به دست آمد. طبق نتایج ارائه‌شده این‌گونه استنباط می‌شود که الگوریتم‌های یادگیرنده زمانی که به‌تنهایی و به‌صورت فردی مدل‌سازی می‌شوند از دقت بالایی در پیش‌بینی مکانی کلاس‌های خاک برخوردار نیستند اما مجموعه این الگوریتم‌ها در قالب مدل تجمعی دقت بالایی در پیش‌بینی مکانی کلاس‌های خاک از خود نشان می‌دهند. ازآنجایی‌که خطای پیش‌بینی مرتبط با هر مدل فردی به‌طور تصادفی توزیع می‌شود، اطلاعات مفیدی که از پیش‌بینی‌های هر مدل فردی به دست می‌آید بسیار محدود است (۱۳)؛ بنابراین مدل تجمعی به دلیل استفاده از اطلاعات همه مدل‌های یادگیری ماشین، پتانسیل بالایی در بهبود پیش‌بینی‌ها نسبت به مدل‌های معمول از خود نشان می‌دهد (۵۳). تقی‌زاده و همکاران (۵۲) در مطالعه نقشه‌برداری رقومی کلاس‌های خاک

1- Jeune *et al.*

2- Random forest and boosted regression trees, RF-BRT

با استفاده از مدل تجمعی شاخص کاپا و صحت کلی را به ترتیب برای مدل‌های فردی جنگل تصادفی ۰/۵۷ و ۸۷ درصد، درخت تصمیم ۰/۴۶ و ۸۱ درصد، رگرسیون لجستیک چندجمله‌ای ۰/۲۳ و ۴۶ درصد و در یک رویکرد تجمعی به ترتیب برابر با ۰/۶۶ و ۹۰ درصد به دست آوردند و به این نتیجه رسیدند مدل تجمعی عملکرد بهتری را در مقایسه با رویکردهای فردی نشان داده است و پیش‌بینی نادرست کلاس‌های خاک را در سطح زیرگروه کاهش داده است. این نتیجه، یک نتیجه مورد انتظار برای این نوع روش مدل‌سازی تجمعی است (۳۴). سیلوین[1] و همکاران (۵۰) نیز در مطالعه خود دریافتند عملکرد پیش‌بینی‌های قطعی به‌دست‌آمده از مدل‌سازی تجمعی بهتر از بسیاری از مدل‌های اجزای جداگانه آن است. ده بوی و همکاران[2] (۵۴) به این نتیجه رسیدند که مدل تجمعی می‌تواند به‌طور قابل‌توجهی دقت پیش‌بینی مدل را بهبود بخشد.

جدول (۵) نتایج دو شاخص صحت کاربر و صحت تولیدکننده برای کلاس‌های خاک در سطح زیرگروه در دو حالت با داده‌های نامتعادل و پس از متعادل‌سازی داده‌ها با رویکرد مدل تجمعی را نشان می‌دهد. مطابق با نتایج اعتبارسنجی دو مدل فردی جنگل تصادفی و درخت تصمیم توسعه‌یافته موفق به پیش‌بینی زیرگروه‌های جیپسیک هاپلوزرپتز و لیتیک زراورتنز که جزء کلاس‌های اقلیت محسوب می‌شوند نشده‌اند؛ اما مدل فردی رگرسیون لجستیک چندجمله‌ای برای دو کلاس کم رخداد (اقلیت) جیپسیک هاپلوزرپتز و لیتیک زراورتنز به ترتیب مقادیر ۳۴ و ۱۰۰ درصد در صحت تولیدکننده و ۲۰ و ۱۰۰ درصد در صحت کاربر را نشان داد. از طرفی نتایج صحت‌سنجی برای مدل تجمعی نشان داد که دو کلاس اقلیت جیپسیک هاپلوزرپتز و لیتیک زراورتنز به ترتیب با مقادیر ۵۰ و ۹۷ درصد در صحت تولیدکننده و ۵۰ و ۱۰۰ درصد در صحت کاربر با بهبود پیش‌بینی همراه است. البته مدل تجمعی بااینکه موفق به پیش‌بینی خوب کلاس‌های اقلیت خاک شده است اما نسبت به مدل فردی رگرسیون لجستیک چندجمله‌ای مقادیر کمتری در پیش‌بینی برخی کلاس‌های اکثریت خاک به‌ویژه دو کلاس تیپیک هاپلوزرپتز و تیپیک زراورتنز نشان داده است (شکل ۳). در واقع این‌گونه استنباط می‌شود که در میان مدل‌های فردی، مدل رگرسیون لجستیک چندجمله‌ای دقت بالاتری در پیش‌بینی کلاس‌های خاک از خود نشان می‌دهد و به‌عنوان مدل قوی‌تر عملکرد بهتری دارد اما در مدل تجمعی به دلیل اینکه دو مدل ضعیف‌تر (جنگل تصادفی و درخت تصمیم توسعه‌یافته) در مدل‌سازی دخیل هستند در برخی موارد باعث کاهش مقادیر صحت تولیدکننده و صحت کاربر نسبت به مدل فردی رگرسیون لجستیک چندجمله‌ای شده است. بررسی مطالعات متعدد نشان داده است که محققان اغلب از روش‌ها یا مدل‌های مختلفی برای پیش‌بینی کلاس‌های خاک بسته به شرایط استفاده می‌کنند. تقریباً همه آن‌ها بیان کردند که هر کدام از مدل‌های یادگیری ماشین عملکرد منحصربه‌فرد خود را دارد و دارای نقاط قوت و ضعف خاص خود است (۳۸، ۵۲ و ۵۴). نتایج پژوهش حاضر نیز نشان داد که توزیع نامتعادل کلاس‌های خاک می‌تواند بر خروجی مدل‌های پیش‌بینی، بدون توجه به نوع الگوریتم استفاده‌شده تأثیر بگذارد و متعادل‌سازی داده‌ها سبب افزایش دقت پیش‌بینی مکانی کلاس‌های خاک و نقشه تولیدشده می‌گردد. از طرفی هر مدل خاص در پیش‌بینی یک کلاس طبقه‌بندی خاک برتری دارد، اما به‌طورکلی مدل‌های تجمعی عملکرد و دقت بالاتری در مقایسه با مدل‌های فردی دارند.

1- Sylvain
2- Tien Bui *et al.*

.

جدول(٤) صحت پیش‌بینی سطح تاکسونومیک زیرگروه قبل و بعد از متعادل‌سازی داده‌ها توسط الگوریتم‌های یادگیرنده

Table(4) Prediction accuracy of the taxonomic level of the subgroup before and after Data balancing by learning algorithms

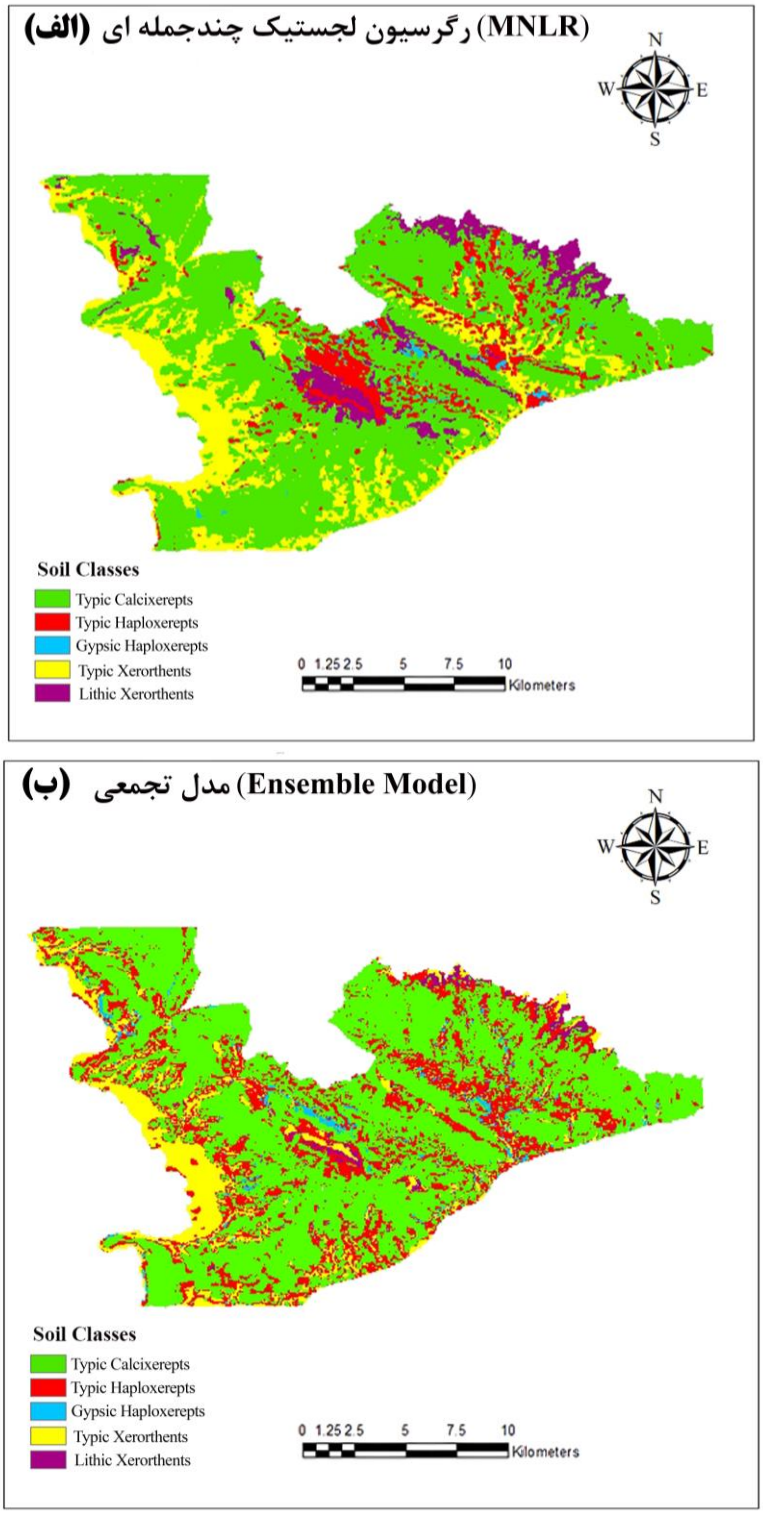
مدل‌های یادگیری ماشین Machine learning models	نوع داده‌ها Data type	شاخص‌های صحت‌سنجی Validation indicators ضریب کاپا Kappa coefficient	 صحت کلی (٪) Overall accuracy (%)
جنگل تصادفی RF	نامتعادل Imbalanced dataset	0.32	65
درخت تصمیم توسعه‌یافته BRT	نامتعادل Imbalanced dataset	0.35	60
رگرسیون لجستیک چندجمله‌ای MNLR	نامتعادل Imbalanced dataset	0.41	66
مدل تجمعی (جنگل تصادفی _ درخت تصمیم توسعه‌یافته) Ensemble Model (RF_BRT)	متعادل Balanced dataset	0.62	70

جدول(٥) صحت تولیدکننده و کاربر برای کلاس‌های خاک در سطح زیرگروه با داده‌های نامتعادل و متعادل بر اساس مدل-های برازش داده‌شده

Table(5) Producer and User accuracy for soil classes at the subgroup level before and after data balancing based on the fitted models

قابلیت اطمینان Validation	صحت تولیدکننده (٪) Producer accuracy (%)				صحت کاربر (٪) User accuracy (%)			
نوع داده‌ها Data type	داده‌های نامتعادل Imbalanced dataset			داده‌های متعادل Balanced dataset	داده‌های نامتعادل Imbalanced dataset			داده‌های متعادل Balanced dataset
کلاس خاک سطح زیرگروه Subgroup of soil	جنگل تصادفی RF	درخت تصمیم توسعه‌یافته BRT	رگرسیون لجستیک چندجمله‌ای MNLR	مدل تجمعی Ensemble model	جنگل تصادفی RF	درخت تصمیم توسعه‌یافته BRT	رگرسیون لجستیک چندجمله‌ای MNLR	مدل تجمعی Ensemble model
Typic Calcixerepts	85	80	94	96	61	62	64	88
Typic Haploxerepts	50	50	67	100	100	67	100	69
Gypsic Haploxerepts	0	0	34	50	NaN	NaN	20	50
Typic Xerorthents	34	40	40	36	65	40	67	65
Lithic Xerorthents	0	0	100	97	NaN	NaN	100	100

*NaN: عدد نیست، هیچ پیش‌بینی برای این کلاس انجام‌نشده است.



شکل(۳) نقشه‌های تولیدشده توسط الگوریتم‌های یادگیری ماشین الف) با داده‌های نامتعادل و ب) بعد از متعادل‌سازی داده‌ها با رویکرد تجمعی

Figure(3) Maps produced by machine learning algorithms a) with Imbalanced dataset and b) after data balancing with Ensemble Model

نتیجه‌گیری

عدم تعادل داده‌ها یکی از رایج‌ترین مشکلاتی است که در نمونه‌های خاک دیده می‌شود و موجب افزایش خطا در نتایج به‌دست‌آمده با استفاده از الگوریتم‌های طبقه‌بندی معمول می‌گردد. به عبارتی توزیع نامتعادل کلاس‌های خاک می‌تواند بر خروجی مدل‌های پیش‌بینی، بدون توجه به نوع الگوریتم استفاده‌شده تأثیر بگذارد و متعادل‌سازی داده‌ها سبب افزایش دقت پیش‌بینی مکانی کلاس‌های خاک و نقشه تولیدشده می‌شود. نتایج نشان داد که پیش‌بینی مکانی کلاس‌های خاک با استفاده از الگوریتم‌های یادگیرنده فردی یا جداگانه موجب کاهش دقت مدل‌سازی می‌شود، اما مجموعه این الگوریتم‌ها در قالب مدل تجمعی دقت بالایی در پیش‌بینی مکانی کلاس‌های خاک از خود نشان می‌دهند. در مقایسه با الگوریتم‌های طبقه‌بندی بررسی‌شده، روش تجمعی به نتایج نسبتاً بهتری هنگام استفاده در مجموعه داده‌های نامتعادل، متشکل از نسبت بالاتری از نمونه‌های اقلیت، دست یافت. در واقع مدل تجمعی پتانسیل بالایی برای بهبود پیش‌بینی کلاس‌های اقلیت خاک نسبت به الگوریتم‌های معمول دارد و اطلاعات دقیق‌تری از خاک برای نظارت بر وضعیت خاک و تغییرات در فضا و زمان با استفاده از نقشه‌برداری رقومی خاک ارائه می‌کند، بنابراین این روش را می‌توان برای مدل‌سازی بهتر کلاس‌های خاک توصیه کرد. با توجه به اینکه مطالعات در زمینه داده‌های نامتعادل خاک در ایران بسیار محدود است پیشنهاد می‌گردد از رویکردهای دیگر متعادل‌سازی داده‌ها همچون یادگیری حساس به هزینه برای نقشه‌برداری رقومی خاک استفاده شود.

References

1. Abeare, S. 2009. Comparisons of boosted regression tree, GLM and GAM performance in the standardization of yellowfin tuna catch-rate data from the Gulf of Mexico lonline [sic] fishery. Louisiana State University and Agricultural and Mechanical College.
2. Adhikari, K., Hartemink, A.E., Minasny, B., Bou Kheir, R., Greve, M.B. and Greve, M.H. 2014. Digital mapping of soil organic carbon contents and stocks in Denmark. PloS one, 9(8), p. e105519.
3. Adhikari, K., Owens, P.R., Ashworth, A.J., Sauer, T.J., Libohova, Z., Richter, J.L. and Miller, D.M. 2018. Topographic controls on soil nutrient variations in a silvopasture system. Agrosystems, Geosciences and Environment, 1(1):.1-15.
4. Artieda, O., Herrero, J., and Drohan, P. J. 2006. Refinement of the differential water loss method for gypsum determination in soils. Soil Science Society of America Journal, 70(6): 1932-1935.
5. Błaszczyński, J. and Stefanowski, J. 2015. Neighbourhood sampling in bagging for imbalanced data. Neurocomputing, 150: 529-542.
6. Bouyoucos, G. J. 1962. Hydrometer method improved for making particle size analyses of soils 1. Agronomy Journal, 54(5): 464-465.
7. Branco, P., Torgo, L. and Ribeiro, R.P. 2016. A survey of predictive modeling on imbalanced domains. ACM Computing Surveys, 49(2): 1-50.
8. Breiman, L. 2001. Random forests. Machine Learning, 45(1): 5-32.
9. Breiman, L., and Cutler, A. 2004. Random Forests. Department of Statistics, University of Berkeley.
10. Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A., and Edwards Jr, T. C. 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma, 239: 68-83.
11. Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D.T., Duan, Z. and Ma, J. 2017. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. Catena, 151: 147-160.
12. Chen, S., Mulder, V.L., Heuvelink, G.B., Poggio, L., Caubet, M., Dobarco, M.R., Walter, C. and Arrouays, D. 2020. Model averaging for mapping topsoil organic carbon in France. Geoderma, 366: 114237.
13. Chen, S., Xue, J. and Shi, Z. 2023. Spectral-guided ensemble modelling for soil spectroscopic prediction. Geoderma, 437: 116594.
14. Congalton, R.G. 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment, 37(1): 35-46.
15. Diks, C.G. and Vrugt, J.A. 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stochastic Environmental Research and Risk Assessment, 24, pp.809-820.
16. Dobarco, M.R., Arrouays, D., Lagacherie, P., Ciampalini, R. and Saby, N.P. 2017. Prediction of topsoil texture for Region Centre (France) applying model ensemble methods. Geoderma, 298: 67-77.
17. Elith, J., Leathwick, J.R. and Hastie, T. 2008. A working guide to boosted regression trees. Journal of Animal Ecology, 77(4): 802-813.
18. Fatehi, Sh., Mohammadi, J., Salehi, M., Momeni, A., Tomanian, T., Jafari, A. 2014. Spatial de-clustering of traditional soil map using multi-class logistic regression and classification trees). Case study: Merck watershed sub-basin in Kermanshah province (14th Congress of Soil Sciences of Iran), Rafsanjan. 208-213. (In Persian)

19. Galar, M., Fernández, A., Barrenechea, E. and Herrera, F. 2013. EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. Pattern Recognition, 46(12): 3460-3471.

20. Garg, K.K., Anantha, K.H., Nune, R., Akuraju, V.R., Singh, P., Gumma, M.K., Dixit, S. and Ragab, R. 2020. Impact of land use changes and management practices on groundwater resources in Kolar district, Southern India. Journal of Hydrology: Regional Studies, 31: 100732.

21. Górecki, T. and Krzyśko, M. 2015. Regression methods for combining multiple classifiers. Communications in Statistics-Simulation and Computation, 44(3): 739-755.

22. Gruszczynski, ´ S., Gruszczynski, ´ W. 2022. Supporting soil and land assessment with machine learning models using the Vis-NIR spectral response. Geoderma 405: 115451.

23. Jafari. A., Finke P.A, Van deWauw, J., Ayoubi, S., and Khademi, H. 2012. Spatial prediction of USDA-great soil groups in the arid Zarand region, Iran: comparing logistic regression approaches to predict diagnostic horizons and soil types. Europian Journal Soil Science, 63(2): 284–298.

24. Jafari, A., Ayoubi, S., Khademi, H., Finke, P.A. and Toomanian, N. 2013. Selection of a taxonomic level for soil mapping using diversity and map purity indices: a case study from an Iranian arid region. Geomorphology, 201: 86-97.

25. Jensen, J.R. 1996. Introductory digital image processing: a remote sensing perspective (No. Ed. 2). Prentice-Hall Inc.

26. Jeune, W., Francelino, M.R., Souza, E.D., Fernandes Filho, E.I. and Rocha, G.C. 2018. Multinomial logistic regression and random forest classifiers in digital mapping of soil classes in western Haiti. Revista Brasileira de Ciência do Solo, 42.

27. Kempen, B., Brus, D.J., Stoorvogel, J.J., Heuvelink, G.B. and de Vries, F. 2012. Efficiency comparison of conventional and digital soil mapping for updating soil maps. Soil Science Society of America Journal, 76(6): 2097-2115.

28. Kleinbaum, A.M. 2018. Reorganization and tie decay choices. Management Science, 64(5): 2219-2237.

29. Koziarski, M. 2020. Radial-based undersampling for imbalanced data classification. Pattern Recognition, 102: 107262.

30. Krawczyk, B., Woźniak, M. and Cyganek, B. 2014. Clustering-based ensembles for one-class classification. Information Sciences, 264: 182-195.

31. Kuhn, M. and Johnson, K. 2013. Applied predictive modeling (Vol. 26, p. 13). New York: Springer.

32. Lanyon, L. E., and Heald, W. R. 1983. Magnesium, calcium, strontium, and barium. Methods of Soil Analysis: Part 2 Chemical and Microbiological Properties, 9: 247-262.

33. Ludwig, B., Murugan, R., Parama, V.R. and Vohland, M. 2019. Accuracy of estimating soil properties with mid- infrared spectroscopy: Implications of different chemometric approaches and software packages related to calibration sample size. Soil Science Society of America Journal, 83(5): 1542-1552.

34. Malone, B.P., Minasny, B., McBratney, A.B., Malone, B.P., Minasny, B. and McBratney, A.B. 2017. Digital Soil Assessments. Using R for Digital Soil Mapping: 245-260.

35. Meng, X.T., Yan, F.G., Cao, B.X., Jin, M. and Zhang, Y. 2022. Efficient real-valued DOA estimation based on the trigonometry multiple angles transformation in monostatic MIMO radar. Digital Signal Processing, 123: 103437.

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

37. Perry Jr, C. R., & Lautenschlager, L. F. 1984. Functional equivalence of spectral vegetation indices. Remote Sensing of Environment, 14(1-3): 169-182.

38. Pourghasemi, H.R., Kariminejad, N., Amiri, M., Edalat, M., Zarafshar, M., Blaschke, T. and Cerda, A. 2020. Assessing and mapping multi-hazard risk susceptibility using a machine learning technique. Scientific Reports, 10(1): 3203.

39. Richards, A. L. (ed). 1954. Diagnosis and improvement of saline and alkaline soils. US Salinity Laboratory Staff. USDA. Handbook, No. 60, Washington DC. USA.

40. Rokach, L. 2010. Ensemble-based classifiers. Artificial Intelligence Review, 33, pp.1-39.

41. Scull, P., Franklin, J., and Chadwick, O.A. 2005. The application of classification tree analysis to soil type prediction in a desert landscape. Ecological Modelling, 181: 1–15.

42. Sharififar, A., Sarmadian, F., Malone, B.P. and Minasny, B. 2019. Addressing the issue of digital mapping of soil classes with imbalanced class observations. Geoderma, 350: 84-92.

43. Soil and Water Research Institute. 2010. Site Selection, Soil Survey and Land Evaluation for Development of Orchards in Zanjan Province, Iran. (In Persian)

44. Soil science division staff. 2017. "Soil survey manual". USDA Handbook 18: 120-131

45. Soil Survey Staff. 2022. Keys to soil taxonomy, 13th edition. USDA Natural Resources Conservation Service.

46. Statistical Yearbook of Zanjan Province. 2019. Land and Climate, National Statistics Organization. (In Persian)

47. Sumner, M. E., and Miller, W. P. 1996. Cation exchange capacity and exchange coefficients. Methods of soil analysis: Part 3 Chemical methods, 5:1201-1229.

48. Supreme Council of Science, Research and Technology. 2013. (In Persian)

49. Swiderski, B., Osowski, S., Kruk, M. and Barhoumi, W. 2016. Aggregation of classifiers ensemble using local discriminatory power and quantiles. Expert Systems with Applications, 46: 316-323.

50. Sylvain, J.D., Anctil, F. and Thiffault, É. 2021. Using bias correction and ensemble modelling for predictive mapping and related uncertainty: a case study in digital soil mapping. Geoderma, 403: 115153.

51. Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B. and Triantafilis, J. 2015. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. Geoderma. 253-254: 67–77.

52. Taghizadeh-Mehrjardi, R., Minasny, B., Toomanian, N., Zeraatpisheh, M., Amirian-Chakan, A. and Triantafilis, J. 2019. Digital mapping of soil classes using ensemble of models in Isfahan region, Iran. Soil Systems, 3(2), p.37.

53. Taghizadeh-Mehrjardi, R., Mahdianpari, M., Mohammadimanesh, F., Behrens, T., Toomanian, N., Scholten, T. and Schmidt, K. 2020. Multi-task convolutional neural networks outperformed random forest for mapping soil particle size fractions in central Iran. Geoderma, 376, p.114552.

54. Tien Bui, D., Shirzadi, A., Shahabi, H., Chapi, K., Omidavr, E., Pham, B.T., Talebpour Asl, D., Khaledian, H., Pradhan, B., Panahi, M. and Bin Ahmad, B. 2019. A novel ensemble artificial intelligence approach for gully erosion mapping in a semi-arid watershed (Iran). Sensors, 19(11): 24-44.

55. Vaysse, K. and Lagacherie, P. 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma, 291: 55-64.

56. Vohland, M., Ludwig, B., Seidel, M., Hutengs, C. 2022. Quantification of soil organic carbon at regional scale: Benefits of fusing vis-NIR and MIR diffuse reflectance data are greater for in situ than for laboratory-based modelling approaches. Geoderma, 405: 115426.

57. Walkley, A. and Black, I.A. 1934. An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. Soil Science, 37(1): 29-38.

58. Xu, Z., Shen, D., Nie, T. and Kou, Y. 2020. A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. Journal of Biomedical Informatics, 107: 103465.

59. Yang, Y., Choi, J.N. and Lee, K. 2018. Theory of planned behavior and different forms of organizational change behavior. Social Behavior and Personality: An International Journal, 46(10): 1657-1671.

60. Zhang, Y. and Hartemink, A.E. 2020. Data fusion of vis–NIR and PXRF spectra to predict soil physical and chemical properties. European Journal of Soil Science, 71(3): 316-333.

61. Zinck, J.A., Metternicht, G., Bocco, G. and Del Valle, H. 2016. Geopedology. An integration of geomorphology and pedology for soils and landscape studies: Springer International Publishing Switzerland, 556p.