نقشه‌برداری رقومی کلاس فامیل خاک با استفاده از رویکرد یادگیری ماشین (مطالعه موردی: اراضی نیمه‌خشک غرب ایران)

زیبا مقصودی[۱]، محمود رستمی‌نیا[۱*]، مرزبان فرامرزی[۲]، علی کشاورزی[۳]، اصغر رحمانی[۳] و سید روح‌اله موسوی[۳]

(تاریخ دریافت: ۱۳۹۸/۵/۲۴ ؛ تاریخ پذیرش: ۱۳۹۸/۸/۱)

چکیده

نقشه‌برداری رقومی خاک همگام با پیشرفت‌های زیرساخت داده‌های مکانی، نقش مهمی را در جهت ارتقاء دانش مطالعات خاکشناسی ایفا می‌کند. لذا تحقیق حاضر با هدف تهیه نقشه رقومی کلاس فامیل خاک با استفاده از مدل‌های جنگل تصادفی و رگرسیون درختی توسعه‌یافته در بخشی از اراضی نیمه‌خشک استان ایلام اجرا شد. متغیرهای محیطی از مدل رقومی ارتفاع با قدرت تفکیک مکانی ۳۰ متر با استفاده از نرم‌افزار SAGAGIS نسخه ۷/۳ استخراج شد. تعداد ۴۶ خاکرخ حفر و ویژگی‌های فیزیکوشیمیایی نمونه‌های خاک اندازه‌گیری و بر اساس سامانه آمریکایی ۲۰۱۴ در سطح فامیل رده‌بندی شد. در محدوده مورد مطالعه سه رده مالی‌سولز، اینسپتی‌سولز و انتی‌سولز شناسایی شد. بر اساس نتایج داده‌کاوی متغیرهای محیطی با استفاده از آنالیز تورم واریانس (VIF)، متغیرهای کمکی ارتفاع، ارتفاع استاندارد شده و شاخص زبری پستی و بلندی بیشترین میزان تغییرپذیری مکانی خاک‌ها را در منطقه مدل‌سازی می‌کند. بهترین پیش‌بینی مکانی کلاس‌های خاک مربوط به فامیل خاک Fine, carbonatic, thermic, Typic Haploxerolls است. همچنین نتایج نشان می‌دهد که مدل‌های جنگل تصادفی و رگرسیون درختی توسعه‌یافته به‌ترتیب صحت عمومی ۰/۸۰ و ۰/۶۴ و شاخص کاپای ۰/۷۰ و ۰/۵۵ را ارائه می‌کند. بنابراین، روش جنگل تصادفی می‌تواند یک روش قابل اعتماد و با دقت مناسب باشد که حتی با تعداد نمونه کم تخمین قابل قبولی را ارائه کند.

واژه‌های کلیدی: پیش‌بینی مکانی، کلاس خاک، رگرسیون درختی توسعه‌یافته، جنگل تصادفی

۱. گروه علوم و مهندسی خاک، دانشکده کشاورزی، دانشگاه ایلام، ایلام
۲. گروه مهندسی مرتع و آبخیزداری، دانشکده کشاورزی، دانشگاه ایلام، ایلام
۳. گروه علوم و مهندسی خاک، دانشکده مهندسی و فناوری کشاورزی، پردیس کشاورزی و منابع طبیعی، دانشگاه تهران، کرج
*: مسئول مکاتبات: پست الکترونیکی: m.rostaminya@ilam.ac.ir

مقدمه

با توجه به محدود بودن میزان اراضی و از طرفی افزایش جمعیت، توجه جدی به موضوع نقشه‌برداری خاک‌ها و شناسایی منابع اراضی برای بهره‌برداری از آنها در کاربری‌های اصلی از قبیل کشاورزی، مراتع، جنگل‌کاری و مهندسی ضروری به‌نظر می‌رسد (۸). از سوی دیگر به‌دلیل محدودیت روش‌های معمول تهیه نقشه‌های خاک و نیاز جدی به ارائه اطلاعات قابل اعتمادتر و به‌هنگام درباره خاک‌ها با هزینه‌های منطقی و ارتقای تفسیر نتایج به‌نحوی که افراد غیرخاکشناس و غیرمتخصص بتوانند از این اطلاعات استفاده کنند، استفاده از روش‌های نوین در تهیه نقشه خاک را ضروری می‌کند. نقشه‌برداری رقومی خاک (Digital soil mapping) می‌تواند جایگزین مناسبی برای روش‌های معمولی نقشه‌برداری خاک شود (۶). نقشه‌برداری رقومی خاک شامل روش‌ها و مدل‌هایی است که بین توزیع خاک (کلاس یا خصوصیات خاک) و داده‌هایی که به‌آسانی و با هزینه کم از طریق روش‌های سنجش از دور و داده‌های ژئومورفومتری به‌دست می‌آیند و تحت عنوان متغیرهای کمکی محیطی به‌کار برده می‌شوند، ارتباط برقرار می‌کند (۱۳). یکی از مهم‌ترین ویژگی‌های نقشه‌برداری رقومی خاک استفاده از مدل‌های کمّی مختلف به‌منظور ساده‌سازی پیچیدگی‌های موجود در سامانه طبیعی خاک است و به بررسی ارتباط بین داده‌های حاصل از مشاهدات خاک و اطلاعات کمکی که نماینده فاکتورهای خاکسازی هستند می‌پردازد. در واقع مدل‌های خاک – زمین‌نما (Soil –landscape modeling) شکل ساده شده‌ای از روابط پیچیده موجود بین الگوی پراکنش خاک‌ها در سیمای اراضی هستند. خروجی این مدل‌ها درنهایت به ارائه نقشه پیش‌بینی خاک‌ها، میزان صحت و گاهی همراه با ارائه عدم قطعیت ختم می‌شود (۱۰ و ۱۴). تا به‌حال مدل‌های مختلفی برای پیش‌بینی کلاس‌های خاک مورد بررسی قرار گرفته است که می‌توان به مدل‌های رگرسیون لاجستیک چندمتغیره (Multinomial logestic regression) ، شبکه‌های عصبی مصنوعی (Artificial neural network)، رگرسیون درختی (Tree regression)، سیستم فازی (Fuzzy system)، درختان تصمیم‌گیری تصادفی (Random Forest) و الگوریتم ژنتیک (Genetic algorithms) اشاره شد. طی مطالعه‌ای در کشور قبرس، نقشه رقومی با استفاده از دو مدل رگرسیون لاجستیک چندمتغیره و جنگل تصادفی تهیه شد. نتایج این تحقیق نشان داد که مدل جنگل تصادفی نسبت به مدل رگرسیون لاجستیک چندمتغیره توانایی و دقت بالاتری برای پیش‌بینی مکانی کلاس‌های خاک منطقه مورد مطالعه داشته و همچنین این پیش‌بینی با خطای کمتری در مقایسه با مدل لاجستیک چندمتغیره همراه بوده است (۵). نتایج عباس‌زاده افشار (۱) در پیش‌بینی مکانی گروه‌های بزرگ با استفاده از سه مدل رگرسیونی لجستیک چندجمله‌ای، رگرسیون درختی توسعه‌یافته و درخت تصمیم در جنوب شرق ایران نشان داد، بهترین پیش‌بینی مربوط به مدل درخت تصمیم بود. همچنین مدل درخت تصمیم نشان داد که ساختار درختی ایجاد بین متغیر هدف و متغیرهای انتخاب شده در مدل، باعث افزایش دقت این مدل نسبت به سایر مدل‌های رگرسیونی مورد استفاده شده است. پهلوان‌راد و همکاران (۲۱) مدل تکنیک درختان تصادفی و رگرسیون لاجستیک را برای تهیه نقشه رقومی گروه‌های بزرگ، زیرگروه‌های خاک و سری خاک در اراضی گلستان مورد استفاده قرار داد. نتایج این تحقیق نشان داد که مدل جنگل تصادفی از دقت بهتر و مطلوب‌تری نسبت به مدل رگرسیون لاجستیک برخوردار است. در مطالعه‌ای دیگر، محققان برای پیش‌بینی توزیع مکانی بافت خاک از دو روش رگرسیون درختی و روش جنگل تصادفی استفاده کرده و مشاهده شدند که روش جنگل تصادفی دارای دقت بالاتری نسبت به رگرسیون درختی بود (۱۸). مصلح و همکاران (۱۹) نقشه‌برداری رقومی کلاس‌های خاک تا سطح فامیل را با استفاده از مدل رگرسیون درختی توسعه‌یافته در دشت شهرکرد انجام دادند. نتایج این مطالعه نشان داد که از بین پارامترهای محیطی مورد استفاده در فرایند مدل‌سازی، پارامترهای مشتق شده از مدل رقومی ارتفاع بیشترین اهمیت در پیش‌بینی مکانی

کلاس‌های خاک را داشتند. سطح رده‌بندی مورد نظر، میزان تنوع خاک‌ها در منطقه مورد مطالعه، تراکم نمونه‌برداری و نوع پارامترهای محیطی مورد استفاده از مهم‌ترین عواملی هستند که می‌توانند صحت پیش‌بینی کلاس‌های خاک را تحت تأثیر قرار دهند. روش رگرسیون درختی توسعه‌یافته را با استفاده از داده‌های کم برای نواحی خشک استفاده شد و به نتایج نشان داد که این روش می‌تواند به‌عنوان یک روش قابل اعتماد برای تخمین کلاس‌های خاک مورد استفاده قرار گیرد (۱۲). همچنین در مطالعه دیگری نیز توسط تاجیک و همکاران (۳۰) به بررسی نقش سه فاکتور پیش‌بینی کننده محیطی، ژئومورفومتری، سنجش از دور و داده‌های خاک روی تنوع جوامع زیستی خاک پرداختند و گزارش کردند که مدل جنگل تصادفی در پیش‌بینی مکانی میزان فراوانی و غنی‌شدگی جوامع زیستی مورد نظر در دو عمق سطحی و زیرسطحی دارای دقت مناسبی است.

با توجه به محدودیت‌های روش نقشه‌برداری سنتی، اطلاعات اندکی از توزیع مکانی کلاس‌های خاک در ایران در مقیاس مدیریت‌پذیر وجود دارد. این محدودیت یک مسئله مشترک در سایر نقاط دنیا است که در آن اطلاعات در زمینه توزیع مکانی خاک‌ها کمیاب و درصورت وجود ناقص هستند. با توجه به نیاز کشور به مجموعه داده‌های رقومی اطلاعات خاک، این مطالعه با هدف کاربرد روش‌های رگرسیون درختی توسعه‌یافته و جنگل تصادفی برای تهیه نقشه رقومی کلاس‌های خاک در سطح فامیل و نیز ارزیابی کارایی این مدل‌ها در اراضی سیوان استان ایلام طراحی شده است.

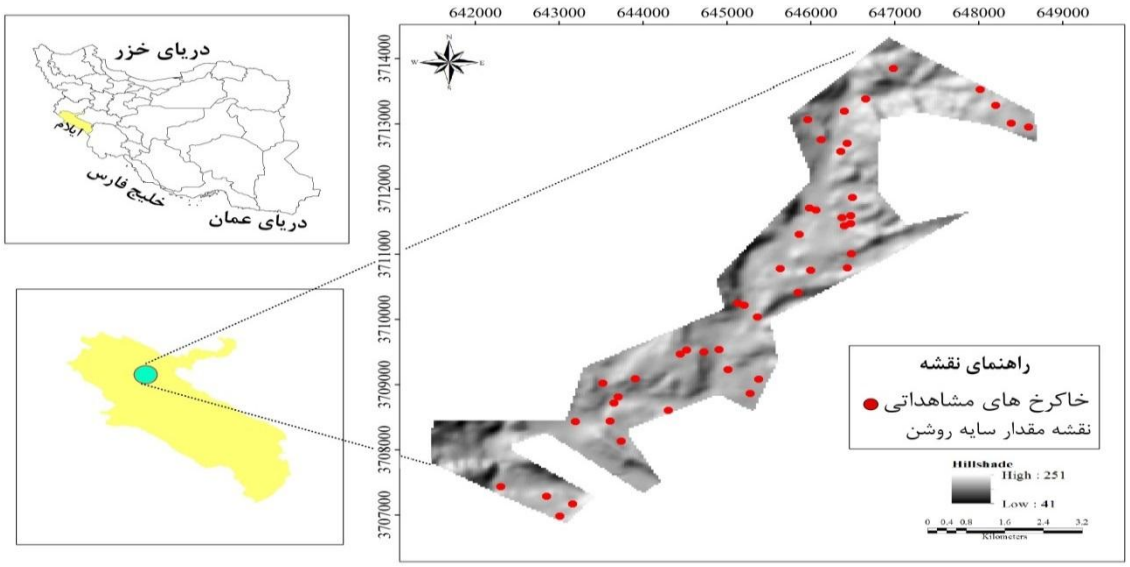
مواد و روش‌ها

منطقه مطالعاتی

منطقه مورد مطالعه با مساحتی حدود ۱۲۰۰ هکتار قسمتی از اراضی دهستان میشخاص از توابع بخش سیوان در شهرستان ایلام است که بین طول‌های جغرافیایی "۳۰ ′۳۹ °۴۶ تا "۳۰ ′۲۷ °۴۶ شرقی و عرض‌های جغرافیایی "۳۰ ′۲۵ °۳۳ تا ′۳۹ °۳۳ شمالی واقع شده است. محدوده مطالعاتی از شمال به کوه‌های سیوان، از جنوب به کوه‌های شمالی شهرستان ملکشاهی (کبیرکوه)، از غرب به ارتفاعات مجاور روستای سربیشه شرق منطقه گاوز و از شرق به ارتفاعات شرقی منطقه محدود می‌شود (شکل ۱). از نظر فیزیوگرافی اراضی مورد مطالعه شامل سه تیپ اراضی کوهستان، تپه و دشت‌های رسوبی رودخانه‌ای هستند. از نظر زمین‌شناسی، مواد مادری تشکیل‌دهنده خاک‌های منطقه شامل آهک‌های مارنی- رسی تیره رنگ، شیل‌های خاکستری رنگ، آهک‌های رسی دانه‌ریز خاکستری رنگ با لایه‌بندی منظم هستند. بر اساس روش طبقه‌بندی اقلیمی آمبرژه منطقه مورد مطالعه جزء مناطق نیمه‌مرطوب معتدل است که دارای زمستان‌های سردومرطوب و تابستان‌های گرم‌وخشک است. میانگین بارندگی سالانه ۴۸۰/۲ میلی‌متر و به‌ترتیب با متوسط، بیشینه و کمینه دمای هوا ۱۶/۸، ۲۲/۶ و ۱۱ درجه سانتی‌گراد، به استناد نقشه رژیم‌های رطوبتی و حرارتی ایران (۲) منطقه مذکور دارای رژیم رطوبتی زریک (Xeric) و رژیم حرارتی ترمیک (Thermic) است که وضعیت مورفولوژیکی خاک‌ها، مشاهدات و بررسی‌های صحرایی نیز مؤید این موضوع است. پوشش گیاهی بومی منطقه شامل خار زرد، کنگر، گیاهان مرتعی، زالزالک و درخت بلوط و کاربری غالب اراضی منطقه شامل کشت دیم گندم و جو و زراعت آبی خیار، گوجه، ذرت و لوبیا و باغ‌های مثمر گردو، زردآلو، هلو، انجیر و انگور است.

عملیات میدانی و آنالیزهای آزمایشگاهی

از تفسیر چشمی تصاویر ماهواره‌ای گوگل ارث، نقشه توپوگرافی و بازدیدهای میدانی به‌منظور تفکیک واحدهای فیزیوگرافی اولیه برای تعیین مناطق نمونه‌برداری منطقه استفاده شد. در مجموع ۴۶ خاکرخ، بر اساس روش آزاد شناسایی خاک حفر شد (شکل ۱). نمونه‌برداری و تشریح کلیه خاکرخ‌های



شکل ۱. موقعیت منطقه مورد مطالعه و نیمرخ‌های مطالعاتی

حفر شده بر اساس راهنمای تشریح و نمونه‌برداری خاک آمریکایی (۲۰۱۲) انجام شد (۲۴). آزمایش‌های فیزیکی و شیمیایی روی نمونه‌های خاک با استفاده از روش‌های استاندارد و متداول انجام گرفت. خاکرخ‌ها بر اساس سیستم رده‌بندی خاک آمریکایی تا سطح فامیل انجام شد (۲۵).

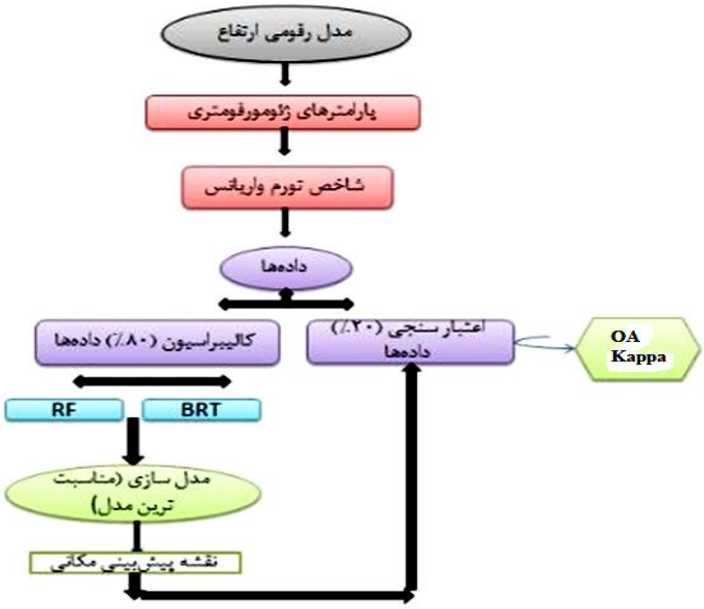
روندنمای کلی پژوهش

در شکل ۲ روندنمای کلی اجرای این پژوهش به‌منظور توصیف فعالیت‌های صورت گرفته ارائه شده است. این مراحل به‌ترتیب شامل موارد زیر است:

۱) جمع‌آوری داده‌های خاک از طریق نمونه‌برداری به‌روش آزاد شناسایی خاک، ۲) تهیه پارامترهای ژئومورفومتریک بر اساس مدل رقومی ارتفاع، ۳) انتخاب بهینه‌ترین متغیرهای محیطی کمکی، ۴) مدل‌سازی مکانی داده‌ها بر اساس دو مدل داده‌کاوی RF و BRT و ۵) تهیه نقشه پیش‌بینی مکانی بر اساس مناسب‌ترین مدل پیش‌بینی مکانی و اعتبارسنجی نتایج مدل‌سازی.

متغیرهای محیطی کمکی

از مدل رقومی ارتفاع با قدرت تفکیک مکانی ۳۰ متر، برای استخراج اطلاعات کمکی خاک استفاده شد. ویژگی‌های اولیه و ثانویه مدل رقومی ارتفاع بر اساس مطالعات انجام شده عبارتند از: جهت شیب (Acpect)، شاخص همگرایی (Convergence Index)، هموار کف دره (multiresolution index of the valley botton flatness)، شاخص بالای پشته با درجه تفکیک بالا (multiresolution index of the ridge top flatness)، ارتفاع شیب (Slope height)، عمق دره (Valley Depth)، ارتفاع نرمال شده (Normalized height)، ارتفاع استاندارد شده (Standardized height)، موقعیت میانی شیب (Mid-slope height)، شاخص موقعیت توپوگرافی (Topographic Position Index)، شاخص زبری پستی‌وبلندی (Terrain Ruggedness Index)، پستی بلندی کاذب (Analytical hillshading)، تابش مستقیم (Direct insloation)، تابش پخشیده (Diffuse insloation)، مساحت حوضه آبخیز (Catchment area)، شاخص خیسی توپوگرافی (Topographic Wetness Index)، فاکتور درصد طول شیب (Ls factor)، موقعیت نسبی شیب (Relative slope position)، انحنای سطحی (Plan curvature) و انحنای نیمرخ (Profile curvature). انتخاب متغیرهای کمکی در این مطالعه



شکل ۲. روندنمای کلی تحقیق در منطقه مورد مطالعه

بر اساس شاخص داده‌کاوی تورم واریانس در محیط نرم‌افزار R انجام شد. فاکتور تورم واریانس (Variance Inflation Factor) در واقع نشان می‌دهد که واریانس ضرایب رگرسیونی برآورد شده تا چه حد بیشتر از متغیرهای تخمینی که همبستگی خطی با هم ندارند، افزایش یافته است. مقدار VIF همواره بیشتر از ۱ است (۲۰). مقدار بیشتر از ۱۰ برای این شاخص نشان از هم‌خطی جدی و باعث عدم اعتماد به نتایج مدل می‌شود (۱۱). بر اساس این شاخص، پارامترهایی محیطی که کمترین همبستگی را با هم داشتند انتخاب و برای مدل‌سازی از آنها استفاده شد. متغیرهای محیطی مورد استفاده در مدل‌سازی شامل: ارتفاع، ارتفاع استاندارد شده و شاخص زبری پستی و بلندی در محیط نرم‌افزارهای SAGAGIS7.3 محاسبه و استخراج شد.

مدل‌سازی خاک- زمین‌نما

مدل جنگل تصادفی

مدل جنگل تصادفی یک روش ناپارامتری است که قادر به پیش‌بینی متغیرهای کمّی یا متغیرهای طبقه‌بندی شده بر اساس مجموعه‌ای از متغیرهای پیش‌بینی کننده کمّی و کیفی است. این روش اولین بار توسط بریمن (Breiman) گسترش یافت و یکی از جدیدترین روش‌های داده‌کاوی است. مدل جنگل تصادفی (RF) در حقیقت مجموعه‌هایی از درختان پیش‌بینی کننده با احتمال و پراکندگی یکسان هستند. اساس این روش بر مبنای انتخاب خطای داده‌های بزرگ‌تر به‌عنوان خطای اصلی و ایجاد همبستگی بین خطاهای دیگر داده‌ها به‌ترتیب است (۴). مدل درختان تصمیم‌گیری تصادفی (RF) مدل توسعه‌یافته از مدل طبقه‌بندی و رگرسیون درختی (Classification and Regression Tree) است. روش CART روشی است که داده‌ها را به‌صورت تکراری برای ایجاد ارتباط بین متغیر پاسخ و متغیرهای مستقل و انجام پیش‌بینی مکانی جداسازی می‌کند (۵). تکنیک RF مجموعه‌ای از شرط‌های منطقی است که به‌صورت یک الگوریتم برای پیش‌بینی کمّی یک متغیر با ساختار درختی به‌کار می‌رود. ایجاد درخت در این روش شامل دو مرحله است، مرحله اول ایجاد و رشد درخت است که شامل پیوند و انشعاب است و مرحله دوم توقف و هرس است که هدف آن به حداقل رساندن خطای پیش‌بینی است.

رگرسیون درختی توسعه‌یافته

رگرسیون درختی توسعه‌یافته از مجموعه روش‌های یادگیری ماشین و ترکیبی از دو تکنیک آماری بوستینگ و رگرسیون درختی است (۱۵). بوستینگ روشی است که در آن مدل‌های درختی به‌صورت تکرارپذیر با زیرمجموعه‌ای از داده‌های آموزشی برازش داده می‌شوند. در برازش رگرسیون درختی توسعه‌یافته باید دو پارامتر، میزان یادگیری و پیچیدگی درختی مشخص شود. میزان یادگیری سهم هر درخت متوالی را در مدل نهایی تعیین می‌کند و پیچیدگی درخت، اثرات اصلی یا اثرات متقابل بین متغیرها را نشان می‌دهد (۹). در تحقیق حاضر، متغیرهای محیطی با اندازه سلول ۳۰ مترمربع و کلاس‌های خاک وارد نرم‌افزار آماری R3.5.1 و نسخه R-Studio8.8.17 و بسته‌های نرم‌افزاری Random Forest و Caret، اجرا شد (۱۶). در مدل RF تعداد درختان در جنگل و تعداد متغیرهای محیطی در گره هر درخت توسط کاربر مشخص می‌شود. مناسب‌ترین مقدار این دو پارامتر با روش سعی و خطا برای به‌دست آوردن کمترین مقدار خطا به‌دست آمد (۲۲). پس از اتمام مدل‌سازی کلاس‌های خاک، نقشه نهایی کلاس‌های خاک برای نمایش نهایی به نرم‌افزار ArcGIS10.6.1 منتقل شد.

ارزیابی دقت مدل‌های پیش‌بینی کننده

برای بررسی صحت مدل مورد استفاده، داده‌ها به‌طور تصادفی به داده‌های آموزشی و اعتبارسنجی تقسیم شد. داده‌های آموزشی، ۸۰ درصد (۳۶ پروفیل) و داده‌های اعتبارسنجی، ۲۰ درصد (۱۰پروفیل) کل داده‌ها را شامل شد. هر مدل با داده‌های آموزشی برازش داده شد و سپس پیش‌بینی برای داده‌های اعتبارسنجی انجام شد. کلاس‌های پیش‌بینی شده با استفاده از ماتریس خطا برحسب درصد بیان می‌شود. پارامترهای استخراج شده از ماتریس خطا شامل صحت کاربر، صحت تولید کننده و معیار صحت کلی نقشه اعتبارسنجی شد.

صحت عمومی

$$OA = \sum_{i=1}^{n} Xij / N \qquad (۱)$$

در رابطه ۱ به‌ترتیب OA صحت کلی و N معرف کل پیکسل‌های طبقه‌بندی شده و $\sum_{i=1}^{n} Xij$ نمایه مجموع پیکسل‌های قطر اصلی ماتریس خطا (پیکسل‌های صحیح طبقه‌بندی شده) است. صحت کلی طبقه‌بندی ارتباط بین همه داده‌های مورد استفاده و داده‌های طبقه‌بندی شده 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.

شاخص کاپا

$$Kappa = N\sum_{i=1}^{n} Xij - \sum_{i-1}^{n} (Xio - Xoi) / N^2 - \sum_{i=1}^{n} (Xio - Xoi) \qquad (۲)$$

آماره کاپا یک شاخص قوی است که نسبت احتمال حضور یا عدم حضور کلاس‌هایی که به‌درستی به‌وسیله مدل پیش‌بینی شدند، را محاسبه می‌کند. بنابراین آماره کاپا همیشه کمتر از خلوص نقشه است. دامنه تغییرات آماره کاپا بین صفر تا یک است. آماره کاپا بالاتر از ۰/۸، ۰/۸-۰/۴و کمتر از ۰/۴ به‌ترتیب نشان‌دهنده توافق قوی، متوسط و ضعیف هستند (۲۹). در رابطه (۲)، n تعداد ردیف‌ها در ماتریس، Xij تعداد مشاهدات در ردیف i و ستون j (درایه‌های قطر اصلی)، Xio و Xoi مجموع حاشیه به‌ترتیب ردیف r و ستون i، N تعداد کل مشاهدات است.

صحت تولید کننده

$$P A = \frac{a_{tt}}{\sum_{i=1}^{N} a_{ik}} \qquad (۳)$$

در رابطه ۳، a_{tt} تعداد پیکسل‌های صحیح طبقه‌بندی شده روی قطر اصلی و $\sum_{i=1}^{N} a_{ki}$ جمع تعداد پیکسل‌هایی است که در آن ستون به‌عنوان نمونه‌های آموزشی طبقه‌بندی شده‌اند. قابلیت

جدول ۱. رده‌بندی خاک‌ها بر اساس سامانه آمریکایی خاک‌ها (۲۰۱۴)

ردیف	Soil Order رده خاک	Family soil class فامیل خاک	Subgroup زیرگروه	خاکرخ شاهد
۱	Inceptisols	Fine, carbonatic, thermic	Typic Calcixerepts	۲۳
۲	Inceptisols	Fine, carbonatic, thermic	Typic Haploxerepts	۳۶
۳	Inceptisols	Fine-loamy over fragmental, carbonatic, thermic	Typic Haploxerepts	۳۹
۴	Mollisols	Fine, carbonatic, thermic	Typic Haploxerolls	۱۱
۵	Inceptisols	Fine-loamy, carbonatic, thermic	Typic Haploxerepts	۱۰
۶	Entisols	Fine-loamy, carbonatic, thermic, shallow	Typic Xerorthents	۹

اطمینان تولید کننده ارتباط بین همه کلاس‌های صحیح پیش‌بینی شده و مجموع کلاس‌های صحیح پیش‌بینی شده کلاس‌های حضور مشاهده شده که به غلط جزء کلاس‌های عدم حضور پیش‌بینی شدند است.

صحت کاربر

$$UA = \frac{a_{ij}}{\sum_{i=1}^{N} a_{ik}} \quad (۴)$$

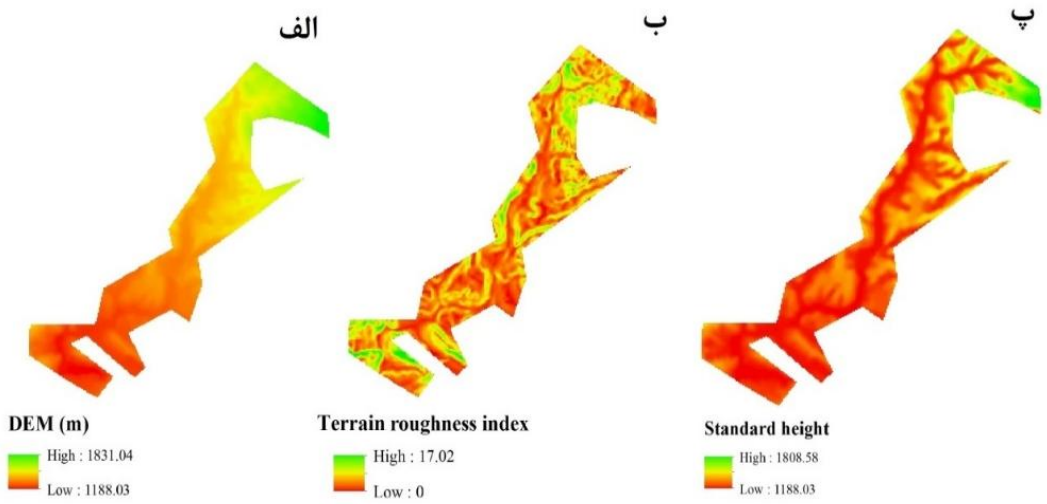
در رابطه ۴، a_{ij} تعداد پیکسل‌های درست طبقه‌بندی شده روی قطر اصلی و $\sum_{i=1}^{N} a_{ki}$ نمایه جمع تعداد پیکسل‌هایی است که در آن ردیف به‌عنوان نمونه‌های آموزشی آن طبقه‌بندی شده‌اند. دامنه تغییرات صحت تولیدکننده و صحت کاربر حد واسط صفر و یک است که درنتیجه مقادیر بالاتر نشان‌دهنده عملکرد مناسب مدل است.

نتایج و بحث

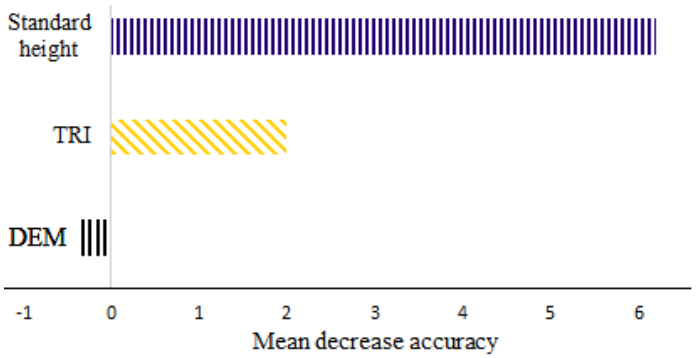
شرایط ژئومورفولوژیک منطقه مطالعاتی روی تعدادی از خصوصیات خاک از جمله بافت خاک، عمق، درصد سنگریزه، میزان ماده آلی و نیز مقدار آهک تجمع یافته در مقاطع خاک بیشترین تأثیر را داشته است که مجموعه این عوامل باعث تمایز خاک‌ها شده است. بر اساس سیستم رده‌بندی خاک آمریکایی، خاک‌های منطقه شامل سه رده مالی‌سولز، اینسپتی‌سولز و انتی‌سولز و ۶ کلاس در سطح فامیل هستند (جدول ۱).

مدل‌سازی مکانی کلاس‌های فامیل خاک به روش RF

مهم‌ترین متغیرهای کمکی مورد استفاده در پیش‌بینی مکانی کلاس‌های فامیل خاک توسط مدل جنگل تصادفی و درخت تصمیم توسعه‌یافته بر اساس نتایج عملکرد شاخص داده‌کاوی تورم واریانس (VIF) نشان داد که از میان ۱۵ متغیر ژئومورفومتری مورد استفاده، متغیرهای کمکی ارتفاع، ارتفاع استاندارد و شاخص زبری پستی بلندی قادر هستند بیشترین میزان تغییرپذیری مکانی خاک‌ها را در منطقه مدل‌سازی کنند (شکل ۳). همچنین در روش RF بر اساس شاخص میانگین حداقل صحت (Mean decrease accuracy) اهمیت متغیرهای کمکی منتخب نشان‌دهنده این است که پارامتر ارتفاع استاندارد (Standard Height) مؤثرترین متغیر محیطی در مدل‌سازی مکانی خاک‌ها در منطقه مطالعاتی بوده است (شکل ۴). ویژگی‌های مدل رقومی ارتفاع هم به‌لحاظ ریاضی و منطقی و همچنین از دیدگاه تجربی دارای ارتباط نزدیکی با ویژگی‌های محیطی و خاک هستند که استفاده از این ویژگی‌ها در شناسایی خاک‌ها تا حد زیادی منجر به صرفه‌جویی در زمان و هزینه مطالعات و افزودن دقت نقشه‌های تولیدی می شود (۲).



شکل ۳. پارامترهای محیطی مؤثر بر پیش‌بینی کلاس‌های فامیل خاک (VIF)



شکل ۴. مقدار اهمیت متغیرهای استفاده شده در پیش‌بینی کلاس‌های فامیل خاک

نتایج پیش‌بینی مکانی کلاس‌های خاک منطقه نشان داد که بیشترین وسعت مربوط به واحد خاک شماره ۴ با کلاس فامیل Fine, carbonatic, thermic, Typic Haploxerolls با ۶۹۸/۳۳ هکتار مساحت که ۵۸/۱۹ درصد کل منطقه را شامل می‌شود و کمترین وسعت مربوط به واحد خاک شماره ۶ با کلاس فامیل Fine-loamy, carbonatic, thermic, shallow Typic Xerorthents با حدود ۱۲/۰۷ هکتار که حدود یک درصد از کل مساحت منطقه را شامل می‌شود (جدول ۲). نتایج اعتبارسنجی مدل‌های RF و BRT برای خاک‌های منطقه (جدول ۳) نشان داد که صحت کلی (OA) و نمایه سازگاری کاپا (K) برای کلاس‌های فامیل خاک با استفاده از مدل جنگل تصادفی و درخت تصمیم توسعه‌یافته به‌ترتیب برابر با ۸۰ و ۷۰ درصد و ۶۴ و ۵۵ درصد حاصل شد. بر اساس ضریب کاپای به‌دست آمده درجه توافق خوب است و می‌تواند نشان‌دهنده این باشد که نتایج مدل‌سازی برای کلاس‌های خاک در منطقه مورد مطالعه با دقت بسیار بالایی مورد پذیرش است (۲۹). نتایج به‌دست آمده از این تحقیق با نتایج سایر پژوهشگران (۵، ۲۳ و ۲۶) از این جهت که مدل جنگل‌های تصادفی قادر به پیش‌بینی مطلوب‌تر کلاس‌های خاک با درصد خطای پایین است مطابقت داشت. با استفاده از مدل جنگل تصادفی نقشه رقومی خاک منطقه آبیک استان قزوین تهیه شد. نتایج نشان داد که مدل‌سازی با استفاده از الگوریتم جنگل تصادفی توانست کلاس‌های خاک منطقه را با دقت بالایی (ضریب کاپای ۸۳٪) پیش‌بینی کند (۱۵). در مقایسه روش‌های شبکه عصبی مصنوعی و درخت

جدول ۲. مساحت و درصد واحدهای کلاس خاک بر اساس روش RF

فامیل خاک	تعداد پروفیل	مساحت (هکتار)	درصد
۱	۴	۵۴/۳۹	۴/۵۳
۲	۱۱	۳۲۰/۸۷	۲۶/۷۳
۳	۴	۷۱/۳۲	۵/۹۴
۴	۲۲	۶۹۸/۳۳	۵۸/۱۹
۵	۳	۴۳/۰۲	۳/۵۸
۶	۲	۱۲/۰۷	۱/۰۳
مجموع		۱۲۰۰	۱۰۰

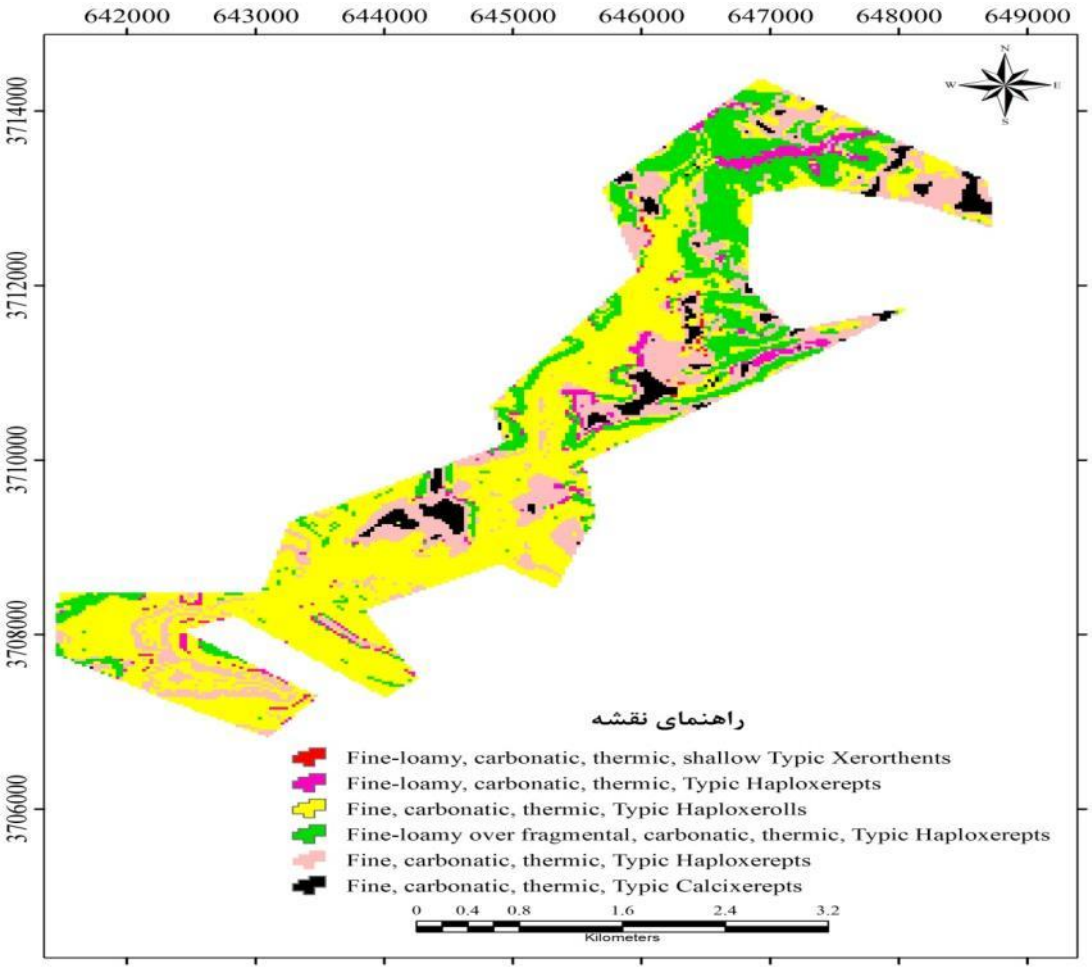
جدول ۳. نتایج اعتبارسنجی پیش‌بینی کلاس‌های فامیل خاک با استفاده از مدل‌های RF و BRT

مدل طبقه‌بندی	صحت عمومی (OA)	نمایه سازگاری کاپا (K)
RF	۸۰	۷۰
BRT	۶۴	۵۵

تصمیم در تهیه نقشه‌های رقومی خاک منطقه اردکان، محققان به این نتایج رسیدند که مدل‌های درختی نسبت به روش‌های شبکه عصبی مصنوعی دارای دقت بالاتری هستند و همچنین تفسیر نتایج درختی در تغییرپذیری خاک‌ها بسیار راحت‌تر است (۲۸). در مطالعه‌ای پیش‌بینی مکانی کلاس‌های خاک را با استفاده از مدل‌های رگرسیونی و درخت تصمیم در منطقه جنوب شرق ایران انجام شد. شاخص‌های ارزیابی مدل‌ها از جمله صحت کلی، شاخص کاپا به‌ترتیب برای مدل رگرسیون MLR، ۰/۷۱، ۰/۶۵، مدل BRT، ۰/۸۵، ۰/۸۱ و مدل DT، ۰/۸۶، ۰/۸۴ محاسبه شد. نتایج مقایسه ارزیابی دقت مطالعه نشان داد که بهترین پیش‌بینی مربوط به مدل درخت تصمیم بود (۱). در این پژوهش نتایج اعتبارسنجی حاکی از این است که مدل رگرسیون درختی توسعه‌یافته دارای مقادیر صحت عمومی و نمایه سازگاری کاپای کمتری نسبت به مدل جنگل تصادفی بود که دلیل آن را می‌توان توانایی روش جنگل تصادفی و بهره گیری از تعداد درخت بیشتر (n tree) در مدل‌سازی کلاس‌های خاک منطقه دانست که سبب می‌شود، فرایند آموزش مدل با کیفیت مناسب‌تری صورت پذیرد و نتایج قابل اعتمادتری تولید کند. در همین راستا، از نظر پژوهشگران مختلفی، فراوانی کلاس‌های خاک در مناطق مورد مطالعه به‌عنوان یک فاکتور مهم در تعیین خلوص نقشه‌های خاک در نقشه‌برداری رقومی است (۱۲، ۲۱ و ۲۷).

تهیه نقشه رقومی با استفاده از مدل RF

با توجه به نقشه پیش‌بینی مکانی کلاس‌های خاک در سطح فامیل (شکل ۵) می‌توان بیان کرد که بهترین پیش‌بینی مربوط به واحد خاک شماره ۴ با کلاس فامیل Fine, carbonatic, thermic, Typic Haploxerolls است که بیشترین مساحت خاک‌های منطقه را نیز به‌خود اختصاص داده است. فراوانی تعداد نقاط مشاهداتی مربوطه به کلاس خاک فامیل ۴ از دلایل پیش‌بینی موفق و بالای مدل جنگل تصادفی بوده است. در تحقیقی بیان شد که کلاس‌های خاکی که دارای نقاط پروفیلی بیشتری بودند، خطای پیش‌بینی کمتری داشتند (۷). محققان کاربرد روش درختان تصادفی را در پیش‌بینی کلاس‌های خاک در اراضی با پستی‌وبلندی کم مورد مطالعه قرار دادند و به این نتایج رسیدند که



شکل ۵. پراکنش مکانی کلاس‌های فامیل خاک با استفاده از مدل RF

کلاس‌های خاکی که تعداد نقاط مشاهداتی بیشتری داشتند دارای خطای پیش‌بینی کمتری بودند و برای افزایش دقت و پیش‌بینی بهتر کلاس‌هایی که دارای دقت کمتری هستند می‌توان با افزایش نقاط نمونه‌برداری، دقت کلاس‌ها و در نتیجه دقت نقشه تولیدی را افزایش داد (۱۹). در بین کلاس‌های فامیل خاک، بهترین پیش‌بینی مربوط به فامیل خاک Fine, carbonatic, thermic, Typic Haploxerolls به‌دست آمد که مقادیر بالای صحت کاربر و تولید کننده در جدول ۴ این نتایج را نشان می‌دهد. همچنین نتایج مطالعات نشان می‌دهد که نواحی با احتمال بالا برای گروه‌های بزرگ خاکی به زمین نماهای کاملاً شناخته شده، منطبق است (۷). صحت کاربر و تولید کننده نشان دهنده تخمین سطح پیش‌برآورد و کم‌برآورد در پیش‌بینی کلاس‌های خاک است (۱۷). در فامیل خاک Fine, carbonatic, thermic, Typic Calcixerepts میزان صحت کاربر بیشتر از صحت تولید کننده است و نشان‌دهنده این است که مدل این واحد خاک را در منطقه کم برآورد کرده است. همچنین در فامیل خاک Typic Haploxerepts Fine, carbonatic, thermic, مقدار صحت تولیدکننده بیشتر از صحت کاربر است که بیانگر رخداد نوعی بیش‌برآورد در پیش‌بینی این کلاس خاک است.

نتیجه‌گیری

نتایج حاصل از مدل‌های جنگل تصادفی و رگرسیون درختی توسعه‌یافته برای پیش‌بینی کلاس‌های خاک نشان داد که مدل جنگل تصادفی در جداسازی کلاس‌های خاک کارایی و دقت

جدول ۴. صحت تولید کننده، صحت کاربر برای کلاس‌های خاک در سطح فامیل

کلاس‌های خاک	صحت تولید کننده (٪PA*)	صحت کاربر (٪UA*)
1	100	50
2	0.67	100
3	NaN	0
4	100	100
5	NaN	NaN
6	0	NaN

*Producer accuracy * Users accuracy

بالاتری دارد. با توجه به اینکه مدل جنگل تصادفی به‌عنوان یکی از مناسب‌ترین روش‌ها در تولید نقشه‌های رقومی خاک شناخته شده است، در این مطالعه نیز این روش قادر به پیش‌بینی مطلوب کلاس‌های فامیل خاک شد و از آن برای تهیه نقشه نهایی خاک استفاده شد. همچنین نتایج نشان داد که ارتفاع، شاخص زبری پستی‌وبلندی و ارتفاع استاندارد شده مهم‌ترین متغیرهای محیطی در این مطالعه هستند. وجود ارتباط قوی بین داده‌های خاک و متغیرهای محیطی و همچنین تعداد داده‌های مشاهده شده از فاکتورهای تأثیرگذار روی صحت مدل‌سازی روابط خاک- زمین‌نما است. به‌طور کلی بر اساس نتایج به‌دست آمده از این مطالعه می‌توان بیان کرد که روش‌های رقومی نقشه‌برداری خاک تلاش می‌کنند تا برآورد دقیق و مطلوب‌تری از خاک‌ها بر اساس داده‌های محیطی ارائه دهند و واحدهای خاک یکنواخت‌تر را جداسازی کنند. لذا پیشنهاد می‌شود برای تهیه نقشه‌های رقومی خاک از سایر مدل‌های مبتنی بر طبقه‌بندی درختی در مطالعات آتی استفاده شود.

منابع مورد استفاده

1. Abaszadeh Afshar, F., Sh. Ayobi and A. Jafari. 2018. Spatial forecasting of large soil groups using regression and decision tree models in the south-east of Iran. *Agricultural Engineering (Scientific Journal of Agriculture)* 41(2): 133-146. (In Farsi).
2. Banaee, M. 1998. Map of Iran's moisture and heat regimes. Iran Soil and Water Research Institute. (In Farsi).
3. Breiman, L. 2001. Random forests. *Machine Learning* 45(1): 5-32.
4. Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone 1984. Classification and Regression Trees (The Wadsworth Statistics/Probability Series) Chapman and Hall. New York, NY 1-358.
5. Camera, C., Z. Zomeni, J. S. Noller, A. M. Zissimos, I. C. Christoforou and A. Bruggeman. 2017. A high resolution map of soil types and physical properties for Cyprus: A digital soil mapping optimization. *Geoderma* 285: 35-49.
6. Cook, S. E., A. Jarvis and J. P. González. 2008. A new global demand for digital soil information. PP. 31-41. *In*: Digital Soil Mapping with Limited Data. Springer, Dordrecht.
7. Debella-Gilo, M. and B. Etzelmüller. 2009. Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modeling integrated in GIS: Examples from Vest fold County, Norway. *Catena* 77(1): 8-18.
8. Eichhorn, M. P. 2016. What is a natural system? Chap. 21, *In*: Natural Systems*:* The Organization of Life. Chichester, John Wiley & Sons, doi:10.1002/9781118905982.ch21.
9. Elith, J., J. R. Leathwick and T. Hastie. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology* 77(4): 802-813.
10. Grunwald, S. 2016. Environmental Soil-Landscape Modeling: Geographic Information Technologies and Pedometrics, CRC Press.
11. Hair, J. F., R. E. Anderson, R. L. Tatham. and W. C. Black. 1998. Multivariate Date Analysis with Readings. Englewood Cliff, NJ: Prentce.

12. Jafari, A., H. Khademi, P. A. Finke, J. Van de Wauw and S. Ayoubi. 2014. Spatial prediction of soil great groups by boosted regression trees using a limited point dataset in an arid region, southeastern Iran. *Geoderma* 232: 148-163
13. Jenny, H. 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, New York.
14. Kempen, B., D. J. Brus, G. B. Heuvelink and J. J. Stoorvogel. 2009. Updating the 1: 50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. *Geoderma* 151(3-4): 311-326.
15. Khamoshi, E., F. Sarmadiyan and A. Keshavarzi. 2018. Digital soil mapping using stochastic forest model in Abik-Qazvin province. *Journal of Soil Research (Soil and Water Sciences)* 71(4): 885-899.
16. Kuhn, M. 2008. Building predictive models in R using the caret package. *Journal of Statistical Software* 28(5): 1-26.
17. Lacoste, M., B. Lemercier and C. Walter. 2011. Regional mapping of soil parent material by machine learning based on point data. *Geomorphology* 133(1-2): 90-99.
18. Ließ, M., B. Glaser and B. Huwe. 2012. Uncertainty in the spatial prediction of soil texture: comparison of regression tree and Random Forest models. *Geoderma* 170: 70-79.
19. Mosleh, Z., M. H. Salehi and A. Jafari. 2017. Digital classification of soil classes at different levels of American classification using extended tree regression model in Shahrekord plain, Iran. *14th* Iranian Soil Science, Congress-Genesis, Classification, University of Rafsanjan, Rafsanjan, Soil and Landscape Evaluation 343-347.
20. O'brien, R. M. 2007. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity* 41(5): 673-690.
21. Pahlavan-Rad, M. R., F. Khormali, N. Toomanian, C. W. Brungard, F. Kiani, C. B. Komaki and P. Bogaert. 2016. Legacy soil maps as a covariate in digital soil mapping: A case study from Northern Iran. *Geoderma* 279: 141-148.
22. Pahlavan-Rad, M. R., F. Kormali, N. Toomanian, F. kiani and Ch. B. Komaki. 2015. Digital soil mapping using Random Forest model in Golestan province. *Journal of Soil and Water Conservation* 6(21): 73-93.
23. Pahlavan-Rad, M. R., N. Toomanian, F. Khormali, C.W. Brungard, C. B. Komaki and P. Bogaert. 2014. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. *Geoderma* 232: 97-106.
24. Schoeneberger, P. J., D. A. Wysocki and E. C. Benham, (Eds.). 2012. Field Book for Describing and Sampling Soils. Version 3.0. Natural Resources Conservation Service, National Soil Survey Center, Lincoln, NE, 36.
25. Soil Survey Staff. 2014. Keys to Soil Taxonomy. 12th edn. USDA- NRCS, Washington, DC.
26. Taghizadeh-Mehrjardi, R., B. Minasny, F. Sarmadian and B. P. Malone. 2014. Digital mapping of soil salinity in Ardakan region, central Iran. *Geoderma* 213: 15-28.
27. Taghizadeh-Mehrjardi, R., K. Nabiollahi, B. Minasny and J. Triantafilis. 2015. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. *Geoderma* 253: 67-77.
28. Taghizadeh-Mehrjardi, R., F. Sarmadian, M. Omid, Gh. Savaghbi, N. Tomaniyan, M. Rosta and M. Rahemiyan. 2013. Comparison of artificial neural network and decision tree methods in digital soil mapping in Ardakan region. *Iranian Journal of Soil and Water Research* 44(2): 173-182 (In Farsi).
29. Taghizadeh-Mehrjardi, R., B. Minasny, A. B. McBratney, J. Triantafilis, F. Sarmadian and N. Toomanian. 2012. Digital soil mapping of soil classes using decision trees in central Iran. PP. 197-202. *In:* Minasny, B., B. P. Malone, and A. McBratney (Eds.), Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping, CRC Press, Sydney.
30. Tajik, S., S. Ayoubi, H. Shirani and M. Zeraatpisheh. 2019. Digital mapping of soil invertebrates using environmental attributes in a deciduous forest ecosystem. *Geoderma* 353: 252-263.

Journal of Water and Soil Science
Vol. 24, No. 2, Summer 2020, Isfahan University of Technology, Isfahan, Iran.

Digital Mapping of Soil Family Class Using the Machine Learning Approach (A Case Study: Semi-Arid lands in the West of IRAN)

Z. Maghsodi[1], M. Rostaminia[1]*, M. Faramarzi[2], A. Keshavarzi[3], A. Rahmani[3] and S. R. Mousavi[3]

(Received: August 15-2019; Accepted: October 23-2019)

Abstract

Digital soil mapping plays an important role in upgrading the knowledge of soil survey in line with the advances in the spatial data of infrastructure development. The main aim of this study was to provide a digital map of the soil family classes using the random forest (RF) models and boosting regression tree (BRT) in a semi-arid region of Ilam province. Environmental covariates were extracted from a digital elevation model with 30 m spatial resolution, using the SAGAGIS7.3 software. In this study area, 46 soil profiles were dug and sampled; after physico-chemical analysis, the soils were classified based on key to soil taxonomy (2014). In the studied area, three orders were recognized: Mollisols, Inceptisols, and Entisols. Based on the results of the environmental covariate data mining with variance inflation factor (VIF), some parameters including DEM, standard height and terrain ruggedness index were the most important variables. The best spatial prediction of soil classes belonged to Fine, carbonatic, thermic, Typic Haploxerolls. Also, the results showed that RF and BRT models had an overall accuracy and of 0.80, 0.64 and Kappa index 0.70, 0.55, respectively. Therefore, the RF method could serve as a reliable and accurate method to provide a reasonable prediction with a low sampling density.

Keywords: Spatial prediction, Soil class, Boosted Regression Tree, Random Forest

1. Science and Soil Engineering Department, College of Agriculture, Ilam University, Ilam, Iran
2. Rangeland and Watershed Management Group, Faculty of Agriculture, Ilam University, Ilam, Iran.
3. Science and Soil Engineering Department, Faculty of Agricultural Engineering and Technology, College of Agriculture & Natural Resources, Tehran University, Kraj, Iran.
*: Corresponding author, Email: m.rostaminya@ilam.ac.ir