نشریه آب و خاک (علوم و صنایع کشاورزی)
جلد ۳۴، شماره ۴، مهر – آبان ۱۳۹۹، ص. ۹۸۷-۹۷۳

Journal of Water and Soil
Vol. 34, No. 4, Sept-Oct 2020, p. 973-987

مقاله علمی-پژوهشی

کارایی روش‌های مختلف انتخاب متغیر کمکی در نقشه‌برداری رقومی کلاس خاک با استفاده از الگوریتم‌های داده‌کاوی

سمیره نظری[۱]- محمود رستمی‌نیا[۲*]- شمس اله ایوبی[۳]- اصغر رحمانی[۴]- سید روح اله موسوی[۵]

تاریخ دریافت: ۱۳۹۹/۰۲/۰۸
تاریخ پذیرش: ۱۳۹۹/۰۵/۰۵

چکیده

تهیه نقشه‌های خاک با صحت مناسب یک ابزار توانمند برای دست یافتن به استفاده پایدار از اراضی در عرصه‌های کشاورزی و منابع طبیعی محسوب می شود. پژوهش حاضر در بخشی از اراضی وَرگَر شهرستان آبدانان در استان ایلام به‌منظور نقشه‌برداری رقومی کلاس‌های خاک با استفاده از مدل‌های جنگل تصادفی و منطق فازی اجرا گردید. در اراضی مورد مطالعه موقعیت ۴۴ خاکرخ تعیین، حفر، تشریح و نمونه‌برداری از کلیه افق‌های ژنتیکی صورت پذیرفت. پس از انجام آزمایش‌های فیزیکوشیمیایی لازم، رده‌بندی خاک‌ها انجام شد. از مدل رقومی ارتفاع ماهواره آلوس پالسار و نرم‌افزار ساگا جی‌آی‌اس برای تهیه متغیرهای کمکی ژئومورفومتری استفاده گردید. سه رویکرد انتخاب متغیر شامل الگوریتم باروتا، شاخص تورم واریانس و میانگین کاهش صحت به‌همراه دو مدل داده‌کاوی جنگل تصادفی و منطق فازی برای مدل‌سازی روابط خاک-زمین‌نما به کار گرفته شد. نتایج نشان داد که رویکرد انتخاب متغیر میانگین کاهش صحت به‌عنوان مناسب‌ترین روش، از تعداد ۳۵ متغیر کمکی ژئومورفومتری منجر به انتخاب شش متغیر گردید. همچنین رویکرد مدل‌سازی جنگل تصادفی-میانگین کاهش صحت، در سطح زیرگروه با صحت عمومی و شاخص کاپای ۸۴ و ۵۷ درصد دارای بالاترین دقت بود. بررسی نتایج حاصل از رویکرد فازی حاکی از این بود که مقادیر شاخص کاپا و صحت عمومی این روش با سه سناریو دیگر مشابه و اختلاف ناچیزی بین صحت نتایج در سطح فامیل خاک مشاهده گردید. به‌طورکلی استفاده از رویکردهای مختلف انتخاب متغیر می‌توانند موجب افزایش دقت تهیه نقشه‌های رقومی خاک گردند. همچنین افزایش تعداد مشاهدات میدانی و استفاده از سایر متغیرهای محیطی تأثیرگذار بر روی تشکیل خاک‌ها را می توان برای پیش‌بینی کلاس‌های خاک دارای صحت پایین به کارگیری نمود.

واژه‌های کلیدی: جنگل تصادفی، متغیر محیطی، منطق فازی، نقشه‌برداری خاک

مقدمه

نقشه‌های خاک با صحت مناسب، به‌عنوان یک ابزار توانمند در راستای مدیریت منابع طبیعی و کشاورزی، مدل‌سازی های هیدرولوژیک، فرسایش خاک و کمّی کردن عملکردهای بیوفیزیکی به همراه ارزیابی زمین‌نما و مخاطرات طبیعی (سیلاب، ریزگرد و ...) که بر کیفیت محیط زیست و زندگی بشر تأثیرگذار می‌باشد، مورد استفاده واقع می‌گردد (۷). پیش‌بینی و ارائه ارتباط کمّی روابط خاک- زمین‌نما برای مشخص نمودن تغییر پذیری مکانی و ارز یابی منابع خاک ضروری می‌باشد. در مطالعات متعدد، تغییرات مکانی خاک‌ها در سطوح مختلف مطالعاتی بر اساس روش‌های معمول تهیه نقشه خاک و تلفیق آن با نقشه‌های موضوعی تا تهیه نقشه‌های رقومی خاک[۵] توسط محققین در سراسر جهان مورد بررسی واقع شده است (۱۹ و ۲۲). طی دو دهه اخیر در ایران و جهان از نقشه‌برداری رقومی برای نقشه‌برداری کلاس‌ها و ویژگی‌های خاک استفاده می‌شود (۲۲). یکی از اجزاء اصلی نقشه‌برداری رقومی خاک مدل ارتباط‌دهنده بین متغیرهای محیطی و خاک می‌باشد، در همین راستا می‌توان از انواع روش‌های داده‌کاوی بهره گرفت. نوع مدلی که در نقشه‌برداری رقومی خاک استفاده می‌شود نقش کلیدی در تغییرات مکانی کلاس‌های خاک دارد (۵).

داده‌کاوی یک دانش میان رشته‌ای است که حوزه‌های مختلف

۱- دانش آموخته کارشناسی ارشد رشته مهندسی علوم خاک، دانشگاه ایلام
۲- استادیار گروه مهندسی آب و خاک، دانشکده کشاورزی، دانشگاه ایلام
(*- نویسنده مسئول: Email: m.rostaminya@ilam.ac.ir)
۳- استاد گروه علوم و مهندسی خاک، دانشکده کشاورزی، دانشگاه صنعتی اصفهان
۴ و ۵- دانشجویان دکتری، گروه علوم و مهندسی خاک، دانشکده مهندسی و فناوری کشاورزی، دانشگاه تهران

DOI: 10.22067/jsw.v34i4.85748

5- Digital soil mapping

پایگاه داده، آمار، یادگیری ماشین و سایر زمینه‌های مرتبط با این علوم را با هم تلفیق می‌نماید، تا اطلاعات و دانش ارزشمند نهفته در حجم بزرگی از داده‌ها استخراج گردد. در واقع هدف اصلی از داده‌کاوی استخراج الگوها یا مدل‌های مخفی در داده‌هاست (۱). روش‌های داده‌کاوی متعددی توسط محققین برای پیش‌بینی مکانی کلاس‌های خاک مورد استفاده واقع گردیده‌است که می‌توان به مطالعات رگرسیون لاجیستیک (۱۷)، شبکه‌های عصبی مصنوعی (۳۴)، طبقه‌بندی درختی (۳۰)، جنگل تصادفی (۲۲، ۱۴ و ۲۸) اشاره نمود. منطق فازی یکی از رویکردهای داده‌کاوی است که به‌طور پیوسته ماهیت مرز و طبقه خاک در واحدهای نقشه‌برداری خاک را حفظ می‌کند (۲۵). مطالعاتی در خصوص استفاده از رویکرد نمونه‌مدار نقشه‌برداری رقومی خاک تحت عنوان منطق فازی صورت پذیرفته است (۲۰ و ۳۳). اگرچه استفاده از روش‌های مبتنی بر نقشه‌برداری رقومی خاک در سال‌های اخیر، به میزان قابل‌توجهی افزایش یافته است، اما بخش غالب مطالعات صورت گرفته بر روی طبقه‌بندی کلاس‌های خاک تا سطح گروه بزرگ یا زیرگروه خاک متمرکز بوده‌اند (۱ و ۵).

بنابراین، چالش اصلی نقشه‌برداری رقومی خاک برآورد تغییرات مکانی خاک در سطوح پایین‌تر رده‌بندی (فامیل و سری) با دقت بالا می‌باشد. از سوی دیگر، نتایج مدل مورد استفاده در تهیه نقشه رقومی کلاس‌های خاک در سطوح پایین رده‌بندی در صورت زیاد شدن تعداد متغیرهای محیطی به انتخاب بهینه متغیرهای محیطی وابستگی بالایی دارد. ما سوا و همکاران (۲۰۱۸) از روش‌های یادگیری ماشینی در پیش‌بینی مکانی کلاس‌های خاک در تانزانیا استفاده نموده و معتقدند مشتقات مدل رقومی ارتفاع[1] می‌تواند در پیش‌بینی مفید باشد. تعداد زیادی از متغیرهای پیش‌بینی کننده در نقشه‌برداری رقومی خاک قابل‌استفاده هستند. با این‌حال، وجود متغیرهای اضافه ممکن است موجب کاهش دقت فرآیند مدل سازی گردد (۱۸). مطالعه‌ای در برزیل توسط کامپوس و همکاران (۲۰۱۹) با هدف مقایسه‌ی سه الگوریتم انتخاب متغیر شامل (دو الگوریتم فیلتر و یک الگوریتم رپر[2]) و نقش متغیرهای منتخب بر روی دقت مدل پیش‌بینی کننده رقومی کلاس‌های خاک صورت پذیرفت. نتایج ایشان نشان داد که الگوریتم رپر با کاهش ۷۰ درصدی از مجموع ۴۰ متغیر کمکی مستخرج از مدل رقومی ارتفاع نسبت به دو الگوریتم انتخاب متغیر دیگر در پیش‌بینی مکانی کلاس‌های خاک نتایج با دقت بالاتری را ارائه نمود (۶). موسوی و همکاران (۲۰۱۹) در مطالعه‌ای دو الگوریتم داده‌کاوی جنگل تصادفی و رگرسیونی درختی توسعه‌یافته و دو روش انتخاب متغیر تجزیه مؤلفه‌های اصلی[3] و فاکتور تورم واریانس[4] را برای نقشه برداری رقومی کلاس خاک به کار گرفتند و بیان نمودند که متغیرهای محیطی که توسط فاکتور تورم واریانس انتخاب و استفاده از جنگل تصادفی در فرآیند پیش‌بینی مکانی کلاس‌های خاک از دقت بالاتری برخوردار می‌باشند (۲۲). در دو دهه گذشته مطالعات متعددی با استفاده از نقشه‌برداری رقومی خاک صورت پذیرفته است اما مناطق با توپوگرافی پیچیده و با شدت پستی و بلندی بالا به‌دلیل وابسته بودن این پیچیدگی به مقیاس و برهم‌کُنش مستقیم تغییرات مکانی خاک‌ها با عوامل خاک‌سازی از جمله توپوگرافی باعث شده تا جنبه‌های مختلفی از زوایای این مطالعات در درک روابط خاک- زمین‌نما ناشناخته مانده و همچنان به‌عنوان یک چالش برای نقشه‌برداری خاک‌ها در این مناطق پابرجا باشد. این تحقیق با هدف استفاده از رویکردهای داده کاوی الگوریتم درختان تصادفی و منطق فازی به همراه سه رویکرد انتخاب متغیر کمکی برای مدل‌سازی مکانی کلاس‌های خاک در دو سطح تاکسونومیک فامیل و زیرگروه در اراضی تپه‌ماهوری منطقه وَرگَر شهر آبدانان در استان ایلام انجام شد.

مواد و روش‌ها

منطقه‌ی مورد مطالعه

منطقه مورد مطالعه محدوده‌ای با مساحت ۱۰۲۷ هکتار از اراضی منطقه ورگر شهرستان آبدانان در استان ایلام می‌باشد (شکل ۱). بر اساس داده‌های ایستگاه هواشناسی سینوپتیک آبدانان، منطقه مورد مطالعه دارای متوسط بارندگی سالیانه ۶۲۸/۶ میلی‌متر و متوسط درجه حرارت سالیانه ۲۲/۶ درجه سانتی‌گراد می‌باشد. رژیم رطوبتی و حرارتی خاک‌های منطقه بر اساس مدل نیوهال (۳۱) و نرم‌افزار jNSM به ترتیب یوستیک و هایپرترمیک می‌باشد. محدوده مورد مطالعه بر روی سه واحد فیزیوگرافی شامل تپه‌ها، دشت‌های دامنه‌ای و دشت‌های آبرفتی رودخانه‌ای واقع‌شده است، جانمایی خاکرخ‌های مطالعاتی در سیستم مختصات متریک[5] نمایش داده شده است (شکل ۱).

نمونه‌برداری و آنالیزهای آزمایشگاهی

پس از تفسیر و بررسی تصاویر ماهواره‌ای و انطباق تغییرات واحدهای شکل زمین با نقشه‌های مدل رقومی ارتفاع، شیب، جهت شیب و زمین‌شناسی، واحدهای فیزیوگرافی منطقه جداسازی گردید. بر اساس استاندارد مطالعات خاکشناسی در سطح تفصیلی (۲۶) موقعیت۴۴ خاکرخ مشاهداتی در هر یک از واحدهای موردنظر بر اساس الگوی نمونه‌برداری طبقه‌بندی تصادفی تعیین و با استفاده از سیستم موقعیت‌یاب جهانی در محل‌های موردنظر حفر، تشریح و از کلیه افق‌های ژنتیکی قابل‌شناسایی نمونه‌برداری گردید. سپس

1- Digital Elevation Model
2- warpper
3- Principal component analysis
4- Variance inflation factor
5- Universal Transfer Mercator

نمونه‌های خاک برای انجام آزمایش‌های فیزیکی و شیمیایی لازم به آزمایشگاه دانشگاه ایلام منتقل گردیدند. در این پژوهش بافت خاک به روش هیدرومتری (۱۰)، کربنات کلسیم معادل به روش حجم‌سنجی (۲۳)، ظرفیت تبادل کاتیونی با روش باور (۲۹)، ماده آلی به روش والکی و بلک (۳۲)، پ‌هاش در گل اشباع و هدایت الکتریکی در عصاره اشباع به روش‌های استاندارد، اندازه‌گیری شد. در نهایت رده‌بندی تمامی خاکرخ‌ها بر اساس سامانه رده‌بندی خاک آمریکایی (۲۶) تا سطح فامیل نهایی گردید.

متغیرهای کمکی

در پژوهش حاضر از مدل رقومی ارتفاع با قدرت تفکیک مکانی ۱۲/۵ متر برگرفته از ماهواره آلوس پالسار[1] (۲۰۱۱) استفاده گردید. ویژگی‌های ژئومورفومتری درصد شیب، جهت شیب، خمیدگی صفحه‌ای، خمیدگی نیم‌رخی، شاخص قدرت جریان، شاخص موقعیت توپوگرافی، شاخص خیسی، پستی و بلندی کاذب، موقعیت میانی شیب، شاخص بالای پشته با درجه کیفیت بالا، شاخص همواری دره با درجه کیفیت بالا، ارتفاع نرمال، موقعیت نسبی شیب، عمق دره، ارتفاع استاندارد شده و ارتفاع شیبدار با استفاده از نرم‌افزار ساگا جی‌آی‌اس نسخه ۷/۳ محاسبه و استخراج گردید. نقش ناهمواری‌ها در فرآیندهای مختلف توسط ویژگی‌های ذکر شده به‌صورت کمّی بیان می‌گردد. روش استخراج تمام پارامترهای مزبور در روش ارائه‌شده توسط هنگل و همکاران تشریح گردیده است (۱۱).

انتخاب متغیر بهینه

برای انتخاب متغیرهای محیطی بهینه از نظر اهمیت آن‌ها در فرآیند مدل‌سازی از سه الگوریتم نظارت‌شده باروتا، شاخص‌های تورم واریانس و میانگین کاهش صحت استفاده گردید.

الگوریتم باروتا[2]: روشی است که بر اساس الگوریتم طبقه‌بندی جنگل تصادفی اجرا می‌شود (۱۶). این الگوریتم هنگامی که مجموعه‌ای از داده‌های چندین متغیر برای ساخت مدل استفاده می‌شود مؤثر می‌باشد. در این الگوریتم با افزودن یک تصادف و جمع‌آوری نتایج از مجموعه نمونه‌های تصادفی می‌تواند تأثیر گمراه کننده نوسانات تصادفی را کاهش داد و همبستگی ویژه‌ای ایجاد کند. الگوریتم انتخاب متغیر باروتا از جمله الگوریتم‌های نظارت شده می‌باشد که مشابه سایر الگوریتم‌ها با رویکرد نظارت شده دارای رویکرد انتخاب متغیر پیش‌رونده[3] و عقب‌رونده[4] می‌باشد (۱۵).

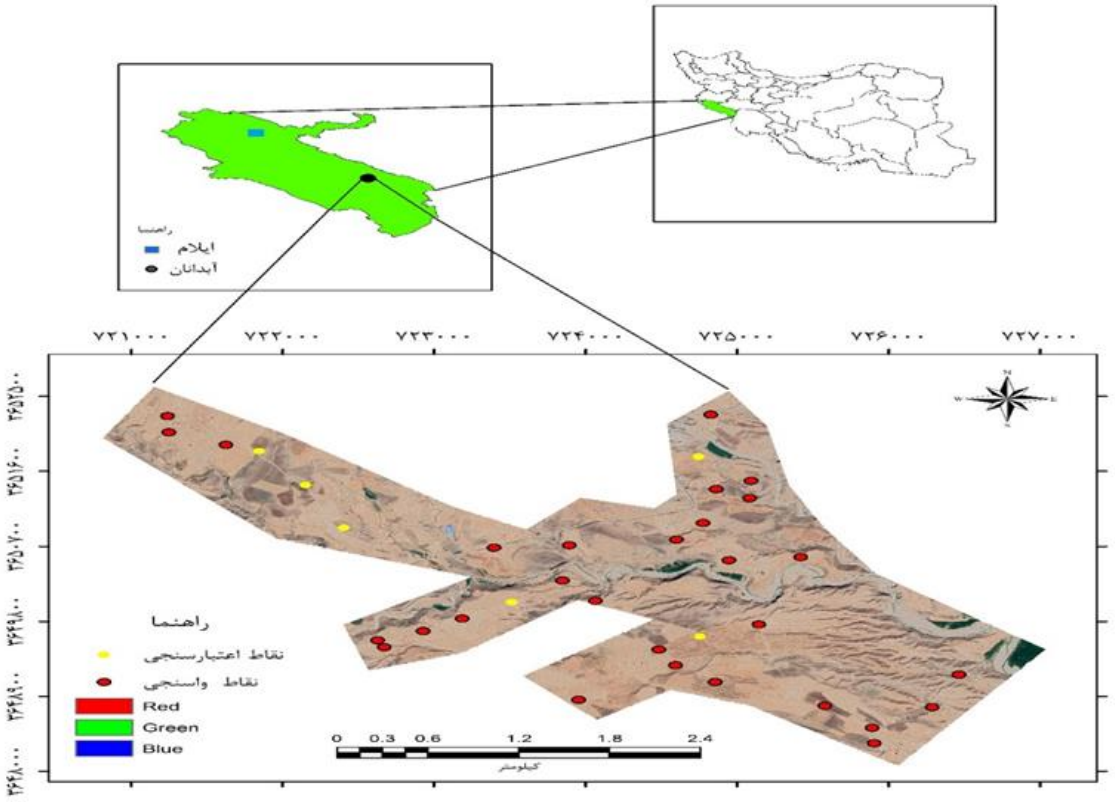
شاخص تورم واریانس[5]: یکی از راهکارهای مدیریت اثر هم‌خطی چندگانه متغیرهای مستقل استفاده از شاخص تورم واریانس می‌باشد. وقتی متغیرهای محیطی همبستگی بالایی با یکدیگر داشته باشند اثر هم‌خطی ایجاد خواهد شد و این خود باعث کاهش دقت فرآیند مدل‌سازی متغیر وابسته می‌گردد. شاخص تورم واریانس طبق جدول ۱ دارای دامنه قابل قبولی برای انتخاب متغیرها می‌باشد (۲). در این مطالعه از مقادیر تورم واریانس برابر ۱ تا کمتر از ۵ به‌عنوان متغیرهای محیطی مستقل برای پیش‌بینی کلاس‌های خاک مورد استفاده واقع گردید (جدول ۱).

الگوریتم کاهش میانگین صحت[6]: مدل جنگل تصادفی بر اساس روش میانگین کاهش دقت، اهمیت متغیرها را مورد بررسی قرار می‌دهد. کاهش میانگین صحت، اهمیت جایگزینی هم نامیده می‌شود چون در این رویکرد متغیرهایی که به‌طور تصادفی برای هر درخت در مدل تولیدشده است با مقادیر درست متغیرها جایگزین می‌شود.

مدل‌سازی‌های مکانی

جنگل تصادفی: جنگل تصادفی به‌عنوان یک روش غیرپارامتریک یادگیری ماشینی متعلق به خانواده روش‌های تلفیقی[7] است. این الگوریتم نوع توسعه‌یافته‌ای از مدل طبقه‌بندی و رگرسیون درختی است و توسط بریمن ارائه شده است (۳). جنگل تصادفی در حقیقت مجموعه‌هایی از درخت‌های پیش‌بینی کننده با احتمال یکسان و دارای پراکندگی یکسانی هستند. در این الگوریتم شمار زیادی درخت تصمیم‌گیری ساخته‌شده برای یک پیش‌بینی معین باهم ترکیب می شوند. این الگوریتم هر دو دسته متغیرهای پیوسته و گسسته را می‌پذیرد. جنگل تصادفی در برابر خطاهای موجود در پیش‌بینی‌ها مقاوم می‌باشد و بنابراین به انتخاب اولیه متغیرهای کمکی نیازی نمی‌باشد.

1- ALOS PALSAR
2- Boruta
3- Forward
4- Backward
5- Variance inflation factor
6- Mean decrease accuracy
7- Ensemble



شکل ۱- موقعیت منطقه مورد مطالعه
Figure 1- Location of the study area

جدول ۱- حدود دامنه تغییرات VIF
Table 1- Range of VIF changes

دامنه تغییرات Variation range	تشریح کلاس Class outline
VIF = 1	فاقد اثر هم‌خطی و همبستگی بین متغیرها No effect of collinearity and correlation between variables
1<VIF < 5	اثر هم‌خطی و همبستگی متوسط Collinearity effect and average correlation
5< VIF <10	دارای اثر هم‌خطی و همبستگی بالا Has high collinearity and high correlation
VIF >10	اثر هم‌خطی و همبستگی خیلی بالا Very high collinearity and correlation effect

روش جنگل تصـــادفی می‌تواند رابطه های غیرخطی بین متغیر وابســته و متغیر پیش‌بینی کننده را مدل‌ســازی نماید، زیرا از تجزیه و تحلیل متعدد مجموعه داده، برای برر سی روابط ا ستفاده می‌کند. مدل جنگل تصادفی در نرم‌افزار R نسخه ۳,۵,۱ با ا ستفاده از بسته‌های " جنگل تصادفی[1] " و "کارت[2]" اجرا گردید.

منطق فازی: در این مطالعه از رویکرد استنباطی خاک–سرزمین[3] در محیط نرم‌افزار سولیم سولوشن[4] نسخه ۲۰۱۵ استفاده شده است. این رویکرد شـــامل مجموعه‌ای از تکنیک‌های اســـتنتاج تحت منطق

1- Random forest
2- Caret
3- Soil landscape inference model
4- SoLIM Solutions

فازی در ترکیب با داده‌های محیطی برای تولید نقشه عضویت طبقات خاک در سراسر منطقه می‌باشد (۳۵). ویژگی مهم مدل رقومی ارتفاع به‌عنوان لایه‌ی پیش‌بینی کننده در پایگاه داده اطلاعات جغرافیایی استفاده شده‌است. پایگاه دانش در این مطالعه، بر مبنای فامیل و زیرگروه خاک با استفاده از موقعیت مکانی خاکرخ‌های مشاهداتی و استدلال نمونه‌مدار[1] تعریف شد. بر اساس توابع شباهت گاور برای متغیرهای پیوسته ژئومورفومتری، یک رابطه فازی رستری برای هر کدام از کلاس‌های زیرگروه و فامیل خاک تعریف گردید و منجر به تهیه نقشه عضویت فازی با استفاده از پایگاه داده‌های جغرافیایی پارامترهای ژئومورفومتری در بخش‌های مختلف از اراضی می گردد، هر پیکسل از نقشه های عضویت فازی عددی بین ۰ تا ۱۰۰ را به خود اختصاص می‌دهد. نقشه‌های عضویت فازی ایجاد شده در نهایت غیر فازی می‌شوند. در فرآیند غیرفازی کردن، خاک با مقادیر بیشینه عضویت فازی در یک موقعیت مکانی به‌عنوان خاک آن موقعیت مکانی انتخاب می‌گردد. فرایند غیرفازی کردن، چندین نقشه عضویت فازی مربوط به خاک‌های مختلف را تبدیل به یک نقشه رقومی می‌کند که در آن هر پیکسل تنها متعلق به یک نوع خاک می‌باشد.

به‌طورکلی در این پژوهش از دو روش جنگل تصادفی و فازی به همراه سه روش انتخاب متغیر شامل الگوریتم میانگین کاهش دقت، الگوریتم تورم واریانس و الگوریتم باروتا استفاده شد و از تلفیق روش جنگل تصادفی با سه الگوریتم انتخاب متغیر ذکر شده، همراه با روش ترکیبی مدل‌سازی فازی خاک-زمین‌نما و رویکرد انتخاب متغیر کاهش میانگین صحت به عنوان مؤثرترین رویکرد انتخاب متغیر، در دو سطح تاکسونومیک فامیل و زیرگروه خاک اقدام به تهیه نقشه رقومی کلاس‌های خاک گردید.

اعتبار سنجی: برای بررسی صحت مدل‌های مورد استفاده در پیش‌بینی مکانی کلاس‌های خاک، اعتبارسنجی بر اساس مقایسه داده‌های مشاهده شده و پیش‌بینی شده صورت گرفت. از شاخص‌های صحت کلی، شاخص کاپا، صحت کاربر و صحت تولیدکننده برگرفته از ماتریس خطا برای ارزیابی دقت استفاده گردید. صحت کلی (رابطه ۱)، ارتباط بین کل داده‌های مورد استفاده و داده‌های طبقه‌بندی شده (ff, tt) را تشریح می‌کند (۳۰).

$$\text{صحت کلی} = \frac{tt+ff}{tt+ft+tf+ff} \quad (۱)$$

صحت کاربر (رابطه ۲)، شامل کلاس‌های حضور صحیح پیش‌بینی‌شده (tt) به مجموع کلاس‌های حضور صحیح پیش‌بینی‌شده و کلاس های عدم حضور که به غلط جزء کلاس های حضور مشاهده‌شده پیش‌بینی شدند (tf) می‌باشند (۳۰).

$$\text{صحت کاربر} = \frac{tt}{tt-ff} \quad (۲)$$

صحت تولید کننده (رابطه ۳)، ارتباط بین همه کلاس‌های صحیح پیش‌بینی‌شده و مجموع کلاس‌های صحیح پیش‌بینی‌شده کلاس‌های حضور مشاهده‌شده که به غلط جزء کلاس‌های عدم حضور پیش‌بینی شدند (ft) می‌باشد (۳۰).

$$\text{صحت تولید کننده} = \frac{tt}{tt-ft} \quad (۳)$$

شاخص کاپا (رابطه ۴)، یک شاخص قوی است که نسبت احتمال حضور یا عدم حضور کلاس را که به‌درستی به‌وسیله مدل پیش‌بینی شدند، محاسبه می‌کند، بنابراین شاخص کاپا همیشه کم‌تر از صحت کلی نقشه است. در شاخص کاپا دقت مشاهده به‌عنوان مجموع مقادیر قطری در ماتریس و شانس خطا شامل مجموع مقادیر غیرقطر ماتریس است. دامنه تغییرات آماره کاپا بین صفر تا یک است. مقادیر این شاخص در دامنه‌های، بیش از ۰/۸، ۰/۸-۰/۴ و کمتر از ۰/۴ به ترتیب نشان‌دهنده توافق قوی‌تر، متوسط و ضعیف می‌باشد (۳۰).

$$\text{شاخص کاپا} = \frac{observed\ accuracy - chance\ agreement}{1 - chance\ agrement} \quad (۴)$$

نتایج و بحث

با توجه به خصوصیات فیزیکوشیمیایی خاک‌رخ های شاهد که در جدول ۳ ارائه شده است، از نظر رده‌بندی، خاک‌ها با سه رده مالی‌سولز، اینسپتی‌سولز و انتی‌سولز، شش کلاس در سطح زیرگروه و ۱۱ کلاس در سطح فامیل شناسایی گردید. زیرگروه خاک تیپیک کلسی‌یوستالز[2] با ۴۸/۹۰ درصد و کلاس فامیل (فاین سیلتی،کربناتیک، هایپرترمیک تیپیک هاپلویوستپز[3]) با ۳۲/۷ درصد از کل اراضی دارای بیشترین مساحت و زیرگروه ۴ (تیپیک هاپلویو ستپز[4]) و فامیل۶ (فاین سیلتی، کربناتیک، هایپرترمیک تیپیک کلسی یوستالز[5]) به ترتیب با ۱/۸۵ و ۰/۱۸ کمترین درصد مساحت را در منطقه شامل می‌شوند (جداول ۲ و ۳). در جدول ۵ نتایج انتخاب متغیرهای محیطی برای مدل‌سازی بر اساس سه روش شاخص تورم واریانس، میانگین کاهش صحت و الگوریتم باروتا ارائه شده است.

از میان ۳۵ متغیر کمکی مورد استفاده در نهایت بر اساس روش تورم واریانس، هشت متغیر، بر اساس روش میانگین کاهش دقت، شش متغیر و بر اساس روش باروتا، سه متغیر به‌عنوان متغیرهای کمکی برای مدل‌سازی مکانی کلاس‌های خاک انتخاب شد. متغیر کمکی شاخص قدرت جریان[6] به صورت مشترک توسط شاخص‌های تورم واریانس و کاهش میانگین صحت انتخاب گردید.

1- Sample-based reasoning
2- Typic Calciustolls
3- Fine-silty, carbonatic, hyperthermic Typic Haplustepts
4- Typic Haplustepts
5- Fine-silty, carbonatic, hyperthermic Typic Calciustolls
6- Stream power index

جدول ۲- مساحت زیرگروه‌های خاک با استفاده از RF

Table 2- Area of soil subgroups using RF

کلاس خاک Soil class	زیرگروه خاک Soil subgroup	مساحت Area (ha)	مساحت Area (%)
1	Typic Haplustolls	150.85	14.68
2	Typic Calciustepts	239.02	23.27
3	Typic Calciustolls	502.22	48.90
4	Typic Haplustepts	19.01	1.85
5	Typic Ustorthents	115.98	11.30
مجموع Total	---	1027.08	100

جدول ۳- خصوصیات فیزیکوشیمیایی و رده‌بندی خاکرخ های شاهد فامیل‌های خاک در اراضی ورگر آبدانان

Table 3- The physiochemical properties and classification of representative profile of soil families in the Vargar lands of Abdanan

ظرفیت تبادل کاتیونی CEC ($cmol_+/kg$)	رنگ خاک Soil Color		هدایت الکتریکی EC ($dS.m^{-1}$)	واکنش خاک pH	کربن آلی O.C	کربنات کلسیم معادل CCE	سنگ و سنگریزه R.F	شن Sand	سیلت Silt	رس Clay	عمق Depth (cm)	افق Horizon
	مرطوب Moist	خشک Dry			درصد (%)							
Coarse-loamy, carbonatic, hyperthermic Typic Calciustolls, p.36												
16.5	7.5YR3/3	7.5YR 4/3	0.104	7.42	1.29	30	12	45.8	30	24.2	0-20	Ap
9.9	7.5YR3/4	7.5YR5/4	0.105	7.69	0.71	42	40	49.16	35.84	15	20-60	Bk
-	-	-	-	−	−	−	>90	-	−	-	>60	C
Fine-loamy, carbonatic, hyperthermic Typic Calciustepts, p.1												
13.3	7.5YR4/4	7.5YR6/4	0.135	7.54	1.2	45	10	45.8	35.8	18.3	0-25	Ap
15.4	7.5YR4/4	7.5YR6/4	0.112	7.72	0.85	48	30	51.7	23.3	25	25-65	Bw
12.8	7.5YR5/4	7.5YR6/4	0.104	7.85	0.2	57	15	52.5	23.3	24.2	65-90	Bk1
12.8	7.5YR 3/4	7.5YR4/4	0.104	7.85	0.2	57	15	52.5	23.3	24.2	90-150	Bk2
Fine silty, carbonatic, hyperthermic Typic Calciustepts, p.24												
11.7	7.5YR5/4	7.5YR6/2	0.34	7. 47	0.74	6	5	45.84	35.82	18.34	0-25	Ap
14.8	7.5YR4/4	7.5YR6/3	0.225	7.8	0.67	49	3	51.66	23.34	25	25-63	Bw
12.1	7.5YR5/4	7.5YR6/4	0.157	8.2	0.33	55	3	52.5	23.34	24.16	63-105	Bk1
11.8	7.5YR5/4	7.5YR6/4	0.157	8.2	0.28	55	3	52.5	23.34	24.16	105-150	Bk2
Clayey, carbonatic, hyperthermic Typic Calciustepts, p.19												
19.6	7.5YR4/4	7.5YR 5/4	0.601	7.5	0.71	25	5	46	24	30	0-15	Ap
19.6	7.5YR5/3	7.5YR6/3	0. 612	7.7	0.62	31	5	43	24	33	15-30	Bw
19.1	7.5YR6/4	7.5YR6/3	0.478	8.2	0.44	65	3	42	21.5	36.5	30-70	Bk1
19.1	7.5YR6/4	7.5YR6/3	0.478	8.2	0.44	65	5	42	21.5	36.5	70-150	Bk2
Coarse-silty, carbonatic, hyperthermic Typic Calciustepts, p.11												
10.8	7.5YR4/3	7.5YR6/3	0.148	7.4	0.97	25	10	62.5	22.5	15	0-20	Ap
11.1	7.5YR5/4	7.5YR6/4	0.099	7.7	0.68	35	5	63.33	19.17	17.5	20-40	Bw
8.6	7.5YR5/4	7.5YR7/4	0.117	8.2	0.44	56	5	66.66	19.18	14.16	40-90	Bk1
8.6	7.5YR5/4	7.5YR7/4	0.117	8.4	0.44	56	8	66.66	19.18	14.16	90-150	Bk2
Fine-silty, carbonatic, hyperthermic Typic Calciustolls, p.14												
18.3	7.5YR4/3	7.5YR5/3	0.144	7.4	1.39	18	5	60	21.67	18.33	0-20	Ap
16.4	7.5YR4/3	7.5YR5/3	0.123	7.6	0.92	25	3	50	29.17	20.83	20-60	Bw
15.2	7.5YR5/3	7.5YR6/3	0.156	7.8	0.43	55	7	45.83	26.67	27.5	60-110	Bk1
15.2	7.5YR6/3	7.5YR6/3	0.156	8.1	0.43	55	5	45.83	26.67	27.5	110-150	Bk2

ادامه جدول ۳- خصوصیات فیزیکوشیمیایی و رده‌بندی خاکرخ‌های شاهد فامیل‌های خاک در اراضی ورگر آبدانان

Continued Table 3- The physiochemical properties and classification of representative profile of soil families in the Vargar lands of Abdanan

ظرفیت تبادل کاتیونی CEC (Cmol+/kg)	رنگ خاک Soil Color		هدایت الکتریکی EC (ds.m-1)	واکنش خاک pH	کربن آلی O.C	کربنات کلسیم معادل CCE	سنگ و سنگریزه R.F	شن Sand	سیلت Silt	رس Clay	عمق Depth (cm)	افق Horizon
	مرطوب Moist	خشک Dry					درصد(%)					
Loamy-skeletal, carbonatic, hyperthermic Typic Calciustolls, p.20												
16.7	7.5YR3/3	7.5YR 4/3	0.14	7.54	1.2	45	10	45.84	35.82	18.34	0-25	Ap
14.4	7.5YR3/4	7.5YR5/4	0.11	7.61	0.5	48	50	51.66	23.34	25.00	25-65	Bw
12.8	7.5YR5/3	7.5YR6/3	0.10	7.82	0.2	56	30	52.50	23.34	24.16	65-90	Bk1
12.8	7.5YR6/3	7.5YR6/3	0.10	7.96	0.2	56	15	52.50	23.34	24.16	90-150	Bk2
Fine-silty, carbonatic, hyperthermic Typic Haplustepts, p.4												
19.1	7.5YR4/4	7.5YR 5/4	1.78	7.62	0.65	20	2	40.0	29.5	29.5	0-25	Ap
17.2	7.5YR6/4	7.5YR6/5	0.95	7.88	0.44	40	3	57.5	9.5	33	25-70	Bw1
17.2	7.5YR6/3	7.5YR6/4	0.95	7.88	0.44	42	5	57.5	9.5	33	70-100	Bw2
17.2	7.5YR6/3	7.5YR6/4	0.95	7.88	0.44	33	5	57.5	9.5	33	100-150	Bw3
Fine-silty, carbonatic, hyperthermic Typic Haplustolls, p.34												
19.6	7.5YR3/3	7.5YR 4/3	1.34	7.14	2.3	35	-	58.0	19.5	22.5	0-20	Ap
16.5	7.5YR4/4	7.5YR5/4	0.51	7.45	0.62	43	-	53.5	16.0	30.5	20-60	Bw1
16.1	7.5YR5/4	7.5YR5/4	0.51	7.78	0.45	43	-	53.5	16.0	30.5	60-100	Bw2
16.1	7.5YR5/4	7.5YR5/4	0.51	7.78	0.45	45	-	53.5	1.06	30.5	100-150	Bw3
Loamy-skeletal, carbonatic, hyperthermic Typic Haplustolls, p.30												
14.8	7.5YR3/3	7.5YR 4/3	0.76	7.44	1.79	35	35	57	20.5	22.5	0-20	Ap
11.3	7.5YR4/4	7.5YR5/4	0.40	7.73	0.77	55	36	61.5	12.5	26	20-55	Bw
7.6	-	-	0.23	8.32	0.5	60	50	62	16	22	>55	CB
Loamy-skeletal, mixed, calcareous, hyperthermic Typic Ustorthents, p.22												
12.4	7.5YR4/4	7.5YR5/4	0.615	7.74	0.61	26	35	61	19	20	0-25	Ap
8.3	7.5YR6/4	7.5YR7/4	0.332	7.85	0.32	38	45	51	16	33	25-100	BC
			-	-	-	-	60	-	-	-	>100	C

نتایج ارزیابی مدل‌ها

مقادیر صحت پیش‌بینی مکانی هر یک از کلاس‌های خاک بر اساس سه آماره صحت کلی و شاخص کاپا و خطای برون کیسه‌ای و همچنین نتایج روش فازی برای دو سطح زیرگروه و فامیل خاک در جدول ۵ ارائه شده‌است. نتایج نشان می‌دهد که بیش‌ترین مقدار شاخص کاپا برای زیرگروه خاک مربوط به روش جنگل تصادفی همراه با رویکرد انتخاب متغیر میانگین کاهش صحت به مقدار ۰/۵۷ می‌باشد. نکته قابل‌توجه این بود که مقادیر شاخص کاپا برای هر سه روش اجرا شده با جنگل تصادفی و منطق فازی در پیش‌بینی فامیل خاک مقدار ۰/۳ برآورد شده است. مدل جنگل‌های تصادفی نسبت به دو روش باروتا و تورم واریانس با صحت کلی و شاخص کاپا، ۸۴ در صد، ۰/۵۷ در سطح زیرگروه دارای صحت بیشتری می‌باشد که تأیید کننده‌ی توانایی چشم‌گیر جنگل تصادفی برای پیش‌بینی الگوی پراکنش زیرگروه‌های خاک در منطقه مطالعاتی است. البته با توجه به اینکه زیرگروه توسط هر سه مدل برازش داده شده با دقت و صحت بالا پیش‌بینی شدند، احتمالاً این موضوع را می‌توان به انتخاب مناسب متغیرهای پیش‌بینی کننده توسط این مدل نسبت داد. این نتایج با نتایج محققانی همچون موسوی و همکاران (۲۰۱۹)، خاموشی و همکاران (۲۰۱۹)، هیونگ و همکاران (۲۰۱۴)، از لحاظ برتری داشتن مدل جنگل تصادفی برای پیش‌بینی مطلوب

کلاس‌های خاک مطابقت داشت (۱۲، ۱۴ و ۲۲). مقدار خارج از سبد[1] در روش تورم واریانس با مقدار ۶۷/۸۶ درصد دارای کمترین مقدار خطای برون‌کیسه‌ای در بین سه روش و دو سطح پیش‌بینی خاک می‌باشد. استام و همکاران (۲۰۱۰) مقدار خطای برون کیسه‌ای را برای پیش‌بینی ۶۷۲ خاکرخ در سطح فامیل خاک، ۵۵/۲ درصد برآورد کردند (۲۸). پهلوان‌راد و همکاران (۲۰۱۴) گزارش نمودند که خطای برون کیسه‌ای وابسته به تعداد کلاس‌ها در هر سطح رده‌بندی بوده و صحت پیش‌بینی کلاس‌های خاک از سطوح بالای رده‌بندی به سمت سطوح پایین‌تر رده‌بندی کاهش می‌یابد. ایشان خطای برون کیسه را برای سطوح گروه بزرگ و زیرگروه و سری خاک به ترتیب ۴۸/۵، ۵۱/۵ و۵۶/۵ درصد محاسبه کردند که با نتایج پژوهش حاضر مطابقت دارد (۲۴). جعفری و همکاران (۲۰۱۳)، فاتحی (۲۰۱۵) و موسوی و همکاران (۲۰۱۹)، نیز به این نتیجه رسیدند که افزایش فراوانی کلاس‌ها در سطح رده‌بندی پایین‌تر تا سطح رده صحت پیش‌بینی مدل را کاهش می‌دهد (۹، ۱۳ و ۲۲). نتایج حاصل از رویکرد فازی حاکی از این بود که مقادیر شاخص کاپا و صحت عمومی این روش با سه سناریو دیگر در سطح فامیل خاک مشابهت دارد. در روش فازی مشاهده شد که مقادیر کاپا و صحت عمومی در سطح زیرگروه نسبت به سناریوهای دیگر دارای مقادیر کمتری می‌باشد، این نتایج حاکی از برتری داشتن روش جنگل نسبت به روش فازی می‌باشد.

جدول ٤- مساحت فامیل‌های خاک با استفاده از مدل RF

Table 4- Area of soil families using the RF model

کلاس خاک Soil class	فامیل خاک Soil family	مساحت Area	
		هکتار (ha)	درصد (%)
1	Coarse-loamy, carbonatic, hyperthermic Typic Calciustolls	48.08	4.68
2	Fine-loamy, carbonatic, hyperthermic Typic Calciustepts	9.39	0.91
3	Fine silty, carbonatic, hyperthermic Typic Calciustepts	107.53	10.46
4	Clayey, carbonatic, hyperthermic Typic Calciustepts	7.69	0.74
5	Coarse-silty, carbonatic, hyperthermic Typic Calciustepts	303.61	29.56
6	Fine-silty, carbonatic, hyperthermic Typic Calciustolls	1.90	0.18
7	Loamy-skeletal, carbonatic, hyperthermic Typic Calciustolls	8.60	0.83
8	Fine-silty, carbonatic, hyperthermic Typic Haplustepts	335.86	32.70
9	Fine-silty, carbonatic, hyperthermic Typic Haplustolls	58.01	5.64
10	hyperthermic Typic Haplustolls Loamy-skeletal, carbonatic,	11.59	1.17
11	Loamy-skeletal, mixed, calcareous, hyperthermic Typic Ustorthents	134.82	13.12
مجموع Total	---	1027.08	100

نتایج صحت پیش‌بینی کلاس‌های خاک نشان داد که عموماً پیش‌بینی مدل برای کلاس‌هایی که فراوانی بیشتری در مشاهدات داشتند با صحت بالاتری همراه است. بر اساس نتایج جدول ۶ زیرگروه شماره ۳ (تیپیک کلسی یوستپز[2]) که ۴۸/۹۰ درصد خاکرخ‌های مشاهداتی را به خود اختصاص داده است دارای بالاترین مقادیر صحت کاربر و تولیدکننده می‌باشد. زیرگروه شماره یک (تیپیک هاپلویوستپز[3]) با ۱/۸۵ درصد فراوانی دارای کمترین صحت و بیشترین خطای پیش‌بینی هستند. برای افزایش صحت کلاس‌هایی که دارای کمترین صحت و صحت نامشخص می‌باشند می‌توان از تعداد نقاط مشاهداتی بیشتری استفاده کرد. دلیل دیگر این صحت نامشخص و کم داده‌های محیطی به کار رفته در پهنه بزرگی از منطقه ا ست که باعث می شود تغییرات خاک با صحت کمتری پیش‌بینی شود. برای بهبود این وضعیت کاربست تصاویر ماهواره‌ای و مدل رقومی ارتفاع با تفکیک مکانی بهتر ممکن است سبب افزایش صحت پیش‌بینی‌ها گردد. دقت کاربر و قابلیت اطمینان به ما اجازه می‌دهد که سطح بیش برآورد و کم برآورد در پیش‌بینی گروه‌های بزرگ خاک را تخمین بزنیم.

در روش میانگین کاهش صحت، مقادیر درست متغیرها با مقادیری که به‌طور تصادفی برای هر درخت تولید شده است جایگزین می‌شود و اگر این جایگزینی اثری روی خطای اندازه‌گیری نداشته باشد اهمیت آن کم می‌باشد و اگر مقدار خطای اندازه‌گیری افزایش یابد، آن متغیر مهم می‌باشد (۲۲). بر اساس شکل ۳ مهم‌ترین متغیری های محیطی در فرآیند مدل‌سازی و پیش‌بینی زیرگروه و فامیل خاک با سناریو جنگل تصادفی-میانگین کاهش صحت شامل شاخص موقعیت توپوگرافی، مدت تابش سالیانه، مساحت سطحی، انحنای شیب‌های بالایی، شاخص طبقه‌بندی ارا ضی پ ست، انحنای اراضی می‌باشد.

1- Out of bag (OOB)

1- Typic Calciustepts

3- Typic Haploustepts

جدول ٤- نتایج حاصل از سه روش انتخاب متغیر کمکی

Table 4- result of three methods for auxiliary variable selection

الگوریتم انتخاب متغیر Variable selection algorithm	پارامتر ژئومورفومتری Geomorphometric parameter	نماد متغیر کمکی (در مطالعه حاضر) Auxiliary variable symbol	تعریف Definition	رفرنس Reference
باروتا Boruta	انحنای محلی Local curvature	Local_Curve	انحنای محلی در موقعیت‌های مختلف زمین نما Local curvature in landcape different positions	(۸)
	ارتفاع نرمال شده Normalized height	Normalized	جابجایی عمودی یک سلول مدل رقومی ارتفاع و یک ناحیه تحت تأثیر مشخص را در نظر می‌گیرد It considered vertical movement a digital elevation model cell and distinct affected area	
	شاخص قدرت جریان Stream power index	*Stream_Pow	بیانگر قدرت جریان ناشی از تغییرات توپوگرافی سطح زمین Represntation of stream power duo to terrain surface topography variation	
شاخص تورم واریانس Variance inflation index	مدل رقومی ارتفاع Digital elevation model	Dem	-	
	مساحت حوزه آبگیر اصلاح شده Modified catchment area	Modified_C	تجمع جریان در سلول‌ها به عنوان مجموع جریان قبلی در حوضه آبخیز Flow accumulation in grid cells as a total previous flow in catchment area	
	ارتفاع شیب‌دار Slope height	Slope_Heig	فاصله عمودی سطح مبنا تا قله Vertical distance from base level to summit	
	بافت سطح عوارض زمین Terrain surface texture	Texture	زبری و نرمی سطح عوارض زمین Rudggeness and softness of terrain surface	
	مساحت آبگیر کل Total catchment Area	Total_Catc	مساحت حوزه آبخیز بالای یک آبراهه Catchment area UpSTREAM	
	فاصله عمودی تا یک سطح اساس شبکه کانال vertical distance to a channel network base level	Vertical_D	فاصله عمودی تا سطح مبنای شبکه آبراهه Vertical distance to channel network base level	
	شاخص همگرایی Convergence Index	Convergenc	یک پارامتر سیمای اراضی که بیانگر مطابقت جهت شیب سلول‌های اطراف با جهت ماتریس تئوری می‌باشد A parameter of landscape that represents concordance aspect f neighberhood cells with matrix direction	
میانگین کاهش صحت Mean decrease accuracy	تابش سالیانه Annual insolation	Diurnal_An	پتانسیل سالانه تابش کل در گام‌های زمانی و ذخیره سری‌های زمانی در یک مجموعه شبکه‌ایی Potential pf annual radiation in time steps and storage of time series in grid set	
	شاخص موقعیت توپوگرافی Topographic position Index	Topographi	شناسایی قسمت‌های بالایی، میانی و پایینی زمین‌نما بر اساس اختلاف ارتفاع در مقیاس‌های کوچک و بزرگ It districts upper, middle and lower portion of landscape based on elevation difference in small and large scales	
	انحنای شیب بالادست Upslope curvature	Upslope_Cu	بیانگر فاصله و جریان متوسط متناسب با انحنای محلی نسبت به ناحیه پایدار سلول در زیر منطقه الگوریتم جهت جریان چندگانه It represents distance and average flow appropriate with local curvature relation to cell stable area in subregion of multiple flow direction algorithm	
	مساحت سطح واقعی Real surface area	Surface_Ar	منطقه سلولی واقعی (پیش بینی نشده) را محاسبه می‌کند It calculated real cell area(unpredicted)	
	تحدب سطح عوارض زمین Terrain surface convexity	Convexity	میزان برآمدگی سطح عوارض زمین Content pf terrain surface convexity	
	شاخص طبقه بندی عوارض زمین برای اراضی پست Terrain classification index for lowlands	TCI_Low	.Terrain Classification Index for Lowlands (TCI Low) شاخص طبقه‌بندی عوارض در اراضی پست	

*این ویژگی به صورت مشترک در دو الگوریتم انتخاب متغیر MDA و Boruta انتخاب گردید.

جدول ۵- پیش‌بینی سطوح تاکسونومیک زیرگروه و فامیل توسط روش‌های یادگیری ماشین

Table 5- Predicting subgroup and family taxonomic levels by machine learning methods

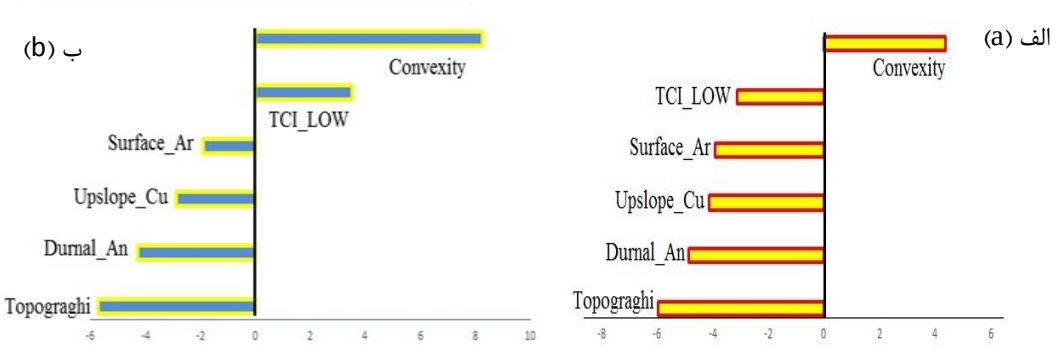
روش یادگیری ماشین Machine learning method	شاخص صحت سنجی Validation index	زیرگروه Subgroup	فامیل Family
RF-MDA	شاخص کاپا Kapa	0.57	0.3
	صحت کلی %OA	84	50
	%OOB	72.42	93.1
RF-VIF	شاخص کاپا Kappa	0.3	0.3
	صحت کلی OA٪	58	50
	%OOB	67.86	93.1
RF-Boruta	شاخص کاپا Kappa	0.55	0.3
	صحت کلی OA٪	67	50
	OOB٪	82.76	86.21
Fuzzy-MDA	شاخص کاپا Kappa	0.18	0.3
	صحت کلی OA٪	50	50

جدول ٦- مقایسه دقت پیش‌بینی زیرگروه خاک توسط مدل‌های برازش شده در داده‌های اعتبار سنجی

Table 6- Comparison of accuracy of soil subgroup prediction by models fitted to validation data

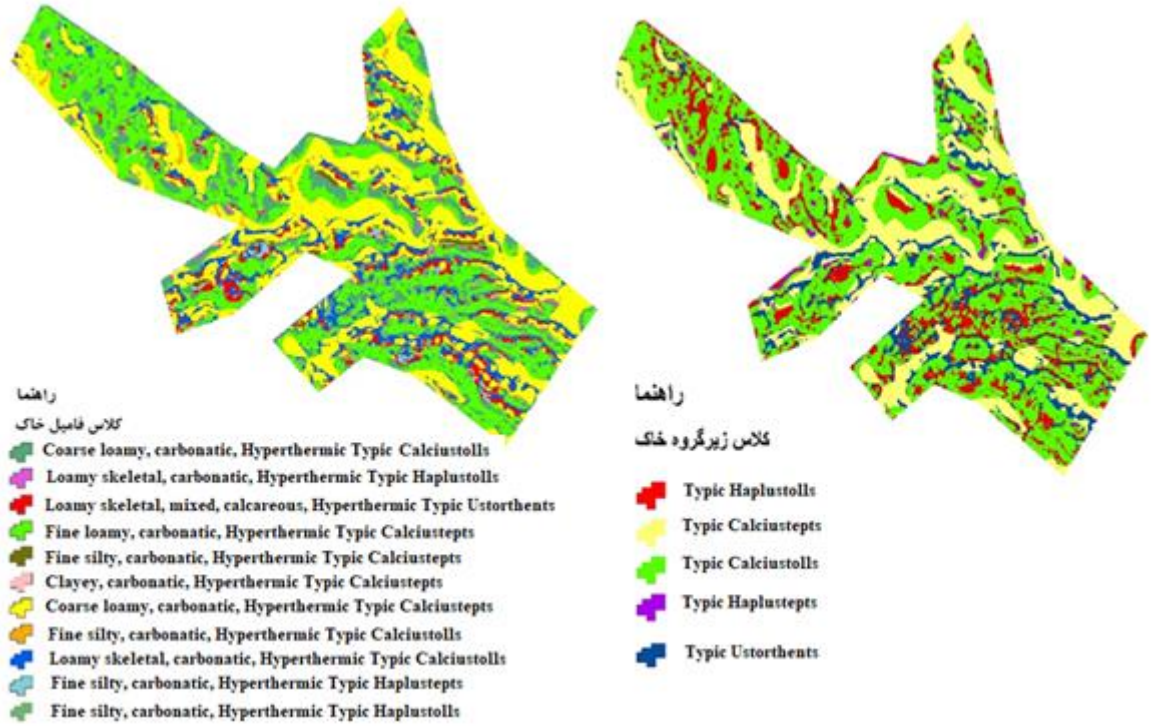
RF-VIF		RF-MDA		RF-Boruta		زیرگروه‌های خاک Soil subgroups
UA	PR	UA	PR	UA	PR	
100	100	100	50	100	50	Typic Calciustepts
75	75	80	100	100	100	Typic Calciustolls
NAN	NAN	NAN	NAN	NAN	NAN	Typic Haplustepts
0	0	NAN	NAN	100	67	Typic Haplustolls
0	0	NAN	NAN	0	NAN	Typic Ustorthents

UA: صحت کاربر PR: صحت تولیدکننده



شکل ۳- آنالیز حساسیت مؤثرترین متغیرهای کمکی در پیش‌بینی کلاس‌های خاک. بترتیب الف: زیرگروه و ب: فامیل خاک

Figure 3- Sensitivity analysis of most influence covariates in the prediction of soil classes (a) Subgroup and (b) Soil Family.



ب (b) الف (a)

شکل ٤- پراکنش مکانی کلاس‌های زیرگروه (الف) و فامیل خاک (ب) با استفاده از مدل RF-MDA

Figure 4- Spatial distribution of classes (a) subgroup and (b) soil family using RF-MDA model

این موضوع می‌تواند مؤید این باشد که در این منطقه، پستی و بلندی تأثیر بسزایی در متمایز کردن کلاس‌های خاک از هم دارد. همان‌طور که در شکل ۳ ملاحظه می‌شود فاکتور انحنای اراضی از بالاترین اهمیت و دارای مقادیر تأثیر مثبت در پیش‌بینی زیرگروه و فامیل خاک می‌باشد.

مدل جنگل تصادفی برای داده‌کاوی در مجموعه داده‌های بزرگ و پر شمار طراحی گردیده است (۱۲ و ۵). از سوی دیگر ادامه رده‌بندی یک پدون تا سطوح پایین‌تر رده‌بندی به معنی جداسازی دقیق‌تر خاکرخ‌های مشاهده شده در منطقه است. این وضع به کاهش فراوانی مشاهدات در هریک از کلاس‌ها می‌انجامد و در پی آن توان پیش‌بینی مدل جنگل تصادفی به شدت کاهش یافته و در نتیجه خطا در پیش‌بینی بسیار بزرگ خواهد شد. با توجه به اینکه مدل جنگل تصادفی بالاترین صحت در پیش‌بینی زیرگروه خاک در منطقه مورد مطالعه را دارا بود، در شکل ۴ نقشه خروجی مدل جنگل تصادفی در سطح زیرگروه و فامیل خاک مورد مطالعه ارائه شده است.

در نقشه پیش‌بینی شده، شش زیرگروه و ۱۱ فامیل خاک موجود در منطقه مورد مطالعه تشخیص داده شد. همان‌طور که در نقشه شکل ۴ ملاحظه می‌شود، خاک‌هایی با تکامل کم (تیپیک یوست اورتنتز[1]) در زمین‌نماهای مرتفع با شیب تند و فرسایش‌پذیری زیاد قرارگرفته‌اند. زیرگروه‌های خاک با تکامل متوسط به بالا (تیپیک کلسی یوستپز، تیپیک کلسی یوستالز[2]) در مناطق با ارتفاع و شیب کم تا متوسط که امکان آبشویی کربنات‌ها در نیمرخ خاک زیاد است تشکیل شده‌اند. بررسی دقیق نقشه پیش‌بینی پراکنش زیرگروه و مقایسه‌ی آن با مشاهدات واقعی انجام‌شده در صحرا تطابق چشم‌گیری نشان داد. این نتایج توانایی مدل جنگل تصادفی در ایجاد ارتباط میان متغیرهای محیطی و کلاس‌های خاک را نشان می‌دهد. زیرگروه تیپیک کلسی یوستالز و تیپیک هاپلویوستپز[3] به ترتیب بیشترین و کمترین مساحت نقشه پیش‌بینی زیرگروه خاک منطقه را تشکیل می‌دهند.

نتیجه‌گیری

نتایج حاصل از تکنیک‌های نقشه‌برداری رقومی خاک در منطقه مورد مطالعه نشان‌دهنده این بود که متغیرهای ژئومورفومتری (شاخص همگرایی/ واگرایی، شاخص طبقه‌بندی زمین در اراضی پست و مساحت واقعی سطح) بیشترین تأثیر را در پیش‌بینی کلاس‌های خاک

1- Typic Ustorthents
2- Typic Calciustolls
3- Typic Haploustepts

دارا بودند. از میان هشت رویکرد مدل سازی رقومی کلاس‌های خاک رویکرد جنگل تصادفی-میانگین کاهش صحت برای تهیه نقشه کلاس‌های خاک در سطح تاکسونومیک زیرگروه دارای بالاترین میزان صحت عمومی و شاخص کاپا بود. بر اساس نتایج این مطالعه، برای ارتقاء دقت و صحت پیش‌بینی مکانی کلاس های خاک مخصوصاً در مورد واحدهایی که دارای دقت پایین‌تری می‌باشند افزایش تعداد مشاهدات میدانی و استفاده از سایر متغیرهای محیطی تأثیرگذار بر روی تشکیل خاک‌ها از قبیل تصاویر ماهواره‌ای، اطلاعات زودیافت خاک و نمایندگان مواد مادری تشکیل دهنده خاک‌های منطقه را می توان به کارگیری نمود. به طورکلی رویکردهای نقشه‌برداری رقومی می‌توانند فرآیند نقشه‌برداری خاک‌ها را در گستره‌ای وسیع و متشکل از هرگونه عوارض طبیعی به یک باره انجام دهند و سرعت عمل و کارآمدی نقشه‌ها را در انتقال داده‌ها و اطلاعات افزایش داده و قابلیت استفاده از آن‌ها را برای بخش وسیعی از کاربران اراضی مقدور سازد.

منابع

1- Abbaszadeh F., Ayubi Sh., and Jafari A. 2018. Spatial forecasting of large soil groups using regression and decision tree models in the southeast region of Iran. Crop Engineering (Journal of Agricultural Science) 41: 123-146. (In Persian with English abstract)
2- Akinwande M., Dikko H., and Samson A. 2015. Variance Inflation Factor: As a condition for the inclusion of suppressor variable(s) in regression analysis. Open Journal of Statistics 5: 754-767.
3- Breiman L. 2001. Random forests. Machine Learning 45(1): 5-32.
4- Breiman L., and Cutler A. 2004. Random Forests, URL: http://www. stat. berkeley. edu/users/breiman. Random Forests/cc_papers. htm.
5- Brungard C.W., Boettiger J.L., Duniway M.C., Wiks S.A., and Edwards T.C. 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239: 68-83.
6- Campos A.R., Giasson E., Costa J.J.F., Machado I.R., Silva E.B.D., and Bonfatti B.R. 2018. Selection of environmental covariates for classifier training applied in digital soil mapping. Revista Brasileira de Ciência do Solo 42.
7- Chen T., Niu R.Q., Li P.X., Zhang L.P., and Du B. 2011. Regional soil erosion risk mapping using RUSLE, GIS, and remote sensing: a case study in Miyun watershed, North China. Environmental Earth Sciences 63(3): 533-541.
8- Conrad O., Bechtel B., Bock M., Dietrich H., Fischer E., Gerlitz L., Wehberg J., Wichmann V., and Böhner J. 2015. System for automated geoscientific analyses (SAGA) v. 2.1.4, Geoscientific Model Development Discussions 8(2).
9- Fatehi Sh. 2015. Scale descending properties and agglomeration of soil classes in part of Karkheh River Watershed in Kermanshah Province. PhD Thesis-Faculty of Agriculture-Shahrekord University.
10- Gee G.W., and Bauder J.W. 1986. Particle-size analysis 1. Methods of soil analysis: Part 1— Physical and mineralogical methods, (methodsofsoilan1), 383-411.
11- Hengel T., Rossiter D.G., and Stein A. 2003. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Geoderma 120: 75-93.
12- Heung B., HO H.C., Zhang J., Knudby A., Bulmer C. E., and Schmidt M.G. 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265: 62-77.
13- 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. European Journal of Soil Science 63(2): 284-298.
14- Khamoshi A., Sarmadian F., and Keshavarzi A. 2019. Digital soil mapping using random forest model in Abyek Region, Qazvin Province. Journal of Soil Research (Soil and Water Sciences) 32: 384. (In Persian with English abstract)
15- Kursa M.B., and Rudnicki W.R. 2010. Feature selection with the Boruta package. Journal of Statistical Software 36(11): 1–13.
16- Liaw A., and Wiener M. 2002. Classification and regression by random Forest. R news 2(3): 18-22.
17- Maghsodi Z., Rostaminia M., Faramarzi M., Keshavarzi A., and Rahmani A. 2018. Spatial forecasting of soil units in geographical information systems environment in Ilam Province. Journal of Soil Research (Soil and Water Sciences) 33: 254-268. (In Persian with English abstract)
18- Massawe B.H., Subburayalu S.K., Kaaya A.K., Winowiecki L., and Slater B.K. 2018. Mapping numerically classified soil taxa in Kilombero Valley, Tanzania using machine learning. Geoderma 311: 143-148.
19- Menezes M.D.D., Silva S.H.G., Mello C.R.D., Owens P.R., and Curi N. 2018. Knowledge-based digital soil mapping for predicting soil properties in two representative watersheds. Scientia Agricola 75(2): 144-153.
20- Minasny B., and McBratney A.B. 2016. Digital soil mapping: a brief history and some lessons. Geoderma 264: 301–311.
21- 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.
22- Mousavi S.R., Sarmadian F., Rahmani A., and Khamoushi S.E. 2019. Digital soil mapping with regression classification approaches by RS and Geomorphometrics covariates in the Qazvin plain, Iran. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences.
23- Nelson R.E. 1982. Carbonate and gypsum. In: Page AL (ed) Methods of soil analysis. American Society of Agronomy, Madison, pp 181–197.
24- Pahlavan Rad M.R., Toomaninan N., Khormali F., Brungard C.W., Bayram Komaki C., and Bogaert P. 2014. Updating soil survey maps using random forest and conditioned Latin hypercube in the loss derived soils of northern Iran. Geoderma 232: 97-106.
25- Rahmani A., Sarmadian F. Mousavi S.R., and Khamoushi S.E. 2019. Digital mapping of some surface soil properties using two random Forest and fuzzy logic approaches (Case Study: part of Kouhin lands, Qazvin Province). 16th Iranian Soil Science Congress. University of Zanjan. Zanjan. September 7th. (In Persian)
26- Soil science division staff. "Soil survey manual". USDA Handbook 18. 2017: 120-131.
27- Soil survey staff. 2014. Keys to soil taxonomy, United States Department of Agriculture. 12nd ed. Natural Resources Conservation Service.
28- Stum A.k., Boettinger J., White M., and Ramse R. 2010. Random forests applied as soil spatial model in arid. In digital soil mapping (pp. 179-190). Springer, Dordrecht.
29- Sumner M.E., and Miller W.P. 1996. Cation exchange capacity and exchange coefficients. Methods of soil analysis part 3—chemical methods, (methodsofsoilan3), 1201-1229.
30- 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: 67–77.
31- Van Wambeke A.R. 2000. The Newhall simulation model for estimating soil moisture and temperature regimes. Department of Crop and Soil Sciences. Cornell University, Ithaca, NY. USA.
32- 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.
33- Yang L., Qi F., Zhu A., Shi J., and An Y. 2016. Evaluation of integrative hierarchical stepwise sampling for digital soil mapping. Soil Science Society of America Journal 80(3): 637-651.
34- Zhao Z., Chow T. L., Rees H. W., Yang Q., Xing Z., and Meng F. 2009. Predict soil texture distributions using an artificial neural network model. Computers and Electronics in Agriculture 65(1):36-48.
35- Zhu A.X., and Band L.E. 1994. A knowledge-based approach to data integration for soil mapping. Canadian Journal of Remote Sensing 20(4): 408-418.

Journal of Water and Soil
Vol. 34, No. 4, Sept-Oct 2020, p. 973-987

نشریه آب و خاک (علوم و صنایع کشاورزی)
جلد ۳۴، شماره ۴، مهر – آبان ۱۳۹۹، ص. ۹۸۷-۹۷۳

The Efficiency of Different Feature Selection Methods in Digital Mapping of Subgroup and Soil Family Classes with Data Mining Algorithms

S. Nazari[1]- M. Rostaminia [2]*- SH. Ayoubi [3]- A. Rahmani [4]- S.R Mousavi [5]

Received: 27-04-2020
Accepted: 26-07-2020

Background and Objectives: High-accuracy of soil maps is a powerful tool for achieving land sustainability in agricultural and natural resources. This study was conducted to determine the effect of different feature selection methods with machine-learning algorithms to prepare digital mapping of soil classes at two taxonomic levels from subgroup down to family in the interest region, i.e. Vargar lands of Abdanan city, related to the Ilam province.

Materials and Methods: Study area is 1027 hectares with 628.6 mm and 22.6 ºC mean annual precipitation and temperature, respectively. Three major physiographic units including Hilland, Piedmont plain and Alluvial plain were considered. Soil moisture and temperature regimes are calculated based on the Newhall model in JNSM 6.1 version software. A total of 44 soil profile observation with random sampling pattern was determined based on standardized soil surveys then digging, description and after sampling from all genetic horizons then soil samples were transferred to the laboratory. Finally, all of the soil profiles were classified based on the soil taxonomy system (2014) down to the family level. Geomorphometric covariates as a representative of soil-forming factors were prepared from the digital elevation model (ALOS PALSAR Satellite,2011) with 12.5 m resolution in SAGA GIS 7.4 version software. Three feature selection approaches included Boruta, Variance inflation factors (VIF) and Mean decrease accuracy (MDA) with two Random forest (RF) and Fuzzy logic data mining algorithms were applied for relating soil-landscape relationship by using "random-forest", "caret" packages in R 3.5.1 and SoLIM solution version 2015 software's. Sample-based project used for predicting soil classes in Fuzzy logic modeling process. In total observation profile split into two data set included 80 percent (n=36) for calibrating and 20 percent for validating (n=8) based on bootstraps sampling algorithm random forest. Internal validation of the random forest algorithm was done based on out of bag error percentage (OOB%). The best model performance was determined based on overall accuracy (OA) and kappa index, also for each individual class user accuracy (UA) and producer accuracy (PA) were applied.

Results: The results have shown that from a number of 40 geomorphometry covariates, six covariates included Terrain classification index for lowlands, Annual insolation, Topographic position Index, Upslope curvature, Real surface area, and Terrain surface convexity were selected by MDA as the best environmental covariates. Also, the RF-MDA method with overall accuracy of 84% and Kappa index of 0.56 had the best performance compared to other methods (RF_VIF, RF-BO ,Fuzzy-MDA) in the subgroup level with 58, 55, 50 and 0.3, 0.67 and 0.18 respectively. Out of bag error results (%OOB) for RF-MDA, RF-VIF and RF-Boruta were obtained that 72.42%, 67.86%, and 82.76% for subgroup level and 93.10%, 93.10% and 86.21% for the family level respectively. while there was little difference between the accuracy of the method at the family taxonomic level and performed similar results in the modeling of soil classes process. The results of the fuzzy approach showed that the kappa index values and overall accuracy of this method were similar to the other three scenarios and there was a slight difference between the accuracy of the results at the soil family level. In the fuzzy method, it was observed that the kappa and overall accuracy values at the subgroup level were lower than the other scenarios. Fuzzy approaches in contrasted to RF modeling prevented continuous spatial variability by generating fuzzy maps for each of the soil classes in the landscape. These results indicate that the random forest method is superior to the fuzzy method in family class mapping and soil subgroups. Based on the MDA sensitivity analysis index, similarly, three geomorphometry covariates included Terrain surface convexity (convexity), Terrain classification index for lowlands (TCI_Low) and Real surface area (Surface_Ar) had the highest importance for predicting soil classes at two taxonomic levels. With regarded to final soil predicted maps area, two classes (Fine-silty, carbonatic, hyperthermic Typic Haplustepts) and Typic Calciustolls with 32.70% and 48.90% and (Fine-silty, carbonatic,

1- M.Sc. Graduated Student of Soil Science, ILam Uiverstity
2- Assistance Professor of Soil and Water Department, Agriculture Faculty, ILam Univerity
(*- Correspond Author Email: m.rostaminya@ilam.ac.ir)
3- Full Professor of Soil Science and Engineering, Agriculture Faculty, Isfahan University of Technology
4 and 5- Ph.D. Students of Soil Science and Engineering, Agricultue and Engineering Technology Faculty, University of Tehran
DOI: 10.22067/jsw.v34i4.85748

hyperthermic Typic Calciustolls) and Typic Haplustepts with 0.18% and 1.85% had the highest and lowest content at family and subgroup maps respectively.

Conclusion: In general, using different variable selection approaches in situations where soil classes have a relatively imbalanced abundance can increase the accuracy of digital mapping in soil studies. Increasing the number of field observations and the use of other environmental variables affecting soil formation can also be used for graduating in prediction low-accuracy soil classes.

Keywords: Random forest, Environmental covariates, Fuzzy logic, Soil mapping