کارایی سه مدل kNN، RF و SVM و مدل به دست آمده از ترکیب آنها به روش GR برای مدل‌سازی بافت خاک

فرشته میرزایی، علیرضا امیریان چکان*، روح‌الله تقی‌زاده مهرجردی و حمیدرضا متین‌فر

دانشجوی دکتری گروه مهندسی علوم خاک، دانشکده کشاورزی، دانشگاه لرستان، خرم آباد، ایران.: f.mirzaie1374@gmail.com

استادیار گروه مهندسی علوم خاک، دانشکده کشاورزی، دانشگاه لرستان، خرم آباد، ایران: amirian.ar@lu.ac.ir

استادیار گروه مرتع و آبخیزداری، دانشکده کشاورزی و منابع طبیعی، دانشگاه اردکان، اردکان، ایران.: rtaghizadeh @ardakan.ac.ir

استاد گروه مهندسی علوم خاک، دانشکده کشاورزی، دانشگاه لرستان، خرم آباد، ایران.: matinfar.h@lu.ac.ir

«مقاله پژوهشی»

دریافت:۱۴۰۲/۹/۱۸ و پذیرش: ۱۴۰۳/۲/۲۲

چکیده

بافت خاک یکی از مهمترین ویژگی‌هایی است که رفتار فیزیکی، شیمیایی و بیولوژکی خاک را کنترل می‌کند. روش‌های مختلفی برای مدل‌سازی بافت خاک استفاده می‌شوند که هر کدام دارای مزایای خاص خود هستند. یکی از راهکارهای سود بردن از مزایای این مدل‌ها ترکیب تخمین آنها است. با توجه به این که بافت خاک یک داده مرکب است، وقتی اجزاء آن جداگانه تخمین زده می‌شوند تضمینی برای برابر ۱۰۰ شدن جمع سه جزء تخمینی وجود ندارد. برای تضمین برابر با ۱۰۰ شدن تخمین‌های سه جزء بافت می‌توان از تبدیل‌های log-ratio استفاده کرد. اطلاعات کمی در خصوص کارآیی مدل‌های ترکیبی در مدل‌سازی داده‌های تبدیل‌شده و نشده بافت وجود دارد و به نظر می‌رسد بر اساس این رویکرد تا کنون مطالعه‌ای روی بافت خاک انجام نشده است. در این بررسی تعداد ۲۰۰ نمونه سطحی از خاک‌های منطقه کوهدشت برداشت شد. سه مدل جنگل تصادفی (RF)، k نزدیکترین همسایه (kNN) و ماشین‌های بردار پشتیبان (SVM) و مدل حاصل از ترکیب آن‌ها به روش Granger-Ramanathan (GR) برای مدل‌سازی، روش‌های additive log-ratio (alr)، centred log-ratio (clr) و isometric log-ratio (ilr) برای تبدیل داده‌ها و داده‌های حاصل از مدل رقومی ارتفاع (DEM) و تصاویر لندست هشت و سنتینل دو به عنوان ورودی مدل‌ها استفاده شدند. نتایج نشان داد متغیرهای استخراج‌شده از DEM اهمیت بیشتری در پیش‌بینی بافت خاک داشتند. به‌طور کلی، هر چهار مدل با استفاده از تبدیل alr منجر به تخمین‌های بهتری نسبت به تبدیل‌های clr و ilr و داده‌های تبدیل‌نشده (UT) گردیدند. مدل ترکیبی(GR) با مقادیر RMSE برابر با ۵/۰۷، ۴/۲۱، ۵/۸۱ و ۶/۰۹ درصد برای رس، مقادیر ۷/۱۱، ۵/۱۵، ۹/۰۴ و ۶/۷۰ درصد برای سیلت و مقادیر ۹/۲۰، ۷/۶۷، ۱۱/۶۹ و ۸/۷۴ درصد برای شن به ترتیب برای داده‌های UT و تبدیل‌های alr، clr و ilr منجر به بهبود تخمین‌ها نگردید. به‌طور کلی، مدل SVM با داده‌های تبدیل شده به روش alr کارآیی کمی بیشتر از سایر مدل‌ها داشت. نتایج نشان داد ترکیب چند مدل یادگیری ماشین الزاما باعث بهبود تخمین‌ها نمی‌گردد و به جای ترکیب نتایج چند الگوریتم که ممکن است باعث پیچیدگی فرایند مدل‌سازی شود، می‌توان از یک مدل مناسب برای برآورد بافت خاک استفاده کرد.

کلمات کلیدی: تبدیل لگاریتمی، جنگل تصادفی، داده مرکب، مدل‌های ترکیبی

* آدرس ایمیل نویسنده مسئول: amirian.ar@lu.ac.ir

مقدمه

بافت خاک یکی از مهم‌ترین خصوصیات فیزیکی خاک است که بیشتر فرآیندهای فیزیکی، شیمیایی، بیولوژیکی و هیدرولوژیکی خاک را کنترل می‌کند و نقش مهمی در حساسیت خاک به تخریب، انتقال و توزیع آب در خاک، کیفیت خاک و بهره‌وری از آن دارد. بافت خاک توانایی نگهداری آب و عناصر غذایی، نفوذپذیری، زهکشی، تهویه، مقدار کربن آلی، ظرفیت بافری، تخلخل و بسیاری از خواص دیگر خاک را تحت تاثیر قرار می‌دهد (اکپا و همکاران، ۲۰۱۴). همچنین در مدل‌سازی‌های هیدرولوژکی، اکولوژیکی و زیست–محیطی، بافت خاک معمولا به عنوان ورودی توابع انتقالی خاک استفاده می‌شود (میناسنی و مک براتنی، ۲۰۱۸؛ ون لوی و همکاران، ۲۰۱۷)؛ بنابراین داشتن داده‌های دقیق و کمی از بافت خاک اهمیت زیادی در مدیریت اراضی زراعی، منابع طبیعی و محیط زیست دارد.

بافت خاک رایج‌ترین داده مرکب[1] در علوم خاک است. داده مرکب داده‌ای است که دارای چند جزء غیر منفی است که مجموع آنها برابر با واحد است (ایتچیسون، ۱۹۸۶). چون بافت خاک یک داده مرکب است، وقتی اجزاء آن (رس، سیلت و شن) توسط روش‌های یادگیری ماشین به صورت جداگانه تخمین زده می‌شوند تضمینی برای این‌که مجموع اجزاء تخمینی برابر با واحد شود وجود ندارد (لارک و بیشاپ، ۲۰۰۷). یک روش رایج برای رفع این مشکل این است که دو جزء تخمین زده شوند و جزء سوم از اختلاف مجموع دو جزء تخمینی از عدد واحد به دست آید. در این صورت بسته به این‌که ابتدا کدام دو جزء تخمین زده می‌شوند، ممکن است برای یک جزء تخمین‌های متفاوتی به‌دست آید. در نتیجه در مدل‌سازی‌هایی که نیاز است اجزاء بافت خاک تخمین زده شوند ممکن است نتایج متفاوتی به‌دست آید. راهکار مناسب دیگر برای تضمین برابر با ۱۰۰ شدن تخمین‌ها استفاده از تبدیل‌های log- ratio معرفی شده توسط ایتچیسون (۱۹۸۶) است. تبدیلات alr[2]، clr[3] و ilr[4] از روش‌های رایج تبدیل داده‌های مرکب هستند (ایتچیسون، ۱۹۸۶؛ اگوزکیو و همکاران، ۲۰۰۳؛ فیلزموسر و همکاران، ۲۰۰۹). با وجود این که از این تبدیلات برای مدل‌سازی بافت خاک استفاده شده است (اوده و همکاران، ۲۰۰۳؛ لارک و بیشاپ، ۲۰۰۷؛ لیو و همکاران، ۲۰۱۲؛ اکپا و همکاران، ۲۰۱۴؛ پوگیو و گیمونا، ۲۰۱۷؛ وانگ و شی، ۲۰۱۷)، ولی مقایسه کارآیی روش‌های مختلف یادگیری ماشین در مدل‌سازی داده‌های تبدیل‌شده فقط در مطالعات کمی انجام گرفته است.

روش‌های مبتنی بر فنون نقشه‌برداری رقومی خاک، از رویکردهای مناسب برای کسب داده های کمی، مکانی و پیوسته از بافت خاک هستند. گسترش تکنیک‌های نقشه‌برداری رقومی خاک و افزایش علاقه به اشتراک داده‌ها اغلب منجر به در دسترس بودن نقشه‌های متعدد برای یک منطقه و یک ویژگی خاص می‌شود و این نقشه‌ها معمولا از نظر مقیاس، وسعت، روش مدل‌سازی و دقت متفاوت هستند (رومن دوبارکو و همکاران، ۲۰۱۷؛ چن و همکاران، ۲۰۲۰). هر کدام از مدل‌های استفاده شده در تهیه این نقشه‌ها دارای ایرادت و مزایای خاص خود است و یک مدل واحد معمولاً در همه شرایط مناسب نیست. یک رویکرد مناسب برای سود بردن از مزایای مدل‌های مختلف و به‌دست آوردن نقشه‌های دقیق‌تر، ترکیب و یا میانگین‌گیری از تخمین‌های آن‌ها است. ایده روش‌های میانگین‌گیری این است که اگر چندین مدل ضعیف را با هم ترکیب کنیم، می‌توانیم یک مدل قوی‌تر ایجاد کنیم (لانتز، ۲۰۱۵)؛ به عبارت دیگر، با میانگین‌گیری از پیش‌بینی‌های مدل‌های متعدد، می‌توان مدلی ترکیبی ایجاد کرد که در مقایسه با هر کدام از مدل‌ها به صورت جداگانه، عملکرد بهتری داشته باشد (دیکز و وروگت، ۲۰۱۰). روش‌های مختلفی برای این منظور به کار می‌رود که از آن جمله می‌توان به روش‌های میانگین‌گیری با وزن‌های برابر، میانگین‌گیری وزنی بر اساس

---

[1]. Compositional data

[2]. Additive log-ratio

[3]. Centred log-ratio

[4]. Isometric log-ratio

واریانس، میانگین‌گیری بر اساس مدل بیزی، میانگین‌گیری گرنجر راماناتان[1] (GR)، میانگین‌گیری بر اساس معیار اطلاعات و میانگین‌گیری مالو اشاره کرد. با وجود این که از این رویکرد برای مدل‌سازی ویژگی‌هایی از جمله شوری خاک (عابدی و همکاران، ۲۰۲۹)، کلاس‌های تاکسونومیک خاک (تقی زاده مهرجردی و همکاران، ۲۰۱۹) و ماده آلی خاک (ملانو و همکاران، ۲۰۱۴؛ چن و همکاران، ۲۰۲۰) استفاده شده است، ولی استفاده از این رویکرد در مدل‌سازی داده‌های خام و تبدیل شده (تبدیلات log-ratio) بافت خاک تا کنون گزارش نشده است.

در کشورهایی (مانند ایران) که داده‌های خاک برای ایجاد نقشه‌های قابل اعتماد محدود است، می‌توان از روش میانگین‌گیری از مدل‌ها برای ترکیب نقشه‌های موجود با استفاده از تعداد محدودی از داده‌های واسنجی استفاده کرد (چن و همکاران، ۲۰۲۰). مطالعات محدودی در زمینه کارآیی رویکردهای میانگین‌گیری از مدل‌ها برای مدل‌سازی خواص خاک در دنیا و ایران انجام شده است و کارآیی این رویکرد در مقایسه با مدل‌های رایج یادگیری ماشین برای مدل‌سازی داده‌های تبدیل‌شده و تبدیل‌نشده بافت خاک تا کنون گزارش نشده است؛ بنابراین در پژوهش حاضر برای نخستین بار کارآیی رویکرد ترکیب تخمین‌های مدل‌های یادگیری ماشین مختلف برای برآورد اجزاء بافت خاک با استفاده از داده‌های خام (تبدیل نشده) و تبدیل شده (تبدیل‌های log-ratio) بر اساس رویکردهای نقشه‌برداری رقومی خاک انجام گرفت. نتایج این پژوهش مشخص می‌کند آیا ترکیب چند مدل که معمولا باعث پیچیدگی فرایند مدل‌سازی و صرف وقت بیشتری می‌شود باعث بهبود معنی‌داری در تخمین اجزاء بافت خاک نسبت به استفاده از هر مدل به صورت جداگانه می‌شود؟ بنابراین هدف این تحقیق به طور مشخص بررسی کارآیی سه روش یادگیری ماشین شامل k نزدیکترین همسایه[2] (kNN)، جنگل تصادفی[3] (RF) و ماشین‌های بردار پشتیبان[4] (SVM) در مدل‌سازی داده‌های تبدیل‌شده و تبدیل‌نشده بافت خاک و مقایسه کارآیی این سه مدل با مدل حاصل از ترکیب آن‌ها (مدل GR) می‌باشد.

مواد و روش‌ها

مشخصات منطقه

منطقه مورد مطالعه بخشی از دشت داوود رشید در شمال‌غربی شهر کوهدشت در استان لرستان به وسعت تقریبی ۲۰۰۰ هکتار می‌باشد که در محدوده طول شرقی ′۵۱ °۴۶ تا ′۵۰ °۴۷ و عرض شمالی ′۹ °۳۳ تا ′۵۶ °۳۳ واقع گردیده است (شکل ۱). بر اساس داده‌های ایستگاه هواشناسی کوهدشت، میانگین دمای سالانه ۱۶ درجه سانتی‌گراد و میزان بارندگی سالیانه ۴۵۰ میلی‌متر می‌باشد. منطقه دارای اقلیم معتدل با تابستان‌های گرم و زمستان‌های سرد است. بـه استناد نقشـه رژیم‌های رطوبتی و حرارتی خاک‌های ایران (بنایی، ۲۰۰۰)، رژیم‌های رطوبتی و حرارتی خاک‌های منطقه به ترتیب زریک و ترمیک می‌باشند. ارتفاع متوسط منطقه از سطح دریا حدود ۱۲۲۵ متر است. واحد ژئومورفولوژی منطقه دشت دامنه‌ای و مواد مادری تشکیل دهنده خاک‌های منطقه اغلب آهکی هستند. منطقه از نظر زمین‌شناسی جزء پهنه زاگرس چین‌خورده محسوب می‌شود. پوشش گیاهی بومی منطقه شامل انواع گیاهان مرتعی، زالزالک، بلوط، بادام و کاربری غالب اراضی منطقه شامل کشت دیم گندم و جو و زراعت آبی چغندرقند، یونجه، ذرت، صیفی‌جات و باغ‌های زردآلـو، انجیـر و انگـور است.

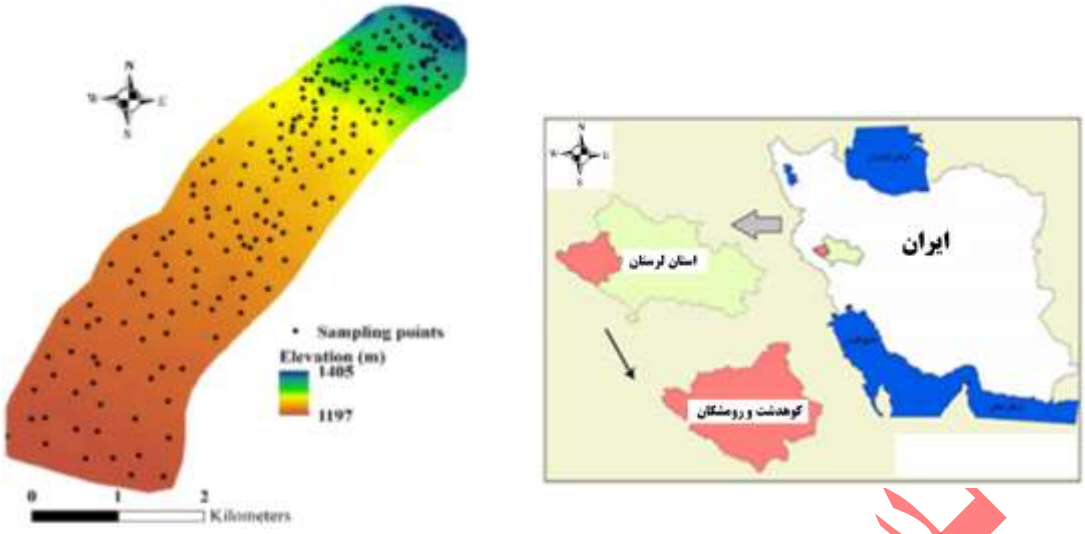
---

[1]. Granger-Ramanathan averaging

[2]. K-nearest neighbor

[3]. Random forest

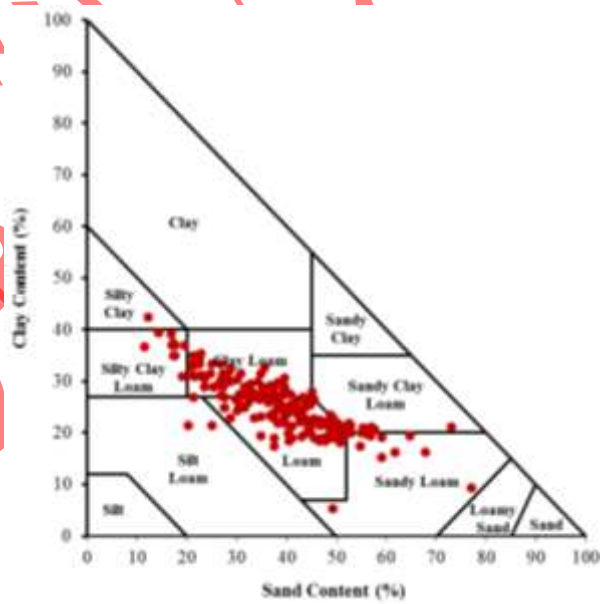
[4]. Support vector machines



شکل۱- موقعیت منطقه مورد بررسی و نقاط نمونه‌برداری

نمونه‌برداری و تعیین بافت خاک

برداشت نمونه‌های خاک و تعیین بافت خاک در آزمایشگاه در سال ۱۴۰۰ انجام شد. برای انتخاب نقاط نمونه‌برداری، محدوده‌ای انتخاب شد که دارای تنوع مناسبی از نظر بافت بود و تعداد ۲۰۰ نمونه خاک از عمق سطحی (۳۰ سانتی‌متری) به صورت تصادفی برداشت گردید. نمونه‌های خاک پس از خشک شدن در هوای آزاد، کوبیده و از الک دو میلی‌متری عبور داده شدند و درصد رس، سیلت و شن آنها به روش هیدرومتر (جی و باودر، ۱۹۸۶) تعیین شد. در (شکل ۲) توزیع کلاس‌های بافت خاک روی مثلث بافت خاک نشان داده شده است. بیشتر نمونه‌ها دارای کلاس‌های لوم و لومی رسی هستند و تعداد کمتری در کلاس‌های لومی رسی شنی، لومی رسی سیلتی و لومی شنی قرار دارند.



شکل۲- توزیع کلاس‌های بافت نمونه‌های خاک بررسی شده

تبدیل داده‌های بافت به روش‌های log-ratio

بردار $x=[x_1, x_2, \ldots, x_D]$ یک داده‌D جزیی وقتی یک داده مرکب در نظر گرفته می‌شود که همه اجزاء آن اعداد حقیقی مثبت و در برگیرنده اطلاعات نسبی باشند. فضای نمونه مناسب برای داده‌های مرکب یک‌Simplex (داده خام) به صورت زیر است (ایتچیسون، ۱۹۸۶):

$$S^D=\{x=[x_1,x_2, \ldots, x_D] \mid x_i>0, i=1, 2, \ldots, D; \sum_{i=1}^{D} x_i=\kappa \quad \text{رابطه ۱}$$

در این رابطه S^D بیانگر بردارهای یک ترکیب D جزیی و K مقدار ثابتی است که معمولا برابر با ۱۰۰ یا یک است. ایتچیسون (۱۹۸۶) برای اطمینان از برابر واحد شدن مجموع اجزاء تخمینی یک داده مرکب، تبدیل‌های لگاریتمی نسبتی را پیشنهاد داد. در این تحقیق داده‌های بافت با استفاده از سه نوع تبدیل رایج شامل alr، clr و ilr تبدیل شدند. سپس داده‌های تبدیل شده با استفاده از سه مدل یادگیری ماشین و یک مدل ترکیبی مدل‌سازی شدند. در نهایت برای به دست آوردن تخمین‌های رس، سیلت و شن، داده‌های تبدیل شده تخمینی مجددا به صورت معکوس تبدیل شدند[1]. جزییات روابط مربوط به تبدیل‌های مذکور و معکوس آن‌ها در امیریان چکان و همکاران (۲۰۱۹) ارائه شده است. تبدیل‌های alr، clr و ilr و معکوس آنها از طریق روابط ۲ تا ۹ و با استفاده از بسته نرم‌افزاری compositions در محیط R انجام شد.

تبدیل alr با استفاده از رابطه زیر، داده‌های خام رابه مختصات alr تبدیل می‌کند:

$$y = alr(x) = [ln\frac{x_1}{x_D}, ln\frac{x_2}{x_D}, \ldots, ln\frac{x_{D-1}}{x_D}] \quad \text{رابطه ۲}$$

معکوس تبدیل alr (تبدیل داده‌های با مختصات alr به داده خام اولیه) با استفاده از رابطه زیر انجام می‌شود:

$$x = agl(y) = [\frac{\exp(y_1)}{1+\sum_{i=1}^{D-1}\exp(y_i)}, \ldots, \frac{\exp(y_{D-1})}{1+\sum_{i=1}^{D-1}\exp(y_i)}, \frac{1}{1+\sum_{i=1}^{D-1}\exp(y_i)}] \quad \text{رابطه ۳}$$

تبدیل clr داده‌های خام را به مختصات clr به صورت زیر تبدیل می‌کند:

$$y = clr(x) = [ln\frac{x_1}{(\prod_{i=1}^{D} x_i)^{\frac{1}{D}}}, ln\frac{x_2}{(\prod_{i=1}^{D} x_i)^{\frac{1}{D}}}, \ldots, ln\frac{x_D}{(\prod_{i=1}^{D} x_i)^{\frac{1}{D}}}] \quad \text{رابطه ۴}$$

معکوس تبدیل clr (clr^{-1}) با استفاده از رابطه زیر انجام می‌شود:

$$x = clr^{-1}(y) = [\frac{\exp(y_1)}{\sum_{i=1}^{D-1}\exp(y_i)}, \frac{\exp(y_2)}{\sum_{i=1}^{D-1}\exp(y_i)}, \ldots, \frac{\exp(y_D)}{1+\sum_{i=1}^{D-1}\exp(y_i)}] \quad \text{رابطه ۵}$$

تبدیل ilr داده‌های خام را به مختصات ilr به صورت زیر تبدیل می‌کند:

$$y = ilr(x) = (y_1, y_2, \ldots, y_{D-1}) \in \mathbb{R}^{D-1} \quad \text{رابطه ۶}$$

در این رابطه y_i به صورت رابطه زیر تعریف می‌شود:

$$y_i = \frac{1}{\sqrt{i(i+1)}} \ln\left(\frac{\prod_{j=1}^{i} x_j}{(x_i+1)^i}\right), (i = 1, 2, \ldots, D-1) \quad \text{رابطه ۷}$$

تبدیل معکوس از مختصات ilr به داده‌های اولیه خاک با استفاده از رابطه زیر انجام می‌شود:

$$x = ilr^{-1}(y) = [\left(1+\frac{\sum_{i=0,i\neq 1}^{D} f(i)}{f(0)}\right)^{-1}, \ldots, \left(1+\frac{\sum_{i=0,i\neq 1}^{D} f(i)}{f(D-1)}\right)^{-1}] \quad \text{رابطه ۸}$$

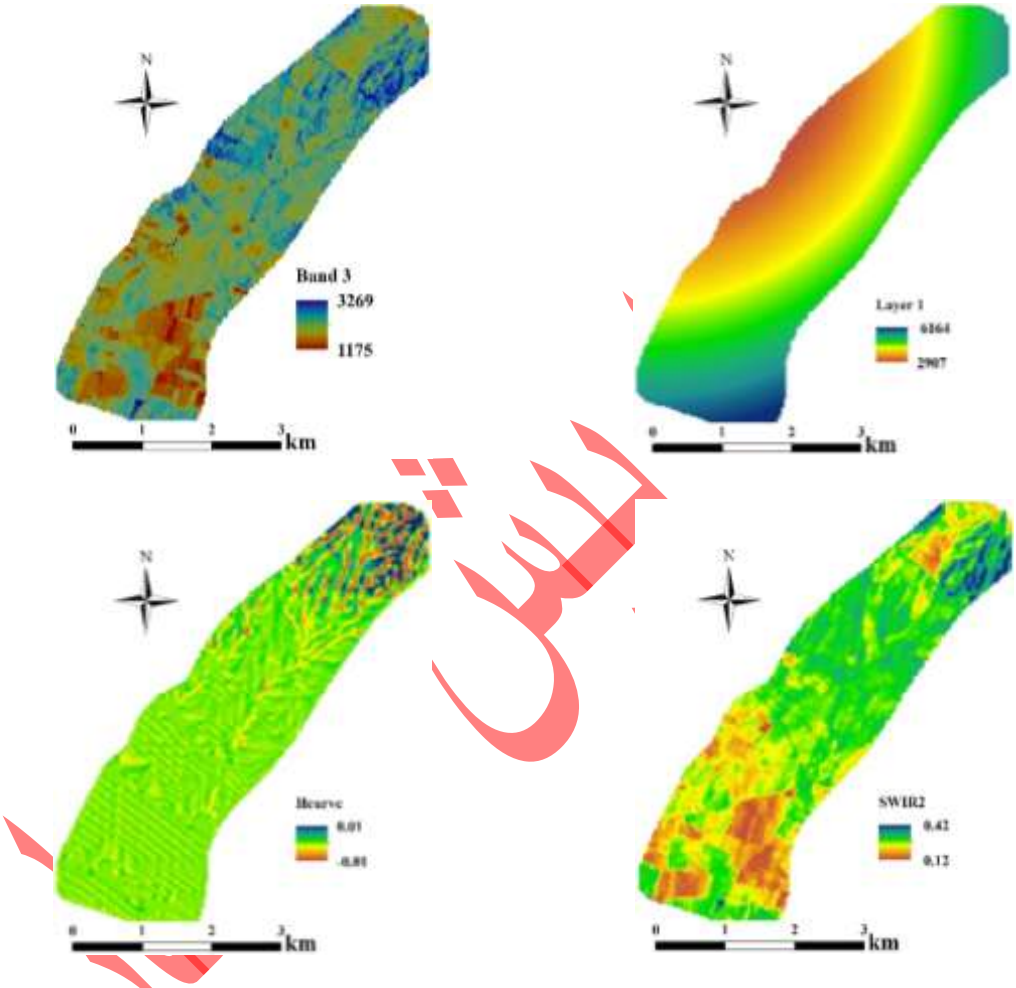
در این رابطه $f(i)$ به صورت زیر تعریف می‌شود:

$$f(i) = (\frac{1}{f(i-1)}\exp(\sqrt{i(i+1)}y))^{-1/i}, f(0) = 1 \quad \text{رابطه ۹}$$

[1]. Back transformation

استخراج متغیرهای کمکی

در این مطالعه داده‌های کمکی به عنوان نماینده فاکتورهای خاک‌سازی از منابع مختلفی از جمله مدل رقومی ارتفاع (DEM)، تصویر ماهواره لندست ۸ و تصویر Sentinel-2 به دست آمدند (جدول ۱). خصوصیات زمین‌نما در محیط نرم افزار SAGA از DEM استخراج گردید (هنگل و همکاران، ۲۰۰۳). همچنین باندهای تصاویر لندست ۸ و Sentinel-2 و ترکیب این باندها به صورت شاخص‌هایی از جمله شاخص‌های رس، روشنایی، پوشش گیاهی و کربنات به عنوان متغیر کمکی استفاده گردید. در (شکل ۳) توزیع مکانی چهار متغیر کمکی به عنوان نمونه نشان داده شده است.



شکل ۳- توزیع مکانی چهار تا از متغیرهای کمکی شامل باند ۳ تصویر لندست ۸ (Band 3)، میدان فاصله اقلیدسی (Layer 1)، انحنای افقی (Hcurve) و مادون قرمز موج کوتاه (SWIR2)

جدول ۱- متغیرهای کمکی مهم استفاده شده در مدل‌سازی اجزاء بافت خاک

منبع متغیر	نام متغیر	علامت اختصاری	رابطه/ نرم افزار	منبع
مدل رقومی ارتفاع	فاصله عمودی تا شبکه کانال	Chnl_Dist	SAGA	کنراد و همکاران (۲۰۱۵)
	انحنای افقی	Hcurve	SAGA	کنراد و همکاران (۲۰۱۵)
	میدان فاصله اقلیدسی	Layer1, …, Layer5	ArcGIS	بهرنس و همکاران (۲۰۱۸)
	مساحت سطح	Surface-area	SAGA	کنراد و همکاران (۲۰۱۵)
	شاخص چندتفکیکی همواری کف دره	MRVBF	SAGA	گالانت و داولینگ (۲۰۰۳)
	مساحت حوضه اصلاح شده	MCA	SAGA	کنراد و همکاران (۲۰۱۵)
	گشودگی مثبت	PosOpen	SAGA	کنراد و همکاران (۲۰۱۵)
	عمق دره	Valley-Depth	SAGA	تقی‌زاده مهرجردی و همکاران (۲۰۱۵)
نصاویر لندست ۸ و سنتینل ۲	باندهای طیفی	B2, B3, B4, B5, B7, B8, B11, B12, blue, green, red, NIR, SWIR1, SWIR2	-	ولدر و همکاران (۲۰۱۶)
	شاخص پوشش گیاهی بهبود یافته	EVI	(NIR–red)/(NIR + C1 × red–C2 × blue + L)	وانگ و همکاران (۲۰۲۳)
	شاخص کربنات	CarbIndex	red/green	تقی‌زاده مهرجردی و همکاران (۲۰۱۶)
	شاخص روشنایی	BrighIndex	$((red)^2+(NIR)^2)^{0.5}$	مترنیچ و زینک (۲۰۰۳)

مدل‌های مورد استفاده

در این مطالعه برای تخمین درصدهای رس، سیلت و شن در نقاط نمونه‌برداری نشده، متغیرهای محیطی استخراج شده از منابع مختلف به عنوان ورودی به مدل‌های kNN، RF، SVM و GR وارد شدند.

در رویکرد kNN، برای تخمین یک نمونه جدید در مسائل رگرسیونی، تعداد k نمونه از نزدیکترین همسایه‌ها از فضای تخمین‌گر انتخاب می‌شود و مقدار پیش‌بینی‌شده نمونه جدید، میانگین مقادیر k همسایه تعیین شده است. فاصله اقلیدسی (رابطه ۱۰) رایج‌ترین معیار تعیین شباهت در الگوریتم kNN است (لانتز، ۲۰۱۵). در این رابطه dist (p,q) فاصله اقلیدسی بین دو نمونه p و q است که هر کدام دارای n ویژگی است؛ p_1 مقدار اولین ویژگی از نمونه p و q_1 مقدار اولین ویژگی از نمونه q است. به عنوان نمونه، برای تخمین درصد رس خاک در یک مکان، از درصد رس تعداد k نمونه مجاور مکان مورد نظر که دارای کمترین فاصله اقلیدسی با آن هستند میانگین گرفته می‌شود. تعداد نقاط مجاور که در تخمین مقدار رس در نقطه مورد نظر استفاده می‌شوند بستگی به مقدار k دارد؛ مثلا اگر k برابر با سه باشد، از درصد رس سه نقطه که دارای کمترین فاصله با نقطه مورد نظر هستند میانگین گرفته می‌شود؛ بنابراین ابتدا فاصله اقلیدسی نقطه مورد نظر با نقاط مجاور محاسبه می‌شود، سپس به تعداد k از آنها برای تخمین نهایی انتخاب می‌شود.

$$dist\ (p,q)=\sqrt{(p_1-q_1)^2+(p_1-q_1)^2+\cdots+(p_n-q_n)^2}$$ رابطه ۱۰)

مدل RF یکی از الگوریتم‌های رایج یادگیری ماشین با توانایی و قابلیت بالا هم در مسائل رگرسیون و هم در مسائل طبقه‌بندی است که توسط بریمن (۲۰۰۱) معرفی گردید. این مدل دارای ساختار درختی است که بر خلاف سایر روش‌های مبتنی بر درخت تصمیم که از یک درخت برای تصمیم‌گیری استفاده می‌کنند، از تعداد زیادی درخت استفاده می‌کند. در این الگوریتم داده‌ها به دو دسته آموزش و آزمون تقسیم می‌شوند. از داده‌های آموزش به تعداد درختان تصمیم موجود در جنگل نمونه Bootstrap برداشته می‌شود و برای هر نمونه، یک درخت تصمیم ایجاد می‌شود. ایجاد درخت تصمیم در RF از گره ریشه شروع می‌شود.

به این صورت که از متغیرهای (در این تحقیق متغیرهای محیطی) که روی ویژگی مورد نظر (مثل رس، سیلت و شن) تاثیر دارند تعدادی زیرمجموعه به صورت تصادفی انتخاب می‌شود و موثرترین متغیر از هر مجموعه که در واقع معیاری برای تفکیک گره به شاخه‌ها است در گره ریشه قرار می‌گیرد. در گره‌های بعدی متغیرهایی با اهمیت کمتر از متغیر گره ریشه قرار می‌گیرند. این روند ادامه پیدا می‌کند تا بر اساس هر زیرمجموعه از متغیرها یک درخت تشکیل شود. هر درخت مدل‌سازی را روی داده‌ها انجام می‌دهد و متغیر هدف بر اساس متغیرهای ورودی به هر درخت تخمین زده می‌شود؛ به عبارت دیگر هر درخت بر اساس متغیرهایی که بر اساس آنها تشکیل شده است، متغیر مورد نظر را تخمین می‌زند. در نهایت در صورتی که از الگوریتم RF برای تخمین یک متغیر کمی پیوسته (مثل درصد رس) استفاده شود، میانگین تخمین همه درختان به‌عنوان تخمین نهایی آن متغیر در نظر گرفته می‌شود.

هدف مدل SVM که توسط وپنیک (۱۹۹۵) ارائه گردید ایجاد صفحه‌ای به نام ابرصفحه[1] است که فضای داده‌ها را طوری تقسیم کند که بخش‌های نسبتا همگنی در هر طرف صفحه به وجود آید؛ به عبارت دیگر SVM صفحه‌ای را پیدا می‌کند که حداکثر جداسازی ممکن را باعث شود که به آن ابرصفحه با حاشیه حداکثر[2] (MMH) گفته می‌شود. بردارهای پشتیبان نزدیک‌ترین نقاط از هر کلاس به MMH هستند و با استفاده از آن‌ها امکان تعیین MMH وجود دارد. رگرسیون بردار پشتیبان[3] (SVR) الگوریتمی است که همانند سایر روش‌های رگرسیونی، الگوی موجود بین متغیر وابسته (در این تحقیق درصد رس، سیلت و شن) و متغیرهای مستقل (مثل متغیرهای محیطی) را بر اساس داده‌های آموزش شناسایی می‌کند و از این الگو برای تخمین متغیر هدف استفاده می‌کند. برای این منظور، SVR با استفاده از یک تابع غیرخطی ($\varphi(x)$)، داده‌های ورودی را به روش حداقل‌سازی به یک فضا با ابعاد بیشتر نگاشت می‌کند (رابطه ۱۱).

$$f(x) = w^T.\varphi(x) + b$$ رابطه ۱۱)

در این رابطه x داده ورودی، w^T بردار اوزان و b مقدار بایاس است.

در این تحقیق برای ترکیب تخمین سه مدل kNN، RF و SVM از روش GR استفاده شد. این روش توسط گرانجر و رامانتان (۱۹۸۴) و بر اساس روش حداقل مربعات معمولی[4] (OLS) ارائه گردید. در این رویکرد یک مدل رگرسیونی چندگانه به داده‌های واقعی و پیش‌بینی‌شده توسط مدل‌های مورد استفاده برازش داده می‌شود. در واقع تخمین‌های مدل‌های مختلف به‌عنوان متغیر مستقل برای تخمین متغیر وابسته مورد نظر استفاده قرار می‌گیرند و ضریب رگرسیونی به‌دست آمده برای تخمین هر مدل، بیانگر وزن آن مدل در تخمین نهایی متغیر مورد نظر است. به‌عنوان نمونه و بر اساس روش ارائه شده توسط گرانجر و رامانتان (۱۹۸۴)، برای تخمین درصد رس از رابطه ۱۲ استفاده می‌شود. به این صورت که هر کدام از سه مدل مورد استفاده بر اساس متغیرهای محیطی، مقدار رس را برای یک مکان تخمین می‌زنند. این سه تخمین به عنوان متغیر مستقل به مدل رگرسیونی مبتنی بر OLS وارد می‌شوند و مقدار رس به عنوان متغیر وابسته در مکان مورد نظر تخمین زده می‌شود.

$$Clay = \sum_{i=1}^{p} (\alpha_i.Clay_i) + \beta$$ رابطه ۱۲)

در این رابطه α_i وزن اختصاص داده‌شده به مدل iام، $Clay_i$ تخمین به‌دست آمده از مدل iام برای رس و β مقداری ثابت است (عرض از مبدا).

---

[1]. Hyperplane

[2]. Maximum margin hyperplane (MMH)

[3]. Support vector regression

[4]. Ordinary least square

برای مدل‌سازی به روش‌های kNN، RF و SVM از بسته نرم‌افزاری caret (کوهن، ۲۰۰۸) در محیط R و برای مدل‌سازی به روش GR از بسته نرم‌افزاری GeomComb در محیط R استفاده گردید.

تجزیه و تحلیل‌های آماری و اعتبارسنجی مدل‌ها

برای بررسی کارآیی مدل‌های مختلف در تخمین اجزاء بافت خاک، از رویکرد ارزیابی متقابل k-fold [1] استفاده شد. در نهایت کارآیی مدل‌های مورد استفاده در تخمین اجزاء بافت خاک تبدیل‌شده و تبدیل‌نشده با استفاده از معیارهایی از جمله میانگین خطای مطلق (MAE[2])، ریشه میانگین مربعات خطا (RMSE[3]) و فاصله Aitchison (AD) بررسی گردید. فاصلهAitchison معیاری برای نشان دادن دقت تخمین‌ها در مدل‌سازی داده‌های مرکب مثل بافت خاک است. در واقع AD دقت تخمین سه جزء رس، سیلت و شن را به صورت یک شاخص نشان می‌دهد. در رابطه ۴، Z_k عناصر ترکیب z (داده مرکب واقعی) و Z_k^* عناصر ترکیب z* (داده مرکب تخمینی) هستند (ایتچیسون، ۱۹۹۲). محاسبه AD بر اساس رابطه ۱۳ و در محیط نرم‌افزار Excel انجام گرفت.

رابطه ۱۳)
$$AD\,(z,z^*)=\sqrt{\sum_{k=1}^{D}\left[\ln\frac{Z_k}{\left(\prod_{k=1}^{D}Z_k\right)^{1/D}}-\ln\frac{Z_k^*}{\left(\prod_{k=1}^{D}Z_k\right)^{1/D}}\right]^2}$$

نتایج

خلاصه آماری داده‌ها

خلاصه آماری رس، سیلت و شن در جدول ۲ ارائه شده است. مقدار شن از ۱۱/۵ تا ۷۷ درصد، مقدار رس از ۵/۳۷ تا ۴۵/۵ درصد و مقدار سیلت از ۵/۷۵ تا ۵۸/۲۵ درصد متغیر است. میانگین درصدهای رس، سیلت و شن به ترتیب برابر با ۲۵/۵۷، ۳۶/۴۲ و ۳۹/۹۹ درصد است. مقادیر انحراف معیار و ضریب تغییرات نشان می‌دهد شن دارای بیشترین تغییرپذیری در منطقه است. بر اساس طبقه‌بندی ارائه شده توسط ویلدینگ (۱۹۸۵)، ضریب تغییرات برای شن در حد متوسط است (بین ۱۵ تا ۳۵ درصد). تغییرات بافت خاک در منطقه بیشتر تحت تاثیر شیب، فاصله از ارتفاعات مشرف به منطقه و کاربری اراضی می‌باشد. در دامنه کوه‌های منطقه که شیب هم بیشتر است، جزء غالب شن و با دور شدن از ارتفاعات و کم شدن شیب جزء غالب سیلت و رس است. در مناطق تحت کشاورزی، شخم خوردن خاک و مخلوط شدن خاک زیرین با خاک سطحی می‌تواند عاملی در تفاوت خاک این مناطق با مناطق بدون کشت منطقه باشد.

جدول ۲- خلاصه آماری درصدهای رس، سیلت و شن

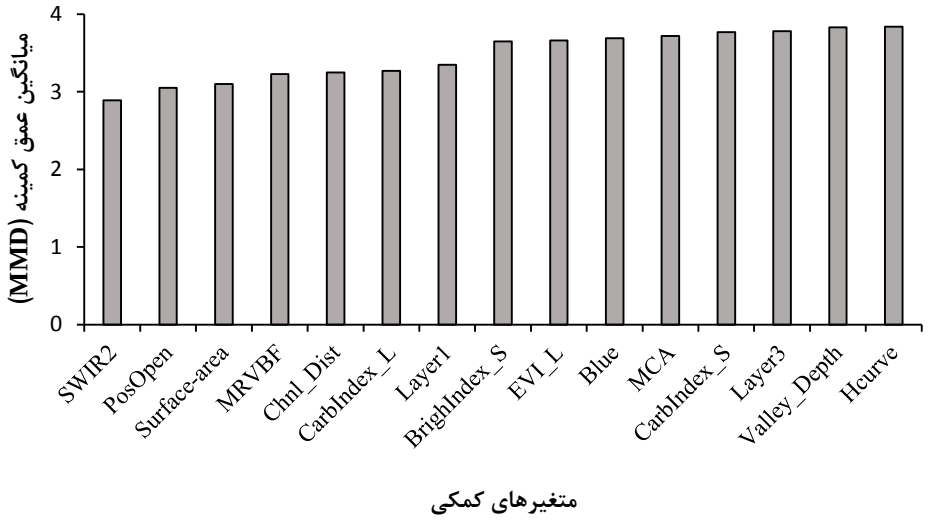
ویژگی	حداقل (٪)	حداکثر (٪)	میانگین (٪)	انحراف معیار (٪)	ضریب تغییرات (٪)
رس	۵/۳۷	۴۵/۵	۲۵/۵۷	۵/۶۱	۲۱/۹۳
سیلت	۵/۷۵	۵۸/۲۵	۳۶/۴۲	۷/۸۴	۲۱/۵۲
شن	۱۱/۵	۷۷	۳۷/۹۹	۱۱/۵۴	۲۹/۷۵

اهمیت نسبی متغیرهای کمکی

[1]. K-Fold cross-validation
[2]. Mean absolute error
[3]. Root mean square error

یکی از مزایای استفاده از مدل‌های درختی مثل جنگل تصادفی، مشخص کردن اهمیت نسبی متغیرهای کمکی است. در این مدل‌ها، متغیرهای مهم در تعداد بیشتری از درختان در گره ریشه قرار می‌گیرند و هر چه متغیر در گره‌های پایین‌تر از گره ریشه قرار گیرد از اهمیت کمتری برخوردار است. یکی از روش‌های با کارآیی بالا، انتخاب مهم‌ترین متغیرها بر اساس میانگین عمق کمینه[1] (MMD) آن‌ها است. MMD عبارت است از تعداد گره‌هایی که در امتداد کوتاه‌ترین مسیر از گره ریشه تا نزدیک‌ترین گره برگ قرار دارد. از این معیار در بررسی اهمیت متغیرهای کمکی در تخمین بافت خاک در شمال ایران استفاده شده است (ملاح و همکاران، ۲۰۲۲).

به طور کلی تاثیر پارامترهای استخراج شده از DEM و تصاویر ماهواره‌ای روی اجزاء بافت متفاوت بود و برخی از آنها همبستگی بیشتری با یک جزء بافت و همبستگی کمتری با سایر جزءها داشتند. در (شکل ۴) اهمیت متغیرهای محیطی مهم در مدل‌سازی بافت خاک بر اساس MMD نشان داده شده است. باند ۷ لندست ۸ (SWIR2) با کمترین عمق (۲/۸۹) دارای بیشترین اهمیت و انحنای افقی (HCurve) با بیشترین عمق (۳/۸۴) دارای کمترین اهمیت است؛ به عبارت دیگر متغیرهایی که دارای MMD کمتری هستند در تعداد بیشتری از درختان موجود در جنگل تصادفی در گره ریشه و یا گره‌های نزدیک ریشه قرار گرفته‌اند و از اهمیت بیشتری برخوردار هستند. به طور کلی، نتایج نشان داد (شکل ۴) متغیرهای به دست آمده از DEM اهمیت بیشتری در مدل‌سازی خاک داشتند (بجز باند ۷ لندست ۸).



شکل ۴- اهمیت نسبی مهم‌ترین متغیرهای کمکی مورد استفاده در مدل‌سازی بافت خاک

SWIR: short wave infrared; PosOpen :positive openness; MRVBF :multi-resolution valley bottom flatness; Chnl_Dist: channel distance; CarbIndex_L: carbonate index (Landsat 8); BrighIndex_S: brightness index (Sentinel 2); EVI_L: enhanced vegetation index (Landsat 8); MCA: modified catchment area; Hcurve: horizontal curvature

مدل‌سازی بافت خاک

کارآیی مدل‌های RF، SVM، kNN و GR در تخمین داده‌های بافت تبدیل شده و تبدیل‌نشده در (جدول ۳) ارائه شده است. کمترین و بیشترین مقادیر RMSE و MAE برای هر چهار مدل در تخمین داده‌های تبدیل‌شده و تبدیل‌نشده به ترتیب مربوط به رس و شن بود. بررسی کارآیی مدل RF نشان داد مقدار RMSE داده تبدیل‌نشده نسبت به داده تبدیل‌شده به روش alr برای اجزای رس، سیلت و شن به ترتیب ۱۹/۸۰، ۳۵/۷۶، ۱۹/۸۷ درصد افزایش یافت (جدول ۳). این مقادیر برای تبدیل clr به

[1]. Mean minimal depth

ترتیب ۲۰/۱۳، ۱۳/۹۷، ۱۷/۲۹ درصد کاهش نشان داد و با تبدیل ilr برای اجزای رس و شن به ترتیب ۲۸/۵۰ و ۲/۴۵ درصد کاهش و برای جز سیلت ۵/۹۷ درصد افزایش نشان داد. به طور کلی و بر اساس معیار RMSE (جدول ۳)، در مقایسه داده تبدیل‌نشده، تبدیل alr باعث بهبود تخمین‌ها گردید، در حالی که تبدیل clr باعث بهبود تخمین‌ها نگردید. تبدیل ilr باعث بهبود تخمین سیلت گردید ولی برای اجزا رس و شن منجر به دقت بیشتر تخمین‌ها نگردید و نسبت به داده‌های تبدیل‌نشده دقت کاهش یافت. به‌طور کلی و بر اساس معیار MAE، نتایج مدل‌سازی بافت خاک با استفاده از RF نشان داد کارآیی استفاده از داده تبدیل‌نشده نسبت به داده تبدیل‌شده به روش alr کمتر و نسبت به تبدیل‌های clr و ilr بیشتر بود (جدول ۳). این روند کلی برای سایر مدل‌های مورد استفاده هم صادق بود. برای هر دو مجموعه داده تبدیل‌شده و تبدیل‌نشده، بیشترین مقدار MAE مربوط به شن (۸/۱۳ درصد) و کمترین مقدار مربوط به رس (۳/۱۱ درصد) بود. مقدار AD برای مدل RF وقتی از داده‌های تبدیل‌نشده استفاده شد نسبت به زمانی که از داده‌های تبدیل‌شده به روش‌های alr، clr و ilr استفاده شد به ترتیب ۴/۰۱، ۲۷/۱۴، ۳۶/۴ درصد کاهش یافت (جدول ۳). همچنین مقدار AD برای هر چهار مدل مورد استفاده با داده‌های تبدیل‌نشده کمتر از داده‌های تبدیل شده بود.

نتایج کارآیی مدل SVM در مدل‌سازی بافت خاک در جدول ۳ ارائه شده است. نتایج نشان داد که مقدار RMSE داده تبدیل‌نشده نسبت به داده تبدیل‌شده به روش alr برای اجزای رس، سیلت و شن به ترتیب ۱۹/۵۷، ۵۷/۳۱ و ۱۶/۰۹ درصد افزایش داشت. این مقادیر برای تبدیل clr، به ترتیب ۲/۶۲، ۱۵/۲۲ و ۲۲/۱۹ درصد کاهش نشان داد. استفاده از داده تبدیل‌شده به روش ilr باعث افزایش ۱۲/۸۸ درصدی در مقدار RMSE در تخمین رس و کاهش ۲۱/۱۱ و ۹/۷۲ درصدی مقدار RMSE برای جز سیلت و شن گردید. نتایج نشان داد (جدول ۳) تبدیل داده‌ها به سه روش مورد استفاده باعث افزایش AD گردید. این مقدار افزایش برای سه تبدیل alr، clr و ilr به ترتیب برابر با ۴/۵۴، ۳۰/۱۹ و ۱۹/۴۸ درصد بود. به طور کلی و بر اساس همه معیارهای خطا، مدل SVM با استفاده از داده‌های تبدیل‌شده به روش alr منجر به بهبود تخمین‌ها گردید ولی برای داده‌های تبدیل‌شده به روش clr کارآیی کمتری نشان داد. کارآیی مدل SVM در استفاده از داده رس تبدیل‌شده به روش ilr کمتر از داده تبدیل‌نشده بود ولی در مدل‌سازی سیلت و شن تبدیل‌شده بیشتر بود.

جدول ۳- کارآیی مدل‌های RF، SVM و kNN و مدل ترکیبی GR در مدل‌سازی داده‌های تبدیل‌نشده (UT) و داده‌های تبدیل‌شده به روش‌های alr، clr و ilr

مدل	نوع داده	RMSE (%)			MAE (%)			AD
		رس	سیلت	شن	رس	سیلت	شن	
RF	UT	۴/۸۴	۷/۶۳	۹/۹۵	۴/۰۷	۵/۹۵	۷/۸۹	۰/۲۶۳
	alr	۴/۰۴	۵/۶۲	۸/۲۰	۲/۱۱	۴/۵۵	۶/۷۰	۰/۲۷۴
	clr	۶/۰۶	۸/۸۷	۱۲/۰۳	۴/۲۳	۰/۰۶	۸/۱۳	۰/۳۶۱
	ilr	۶/۷۷	۷/۲۰	۱۰/۲۰	۴/۸۱	۵/۵۷	۸/۰۶	۰/۳۶۴
SVM	UT	۵/۰۷	۷/۶۳	۸/۸۰	۴/۲۹	۵/۲۸	۶/۹۷	۰/۲۵۲
	alr	۴/۲۴	۴/۸۵	۷/۵۸	۳/۴۳	۳/۷۶	۶/۲۹	۰/۲۶۴
	clr	۵/۷۲	۹/۰۰	۱۱/۳۱	۴/۰۵	۶/۱۰	۷/۸۳	۰/۳۶۱
	ilr	۵/۸۲	۶/۳۰	۸/۰۲	۴/۱۴	۴/۷۹	۶/۳۶	۰/۳۱۳
kNN	UT	۴/۸۶	۶/۸۴	۹/۵۵	۴/۰۸	۵/۵۰	۷/۴۰	۰/۲۵۳
	alr	۳/۹۵	۵/۴۶	۷/۶۳	۳/۱۶	۴/۲۰	۵/۸۷	۰/۲۵۴
	clr	۶/۱۱	۹/۴۸	۱۲/۲۹	۴/۳۵	۶/۳۰	۸/۰۵	۰/۳۷۰

	ilr	۶/۴۴	۶/۷۳	۹/۶۵	۴/۵۴	۵/۵۴	۷/۷۲	۰/۳۵۲
GR	UT	۵/۰۷	۷/۱۱	۹/۲۰	۴,۲۸	۵/۳۶	۷/۷۳	۰/۲۵۶
	alr	۴/۲۱	۵/۱۵	۷/۶۷	۳/۳۲	۴/۱۰	۶/۲۰	۰/۲۶۶
	clr	۵/۸۱	۹/۰۴	۱۱/۶۹	۴/۱۶	۶/۰۰	۷/۵۷	۰/۲۵۳
	ilr	۶/۰۹	۶/۷۰	۸/۷۴	۴/۲۴	۵/۲۳	۶/۸۵	۰/۲۲۵

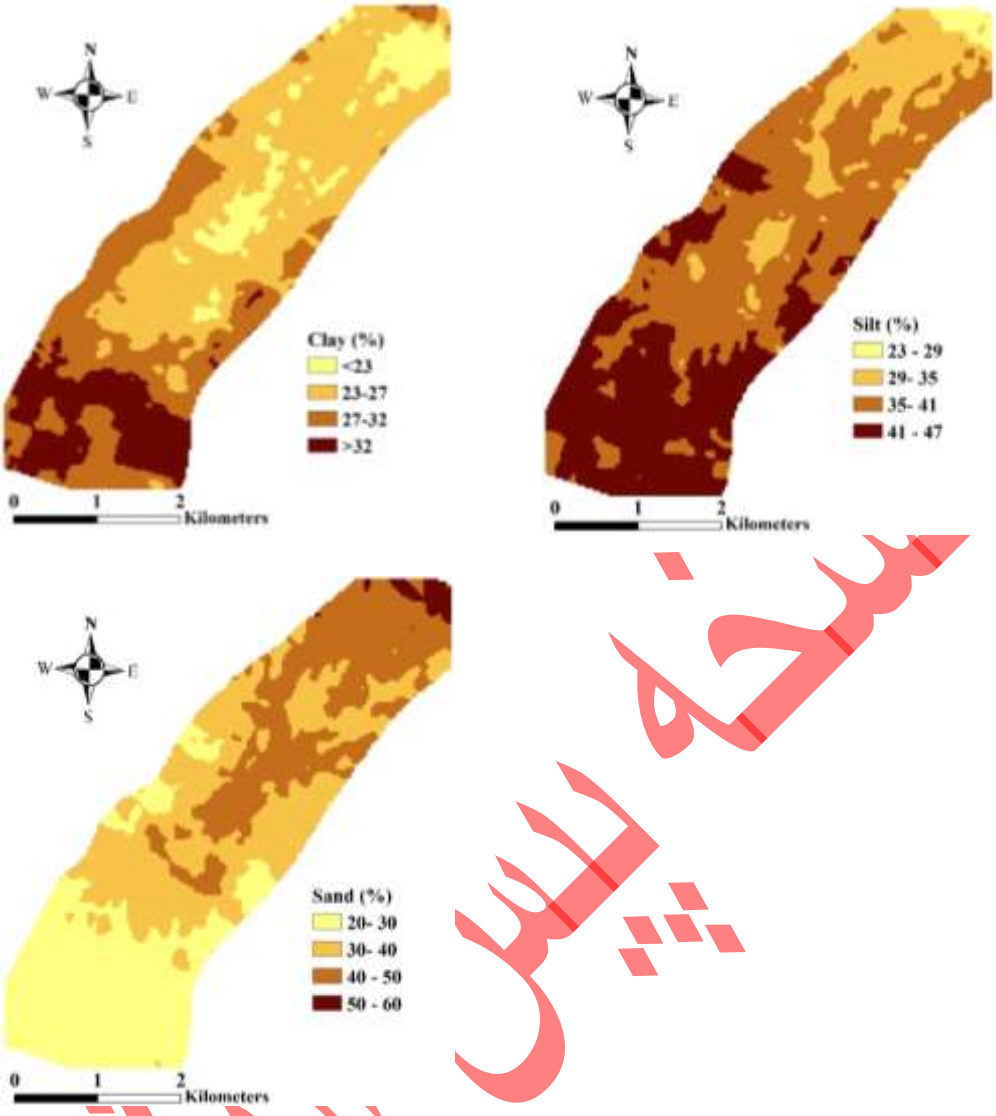
AD: فاصله Aitchison، RMSE: میانگین ریشه مربعات خطا؛ MAE: میانگین خطای مطلق؛alr، clr و ilr به ترتیب تبدیل‌هایadditive log-ratio ، centred-log-ratioو isometric log-ratio؛ RF، SVM و kNN به ترتیب مدل‌های Random Forest، Support Vector Machine وneighbor K- nearest؛ GR: مدل ترکیبی Granger-Ramanathan

مدل‌سازی داده‌های تبدیل‌شده و تبدیل‌نشده با استفاده از مدل kNN نشان داد تبدیل داده‌های رس، سیلت و شن به روش alr به ترتیب باعث کاهش ۲۳/۰۳، ۲۵/۲۷ و ۲۴/۹۰ درصد در RMSE گردید (جدول ۳). این مقادیر برای تبدیل clr روند افزایشی داشت به طوری که تبدیل داده‌های رس، سیلت و شن به ترتیب باعث افزایش ۲۰/۴۵، ۲۷/۸۴ و ۲۲/۴۵ درصد در مقدار RMSE مدل kNN گردید. نتایج استفاده از داده‌های تبدیل‌شده به روش ilr تا حدودی متفاوت بود، به طوری که با تبدیل داده‌های رس و شن مقدار RMSE به ترتیب ۲۴/۵ و ۱/۲۴ درصد افزایش ولی برای جز سیلت کاهش یافت (۱/۶۳ درصد). مقدار AD به دست آمده از مدل‌سازی با kNN بیانگر بالاتر بودن کارآیی این مدل در مدل‌سازی داده‌های تبدیل‌شده نسبت به داده‌های تبدیل‌نشده بود (جدول ۳). این نتیجه برای هر سه تبدیل مورد استفاده صادق بود. به طور کلی نتایج مدل‌سازی با kNN نشان داد مدل‌سازی اجزاء بافت خاک با استفاده از داده‌های تبدیل‌شده به روش alr باعث بهبود تخمین‌ها گردید ولی تبدیل clr منجر به نتایج بهتری نسبت به داده‌های تبدیل‌نشده نگردید. کارآیی مدل kNN در استفاده از داده تبدیل‌شده سیلت به روش ilr بیشتر از داده تبدیل‌نشده بود ولی برای دو جزء رس و شن نتیجه برعکس بود.

نتایج مدل‌سازی داده‌های تبدیل شده و تبدیل نشده بافت خاک با استفاده از مدل GR در (جدول ۳) ارائه شده است. در استفاده از داده‌های تبدیل‌شده به روش alr، مقدار RMSE برای هر سه جزء رس، سیلت و شن نسبت به داده تبدیل‌نشده روند کاهشی نشان داد (به ترتیب ۲۰/۴۲، ۳۸/۰۵ و ۱۹/۹۴ درصد). این روند برای تبدیل clr برعکس بود و باعث افزایش RMSE برای سه جزء رس، سیلت و شن گردید (به ترتیب ۱۲/۷، ۲۱/۳۴ و ۲۱/۳۰ درصد). تبدیل ilr باعث بهبود تخمین‌های دو جزء سیلت و شن گردید (مقدار RMSE به ترتیب ۶/۱۱ و ۵/۲۶ درصد کاهش نشان داد) ولی باعث کاهش کارآیی مدل GR در تخمین رس گردید (RMSE به میزان ۱۶/۷۴ درصد افزایش نشان داد). نتایج کلی نشان داد کارآیی مدلGR با داده‌های تبدیل‌شده به روش alr بیشتر از داده‌های تبدیل‌نشده بود ولی برای تبدیل clr نتایج برعکس بود. برای تبدیل ilr کارآیی مدل GR در مدل‌سازی داده‌های تبدیل‌نشده رس بیشتر ولی برای دو جزء تبدیل‌نشده سیلت و شن کمتر بود.

پهنه‌بندی اجزاء بافت خاک

با توجه به این که اهمیت نسبی متغیرها به روش RF به دست آمد و اختلاف زیادی بین کارآیی این مدل با مدل SVM که بیشترین کارآیی را داشت وجود نداشت، نقشه توزیع مکانی اجزاء بافت با استفاده از مدل RF به دست آمد (شکل ۵). بیشترین مقادیر رس و سیلت در قسمت‌های جنوبی‌تر منطقه و کمترین مقادیر آن‌ها در قسمت‌های مرکزی و شمال‌شرقی منطقه دیده می‌شود. توزیع مکانی شن در منطقه روندی نسبتا معکوس با رس و سیلت داشت و بیشترین مقادیر آن در بخش‌های شمال‌شرقی منطقه مشاهده گردید. این توزیع مکانی تا حدود زیادی متاثر از ارتفاع و شیب منطقه است. در قسمتهای شمالی و شمال‌شرقی شیب و ارتفاع بیشتر (بیشترین ارتفاع ۱۳۸۲ متر) و به تدریج به سمت جنوب منطقه از مقدار آن کاسته می‌شود (کمترین ارتفاع ۱۱۶۷ متر).



شکل ۵- توزیع مکانی ذرات رس، سیلت و شن به دست آمده از مدل RF

بحث

متغیرهای کمکی

داده‌های سنجش از دور بیشتر بیانگر عوارض سطحی و عامل خاک‌سازی موجودات زنده هستند. هر عارضه‌ای در سطح زمین می‌تواند بازتاب طیفی خاص خود را داشته باشد که این بازتاب توسط سنجنده‌ها ثبت می‌شود. خاک‌های مختلف معمولا بازتاب‌های طیفی متفاوتی دارند که با خواص آن‌ها در ارتباط هستند. پوشش گیاهی و کاربری اراضی از جمله عوارض سطحی هستند که هر دو روی بافت خاک تاثیرگذار هستند. همچنین رطوبت خاک که تا حد زیادی تابع بافت خاک است از عوامل مهم تاثیرگذار در بازتاب طیفی است. در تحقیق حاضر SWIR2 مهمترین متغیر کمکی در توصیف تغییرات اجزاء بافت خاک بود (شکل ۴). بازتابهای طیفی و ترکیب باندها بیانگر عوارض سطح زمین هستند. برخی از این عوارض به طور مستقیم و غیرمستقیم با بافت خاک در ارتباط هستند. طیف‌های مادون قرمز و مادون قرمز نزدیک هم به ترکیبات آلی و هم ترکیبات معدنی خاک حساس هستند و در نتیجه برای مطالعات خواص خاک در کشاورزی و علوم محیطی مفید هستند (ژیومی و جیانشی، ۲۰۱۳). طیف سنجی مریی و مادون قرمز نزدیک ساختار و ترکیب مواد و مولکول‌ها را در طول موج ۴۰۰ تا ۲۵۰۰ را انعکاس می‌دهد؛ بنابراین با توجه به تفاوت ترکیب و ساختار رس، سیلت و شن و عوامل مرتبط با بافت خاک مثل مقدار ماده آلی و درصد

رطوبت، می‌توان از این نوع طیف‌سنجی در نشان دادن توزیع مکانی اجزاء بافت خاک استفاده کرد. ویسر و همکاران (۲۰۰۷) از تکنیک‌های طیف‌سنجی برای تعیین مقدار رس خاک استفاده کردند.

از پنج متغیر کمکی مهم دربرآورد اجزاء بافت خاک، چهار متغیر بیانگر توپوگرافی منطقه بود. این در حالی است که از ۱۵ متغیر مهم در مدل‌سازی بافت خاک، کمترین اهمیت مربوط به انحنای افقی یا مماسی (Hcurve) بود که یک پارامتر توپوگرافی است (شکل ۴). این نتایج نشان می‌دهند اهمیت متغیرهای استخراج شده از DEM در برآورد اجزاء بافت خاک می‌تواند بسیار متفاوت باشد. انحنای افقی در واقع نیمرخ شیب در جهت عمود بر شیب کلی یک منطقه است. در واقع این نوع انحنا میزان محدب یا مقعر بودن سطح زمین در امتداد عمود بر شیب کلی را نشان می‌دهد. در مناطقی که شیب کلی وجود دارد ولی شیب جانبی کم است و یا وجود ندارد این انحنا هم کم است و یا وجود ندارد. در نتیجه تغییرات Hcurve در منطقه کم و عامل مهمی در کنترل تغییرات مکانی خواص خاک نیست. در منطقه مورد بررسی شیب ملایم کلی از دامنه ارتفاعات تا مرکز دشت وجود دارد ولی شیب جانبی کم و در برخی جاها وجود ندارد. پارامترهای استخراج شده از DEM نماینده توپوگرافی، یکی از مهم‌ترین فاکتورهای خاک‌سازی هستند (گالانت و اوستین، ۲۰۱۵). این پارامترها با اثر مستقیم و غیرمستقیم روی حرکت آب، توزیع آب و فرایندهای فرسایش و رسوب‌گذاری می‌توانند در بافت خاک موثر باشند. به عنوان نمونه، MRVBF شاخصی از مناطق رسوب‌گذاری است (گالانت و داولینگ، ۲۰۰۳) و مقادیر بالای آن بیانگر مکان‌های رسوب‌گذاری است که معمولا ذرات ریزتر تجمع می‌یابند. همچنین بین شیب و مقدار ذارت ریز و درشت به ترتیب رابطه منفی و مثبت وجود دارد (یومالی و همکاران، ۲۰۱۲).

نتایج بررسی انجام شده در دانمارک نشان داد ارتفاع با کنترل حرکت آب و رسوب بیشترین تاثیر را روی توزیع رس، شن ریز و شن درشت و درجه شیب با تاثیر روی سرعت جریان سطحی و زیرسطحی بیشترین تاثیر را روی توزیع سیلت داشت (گریو و همکاران، ۲۰۱۲). اکپا و همکاران (۲۰۱۴) در مدل‌سازی بافت خاک‌های کشور نیجریه از مدل جنگل تصادفی و متغیرهای کمکی از جمله باندهای طیفی لندست، پارامترهای استخراج شده از DEM، میانگین دما و بارندگی، کاربری اراضی و نوع خاک استفاده کردند. نتایج آن‌ها نشان داد اهمیت نسبی متغیرهای مورد استفاده بسته به عمق خاک و جزء بافت متفاوت بود ولی همانند تحقیق حاضر، پارامترهای استخراج شده از DEM جزء متغیرهای مهم در مدل‌سازی بافت خاک بودند؛ زیرا این پارامترها هم روی حرکت عمودی و هم حرکت جانبی ذرات و آب تاثیر دارند. چاگاس و همکاران (۲۰۱۶) در مدل‌سازی بافت خاک با استفاده از مدل RF در یک منطقه با پستی و بلندی کم نشان دادند بجز باند یک و نسبت باند پنج به هفت لندست ۵، سایر باندهای طیفی و نسبت‌های آن‌ها همبستگی معنی‌داری با اجزاء بافت خاک داشتند. بر خلاف نتایج تحقیق حاضر، این نتایج بیانگر اهمیت بالای داده‌های طیفی در مدل‌سازی بافت خاک بود. به طور کلی در مناطق نسبتا مسطح و با پستی و بلندی کم، پارامترهای استخراج شده از DEM اهمیت کمتری در توزیع مکانی خاک‌ها دارند و در این مناطق معمولا داده‌های سنجش از دور دارای اهمیت بیشتری هستند. در بررسی انجام شده در خاک‌های واقع در دشت سیستان، پهلوان‌راد و اکبری‌مقدم (۲۰۱۸) از مدل جنگل تصادفی و پارامترهای استخراج شده از DEM و تصاویر ماهواره‌ای برای مدل‌سازی بافت خاک استفاده کردند. مطابق با نتایج تحقیق حاضر، نتایج آن‌ها نشان داد برای سه جزء رس، سیلت و شن موثرترین متغیرهای کمکی پارامترهای استخراج شده از DEM بودند. همچنین مهربانی قربانی و همکاران (۲۰۱۹) در مدل‌سازی بافت خاک در منطقه زرند کرمان به این نتیجه رسیدند که داده‌های طیفی، MRVBF و شاخص خیسی (WI) قابلیت بالایی در تخمین بافت خاک داشتند. در تحقیق انجام شده توسط امیریان چکان و همکاران (۲۰۱۹)، نشان داده شد MRVBF، NDVI، شیب و باند ۳ تصویر لندست ۸ بیشترین اهمیت را در تخمین اجزاء بافت خاک داشتند. همچنین نتایج آن‌ها نشان داد به طور کلی اهمیت متغیرهای استخراج شده از تصویر ماهواره‌ای در مدل‌سازی بافت خاک کمتر از متغیرهای استخراج شده از DEM بود که با نتایج تحقیق حاضر

هم‌خوانی دارد. در بررسی انجام شده در شمال غرب ترکیه سه مدل یادگیری ماشین شامل درخت تصمیم، RF و SVM برای مدل‌سازی بافت خاک با استفاده از داده‌های به دست آمده از DEM و سنجش از دور استفاده گردید (کایا و همکاران، ۲۰۲۲). نتایج نشان داد مهم‌ترین متغیر پیش‌بینی کننده شاخص خیسی به دست آمده از DEM و پس از آن داده‌های سنجش از دور بود. این نتایج همانند نتایج حاصل تحقیق حاضر، اهمیت بالای پارامترهای استخراج شده از DEM را نشان می‌دهد.

کارآیی مدل‌های مورد استفاده

به طور کلی و بر اساس مقادیرRMSE و MAE، کارآیی هر سه مدل RF، SVM و kNN در تخمین داده‌های تبدیل‌نشده از داده‌های تبدیل شده به روش‌های clr و ilr بیشتر و از روش alr کمتر بود؛ در صورتی که کارایی مدل حاصل از ترکیب آن‌ها برای تبدیل‌های alr و ilr بیشتر از داده‌های تبدیل‌نشده بود (جدول ۳). نتایج کلی بیانگر کارآیی بالاتر مدل‌های مورد استفاده در مدل‌سازی داده‌های تبدیل‌شده به روش alr بود.

باتوجه به نتایج به دست آمده (جدول ۳) کمترین مقدار RMSE داده‌های تبدیل شده برای سیلت و شن مربوط به مدل SVM بود که نشان‌دهنده کارایی بالای این مدل در تخمین اجزا سیلت و شن بود. در تخمین رس نیز کمترین RMSE برای داده تبدیل‌شده به روش clr وilr مربوط به مدل SVM بود. به‌طور کلی مدل GR نسبت به سه مدل دیگر منجر به تخمین‌های بهتری از اجزاء بافت خاک نگردید (جدول ۳)؛ بنابراین و بر خلاف انتظار، مدل‌های ترکیبی ممکن است همیشه منجر به نتایج بهتری نشوند. در بررسی انجام شده در جنوب‌غربی چین توسط ویو و همکاران (۲۰۱۸)، از سه روش یادگیری ماشین شامل SVM، شبکه‌های عصبی مصنوعی و درخت طبقه‌بندی برای مدل‌سازی کلاس‌های بافت خاک استفاده شد. به‌طور کلی نتایج آنها نشان داد مدل SVM نسبت به دو مدل دیگر دارای کارآیی بهتری در تخمین اجزاء بافت خاک بود.

الگوریتم به کار رفته در SVM با تلفیق جنبه‌هایی از روش‌های رگرسیونی و روش‌های k نزدیکترین همسایه توانایی مدل‌سازی روابط پیچیده را دارد و در نتیجه تشکیل یک مدل یادگیری ماشین قوی را می‌دهد (لانتز، ۲۰۱۵). در نقشه‌برداری رقومی خاک مدل‌سازی خاک بر اساس ارتباط متغیرهای کمکی با ویژگی‌های خاک است. وقتی داده‌های بافت خاک به روش‌های log-ratio تبدیل می‌شوند متغیرهای جدیدی به وجود می‌آیند و مدل‌سازی با استفاده از این متغیرهای جدید انجام می‌شود. همبستگی این متغیرهای جدید با متغیرهای محیطی نسبت به داده های اولیه (تبدیل نشده) ممکن است بیشتر یا کمتر باشد. به طور کلی نتایج این پژوهش بیانگر بهبود تخمین‌ها با استفاده از تبدیل alr بود؛ بنابراین دلیل این بهبود را می‌توان ایجاد متغیرهای جدید با استفاده از تبدیل alr دانست که همبستگی بیشتری با متغیرهای کمکی داشتند.

به‌نظر می رسد اولین تحقیق انجام شده که در آن از تبدیل‌های log-ratio برای تضمین برابر ۱۰۰ درصد شدن تخمین‌های رس، سیلت و شن استفاده گردید، مربوط به اوده و همکاران (۲۰۰۳) است. آن‌ها از تبدیل alr قبل از مدل‌سازی به روش کریجینگ استفاده کردند و بر اساس مقادیر RMSE و میانگین خطا (ME)، نشان دادند تبدیل داده‌ها باعث بهبود کارآیی مدل در تخمین اجزاء بافت خاک گردید که با نتایج تحقیق حاضر هم‌خوانی دارد. در بررسی انجام شده توسط وانگ و شی (۲۰۱۷) از چهار روش تبدیل داده شامل alr، clr، alr و slr[1] برای مدل‌سازی بافت خاک استفاده گردید. نتایج نشان داد از چهار تبدیل مورد استفاده، روش تبدیل clr منجر به کارآیی بالاتری گردید که با نتایج تحقیق حاضر هم‌خوانی ندارد. در تحقیق انجام شده توسط امیریان چکان و همکاران (۲۰۱۹) کارآیی مدل RF در مدل‌سازی بافت خاک با استفاده از داده‌های تبدیل‌شده به سه روش alr، clr و ilr و داده‌های تبدیل‌نشده بررسی گردید. بر اساس RMSE، نتایج آن‌ها نشان داد که هر چند مدل RF کارآیی نسبتا مشابهی در تخمین داده‌های تبدیل‌شده و تبدیل‌نشده داشت، ولی تخمین‌های به دست آمده از داده‌های تبدیل‌شده تا حدودی اریب بود. همچنین نتایج آن‌ها نشان داد هر چند که مقادیر RMSE برای هر سه تبدیل تقریبا مشابه بود، ولی استفاده

[1]. Symmetry log-ratio

از داده‌های تبدیل شده به روش clr منجر به نتایج بهتری گردید که با نتایج تحقیق حاضر همخوانی ندارد. ژانگ و همکاران (۲۰۲۰) از پنج مدل یادگیری ماشین و سه روش تبدیل alr، clr و ilr برای مدل‌سازی بافت خاک در چین استفاده کردند. نتایج آن‌ها بیانگر کارآیی بالاتر مدل RF و استفاده از داده‌های تبدیل‌شده به روش ilr بود که از نظر بهترین مدل و بهترین تبدیل با نتایج تحقیق حاضر همخوانی ندارد. وانگ و همکاران (۲۰۲۱) از دو روش تبدیل alr و ilr و دو مدل یادگیری ماشین شامل RF و Cubist برای مدل‌سازی بافت خاک استفاده کردند. برخلاف تحقیق حاضر که تبدیل alr منجر به بهترین تخمین‌ها شد، نتایج بررسی آن‌ها بیانگر کارآیی بالای مدل RF و تبدیل ilr بود.

مطالعات انجام شده در زمینه استفاده از مدل‌های ترکیبی از جمله GR در مدل‌سازی خواص خاک نسبتا محدود و استفاده از این مدل‌ها در مدل‌سازی داده‌های تبدیل شده و تبدیل نشده بافت خاک گزارش نشده است. ملانو و همکاران (۲۰۱۴) از چهار مدل ترکیبی برای مدل‌سازی pH تا عمق دو متری خاک در کشور استرالیا استفاده کردند. نتایج آنها نشان داد مدل‌های ترکیبی باعث بهبود تخمین pH خاک در همه عمق‌های مورد بررسی گردید که با نتایج تحقیق حاضر هم‌خوانی ندارد. با توجه به کارآیی بالاتر روش GR، آن‌ها این روش را برای مدل‌سازی در مطالعات نقشه‌برداری رقومی خاک پیشنهاد دادند. رومن دوبارکو و همکاران (۲۰۱۷) در مدل‌سازی بافت خاک در کشور فرانسه از دو مدل ترکیبی GR و Bates-Granger استفاده کردند. آن‌ها نشان دادند هر دو مدل فقط باعث بهبود تخمین‌های رس گردیدند و مدل GR نسبت به مدل Bates-Granger منجر به RMSE کمتر و تخمین بهتر عدم قطعیت شد. در بررسی انجام شده توسط چن و همکاران (۲۰۲۰) پنج روش ترکیب مدل برای مدل‌سازی و تخمین ماده آلی در کشور فرانسه مورد بررسی قرار گرفت. نتایج نشان داد هر پنج روش ترکیب مدل‌ها باعث بهبود تخمین ماده آلی خاک گردید که با یافته‌های تحقیق حاضر همخوانی ندارد. عابدی و همکاران (۲۰۲۱) مدل‌سازی شوری خاک در منطقه داراب استان فارس را با استفاده از شش مدل یادگیری ماشین و مدل حاصل از ترکیب آن‌ها انجام دادند. به طور کلی، نتایج آن‌ها نشان داد مدل ترکیبی GR نسبت به شش مدل مورد استفاده باعث کاهش RMSE و افزایش R^2 گردید. این نتایج برخلاف نتایج تحقیق حاضر هستند که نشان داد مدل GR برتری خاصی نسبت به مدل‌های دیگر نداشت. سواین و همکاران (۲۰۲۱) از سه مدل رگرسیون حداقل مربعات جزیی، RF و SVM برای مدل‌سازی اجزاء بافت خاک با استفاده از داده‌های طیفی سنتینل دو و طیف‌سنجی در آزمایشگاه استفاده کردند. نتایج آن‌ها نشان داد ترکیب مدل‌ها به روش میانگین‌گیری ساده باعث بهبود زیادی در دقت تخمین‌های هر سه جزء رس، سیلت و شن گردید.

مقایسه نتایج به‌دست آمده از تحقیق حاضر با نتایج سایر مطالعات انجام‌شده نشان می‌دهد کارآیی مدل‌های مورد استفاده و روش‌های تبدیل داده در مطالعات مختلف تا حدودی متفاوت بود. این تفاوت‌ها می‌تواند مربوط به تفاوت در نوع داده‌ها، تعداد داده، نوع مدل، مقیاس مطالعه و شرایط منطقه مورد مطالعه باشد؛ بنابراین، با وجود این‌که به‌طور کلی در تحقیق حاضر مدل SVM و تبدیل alr منجر به نتایج بهتری گردید، ولی نمی‌توان این مدل و این روش تبدیل داده را برای مدل‌سازی بافت خاک در همه مناطق و شرایط پیشنهاد داد؛ به عبارت دیگر برای هر منطقه بهتر است نوع تبدیل و مدل مناسب را پیدا کرد. همچنین، با وجود این‌که در اکثر مطالعات انجام شده استفاده از ترکیب مدل‌ها نسبت به استفاده از مدل‌ها به صورت جداگانه منجر به بهبود تخمین‌ها گردید، ولی در مطالعه حاضر استفاده از مدل ترکیبی GR منجر به بهبود قابل توجهی در دقت تخمین‌ها نگردید. مدل‌های مختلف معمولا دارای ساختارها و مکانیسم‌های متفاوتی هستند (تجربی، آماری، ریاضیاتی، شبیه‌سازی، ...) و هر کدام ممکن است فرضیاتی برای ساده‌سازی داشته باشند. همچنین برخی مدل‌ها متغیر هدف را بیش و برخی کم برآورد می‌کنند. علاوه بر این، تخمین‌های مدل‌های مختلف ممکن است اختلاف زیادی با هم داشته باشند؛ بنابراین، در برخی شرایط ممکن است ترکیب تخمین‌های مدل‌هایی مختلف بر اساس رویکر مورد استفاده در مدل GR باعث بهبود کارآیی مدل‌ها نشود.

توزیع مکانی اجزاء بافت خاک

نقشه‌های پهنه‌بندی اجزاء بافت خاک نشان داد توزیع ذرات رس، سیلت و شن در قسمت‌های مختلف منطقه متفاوت است (شکل ۵). شیب روی میزان و قدرت رواناب و در نتیجه توانایی حمل ذرات توسط آن تاثیر دارد (یومالی و همکاران، ۲۰۱۲). ذرات ریزتر رس و سیلت به دلیل سبک‌تر بودن راحت‌تر توسط جریان آب از شیب‌های بیشتر به سمت شیب‌های کمتر منتقل می‌شوند ولی ذرات شن به علت درشت بودن و وزن بیشتر کمتر توسط آب‌های سطحی منتقل می‌شوند. در بررسی مطالعات انجام شده روی مدل‌سازی ویژگی‌های خاک با استفاده از روش‌های نقشه‌برداری رقومی، چن و همکاران (۲۰۲۲) بیان کردند در مقیاس گسترده، بعد از مواد مادری مهم‌ترین عامل موثر در توزیع مکانی اجزاء بافت خاک ارتفاع بود. اهمیت ارتفاع در توزیع مکانی اجزاء بافت خاک در مطالعات دیگری هم گزارش شده است (گریو و همکاران، ۲۰۱۲؛ ویو و همکاران، ۲۰۱۸).

نتیجه‌گیری

در این مطالعه کارایی سه مدل kNN، RF و SVM و مدل به‌دست آمده از ترکیب آن‌ها (مدل GR) برای مدل‌سازی بافت خاک مورد بررسی قرار گرفت. در این مطالعه برای اولین بار از یک مدل ترکیبی برای مدل‌سازی داده‌های تبدیل‌شده (تبدیلات log-ratio) و تبدیل‌نشده بافت خاک استفاده گردید و نتایج با تخمین‌های به دست آمده از سه روش رایج یادگیری ماشین مقایسه گردید.

هرچند که کارآیی مدل‌های مختلف برای اجزاء مختلف و تبدیل‌های مختلف متفاوت بود، ولی به‌طور کلی کارآیی هر سه مدل در مدل‌سازی داده‌های تبدیل‌شده به روش alr بیشتر از کارآیی آن‌ها در مدل‌سازی داده‌های تبدیل‌نشده و داده‌های تبدیل‌شده به روش clr و ilr بود. همچنین مدل GR در مدل‌سازی داده‌های تبدیل‌شده به روش alr و ilr کارآیی بیشتری نسبت به داده‌های تبدیل‌نشده نشان داد. مقایسه نتایج این پژوهش با مطالعات انجام شده نشان داد در شرایط و مناطق مختلف و برای اجزاء مختلف بافت خاک، بهترین مدل و بهترین روش تبدیل متفاوت بود و نمی‌توان یک مدل و یک روش تبدیل را برای مدل‌سازی بافت خاک در همه شرایط پیشنهاد داد و بهتر است مدل و روش تبدیل مناسب با توجه به شرایط و جزء مورد نظر (رس، سیلت یا شن) تعیین شود.

برخلاف انتظار، ترکیب تخمین سه مدل به روش GR منجر به تخمین‌های بهتری از اجزاء بافت خاک نگردید. این نتیجه از این نظر اهمیت دارد که برای مدل‌سازی بافت خاک شاید بتوان به جای ترکیب چند مدل که ممکن است باعث پیچیدگی بیشتر مدل‌سازی شود و همچنین منجر به صرف زمان بیشتری برای مدل‌سازی شود، از یک مدل استفاده کرد.

با توجه به این که نتایج متفاوتی برای مدل‌ها و تبدیل‌های مختلف به دست آمد، می‌توان نتیجه گرفت که با افزایش تعداد مدل‌های مورد استفاده و استفاده از سایر روش‌های ترکیب تخمین مدل‌ها، شاید بتوان نتایج بهتری کسب کرد. با توجه به وسعت نسبتا کم منطقه مطالعاتی و تغییرات نسبتا کم عوامل خاک‌سازی، پیشنهاد می‌شود تحقیقات بیشتری در این زمینه در مناطق وسیع‌تر و با تنوع بیشتر فاکتورهای خاک‌سازی انجام شود.

References

1. Abedi, F., Amirian-Chakan, A., Faraji, M., Taghizadeh-Mehrjardi, R., Kerry, R., and Razmjoue, D., and Scholten, T. 2021. Salt dome related soil salinity in southern Iran: Prediction and mapping with averaging machine learning models. Land Degradation and Development, 32: 1540-1555.
2. Adhikari, K., Minasny, B., Greve, M.B., and Greve, M.H. 2014. Constructing a soil class map of Denmark based on the FAO legend using digital techniques. Geoderma, 214: 101-113.
3. Aitchison, J. 1986. The statistical analysis of compositional data. London: Chapman and Hall.

4. Aitchison, J. 1992. On criteria for measures of compositional difference. Mathematical Geology, 24: 365–379.
5. Akpa, S.I.C., Odeh, I.O.A., and Bishop, T.F.A. 2014. Digital mapping of soil particle-size fractions for Nigeria. Soil Science Society of America Journal, 78: 1953-1966.
6. Banaei, M. 2000. Soil resources and use potentiality map of Iran, 1: 1000000. Karaj: Soil and Water Research Institute.
7. Behrens, T., Schmidt, K., Viscarra Rossel, R.A., Gries, P., Scholten, T., and MacMillan, R.A. 2018. Spatial modelling with Euclidean distance fields and machine learning. European Journal of Soil Science, 69: 757-770.
8. Breiman, L. 2001. Random forests. Machine Learning, 45: 5-32.
9. Chagas, C.S., Junior, W.C., Bhering, S.B., and Filho, B.C. 2016. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. Catena, 139: 232-240.
10. Chen, S., Arrouays, D., Mulder, V.L., Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., Hannam, J., Meersmans, J., Richer-de-Forges, A.C., and Walter, C. 2022. Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. Geoderma, 409, 115567.
11. Chen, S., Mulder, V.L., Heuvelink, G.B.M., Poggio, L., Cabuet, M., Román Dobarco, M., and Arrouays, D. 2020. Model averaging for mapping topsoil organic carbon in France. Geoderma, 366, 114237.
12. Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Whichmann, V., and Bohner, J. 2015. System for automated geoscientific analyses (SAGA) v. 2.1.4. Geoscientific Model Development, 8: 1991-2007.
13. Diks, C.G.H., and Vrugt, J.A. 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stochastic Environmental Research and Risk Assessment, 24: 809-820.
14. Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C. 2003. Isometric logratio transformations for compositional data analysis. Mathematical Geology, 35: 279-300.
15. Filzmoser, P., Hron, K., and Reimann, C. 2009. Principal component analysis for compositioal data with outliers. Evironmetrics, 20: 621-632.
16. Gallan, J.C., and Austin, J.M. 2015. Derivations of terrain covariates for digital soil mapping in Australia. Soil Research, 53: 895-906.
17. Gallant, J.C., and Dowling, T.I. 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resources Research, 39: 1347-1359.
18. Gee, G.W., and Bauder, J.W. 1986. Particle size analysis. In A. Klute (ed). Methods of soil analysis: Part 1. American Society of Agronomy, Madison.
19. Granger, C.W., and Ramanathan, R. 1984. Improved methods of combining forecasts. Journal of Forecasting, 32: 197-204.
20. Greve, M.H., Kheir, R.B., Greve, M.B., and Bøcher, P.K. 2012. Quantifying the environmental parameters to predict soil texture fractions using regression-tree model with GIS and LIDAR data: The case study of Denmark. Ecological Indicators, 18: 1-10.
21. Hengl, T., Rossiter D.G., and Stein, A. 2003. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Australian Journal of Soil Research, 418: 1403-1422.
22. Kaya, F., Başayiğit, L., Keshavarzi, A., and Francaviglia, R. 2022. Digital mapping for soil texture class prediction in northwestern Türkiye by different machine learning algorithms. Geoderma Regional, 31, e00584.
23. Lantz, B. 2015. Machine learning with R. Packt Publishing Ltd., Birmingham.
24. Lark, R.M., and Bishop, T.F.A. 2007. Cokriging particle size fractions of the soil. European Journal of Soil Science, 583: 763-774.
25. Liu, F., Geng, X., Zhu, A.X., Fraser, W., and Waddell, A. 2012. Soil texture mapping over low relief areas using land surface feedback dynamic patterns extracted from MODIS. Geoderma, 171- 172: 44-52.

26. Mallah, S., Delsouz Khaki, B., Davatgar, N., Scholten, T., Amirian-Chakan, A., Emadi, M., Kerry, R., Mosavi, A.H., and Taghizadeh-Mehrjardi, R. 2022. Predicting soil textural classes using random forest models: learning from imbalanced dataset. Agronomy, 2613.
27. Malone, B.P., Minasny, B., Odgers, N.P., and McBratney, A.B. 2014. Using model averaging to combine soil property rasters from legacy soil maps and from point data. Geoderma, 232: 34-44.
28. Mehrabi-Gohari, E., Matinfar, H.R., Jafari, A., Taghizadeh-Mehrjardi, R., and Triantafilis, J. 2019. The spatial prediction of soil texture fractions in arid regions of Iran. Soil Systems, 3: 65.
29. Metternicht, G.I., and Zinck, J.A. 2003. Remote sensing of soil salinity: potentials and constraints. Remote Sensing of Environment, 85: 1–20.
30. Minasny, B., and McBratney, A.B. 2018. Limited effect of organic matter on soil available water capacity. European Journal of Soil Science, 69: 39-47.
31. Odeh, I.O.A., Todd, A.J., and Triantafilis, J. 2003. Spatial prediction of soil particle-size fractions as compositional data. Soil Science, 168: 501-515.
32. Pahlavan-Rad, M.R., and Akbarimoghaddam, A. 2018. Spatial variability of soil texture fractions and pH in a flood plain (case study from eastern Iran). Catena, 160: 275-281.
33. Poggio, L., and Gimona, A. 2017. 3D mapping of soil texture in Scotland. Geoderma Regional, 9: 5-16.
34. Román Dobarco, M., Arrouays, D., Lagacherie, P., Ciampalini, R., and Saby, N.P.A. 2017. Prediction of topsoil texture for Region Centre (France) applying model ensemble methods. Geoderma, 298: 67-77.
35. Sun, Y., Wong, A.K.C., and Kamel, M.S. 2009. Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence, 23: 687-719.
36. Swain, S.R., Chakraborty, P., Panigrahi, N., Vasava, H.B., Reddy, N.N., Roy, S., Majeed, I., and Das, B. S. 2021. Estimation of soil texture using Sentinel-2 multispectral imaging data: An ensemble modeling approach. Soil and Tillage Research, 213: 105134.
37. 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: 37.
38. 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
39. Taghizadeh-Mehrjardi, R., Toomanian, N., Khavaninzadeh, A.R., Jafari, A., and Triantafilis, J. 2016. Predicting and mapping of soil particle-size fractions with adaptive neuro-fuzzy inference and ant colony optimization in central Iran. European Journal of Soil Science, 67: 707–725.
40. Umali, B.P., Oliver, D.P., Forrester, S., Chittleborough, D.J., Hutson, J.L., Kookana, R.S., and Ostendorf, B. 2012. The effect of terrain and management on the spatial variability of soil properties in an apple orchard. Catena, 93: 38-48.
41. Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Montzka, C., Nemes, A., Pachepsky, Y.A., Padarian, J. and Schaap, M.G. 2017. Pedotransfer functions in Earth system science: Challenges and perspectives. Reviews of Geophysics, 55: 1199-1256.
42. Vapnik, V.N. 1995. The nature of statistical learning theory. Springer, New York.
43. Waiser, T.H., Morgan, C.L.S., Brown, D.J., Hallmark, C.T. 2007. In situ characterization of soil clay content with visible near-infrared diffuse reflectance spectroscopy, Soil Science Society of America Journal, 71: 389–396.
44. Wang, C., Zhao, L., Fang, H., Wang, L., Xing, Z., Zou, D., Hu, G., Wu, X., Zhao, Y., Sheng, Y., Pang, Q., Du, E., Liu, G., and Yun, H. 2021. Mapping surficial soil particle size fractions in alpine permafrost regions of the Qinghai–Tibet plateau. Remote Sensing, 13: 1392.

45. Wang, D., Yang, H., Qian, H., Gao, L., Li, C., Xin, J., Tan, Y., Wang, Y., and Li, Z. 2023. Minimizing vegetation influence on soil salinity mapping with novel bare soil pixels from multi-temporal images. Geoderma, 439, 116697
46. Wang, Z., and Shi, W. 2017. Mapping soil particle-size fractions: A comparison of compositional kriging and logratio kriging. Journal of Hydrology, 546: 526-541.
47. Wilding, L.P. 1985. Spatial variability: Its documentation, accommodation and implication to soil survey. P. 166–189. In D.R. Nielsen and J. Bouma (ed). Soil spatial variability. Pudoc, Wagenigen.
48. Wu, W., Li, A.D., He, X.U., Ma, R., Liu, H.B., and Lv, J.K. 2018. A comparison of support vector machines, artificial neural network and classification tree for identifying soil texture classes in southwest China. Computers and Electronics in Agriculture, 144: 86-93.
49. Wulder, M.A., White, J.C., Loveland, T.R., Woodcock, C.E., Belward, A.S., Cohen, W.B., Fosnight, E.A., Shaw, J., Masek, J.G., and Roy, D.P. 2016. The global Landsat archive: status, consolidation, and direction. Remote Sensing of Environment, 185: 271–283.
50. Xuemei, L., and Jianshe, L. 2013. Measurement of soil properties using visible and short wave-near infrared spectroscopy and multivariate calibration. Measurement, 46: 3808–3814.
51. Zhang, M., Shi, W., and Xu, Z. 2020. Systematic comparison of five machine-learning models in classification and interpolation of soil particle size fractions using different transformed data. Hydrology and Earth System Sciences, 24: 2505-2526.

Using kNN, RF and SVM and their Combination Using GR for Soil Texture Modeling

F. Mirzaei, A. Amirian-Chakan*, R. Taghizadeh-Mehrjardi and H.R Matinfar

Ph.D Student, Department of Soil Science, Faculty of Agriculture, Lorestan University, Khorramabad, Iran. E-mail: f.mirzaie1374@gmail.com,
Assitant Professor, Department of Soil Science, Faculty of Agriculture, Lorestan University, Khorramabad, Iran. E-mail: amirian.ar@lu.ac.ir
Assitant Professor, Department of Rangeland and Watersghed Management, Faculty of Agriculture and Natural Resources, University of Ardakan, Ardakan, Iran. E-mail: rtaghizadeh @ardakan.ac.ir
Professor, Department of Soil Science, Faculty of Agriculture, Lorestan University, Khorramabad, Iran. E-mail: matinfar.h@lu.ac.ir

Received: December 9, 2023 and Accepted: May 11, 2024

Abstract

Soil texture is one of the most important soil properties that governing soil physical, chemical and biological behaviors. In modeling soil textural fractions different models are used, each has its own advantages. To combine the benefit from different models, one approach is combining their predictions. Since soil texture is a compositional data, when its fractions are estimated separately there is no guarantee that the estimates will sum to 100. Log-ratio transformations before modeling are ways to deal with the problem. Little is known about modeling transformed and untransformed (UT) soil texture data using a combination of different models. In present study, 200 surface soil samples (0-30 cm) were collected from Kuhdasht region. Random forest (RF), k-nearest neighbors (kNN) and support vector machines (SVM) and their combination using Granger-Ramanathan (GR) method were used to model soil texture data. Additive log-ratio (alr), centroid log-ratio (clr) and isometric log-ratio (ilr) transformations were used to transform texture data. Environmental variables derived from Landsat 8 and Sentinel-2 images and a digital elevation model (DEM) were used as input for all models. Results indicated that covariates derived from DEM were more important in modeling soil texture. All models improved the estimates of soil texture fractions when used alr transformed data compare to when using UT, clr and ilr transformed data. The combined model (i.e. GR) did not show superiority over other models. Using GR model RMSE values for alr, clr, ilr transformed clay data and UT clay data were 5.07%, 4.21%, 5.81% and 6.09%, respectively. For silt RMSE values (in the same order as clay) were 7.11%, 5.15%, 9.04% and 6.70%, and for sand were 9.20%, 7.67%, 11.69% and 8.74%, respectively. Generally, SVM using alr transformed data showed a slightly higher potential for modeling soil texture. To sum up, results indicated that combining different machine learning algorithms does not necessarily improved the estimates. Therefore, instead of using a model combination approach that may result in more complexity, it is possible to use a single appropriate model for modeling soil texture.

Keywords: Compositional data, Ensemble model, Log-ratio transformation, Random forest

* Corresponding author's email: amirian.ar@lu.ac.ir