ارائه مدلی غیرپارامتریک با استفاده از تکنیک k- نزدیک‌ترین همسایه در برآورد جرم مخصوص ظاهری خاک

وحیدرضا جلالی[۱] و مهدی همایی[۲*]

(تاریخ دریافت: ۱۳۸۹/۱/۱۴ ؛ تاریخ پذیرش: ۱۳۸۹/۱۱/۲)

چکیده

در مدل‌هایی که به پیش‌بینی فرآیندهای حاکم در محیط خاک می‌پردازند، دانستن جرم ویژه ظاهری خاک به عنوان یک پارامتر ورودی، لازم است. رویکردهای غیرپارامتریک در جنبه‌های مختلفی برای تخمین متغیرهای پیوسته به کار رفته‌اند. در این پژوهش نوعی از الگوریتم‌های غیرپارامتریک از نوع یادگیرنده‌های تنبل موسوم به *k*- نزدیک‌ترین همسایه، برای تخمین جرم ویژه ظاهری خاک با استفاده از دیگر ویژگی‌های کمکی آن شامل توزیع اندازه ذرات خاک، pH خاک، هدایت الکتریکی عصارهٔ اشباع خاک (EC)، درصد اشباع خاک (SP)، درصد کربن آلی خاک (OC) و مقدار آهک به کار گرفته شد. بر اساس تکنیک cross validation برای تخمین جرم ویژه ظاهری هر نمونه خاک هدف، هشت نمونه خاک که حداکثر تشابه به خاک هدف را داشتند، از بانک مرجع که حاوی ۱۳۶ نمونه خاک بود، انتخاب و مقدار جرم ویژه ظاهری آنها برآورد شد. استفاده از آماره‌های ضریب هم‌بستگی پیرسون (r=۰/۸۶)، خطای ماکزیمم (ME=۰/۱۵)، ریشه میانگین مربعات خطا (RMSE=۲/۵)، ضریب تبیین (CD=۱/۳)، کارآیی مدل (EF=۰/۷۵) و ضریب جرم باقی‌مانده (CRM=۰/۰۰۱) نشان داد که در اکثر موارد این تکنیک به صورت قابل‌قبول توانمند است. بر این اساس، می‌توان نتیجه‌گیری کرد که استفاده از این تکنیک به عنوان روشی جایگزین برای اشتقاق توابع انتقالی خاک، به ویژه زمانی که فراهمی داده‌های جدید؛ نیاز به اشتقاق مجدد این توابع را الزام‌آور می‌کند، می‌تواند به کار رود.

واژه‌های کلیدی: تکنیک k- نزدیک‌ترین همسایه، جرم ویژه ظاهری خاک، مدل‌سازی

۱. دانشجوی سابق دکتری خاک‌شناسی، دانشکده کشاورزی، دانشگاه تربیت مدرس تهران و در حال حاضر استادیار خاک‌شناسی، دانشکده کشاورزی، دانشگاه شهید باهنر کرمان

۲. استاد خاک‌شناسی، دانشکده کشاورزی، دانشگاه تربیت مدرس، تهران

* : مسئول مکاتبات، پست الکترونیکی: mhomaee@hotmail.com

مقدمه

جرم ویژه ظاهری خاک، در تبدیل روابط جرمی- حجمی، تخمین ظرفیت خاک در ذخیره‌سازی کربن و تعیین میزان انبارش مواد مغذی در خاک بسیار مورد نیاز است (۲۸). جرم ویژه ظاهری خاک هم‌چنین در برآورد خصوصیات نگهداشت آب در خاک موردنیاز بوده و دانستن آن به عنوان یک پارامتر ورودی در مدل‌هایی که به بررسی حرکت آب، املاح و رسوبات می‌پردازند؛ اجتناب‌ناپذیر است. علاوه بر این، جرم ویژه ظاهری به عنوان شاخصی برای آگاهی از میزان تراکم خاک، تخلخل و باروری موضعی آن به کار برده می‌شود (۲۳). در نهایت می‌توان با استفاده از جرم ویژه ظاهری خاک به عنوان یک ابزار کلیدی، به توصیف ساختمان خاک پرداخت. بنابراین، آگاهی داشتن از میزان جرم ویژه ظاهری خاک در بسیاری جنبه‌های مطالعاتی علوم خاک اجتناب‌ناپذیر است.

با وجود پیشرفت‌های تکنیکی و بهبود ابزارآلات به کار رفته در اندازه‌گیری مستقیم جرم ویژه ظاهری خاک، این روش‌ها همچنان زمان‌بر و همراه با خطا هستند. بنابراین در اکثر گزارش‌های تهیه شده، از این ویژگی صرف‌نظر شده و یا در صورت اندازه‌گیری، هر یک به روش جداگانه‌ای اقدام نموده‌اند (۹). بنابراین پژوهشگران برای حل این مشکل، روش‌های غیر مستقیم را مورد توجه قرار داده‌اند. اشتقاق توابع انتقالی خاک یکی از این روش‌هاست که با استفاده از ویژگی‌های زودیافت خاک، ویژگی‌های دیریافت آن را برآورد می‌کند (۴). روش‌های رگرسیونی و اخیراً، شبکه‌های عصبی مصنوعی دو روش معمول در توسعه توابع انتقالی خاک هستند (۱۰ و ۱۷).

اسکاپ و همکاران (۲۵) با استفاده از شبکه‌های عصبی مصنوعی، توابعی انتقالی با نام تجاری *Rosetta* برای تخمین و برآورد خصوصیات هیدرولیکی خاک ارائه نمودند. وجه تشابه بین اکثر توابع انتقالی موجود، در اشتقاق آنها بر مبنای رویکرد پارامتریک است. بدین معنی که همه این توابع متشکل از پارامترهایی هستند که از برازش یکسری توابع معین بر داده‌ها به دست آمده‌اند، که این رویکرد، خود کاستی‌هایی به همراه دارد.

تعیین رابطه صحیح و حصول اطمینان از یک‌نواختی توزیع تابع احتمال خطا در بین داده‌ها، معمولاً کار ساده‌ای نیست. هم‌چنین هنگامی که بانک داده از تعدادی اندک تشکیل شده باشد، تخمین‌های شکل گرفته بر اساس رویکرد مذکور، بسیار ناپایدار خواهد بود و از طرفی، در مواردی که داده‌های جدید (در مقیاس زمانی و مکانی متفاوت با داده‌های موجود در بانک مرجع) مهیا شد، بازنگری کلی در روابط قبلی و توسعه مجدد آنها الزام‌آور خواهد شد. به همین دلیل کاربران به راحتی قادر به اضافه نمودن داده‌های محلی خود برای بهبود تخمین این توابع نیستند (۱۹).

استفاده از روش‌های غیرپارامتریک می‌تواند به عنوان رویکردی جایگزین، برای اینچنین تخمین‌هایی به کار گرفته شود. این تکنیک‌ها، به جای برازش دادن یکسری توابع معین بر داده‌ها، بر اساس تشخیص الگو و استفاده از اصل تشابهات بنا نهاده شده‌اند. به عنوان نمونه تیم تحقیقاتی اسکاپ و همکاران که در سال ۲۰۰۱ نرم‌افزار *Rosetta* را بر اساس رویکرد پارامتریک و با استفاده از شبکه‌های عصبی مصنوعی ابداع نموده بودند، در پژوهشی نوین در سال ۲۰۰۹ با استفاده از همان پایگاه داده‌ای که در اشتقاق توابع هیدرولیکی خاک به کار برده بودند، به این نتیجه رسیدند که استفاده از تکنیک‌های غیرپارامتریک کارآیی چشمگیری در بهبود تخمین‌های صورت-گرفته خواهد داشت (۲۹). به همین ترتیب نمس و همکاران (۱۸) توابع انتقالی که توسط رولز و همکاران در سال ۱۹۸۲ به روش رگرسیون خطی اشتقاق یافته بودند را بررسی نمودند و دریافتند که این توابع انتقالی از دقت کافی جهت استفاده آنها در مقیاس ایالات متحده آمریکا برخوردار نبوده و با ارزیابی روش غیرپارامتریک *k- نزدیک‌ترین همسایه*، بیان نمودند که روش مذکور از توانایی بالاتری برای تخمین توابع هیدرولیکی در مقیاس کل ایالات متحده آمریکا برخوردار است (۱۸).

استفاد از الگوریتم‌های غیرپارامتریک در مواردی که نحوه ارتباط بین ورودی و خروجی از قبل به طور کامل مشخص نباشد، سودمند و مؤثر خواهد بود (۲۶ و ۳۱). رویکرد *k* –

نزدیک‌ترین همسایه، یکی از مهم‌ترین و توسعه یافته‌ترین رویکردهای غیرپارامتریک می‌باشد که در بسیاری از پژوهش‌های نوین جهت تشخیص الگو و کلاسه‌بندی‌های آماری به کار گرفته شده است (۶). کاربردهای وسیع تکنیک‌های غیرپارامتریک، در جنبه‌های مختلف، کارآیی بالای این تکنیک‌ها را به اثبات رسانده‌اند. شبیه‌سازی و تفریق جریان روانآبهای سطحی (۱۴، ۲۶ و ۲۷)، شبیه‌سازی بارش با استفاده از مدل غیرهمگن زنجیره مارکوف (۱۶)، پخش سیلاب (۲۴)، شبیه‌سازی آب و هوایی با استفاده از تکنیک *k*- نزدیک‌ترین همسایه (۳)، نمونه‌هایی از این کاربردها در هیدرولوژی است. لوپز و همکاران (۱۵) در تحقیقی به ارزیابی تراکم گونه‌های مختلف جنگلی و ایجاد نقشه پوشش گیاهی با استفاده از تکنیک *k- نزدیک‌ترین همسایه* پرداختند و نتیجه‌گیری نمودند که استفاده از تکنیک فوق، علاوه بر سهولت محاسبات، از دقت بالایی نیز برخوردار بوده است.

تحقیق نمس و همکاران (۲۱) در رابطه با تفسیر توزیع اندازه ذرات خاک با کمک تکنیک *k- نزدیک‌ترین همسایه*، در واقع یکی از اولین موارد استفاده این تکنیک در علوم خاک بوده است. پس از اثبات توانایی روش مذکور، نمس و همکاران (۲۱) با استفاده از تکنیک *k- نزدیک‌ترین همسایه*، به تخمین ویژگی‌های هیدرولیکی خاک پرداختند و بیان نمودند که علی‌رغم دقت یکسان تکنیک مذکور و روش شبکه‌های عصبی مصنوعی در اشتقاق و تخمین توابع هیدرولیکی، قابلیت روش *k- نزدیک‌ترین همسایه* جهت وارد نمودن داده‌های محلی، ارجحیتی نسبی برای این روش ایجاد می‌نماید.

تکنیک *k*- نزدیک‌ترین همسایه

بر خلاف توابع انتقالی کلاسیک، تکنیک *k- نزدیک‌ترین همسایه* از هیچ تابع ریاضیاتی از پیش تعریف‌شده‌ای جهت تخمین متغیرهای مختلف استفاده نمی‌نماید. در این رویکرد، یک بانک داده مرجع(‘reference’ data set) – همانند بانک داده‌ای که در آموزش و توسعه توابع انتقالی کلاسیک به کار می‌رود- جهت یافتن نزدیک‌ترین (مشابه‌ترین) خاک به خاک هدف، مورد جستجو واقع می‌شود. نخستین گام در این زمینه، تعیین فاصله بین نمونه هدف با هر یک از داده‌های موجود در بانک داده است. در اکثر مطالعات صورت گرفته در این زمینه، جهت اندازه‌گیری فاصله بین نمونه مجهول (هدف) و نمونه خاک‌های بانک مرجع، از روابط کلاسیک محاسبه فاصله اقلیدسی نمونه هدف تا هر یک از نمونه‌های موجود در بانک مرجع استفاده می‌شود. به طور نمونه برای حالتی که میزان فاصله بین یک نمونه خاک از بانک مرجع با نمونه هدف مدنظر باشد، بر اساس گزارش جاگتاب و همکاران (۱۲)، می‌توان از شکل کلّی رابطه فیثاغورث استفاده نمود(رابطه ۱).

$$D(X,Y)=\sqrt{\sum_{i=1}^{nf}(x_i-y_i)^2} \qquad [1]$$

که در آن X: نماینده خاکی با چند پارامتر مشخص (x_1 تا x_n) (مانند درصد اندازه ذرات، EC، pH، OC و...) از بانک داده مرجع بوده و Y: نمونه خاک هدف با همان تعداد پارامتر (y_1, تا y_n) است.

$X=(x_1, x_2, x_3,\ldots x_n)$

$Y=(y_1,y_2, y_3,\ldots y_n)$

بدین ترتیب نمونه خاک‌های بانک داده به ترتیب صعودی از کمترین (حداکثر تشابه) تا بیشترین فاصله (حداقل تشابه) از نمونه مورد نظر دسته‌بندی و ارزشگذاری خواهند شد. مرحله دوم که باید به آن پرداخته شود، تعداد خاک‌هایی (*K*) است که از فهرست فوق جهت تخمین ویژگی‌های خاک هدف از آنها باید استفاده گردد. به عبارت دیگر برای برآورد ویژگی‌های خاک هدف، چند نمونه خاک از بانک مرجع باید انتخاب گردد؟

پر واضح است که میزان کارایی این روش به طور قابل ملاحظه‌ای به کیفیت انتخاب نزدیک‌ترین (مشابه‌ترین) نمونه‌ها از بانک مرجع با خاک هدف وابسته است. زیرا به راحتی نمی‌توان بیان داشت که برای نیل به بیشترین دقت در تخمین‌ها، استفاده از نزدیک‌ترین نمونه در بانک مرجع با کمترین فاصله، استفاده از دو نمونه نزدیک، پنج نمونه نزدیک، ده نمونه نزدیک

و حتی بیشتر، تعداد نقاط منجر به تخمینی با دقت قابل قبول خواهد شد. برای تعیین تعداد بهینه k در تخمین نمونه هدف، استفاده از تکنیک Leave-One-Out Cross Validation پیشنهاد شده است (۲۱). ایشان در تحقیقات وسیع خود در شرایط مختلف رابطه $k = n^{1/2}$ *for* $n \succ ۱۰۰$ را ارائه نمودند که در آن n، تعداد نمونه در بانک مرجع است. از آنجا که هیچ تحقیقی مشابه در این زمینه صورت نگرفته و هیچ اطلاعات اولیه‌ای در زمینه بهینه‌ترین تعداد k جهت تخمین جرم ویژه ظاهری خاک وجود ندارد، بهینه‌سازی تعداد k نیز بخشی از این تحقیق بوده است.

علی‌رغم کاربرد وسیع و مؤثر رویکردهای غیرپارامتریک بطور عام و رویکرد k *نزدیک‌ترین همسایه* به طور خاص، در زمینه‌های مختلف علوم محیطی و اثبات توانایی‌های این روش، تاکنون پژوهشی در زمینه استفاده از تکنیک k– *نزدیک‌ترین همسایه* جهت برآورد جرم ویژه ظاهری خاک در سطح دنیا صورت نگرفته است. هدف از انجام این تحقیق، بررسی توانایی تکنیک مورد نظر و ارائه مدلی غیرپارامتریک برای برآورد میزان جرم ویژه ظاهری خاک بوده است.

مواد و روش‌ها

منطقهٔ مورد مطالعه، دشت دامنه‌ای قره‌میدان واقع در ۷۰ کیلومتری شمال‌غرب بجنورد واقع شده است. وسعت منطقه بیش از ۳۰۰ هکتار و شیب عمومی آن حدود ۱۵ درصد است. درآغاز پژوهش، با استفاده از نرم‌افزارArcGIS و دستگاه GPS، کل منطقه به شبکه‌هایی با طول مساوی ۱۵۰ متر تقسیم‌بندی شد. از بخش‌های مذکور از عمق ۲۵-۰ سانتی‌متری نمونه‌برداری خاک انجام شد و تعداد ۱۳۶ نمونه خاک تهیه گردید. جرم ویژهٔ ظاهری نمونه‌ها به روش کلوخه تعیین شد (۵). نمونه‌ها در مجاورت هوای آزاد خشک و از الک ۲ میلی‌متری عبور داده شدند. فراوانی نسبی اندازهٔ ذرات به روش هیدرومتری اندازه‌گیری و کلاس بافتی خاک‌ها تعیین گردید (۷). هم‌چنین ویژگی‌های شیمیایی خاک شامل واکنش خاک، pH (۲۲)، هدایت الکتریکی عصارهٔ اشباع خاک، EC (۲۲)، درصد مواد آلی خاک، OC (۳۰)، درصد رطوبت اشباع، SP (۵) و میزان آهک خاک نیز تعیین شد (۱۳).

تکنیک k– *نزدیک‌ترین همسایه* و سایر مشتقات آن متعلق به گروه الگوریتم‌های یاد گیرنده تنبل (lazy learning algorithms) می‌باشند. این الگوریتم، داده‌های در حال توسعه را به صورت غیرفعال (Passive) فقط ذخیره می‌نماید و تا هنگامی که نیاز به تخمین جدید نباشد، هیچ‌گونه فرآیند یادگیری و آموزش صورت نخواهد پذیرفت. به همین دلیل اصطلاح تنبل برای این گونه الگوریتم‌ها به کار برده می‌شود. استفاده از این تکنیک به مفهوم شناسایی و بازیابی نزدیک‌ترین (مشابه‌ترین) حالت نمونه در بانک داده به نمونه هدف بود.

ذکر این نکته در اینجا ضروری است که منظور از داده هدف، نمونه خاکی است که تمام متغیرهای فیزیکی و شیمیایی آن به جز متغیر موردنظر (جرم ویژه ظاهری خاک) مشخص است. در این حالت الگوریتم K.n.n از میان k تعداد از مشابه‌ترین خاک‌های موجود در بانک داده (۱۳۶ نمونه خاک با متغیرهای مشخص) به خاک هدف، به تخمین مقدار جرم ویژه ظاهری خاک هدف (مجهول) می‌پردازد.

به همین منظور، با استفاده از رابطه لال و شارما (۱۴) تخمینی اولیه از تعداد k نمونه جهت وارد نمودن در محاسبات به عمل آمد و سپس با استفاده از تکنیک Leave-One-Out Cross Validation، تعداد دقیق k نمونه خاک محاسبه شد.

تکنیک Cross Validation روشی آماری است که در واقع کیفیت تخمین‌های یک مدل را بر اساس تعداد و نوع داده‌های ورودی مشخص می‌کند. نام دیگر این تکنیک، تخمین چرخشی (Rotation estimation) است، زیرا با جایگذاری n تعداد از داده‌های ورودی، به تخمین متغیر مجهول پرداخته و میزان اختلاف داده تخمینی و مشاهده‌ای را ثبت می‌کند، در گردش بعدی میزان اختلاف بین داده تخمینی و مشاهده‌ای را برای n +۱ تعداد از داده‌های ورودی محاسبه می‌کند. این چرخش تا

زمانی ادامه می‌یابد تا سرانجام، ترکیب مورد نظر که دارای کمترین خطا بین مقادیر تخمینی و مشاهده‌ای است، به دست آید.

پس از تعیین تعداد k *نزدیک‌ترین همسایه* جهت ورود به محاسبات، برنامه موردنظر جهت ورود ویژگی‌های هر خاک شامل مختصات جغرافیای هر نقطه در سیستم متریک، درصد ذرات شن، سیلت و رس، درصد آهک، هدایت الکتریکی عصاره اشباع خاک (EC)، واکنش شیمیایی خاک (pH) و درصد اشباع آن (SP) در محیط برنامه‌نویسی R اجرا شد. فواصل اقلیدسی داده هدف با هر یک از داده‌های بانک مرجع محاسبه و ذخیره گردید. مقدار تخمینی ρ_b برای هر کدام از نمونه‌های هدف بر اساس میانگین وزنی k تعداد از نزدیک‌ترین همسایه‌های از پیش تعیین‌شده، به دست آمد. در نهایت اقدام به ارزیابی و اعتبارسنجی عملکرد مدل با استفاده از یکسری شاخص‌های آماری شد. یکی از شاخص‌های آماری که برای ارزیابی مدل‌ها از آن استفاده می‌شود، ضریب هم‌بستگی پیرسون می‌باشد که توسط رابطه زیر تعریف می‌شود (رابطه ۲).

$$r=\frac{n\left(\sum_{i=1}^{n}(P_i)(O_i)\right)-\left(\sum_{i=1}^{n}P_i\right)\left(\sum_{i=1}^{n}O_i\right)}{\sqrt{\left[n\sum_{i=1}^{n}(P_i)^2-\left(\sum_{i=1}^{n}(P_i)\right)^2\right]\left[n\sum_{i=1}^{n}(O_i)^2-\left(\sum_{i=1}^{n}(O_i)\right)^2\right]}}, \quad -1\le r\le 1 \qquad [2]$$

در این رابطه r: ضریب هم‌بستگی، P_i: مقدار پیش‌بینی شده برای نمونه iام و O_i: مقدار مشاهده شده برای نمونه iام می‌باشد. از آنجا که مقادیر ضریب همبستگی همواره در بازه [۱ و ۱-] قرار می‌گیرند، قضاوت از روی این ضریب ساده بوده و ممکن است به نظر برسد که ضریب هم‌بستگی می‌تواند معیار مناسبی در ارزیابی مدل باشد. با این حال باید توجه داشت که ضریب هم‌بستگی نمی‌تواند به تنهایی شاخص مناسبی برای ارزیابی مدل باشد. زیرا ممکن است در یک مدل فرضی مقادیر پیش‌بینی و مشاهده شده دارای اختلافی فاحش باشند ولی این اشتباهات به گونه‌ای باشد که از یک روند یک‌نواخت پیروی نماید. بنابراین اگرچه ضریب هم‌بستگی به خوبی نشان‌دهنده میزان هم‌آهنگی روند تغییرات مقادیر مشاهده شده نسبت به مقادیر پیش بینی شده است ولی گویای تطابق آنها نیست (۸).

شاخص‌های کمّی دیگری که می‌توان در برآورد دقت مدل از آنها استفاده نمود، عبارت‌اند از آماره‌های خطای ماکزیمم (Maximum Error)(*ME*)، ریشه میانگین مربعات خطا (Root Mean Square Error)(*RMSE*)، ضریب تبیین (Coefficient of Determination)(*CD*)، کارآیی مدل (Efficiency of model)(*EF*) و ضریب جرم باقی‌مانده (Coefficient of Residual Mass)(*CRM*) است. بیان ریاضی آماره‌های مذکور به صورت زیر است (۱۱):

$$ME=\max\left|P_i-O_i\right|_{i=1}^{n} \qquad [3]$$

$$RMSE=\left[\frac{\sum_{i=1}^{n}(P_i-O_i)^2}{n}\right]^{\frac{1}{2}}\frac{100}{\overline{O}} \qquad [4]$$

$$CD=\frac{\sum_{i=1}^{n}(O_i-\overline{O})^2}{\sum_{i=1}^{n}(P_i-\overline{O})^2} \qquad [5]$$

$$EF=\frac{\sum_{i=1}^{n}(O_i-\overline{O})^2-\sum_{i=1}^{n}(P_i-O_i)^2}{\sum_{i=1}^{n}(O_i-\overline{O})^2} \qquad [6]$$

$$CRM=\frac{\sum_{i=1}^{n}O_i-\sum_{i=1}^{n}P_i}{\sum_{i=1}^{n}O_i} \qquad [7]$$

که در آنها P_i مقادیر برآورد شده، O_i مقادیر اندازه‌گیری شده و n تعداد نمونه است. کمترین مقدار برای *ME*، *RMSE* و *CD* صفر است. مقدار *ME* نمایانگر بدترین حالت برآورد مدل است. در حالی‌که مقدار *RMSE* نشان‌دهنده بیش‌برآورد (Overestimate) یا کم‌برآورد (Underestimate) است. *CD* نسبت بین پراکنش مقادیر برآورد شده و اندازه‌گیری شده را نشان می‌دهد. بیشترین مقدار برای *EF* یک است. مقادیر *EF* و *CRM* می‌توانند منفی باشند. *EF* مقادیر برآورد شده را نسبت به مقدار میانگین اندازه‌گیری‌ها مقایسه می‌کند. مقدار منفی *EF* دلالت بر آن دارد که میانگین مقادیر اندازه‌گیری شده تخمین بهتری را نسبت به مقادیر برآورد شده ارائه می‌دهند. *CRM* بیان کننده گرایش مدل به تخمین بیشتر و یا کمتر از مقادیر

اندازه‌گیری شده است. به دست آوردن مقدار منفی *CRM* برای یک مدل تمایل مدل را برای بیش‌برآورد اندازه‌گیری‌ها نشان می‌دهد. اگر تمامی داده‌های برآورد شده و اندازه‌گیری شده یکسان باشند، نتایج آماره‌ها به صورت $ME = 0$، $RMSE = 0$، $CD = 1$، $EF = 1$ و $CRM = 0$ خواهد بود (۲).

نتایج و بحث

جدول ۱، توصیف آماری ویژگی‌های انتخابی خاک‌های موجود در بانک داده را نشان می‌دهد.

همان‌طور که اشاره شد رابطه لال و شارما (۲) تخمینی اولیه از مقدار k تعداد بهینه از نزدیک‌ترین همسایه‌ها از بانک مرجع به داده هدف ارائه می‌دهد. با توجه به این‌که تعداد داده موجود در بانک مرجع ۱۳۶ (n>100) عدد بوده لذا خواهیم داشت:

$k = n^{1/2} \quad for\ n \succ ۱۰۰$

$k = ۱۳۶^{1/2} = ۱۱/۷$

عدد به دست آمده حدود تقریبی میزان k بهینه را نشان می‌دهد ولی برای تعیین دقیق عدد k، از تکنیک Cross Validation استفاده شد (شکل ۱).

شکل ۱ میزان دقت را در تکنیک Cross Validation جهت تعیین تعداد k بهینه را بر اساس آماره مجموع مربعات خطا (Sum of Square Error)(SSE) نشان می‌دهد. همان‌طور که مشاهده می‌شود، میزان خطا در کمترین تعداد همسایگی یعنی K =۱ حداکثر بوده و با افزایش این تعداد از میزان خطا کاسته شده است. این روند تا K=۸ ادامه یافته ولی پس از آن دوباره میزان خطا افزایش یافته و یا به عبارت دیگر دقت تخمین‌ها کاهش یافته است. پس در واقع در دامنه مورد نظر(۱ تا ۱۵) تعداد K=۸ بهینه‌ترین تعداد همسایگی جهت انجام تخمین‌ها بوده است.

در مرحله بعد الگوریتم موردنظر در محیط برنامه R نوشته شد تا خود برنامه به شکل هوشمند از بانک داده، نزدیک‌ترین داده‌ها را به داده هدف انتخاب نموده و پس از مرتب‌سازی آنها بر اساس کمترین فاصله اقلیدسی، ۸ داده نزدیک به داده هدف را جدا، و با میانگین وزنی، مقدار جرم ویژه ظاهری خاک هدف را تخمین بزند. شکل ۲ مقادیر اندازه‌گیری و برآورد شده جرم ویژه ظاهری خاک را در مقابل هم به تصویر کشیده است و در شکل ۳، نقشه پراکنش میزان جرم ویژه ظاهری اندازه‌گیری شده و برآورد شده توسط مدل به تصویر کشیده شده است.

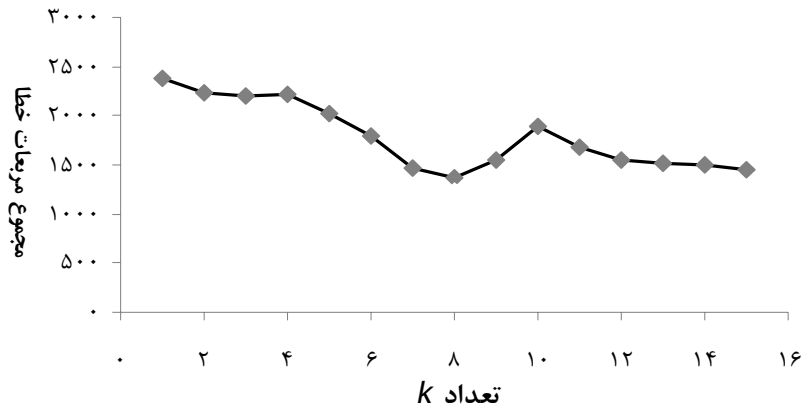
با توجه به شکل فوق می‌توان نتیجه گرفت که روند تغییرات مقادیر برآوردی توسط مدل با مقادیر اندازه‌گیری شده، هم‌آهنگ بوده و یا به عبارت دیگر مدل مورد نظر توانسته است با دقت نسبتاً خوبی مقادیر جرم ویژه ظاهری خاک را برآورد نماید (R^2=۰/۷۴۸). همان‌طور که قبلاً نیز اشاره شد، در تکنیک‌های رگرسیون خطی ضریب هم‌بستگی و یا توان دوم آن که به ضریب تبیین مشهور است، جهت تشخیص هم‌روندی دو متغیر به کار می‌روند. بنابراین، جهت تشخیص میزان دقت و توانایی یک مدل نیاز به یکسری پارامترهای آماری دیگر می‌باشد (۲).

بنابراین در این پژوهش علاوه بر محاسبه ضریب هم‌بستگی، از آماره‌های خطای ماکزیمم (*ME*)، ریشه میانگین مربعات خطا (*RMSE*)، ضریب تبیین (*CD*)، کارآیی مدل (*EF*) و ضریب جرم باقی‌مانده (*CRM*) به عنوان شاخص‌هایی کمّی جهت تعیین میزان کارایی مدل *k نزدیک‌ترین همسایه* استفاده شد. جدول ۲ میزان هر کدام از این آماره‌ها را نشان می‌دهد.

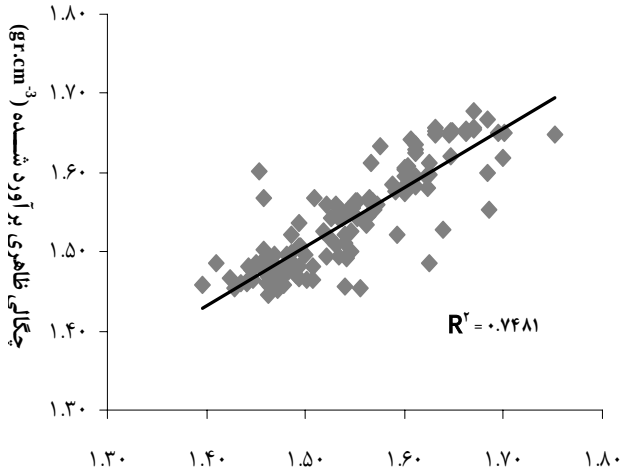
همان‌طور که از داده‌های جدول برمی‌آید، علاوه بر هم‌روندی مقادیر تخمینی با مقادیر اندازه‌گیری شده که از روی ضریب هم‌بستگی نسبتاً بالا (r = ۰/۸۶) استنباط می‌شود، مقدار بالای آماره EF (EF=۰/۷۵) نیز میزان کارآیی مدل *k – نزدیک‌ترین همسایه* را نشان می‌دهد. آماره CD، تمایل مدل را به بیش‌برآورد و یا کم‌برآورد کردن مدل نشان می‌دهد. هرچه مقدار این آماره از یک بیشتر باشد، با توجه به شکل رابطه (رابطه ۵)، تمایل مدل را به بیش‌برآورد نمودن مقادیر تخمینی نشان می‌دهد. همان‌طور که از مقدار آماره CD از جدول برمی‌آید (CD=۱/۳)، مدل اندکی تمایل به بیش‌برآورد نمودن

جدول۱. خلاصه‌ای از آماره‌های توصیفی پارامترهای خاکی به کار رفته جهت تخمین k *نزدیک‌ترین همسایه*

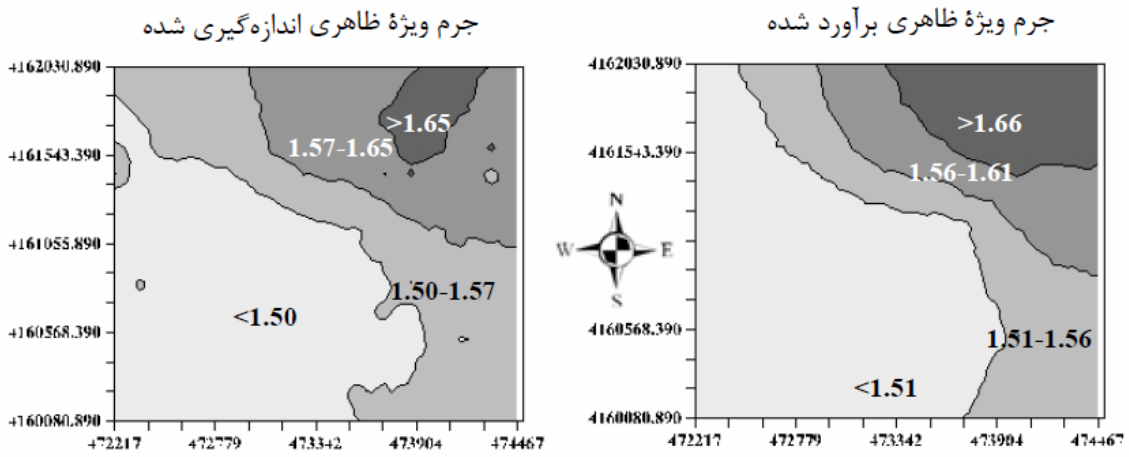
ویژگی	واحد	دامنه	کمینه	بیشینه	میانگین	انحراف معیار	واریانس
جرم ویژهٔ ظاهری	$g.cm^{-3}$	۰/۵	۱/۲۶	۱/۷۵	۱/۵۳	۰/۰۹	۰/۰۰۷
شن	$g.g^{-1}$	۰/۷	۰/۰۴	۰/۷۵	۰/۲۲	۰/۰۸۴	۰/۰۰۷
سیلت	$g.g^{-1}$	۰/۵	۰/۱۳	۰/۶۲	۰/۴۹	۰/۰۵۴	۰/۰۰۳
رس	$g.g^{-1}$	۰/۳	۰/۱۲	۰/۴۴	۰/۳	۰/۰۴۵	۰/۰۰۲
کربن آلی	%	۱/۷	۰/۲۱	۱/۹۱	۰/۸۸	۰/۲۲	۰/۰۴۹
آهک	%	۴۲	۵/۷۵	۴۷/۷۵	۲۱/۹۹	۵/۳۶	۲۸/۷۷
شوری	$dS.m^{-1}$	۳/۵	۰/۲۷	۳/۸۱	۱/۷۴	۰/۴۸	۰/۲۳
pH	-	۱/۹	۶/۴۴	۸/۳۳	۷/۳۶	۰/۲۹	۰/۰۸۸
درصد اشباع	%	۲۳	۴۳	۶۶	۵۱/۳۲	۴/۲۰	۱۷/۶۶



شکل۱.تعیین میزان k بهینه بر اساس آماره مجموع مربعات



شکل۲. جرم ویژه ظاهری اندازه‌گیری شده و برآورد شده توسط تکنیک K- نزدیک‌ترین همسایه



شکل ۳. نقشه پراکنش میزان جرم ویژه ظاهری اندازه‌گیری شده و برآورد شده توسط مدل

جدول۲. آماره‌های محاسباتی جهت تعیین میزان قابلیت *تکنیک k نزدیک‌ترین همسایه* در برآورد جرم ویژهٔ ظاهری خاک

CRM	ME	RMSE	CD	EF	r
۰/۰۰۱	۰/۱۵	۲/۵	۱/۳	۰/۷۵	۰/۸۶

تخمین‌های خود نشان می‌دهد. آماره‌های RMSE، ME و CRM نیز هرکدام به نوعی نشان‌دهنده میزان خطای مدل در انجام برآوردها بوده است که هر چه مقادیر این آماره‌ها اندک و به صفر نزدیک‌تر باشد، دقت مدل بالا بوده و تخمین‌های آن به واقعیت نزدیک‌تر است. با در نظر گرفتن مقادیر ۲/۵، ۰/۱۵ و ۰/۰۰۱ به ترتیب برای آماره‌های RMSE، ME و CRM می‌توان نتیجه‌گیری نمود که مدل *k- نزدیک‌ترین همسایه* در انجام تخمین‌های خود از مقدار خطای اندکی برخوردار بوده و مجدداً می‌توان از توان بالا و دقت قابل قبول این مدل در انجام تخمین جرم ویژه ظاهری خاک اطمینان حاصل نمود.

نتیجه‌گیری

در این پژوهش برای نخستین بار تکنیک *k- نزدیک‌ترین همسایه* به منظور تخمین جرم ویژه ظاهری خاک از روی سایر خصوصیات آن شامل توزیع اندازه ذرات خاک، واکنش خاک (pH)، هدایت الکتریکی عصارهٔ اشباع خاک (EC)، درصد مواد آلی خاک و درصد آهک آن به کار گرفته شد. انجام این روش سه مرحله دارد که به طور خلاصه عبارت‌اند از محاسبه فاصله رکورد هدف (مجهول) با هر یک از رکوردهای آموزش، تعیین k تعداد از نزدیک‌ترین همسایه‌های رکورد مجهول و در نهایت استفاده از برچسب‌های کلاس نزدیک‌ترین همسایه‌های تعیین شده به منظور تعیین برچسب رکورد مجهول (با توجه به اکثریت آراء). پس از تعیین K=۸ به عنوان نزدیک‌ترین (مشابه‌ترین) خاک‌ها از بانک مرجع با خاک هدف، مقدار جرم ویژه ظاهری خاک هدف از طریق الگوریتمی که در محیط نرم‌افزاری R نوشته شده بود، تخمین زده شد. پارامترهایی آماری برای تعیین میزان دقت و توانایی رویکرد مذکور به کار گرفته شد و مشخص گردید که این تکنیک حساسیتی قابل قبول در تخمین جرم ویژه

ظاهری خاک داشته و پیشنهاد می‌گردد تا توانایی این شیوه در تخمین سایر ویژگی‌های خاکی و با سایر بانک‌های داده به کار گرفته شود.

منابع مورد استفاده

۱. جلالی و. ر.، م. همایی و س. خ. میرنیا. ۱۳۸۷. مدلسازی واکنش کلزا به شوری طی مراحل مختلف رشد زایشی. علوم و فنون کشاورزی و منابع طبیعی۱۲(۴۴): ۱۱۱-۱۲۱.

۲. جلالی و. ر.، م. همایی و س. خ. میرنیا. ۱۳۸۶. مدلسازی واکنش کلزا به شوری طی مراحل مختلف رشد رویشی. تحقیقات مهندسی کشاورزی ۸(۴): ۹۵-۱۱۲ .

3. Bannayan, M. and G. Hoogenboom. 2008. Weather Analogue: A tool for lead time simulation of daily weather data based on modified K-nearest-neighbor approach. Env. Model. and Software 23: 703-713.
4. Bouma, J. 1989. Using soil survey data for quantitative land evaluation. Adv. in Soil Sci. 9:177–213.
5. Carter, M. R. 1993. Soil Sampling and Methods of Analysis. Lewis Pub., USA.
6. Dasarathy, B.V. 1991. Nearest neighbor (NN) Norms: NN pattern classification techniques. IEEE Computer Society Press, Los Alamitos, CA.
7. Gee, G.W. and J.W. Bauder. 1986. Particle size analysis. PP. 383-411. *In*: Klute (Ed.), Methods of Soil Analysis. Part 1, 2nd ed., America Society of Agronomy, Madison, WI.
8. Ghorbani Dashtaki, S., M. Homaee, M. H. Mahdian and M. Kouchakzadeh. 2009. Site-dependence performance of infiltration models. Water Resour. Manag. 23: 2777-2790.
9. Heuscher, S.A., C.C. Brandt and P. M. Jardine. 2005. Using soil physical and chemical properties to estimate bulk density. Soil Sci. Soc. Amer. J. 69:51–56.
10. Homaee, M. and A. Farrokhian Firouzi. 2008. Deriving point and parametric pedotransfer function of some gypsiferous soils. Aust. J. Soil Res. 46: 219-2277.
11. Homaee, M., C. Dirksen and R. A. Feddes. 2002. Simulation of root water uptake. I. Non-uniform transient salinity using different macroscopic reduction functions. Agric. Water Manag. 57: 89-109.
12. Jagtap, S.S., U. Lall, J.W. Jones, A.J. Gijsman and J.T. Ritchie. 2004. Dynamic nearest-neighbor method for estimating soil water parameters. Trans. ASAE 47:1437–1444.
13. Klute, A. 1986. Ed. Methods of soil analysis, Part 1: Physical and Mineralogical Methods. 2nd ed., Monogr. 9. ASA. and SSSA, Madison,WI.
14. Lall, U. and A. Sharma. 1996. A nearest-neighbor bootstrap for resampling hydrologic time series. Water Resour. Res. 32:679–693.
15. Lopez, H.F., A.R. Ek and M.E. Bauer. 2001. Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. Remote Sens. Environ. 77: 251– 274.
16. Marshall, L., D. Nott and A. Sharma. 2004. A comparative study of Markov chain Monte Carlo methods for conceptual rainfall-runoff modeling. Water Resour. Res. 40:2501-2512.
17. Minasny, B. and A.B. McBratney. 2002. The Neuro-m method for fitting neural network parametric pedotransfer functions. Soil Sci. Soc. Amer. J. 66:352–361.
18. Nemes, A., D.J. Timlin, A. Pachepsky Ya and W.J. Rawls. 2009. Evaluation of the Rawls et al. (1982) Pedotransfer Functions for their Applicability at the U.S. National Scale. Soil Sci. Soc. Amer. J. 73:1638-1645.
19. Nemes A., R.T. Roberts, W.J. Rawls, Ya.A. Pachepsky and M.Th. van Genuchten. 2008. Software to estimate _33 and _1500 kPa soil water retention using the non-parametric k-Nearest Neighbor technique. Environ. Model. and Software 23: 254-255.
20. Nemes A., W.J. Rawls and Ya. Pachepsky. 2006. Use of the Nonparametric Nearest Neighbor Approach to Estimate Soil Hydraulic Properties. Soil Sci. Soc. Amer. J. 70:327–336.
21. Nemes, A., J.H.M. Wösten, A. Lilly and J.H. Oude Voshaar. 1999. Evaluation of different procedures to interpolate the cumulative particle-size distribution to achieve compatibility within a soil database. Geoderma 90:187–202.
22. Page, A. L., R. H. Miller and D. R. Keeney. 1982. Methods of Soil Analysis, Part II, Physical Properties. ASA, SSSA, Madison, WI.
23. Salifu, K.F., W.L. Meyer and H.G. Murchison. 1999. Estimating soil bulk density from organic matter content, pH, silt and clay. J. Tropic. For. 15:112–120.
24. Sankarasubramanian, A. and U. Lall. 2003. Flood quantiles in a changing climate: Seasonal forecasts and causal

relations. Water Resour. Res. 39(5):1134-1144.
25. Schaap MG, Leij FJ, van Genuchten MTh 2001. Rosetta: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. J. Hydrol. 251: 163–176.
26. Sharma, A. and R. O'Neill. 2002. A nonparametric approach for representing interannual dependence in monthly streamflow sequences.Water Resour. Res. 38:5-15.
27. Souza Filho, F.A. and U. Lall. 2003. Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: Applications of a multivariate, semiparametric algorithm. Water Resour. Res. 39:1307-1317.
28. Tamminen, P. and M. Starr. 1994. Bulk density of forested mineral soils. Silva Fennica 28:53–60.
29. Twarakavi, N.K.C., J. Šimůnek, and M.G. Schaap. 2009. Development of Pedotransfer Functions for Estimation of Soil Hydraulic Parameters using Support Vector Machines. Soil Sci. Soc. Amer. J. 73:1443-1452.
30. Walkley, A. and I.A. Black. 1934. An examination of the Degtjareff for determining soil organic matter and a proposed modification of the chromic acid titration method. Soil Sci. 37: 29-38.
31. Yakowitz, S. 1993. Nearest-neighbor estimation for null-recurrent Markov time series. Stoch. Proc. Appl. 48:311–318.