نشریه آب و خاک (علوم و صنایع کشاورزی)
جلد ۲۵، شماره ۲، خرداد – تیر ۱۳۹۰، ص. ۳۵۵-۳۴۷

Journal of Water and Soil
Vol. 25, No.2, May-Jun 2011, p. 347-355

ارائه مدلی غیرپارامتریک برای برآورد هدایت هیدرولیکی اشباع خاک با استفاده از روش k- نزدیکترین همسایه

وحیدرضا جلالی[۱]- مهدی همایی[۲*]

تاریخ دریافت:۸۹/۴/۲۶
تاریخ پذیرش:۸۹/۸/۱۶

چکیده

هدایت هیدرولیکی خاک در بسیاری از مطالعات مربوط به حرکت آب و انتقال املاح موردنیاز می‌باشد، لیکن در بیشتر موارد به علت محدودیت‌های عملی و یا هزینه‌ای، اندازه‌گیری آن با دشواری همراه است. رویکردهای غیرپارامتریک از جنبه‌های مختلفی برای تخمین متغیرهای پیوسته بکار رفته‌اند. در این پژوهش نوعی از الگوریتم‌های غیرپارامتریک از نوع یادگیرنده‌های تنبل موسوم به k – نزدیکترین همسایه، برای تخمین هدایت هیدرولیکی اشباع خاک با استفاده از دیگر ویژگیهای کمکی خاک شامل توزیع اندازه ذرات، هدایت الکتریکی عصارهٔ اشباع (EC_e)، رطوبت اشباع (θ_s)، درصد کربن آلی (OC)، مقدار مواد خنثی شونده (TNV) و جرم ویژهٔ حقیقی و ظاهری بکار گرفته شد. بر اساس تکنیک ارزیابی تقاطعی برای تخمین هدایت هیدرولیکی اشباع هر نمونه خاک هدف، تعداد ۱۰ نمونه خاک که حداکثر تشابه با خاک هدف را داشتند، از بانک مرجع که حاوی ۱۵۱ نمونه خاک بود، انتخاب و مقدار هدایت هیدرولیکی اشباع آنها برآورد گردید. استفاده از آماره‌های ضریب همبستگی پیرسون(r=۰/۸۰۱)، خطای ماکزیمم (ME=۱۲۰/۴)، ریشه میانگین مربعات خطا (RMSE=۷۱/۵)، ضریب تبیین (CD=۱/۳۲)، کارآیی مدل (EF=۰/۶۵) و ضریب جرم باقیمانده (CRM=-۰/۰۴۶) نشان داد که در بیشتر موارد، این تکنیک بصورتی قابل‌قبول توانایی تخمین کمّیت موردنظر را دارد. بر این اساس، می‌توان نتیجه‌گیری کرد که استفاده از این تکنیک به عنوان روشی جایگزین برای اشتقاق توابع انتقالی خاک، به‌ویژه هنگامی که فراهمی داده‌های جدید؛ نیاز به اشتقاق مجدد این توابع را الزام‌آور می‌کند، می‌تواند بکار رود.

واژه‌های کلیدی: تکنیک k–نزدیکترین همسایه، مدل سازی، هدایت هیدرولیکی اشباع خاک

مقدمه

آگاهی از هدایت آبی اشباع خاک برای درک و مدل‌سازی بسیاری از فرآیندهای فیزیکی خاک ضروری است. ایجاد تمایز بین روانآب سطحی و نفوذ به درون خاک، ماندگاری موقتی آب در محیط ریشه، نرخ انتقال املاح و بسیاری دیگر از فرایندهای کشاورزی و زیست‌محیطی وابسته به میزان هدایت هیدرولیکی اشباع خاک (K_s) می‌باشد (۲۶). با وجود پیشرفت‌های تکنیکی و بهبود ابزارآلات بکار رفته در اندازه‌گیری مستقیم این ویژگی خاک، این روش‌ها همچنان زمان‌بر و همراه با خطا می‌باشند. بنابراین پژوهشگران جهت حل این مشکل، روشهای غیر مستقیم را مورد توجه قرار داده‌اند. اشتقاق توابع انتقالی خاک یکی از این روشهاست که از روی ویژگیهای زودیافت خاک، ویژگیهای دیریافت آن را برآورد می‌کند (۵). تکنیکهای رگرسیونی و اخیراً، شبکه‌های عصبی مصنوعی دو روش معمول در توسعه توابع انتقالی خاک می‌باشند (۸و۱۵). اسکاپ وهمکاران (۲۲) با استفاده از شبکه‌های عصبی مصنوعی، توابعی انتقالی با نام تجاری *Rosetta* جهت تخمین و برآورد ویژگیهای هیدرولیکی خاک ارائه نمودند.

وجه تشابه بین اکثر توابع انتقالی موجود، در اشتقاق آنها بر مبنای رویکرد پارامتریک می‌باشد. بدین معنی که همه این توابع متشکل از پارامترهایی هستند که از برازش یکسری توابع معین بر داده‌ها بدست آمده‌اند که این رویکرد، خود کاستی‌هایی به همراه دارد (۱۸).

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

۱و۲- دانشجوی دکتری و استاد گروه خاکشناسی، دانشکده کشاورزی، دانشگاه تربیت مدرس
(*- نویسنده مسئول Email: mhomaee@modares.ac.ir)

بود. و از طرفی، در مواردی که داده‌های جدید (در مقیاس زمانی و مکانی متفاوت با داده‌های موجود در بانک مرجع) مهیا گردد، 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 جهت بهبود تخمین این توابع نیستند (۱۷).

استفاده از تکنیکهای غیرپارامتریک می‌تواند بعنوان 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. بعنوان نمونه تیم تحقیقاتی اسکاپ وهمکاران 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 بکار برده بودند، به این نتیجه رسیدند که استفاده از تکنیکهای غیرپارامتریک کارآیی چشمگیری در بهبود تخمین‌های صورت‌گرفته خواهد داشت (۲۵). به همین ترتیب نمز و همکاران (۱۶) توابع انتقالی که توسط Rawls 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 جهت تخمین توابع هیدرولیکی در مقیاس کل ایالات متحده آمریکا برخوردار است.

کاربردهای وسیع تکنیک‌های غیرپارامتریک، در جنبه‌های مختلف، کارآیی بالای این تکنیک‌ها را به اثبات رسانده‌اند. شبیه‌سازی و تفریق جریان روانآبهای سطحی (۲۳و۲۴)، شبیه‌سازی بارش با استفاده از مدل غیرهمگن زنجیره مارکوف (۱۴)، پخش سیلاب (۲۱)، شبیه‌سازی آب و هوایی با استفاده از تکنیک k- نزدیکترین همسایه (۳) نمونه‌هایی از این کاربردها در هیدرولوژی می‌باشد. بنیان و هوگنبوم (۴)، در تحقیقی نوین و با استفاده از اصل تشابهات به تخمین ضرایب گیاهی در مدلهای شبیه‌ساز زراعی شامل دو مدل DSSAT و CSM-CERES-Maize پرداختند. لوپز و همکاران (۱۲) و جردسن (۷)، در تحقیقاتی جداگانه به ارزیابی تراکم گونه‌های مختلف جنگلی و ایجاد نقشه پوشش گیاهی با استفاده از تکنیک k- نزدیکترین همسایه پرداختند.

گنوسن و همکاران (۲۰۰۹)، به بررسی توانایی تکنیک k- نزدیکترین همسایه در مطالعهٔ داده‌های ماهواره‌ای مربوط به سطوح جنگلی پرداختند. و به این نتیجه رسیدند که استفاده از تکنیک k- نزدیکترین همسایه در بهم ارتباط دادن داده‌های ماهواره‌ای و پدیده‌های زمینی و در نهایت تهیه نقشهٔ متغیرهای جنگلی، بسیار جذاب و کارا بوده است.

تحقیق نمز و همکاران (۱۹) در رابطه با تفسیر توزیع اندازه ذرات خاک با کمک تکنیک k –نزدیکترین همسایه، در واقع یکی از اولین موارد استفادهٔ این تکنیک در علوم خاک بوده است. پس از اثبات توانایی روش مذکور، نمز و همکاران (۱۸) 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 می نماید.

سگال وهمکاران (۲۰۰۸) نیز به مطالعه حرکت و انتقال آب و املاح در مقیاس ناهمگون مزرعه‌ای پرداختند. این محققین در پژوهش خود از تکنینک k- نزدیکترین همسایه به عنوان الگوریتم میان‌یابی نام برده و بطور موفقیت آمیزی توانستند با استفاده از این رویکرد، درک عمیقی از تغییرپذیری ویژگیهای هیدرولیکی خاک در مقیاس مزرعه‌ای بدست آورند.

علی‌رغم کاربرد وسیع و مؤثر رویکردهای غیرپارامتریک بطور عام و رویکرد k- نزدیکترین همسایه بطور خاص، در زمینه‌های مختلف علوم محیطی و اثبات توانایی‌های این روش، تاکنون پژوهشی مستقل در زمینه استفاده از تکنیک k- نزدیکترین همسایه جهت برآورد هدایت هیدرولیکی اشباع خاک در سطح کشور صورت نگرفته است. هدف از انجام این تحقیق، بررسی توانایی تکنیک مورد نظر و ارائه مدلی غیرپارامتریک برای برآورد میزان هدایت هیدرولیکی اشباع خاک بوده است.

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[1] – همانند بانک داده‌ای که در آموزش و توسعه توابع انتقالی کلاسیک بکار 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، برای 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. بطور نمونه برای حالتی که میزان فاصله بین یک نمونه خاک از بانک مرجع با نمونه هدف مدنظر باشد، بر اساس گزارش جگتاپ و همکاران (۱۰) می‌توان از شکل کلی رابطه فیثاغورث استفاده نمود (رابطه ۱).

1 -‘reference’ data set

$$D(X,Y)=\sqrt{\sum_{i=1}^{nf}(x_i-y_i)^2} \qquad (۱)$$

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

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

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

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 بطور قابل ملاحظه‌ای به کیفیت انتخاب نزدیکترین (مشابه‌ترین) نمونه‌ها از بانک مرجع با خاک هدف وابسته است. جهت تعیین تعداد بهینه k در تخمین نمونه هدف، لال و شارما (۱۱) استفاده از تکنیک ارزیابی تقاطعی[1] را پیشنهاد نموده‌اند. ایشان در تحقیقات وسیع خود در شرایط مختلف رابطه $k=n^{1/2}\ \ for\ n>100$ را ارائه نمودند که در آن n، تعداد نمونه در بانک مرجع می‌باشد. ازآنجا که هیچ تحقیقی مشابه در این زمینه صورت نگرفته و هیچ اطلاعات اولیه‌ای در زمینه بهینه‌ترین تعداد k جهت تخمین هدایت هیدرولیکی اشباع خاک وجود ندارد، بهینه‌سازی تعداد k نیز بخشی از این تحقیق بوده است.

مواد و روش‌ها

منطقهٔ مورد مطالعه، دشت دامنه‌ای قره‌میدان واقع در ۷۰ کیلومتری شمال‌غرب بجنورد می‌باشد. وسعت منطقه بیش از ۳۰۰ هکتار وشیب عمومی آن حدود ۱۵ درصد می‌باشد. درآغاز پژوهش، با استفاده از نرم‌افزار ArcGIS و دستگاه GPS، کل منطقه به شبکه‌هایی با طول مساوی ۱۵۰ متر تقسیم‌بندی شد. از بخشهای مذکور نیز از عمق ۲۵-۰ سانتی‌متری نمونه‌برداری خاک انجام شد و تعداد ۱۵۱ نمونه خاک انتخاب گردید.

در این پژوهش از دستگاه نفوذسنج گلف برای تعیین ضریب آبگذری اشباع خاک استفاده گردید. روش نفوذسنج گلف (۲۰) یکی از روشهای اندازه‌گیری نفوذپذیری تحت بار ثابت می‌باشد. در این روش دبی ثابت آب خروجی از چاهک به خاک اطراف تحت بار آبی ثابت اندازه‌گیری می‌شود. جرم ویژهٔ ظاهری نمونه‌ها به روش کلوخه و جرم ویژهٔ حقیقی از طریق پیکنومتر تعیین شد. نمونه‌ها در مجاورت هوای آزاد خشک و از الک ۲ میلی‌متری عبور داده شدند. فراوانی نسبی اندازهٔ ذرات به روش هیدرومتری تعیین گردید. همچنین سایر ویژگیهای خاک شامل، هدایت الکتریکی عصارهٔ اشباع خاک (EC_e)، درصد مواد آلی خاک (OC)، رطوبت اشباع خاک (θ_s) و میزان مواد خنثی شونده آن (TNV) تعیین گردید.

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 بصورت غیرفعال (passively) 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 بکار برده می‌شود. استفاده از این تکنیک به مفهوم شناسایی و بازیابی نزدیکترین (مشابه‌ترین) حالت نمونه در بانک داده به نمونه هدف است.

به همین منظور، با استفاده از رابطه لال و شرما (۱۱) تخمینی اولیه از تعداد k نمونه جهت وارد نمودن در محاسبات به عمل آمد و سپس با استفاده از تکنیک ارزیابی تقاطعی، تعداد دقیق k نمونه خاک محاسبه گردید. پس از تعیین تعداد k- نزدیکترین همسایه جهت ورود به محاسبات، برنامه موردنظر جهت ورود ویژگیهای هر یک خاک شامل مختصات جغرافیای هر نقطه در سیستم متریک، درصد ذرات شن، سیلت و رس، درصد مواد خنثی‌شونده (TNV)، هدایت الکتریکی عصاره اشباع خاک (EC_e)، مقدار رطوبت اشباع آن (θ_s) و جرم ویژهٔ ظاهری و حقیقی خاکها، در محیط برنامه‌نویسی R اجرا گردید. فواصل اقلیدسی داده هدف با هر یک از داده‌های بانک مرجع محاسبه و ذخیره گردید. مقدار تخمینی برای هر کدام از نمونه‌های هدف بر اساس میانگین وزنی k تعداد از نزدیکترین همسایه‌های از پیش تعیین‌شده، بدست 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).

$$r=\frac{n\left(\sum_{i=1}^{n}(P_i)(O_i)\right)-\left(\sum_{i=1}^{n}P_i\right)\left(\sum_{i=1}^{n}O_i\right)}{\sqrt{\left[n\sum_{i=1}^{n}(P_i)^2-\left(\sum_{i=1}^{n}(P_i)\right)^2\right]\left[n\sum_{i=1}^{n}(O_i)^2-\left(\sum_{i=1}^{n}(O_i)\right)^2\right]}},\quad -1\le r\le 1 \qquad (۲)$$

در این رابطه r: ضریب همبستگی، P_i: مقدار پیش بینی شده برای نمونه iام، و O_i: مقدار مشاهده شده برای نمونه iام می‌باشد. از آنجا که مقادیر ضریب هم بستگی همواره در بازه [۱ و ۱-] قرار می‌گیرند، قضاوت از روی این ضریب ساده است و ممکن است به نظر برسد که ضریب همبستگی می‌تواند معیار مناسبی در ارزیابی مدل باشد. با این حال بایستی توجه داشت که ضریب همبستگی نمی‌تواند به تنهایی

1- Leave-One-Out Cross Validation

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 می‌باشد اما گویای تطابق آنها نیست (۶).

شاخصهای کمّی دیگری که می‌توان در برآورد دقت مدل از آنها استفاده نمود، عبارتند از آماره‌های خطای ماکزیمم[1](ME)، ریشه میانگین مربعات خطا[2](RMSE)، ضریب تبیین[3](CD)، کارآیی مدل[4](EF) و ضریب جرم باقیمانده[5](CRM). بیان ریاضی آماره‌های مذکور به صورت زیر است (۹):

$$ME = \max|P_i - O_i|_{i=1}^{n} \quad (۳)$$

$$RMSE = \left[\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}\right]^{\frac{1}{2}} \frac{100}{\overline{O}} \quad (۴)$$

$$CD = \frac{\sum_{i=1}^{n}(O_i - \overline{O})^2}{\sum_{i=1}^{n}(P_i - \overline{O})^2} \quad (۵)$$

$$EF = \frac{\sum_{i=1}^{n}(O_i - \overline{O})^2 - \sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2} \quad (۶)$$

$$CRM = \frac{\sum_{i=1}^{n}O_i - \sum_{i=1}^{n}P_i}{\sum_{i=1}^{n}O_i} \quad (۷)$$

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 اندازه گیری‌ها 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 اندازه گیری شده است. بدست آوردن مقدار منفی CRM برای یک مدل تمایل مدل را برای بیش‌برآورد اندازه گیری‌ها 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، توصیفی آماری از ویژگیهای انتخابی خاکهای موجود در بانک داده را نشان می‌دهد.

همانگونه که اشاره شد، رابطه لال و شارما (۱۱) تخمینی اولیه از مقدار k تعداد بهینه از نزدیکترین همسایه‌ها از بانک مرجع به داده هدف ارائه می‌دهد. با توجه به اینکه تعداد داده‌های موجود در بانک مرجع ۱۵۱ (n>100) عدد بود، لذا:

$$k = n^{1/2} \quad for \ n \succ 100$$

$$k = 151^{1/2} = 12.3$$

عدد بدست آمده حدود تقریبی میزان k بهینه را نشان می‌دهد، لیکن برای تعیین دقیق عدد k، از تکنیک ارزیابی تقاطعی استفاده شد.

شکل ۱ مقدار دقت در تکنیک ارزیابی تقاطعی (Cross Validation) برای تعیین تعداد k بهینه را بر اساس آماره مجموع مربعات خطا[6] (SSE) نشان می‌دهد. همانطور که مشاهده می‌شود، مقدار خطا در کمترین تعداد همسایگی یعنی K=1 حداکثر بوده و با افزایش این تعداد از میزان خطا کاسته شده است. این روند تا K=۱۰ 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 موردنظر(۱ تا ۱۵) تعداد K=۱۰ 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، ۱۰ داده نزدیک به داده هدف را جدا، و با میانگین وزنی، مقدار هدایت هیدرولیکی اشباع خاک هدف را تخمین بزند. شکل ۲ مقادیر اندازه‌گیری و برآورد شده هدایت هیدرولیکی اشباع خاک را در مقابل هم به تصویر کشیده است و در شکل ۳، نقشه پراکنش میزان هدایت هیدرولیکی اشباع اندازه‌گیری شده و برآورد شده توسط مدل نشان داده شده است.

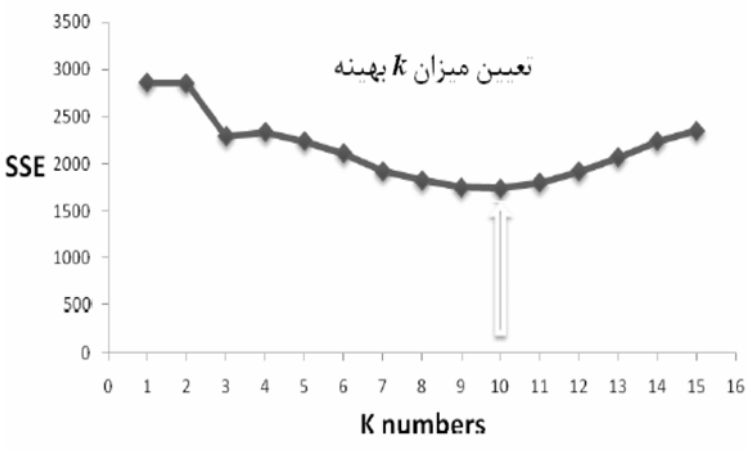
با توجه به شکل۲ می‌توان نتیجه گرفت که روند تغییرات مقادیر برآوردی توسط مدل با مقادیر اندازه‌گیری شده، هم‌آهنگ بوده و یا بعبارت دیگر مدل موردنظر توانسته است با دقت نسبتاً خوبی مقادیر هدایت هیدرولیکی اشباع خاک را برآورد نماید (R^2=0.667).

1 - Maximum Error
2- Root Mean Square Error
3 - Coefficient of Determination
4 - Efficiency of model
5 - Coefficient of Residual Mass

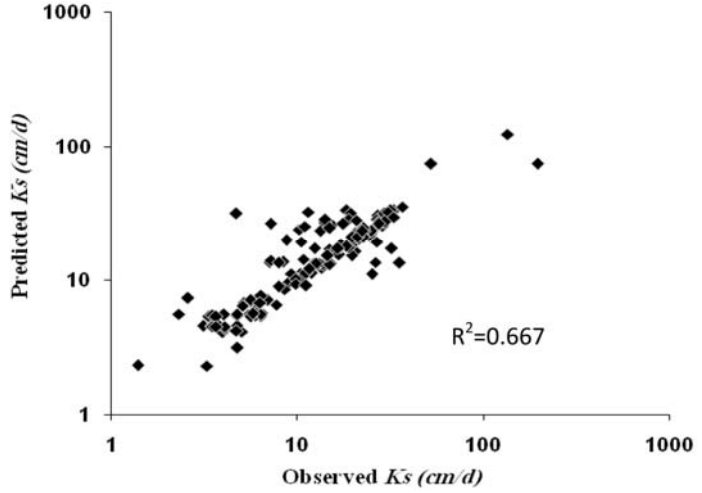
6- Sum of Square Error

جدول۱- خلاصه‌ای از آماره‌های توصیفی پارامترهای خاکی بکار رفته جهت تخمین k نزدیکترین همسایه

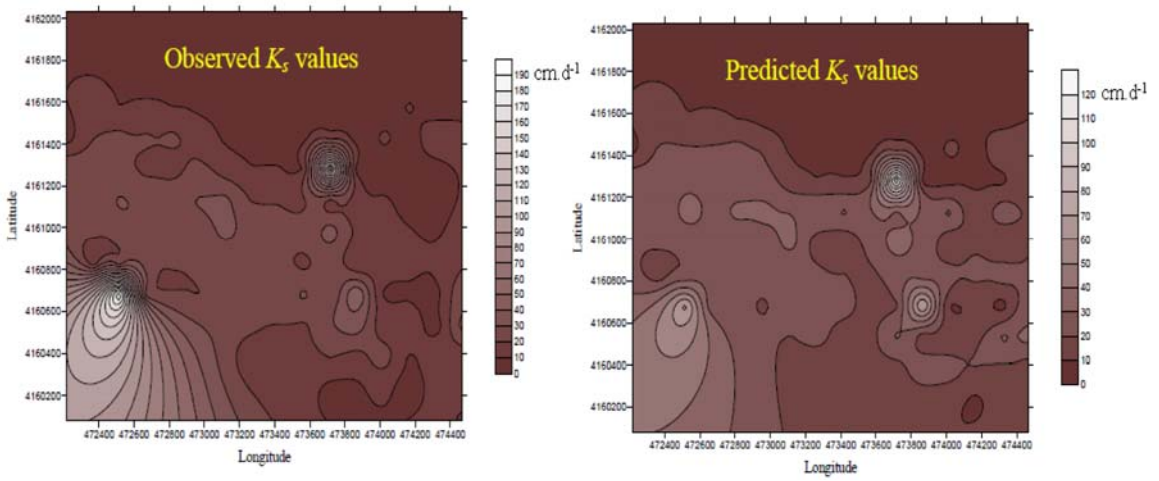
ویژگی	واحد	دامنه	کمینه	بیشینه	میانگین	انحراف معیار	ضریب تغییرات
جرم ویژهٔ ظاهری	$g.cm^{-3}$	۰/۴۹	۱/۲۶	۱/۷۵	۱/۵۳	۰/۰۸۶	۵/۶۲
جرم ویژهٔ حقیقی	$g.cm^{-3}$	۰/۵۹	۲/۱۹	۲/۷۸	۲/۵۶	۰/۰۸۸	۳/۴۳
شن	%	۷۱	۴	۷۵	۲۲	۸/۳۹	۳۹/۷۷
سیلت	%	۴۹	۱۳	۶۲	۴۹	۵/۴۱	۱۱/۱۰
رس	%	۳۲	۱۲	۴۴	۳۰	۴/۵	۱۵
کربن آلی	%	۱/۷۰	۰/۲۱	۱/۹۱	۰/۸۸	۰/۲۲	۲۴/۹
مواد خنثی شونده	%	۴۲	۵/۷۵	۴۷/۷۵	۲۱/۹۹	۵/۳۶	۲۴/۳۸
شوری	$dS.m^{-1}$	۳/۵۴	۰/۲۷	۳/۸۱	۱/۷۴	۰/۴۸	۲۸
رطوبت اشباع (θ_s)	$m^3.m^{-3}$	۰/۰۹	۰/۳۸	۰/۴۶	۰/۴۰۷	۰/۰۱۲۶	۳/۱
هدایت هیدرولیکی اشباع خاک (k_s)	$cm.d^{-1}$	۱۹۳/۲۲	۱/۳۹	۱۹۴/۶۲	۱۷/۱۲	۱۹/۸۵	۱۱۶



شکل۱- تعیین میزان k بهینه بر اساس آماره مجموع مربعات خطا (SSE)



شکل۲- هدایت هیدرولیکی اشباع اندازه‌گیری شده و برآورد شده توسط تکنیک k-نزدیکترین همسایه



شکل ۳- نقشه پراکنش میزان هدایت هیدرولیکی اندازه‌گیری شده و برآورد شده توسط مدل

همانگونه 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 متغییر به کار می‌روند لیکن 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) بعنوان شاخصهایی کمّی جهت تعیین میزان کارایی مدل k –نزدیکترین همسایه استفاده شد. جدول شماره ۲ میزان هر کدام از این آماره‌ها را نشان می‌دهد.

جدول۲- آماره‌های محاسباتی جهت تعیین میزان قابلیت تکنیک k نزدیکترین همسایه در برآورد هدایت هیدرولیکی خاک

CRM	ME	RMSE	CD	EF	r
-۰/۰۴۶۲	۱۲۰/۴۷	۷۱/۵۳	۱/۳۲	۰/۶۵۵۱	۰/۸۰۱۵

همانطور که از داده‌های جدول برمی‌آید، افزون بر هم‌روندی مقادیر تخمینی با مقادیر اندازه‌گیری شده که از روی ضریب همبستگی نسبتاً بالا (r=۰/۸۰۱۵) استنباط می‌شود، مقدار آماره EF (EF=۰/۶۵) نیز میزان کارآیی نسبی مدل k -نزدیکترین همسایه را نشان می‌دهد. آماره CD، تمایل مدل را به بیش‌برآوردی و یا کم-برآوردی 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=۱/۳۲)، مدل تا حدی تمایل به بیش‌برآورد نمودن تخمینهای خود نشان می‌دهد.

آماره CRM نیز بیان کننده گرایش مدل به تخمین بیشتر و یا کمتر از مقادیر اندازه‌گیری شده است. بدست آوردن مقدار منفی (CRM=-۰/۰۴۶۲) برای این مدل، تمایل مدل را برای بیش‌برآورد اندازه‌گیری‌ها نشان می‌دهد.

مقدار ME نمایانگر بدترین حالت برآورد مدل است که به حداکثر اختلاف بین مقادیر اندازه‌گیری شده و برآورد شده اشاره دارد. آماره RMSE، نیز به نوعی نشان‌دهنده میزان خطای مدل در انجام برآوردها بوده است که هر چه مقدار این آماره اندک باشد، دقت مدل بالا بوده و تخمینهای آن به واقعیت نزدیکتر است. با توجه به اندک بودن بانک داده در این پژوهش مقدار آماره RMSE تا حدی بالا بوده است که می‌توان آن را به پایین بودن قدرت انتخاب شبیه‌ترین نمونه-ها به خاک هدف از بین بانک داده نسبت داد.

بطور کلی با درنظر گرفتن مقادیر آماره‌های فوق می‌توان نتیجه-گیری نمود که مدل k –نزدیکترین همسایه در انجام تخمینهای خود از مقدار خطای قابل قبولی برخوردار بوده و به شرط فراهمی یک بانک داده مناسب، می‌توان از توان بالا و دقت قابل قبول این مدل در تخمین ویژگیهای هیدرولیکی خاک استفاده نمود.

نتیجه‌گیری

در این پژوهش، تکنیک k– نزدیکترین همسایه به منظور تخمین هدایت هیدرولیکی خاک با استفاده از دیگر ویژگیهای آن شامل توزیع اندازه ذرات خاک، مقدار رطوبت اشباع خاک (θs)، هدایت الکتریکی عصارهٔ اشباع خاک (ECe)، درصد مواد آلی، درصد آهک، جرم ویژهٔ ظاهری و حقیقی آن بکار گرفته شد. پس از تعیین K=۱۰ با استفاده از تکنیک Cross Validation بعنوان نزدیکترین (مشابه‌ترین) خاکها

از بانک مرجع با خاک هدف، مقدار هدایت هیدرولیکی خاک هدف از طریق الگوریتمی که در محیط نرم‌افزاری R نوشته شده بود، تخمین زده شد. برخی آماره‌ها برای تعیین مقدار دقت و توانایی رویکرد مذکور بکار گرفته شد و مشخص گردید که این تکنیک توانمندی قابل قبولی در تخمین هدایت هیدرولیکی خاک داشته و پیشنهاد می‌گردد تا توانایی این شیوه در تخمین سایر ویژگیهای خاکی نیز بکار گرفته شود.

منابع

۱- جلالی و.ر.، همایی م. و میرنیا س.خ. ۱۳۸۷. مدلسازی واکنش کلزا به شوری طی مراحل مختلف رشد زایشی. علوم و فنون کشاورزی و منابع طبیعی. سال دوازدهم شماره ۴۴ تابستان ۱۳۸۷ ص ۱۲۱-۱۱۱.

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

3- Bannayan M., and Hoogenboom G. 2008. Weather Analogue: A tool for lead time simulation of daily weather data based on modified K-nearest-neighbor approach. Env. Modeling and Software 23: 703-713.

4- Bannayan M., and Hoogenboom G. 2009. Using pattern recognition for estimating cultivar coefficients of a crop simulation model. Field Crops Research 111: 290–302. doi:10.1016/j.fcr.2009.01.007.

5- Bouma J. 1989. Using soil survey data for quantitative land evaluation. Advanced Soil Science. 9:177–213.

6- Ghorbani Dashtaki S., Homaee M., Mahdian M.H., and Kouchakzadeh M. 2009. Site-Dependence Performance of Infiltration Models. Water Resource Management. DOI 10.1007/s11269-009-9408-3.

7- Gjertsen A.K. 2007. Accuracy of forest mapping based on Landsat TM data and a kNN-based method. Remote Sensing of Environment 110: 420–430. doi:10.1016/j.rse.2006.08.018.

8- Homaee M., and Farrokhian Firouzi A. 2008. Deriving point and parametric pedotransfer function of some gypsiferous soils. Australian Journal of Soil Research. 46: 219-2277.

9- Homaee M., Dirksen C., and Feddes R.A. 2002. Simulation of root water uptake. I. Non-uniform transient salinity using different macroscopic reduction functions. Agricultural Water Management. 57, 89-109.

10- Jagtap S.S., Lall U., Jones J.W., Gijsman A.J., and. Ritchie J.T. 2004. Dynamic nearest-neighbor method for estimating soil water parameters. Trans. ASAE 47:1437–1444.

11- Lall U., and Sharma A. 1996. A nearest-neighbor bootstrap for resampling hydrologic time series. Water Resource. Research. 32:679–693.

12- Lopez H.F., Ek A.R., and Bauer M.E. 2001. Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. Remote Sensing of Environment 77: 251– 274.

13- Magnussen S., McRoberts RE., and Tomppo E.O. 2009. Model-based mean square error estimators for k-nearest neighbour predictions and applications using remotely sensed data for forest inventories. Remote Sensing of Environment 113: 476–488. doi:10.1016/j.rse.2008.04.018.

14- Marshall L., Nott D., and Sharma A. 2004. A comparative study of Markov chain Monte Carlo methods for conceptual rainfall-runoff modeling. Water Resource. Research. 40:W02501 10.1029/2003WR002378.

15- Minasny B., and Mc Bratney A.B. 2002. The Neuro-m method for fitting neural network parametric pedotransfer functions. Soil Science Society of American Journal. 66:352–361.

16- Nemes A., Timlin D.J., Pachepsky Ya.A., and Rawls W.J. 2009. Evaluation of the Rawls et al. (1982) Pedotransfer Functions for their Applicability at the U.S. National Scale. Soil Science Society of American Journal. 73:1638-1645. DOI: 10.2136/sssaj2008.0298.

17- Nemes A., Roberts R.T., Rawls W.J., Pachepsky Ya.A, and Van Genuchten M.Th. 2008. Software to estimate _33 and _1500 kPa soil water retention using the non-parametric k-Nearest Neighbor technique. Environmental Modelling and Software 23: 254-255. doi:10.1016/j.envsoft.2007.05.018.

18- Nemes A., Rawls W.J., and Pachepsky Y. 2006. Use of the Nonparametric Nearest Neighbor Approach to Estimate Soil Hydraulic Properties. Soil Science Society of American Journal. 70:327–336 (2006). doi:10.2136/sssaj2005.0128.

19- Nemes A., Wösten J.H.M., Lilly A., and Oude Voshaar J.H. 1999. Evaluation of different procedures to interpolate the cumulative particle-size distribution to achieve compatibility within a soil database. Geoderma 90:187–202.

20- Reynolds W.D., and Elrick D.E. 1987. Laboratory and numerical assessment of the Guelph permeameter method. Soil Science. 144: 244-282.

21- Sankarasubramanian A., and Lall U. 2003. Flood quantiles in a changing climate: Seasonal forecasts and causal relations. Water Resource Research. 39(5):1134 10.1029/2002WR001593.

22- Schaap M.G., Leij F.J., Van Genuchten M.Th. 2001. Rosetta: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. Journal of Hydrology 251, 163–176. doi: 10.1016/S0022-

1694(01)00466-8.
23- Sharma A., and O'Neill R. 2002. A nonparametric approach for representing interannual dependence in monthly streamflow sequences. Water Resource Research. 38:510.1029/2001WR000953.
24- Souza Filho F.A., and Lall U. 2003. Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: Applications of a multivariate, semiparametric algorithm. Water Resource Research. 39:1307 doi:10.1029/2002WR001373.
25- Twarakavi N.K.C., Šimůnek J., and Schaap M.G. 2009. Development of Pedotransfer Functions for Estimation of Soil Hydraulic Parameters using Support Vector Machines. Soil Science Society of American Journal. 73:1443-1452. DOI: 10.2136/sssaj2008.0021.
26- Zeleke T. B., and Si B.C. 2005. Scaling Relationships between Saturated Hydraulic Conductivity and Soil Physical Properties. Soil Science Society of America Journal. 69:1691–1702.

Journal of Water and Soil
Vol. 25, No.2, May-Jun 2011, p. 347-355

نشریه آب و خاک (علوم و صنایع کشاورزی)
جلد ۲۵، شماره ۲، خرداد – تیر ۱۳۹۰، ص. ۳۵۵-۳۴۷

A Nonparametric Model by Using k-nearest neighbor Technique for Predicting Soil Saturated Hydraulic Conductivity

V.R Jalali[1]- M. Homaee[2*]
Received:17-7-2010
Accepted:7-11-2010

Abstract

Saturated hydraulic conductivity (K_s) is needed for many studies related to water and solute transport, but often cannot be measured because of practical and/or cost-related reasons. Nonparametric approaches are being used in various fields to estimate continuous variables. One type of the nonparametric lazy learning algorithms, a *k*-nearest neighbor (k-NN) algorithm, was introduced and tested to estimate saturated hydraulic conductivity (K_s) from other soil properties including soil textural fractions, EC, pH, SP, OC, TNV, ρ_s and ρ_b. A number of 10 nearest neighbors, based on Cross Validation technique were selected to perform saturated hydraulic conductivity prediction from 151 soil sample attributes. The nonparametric k-NN technique performed mostly equally well, in terms of Pearson correlation coefficient (r=0.801), modeling efficiency (EF=0.65), root-mean-squared errors (RMSE=71.15) maximum error (ME=120.47), coefficient of determination (CD=1.32) and coefficient of residual mass (CRM=-0.046) statistics. It can be concluded that the k-NN technique is a competitive alternative to other techniques such as pedotransfer functions (PTFs) to estimate saturated hydraulic conductivity.

Keywords: *k*-nearest neighbor (k-NN), Modeling, Saturated hydraulic conductivity

1-PhD Student and Professor, Department of Soil Science, Faculty of Agriculture, Tarbiat Modares University of Tehran
(*- Corresponding Author Email: mhomaee@ modares.ac.ir)