مجله مدیریت خاک و تولید پایدار
جلد سوم، شماره اول، ۱۳۹۲
http://ejsms.gau.ac.ir

مقایسه روش‌های K– نزدیک‌ترین همسایگی و شبکه عصبی مصنوعی در برآورد ظرفیت تبادل کاتیونی خاک

*علی‌اصغر ذوالفقاری[۱]، محمدتقی تیرگر سلطانی[۲]، تورج افشاری‌بدرلو[۳] و فریدون سرمدیان[۴]
[۱]دانشجوی دکتری گروه مهندسی علوم خاک، دانشگاه تهران، [۲]کارشناس‌ارشد گروه مهندسی علوم خاک، دانشگاه تهران،
[۳]دانشجوی کارشناسی‌ارشد گروه مهندسی علوم خاک، دانشگاه صنعتی شاهرود، [۴]دانشیار گروه مهندسی علوم خاک، دانشگاه تهران
تاریخ دریافت: ۹۱/۳/۹ ؛ تاریخ پذیرش: ۹۱/۹/۲۵

چکیده

اندازه‌گیری ظرفیت تبادل کاتیونی خاک در سطوح وسیع، معمولاً بسیار پرهزینه و وقت‌گیر است. تخمین این کمیت به‌وسیله ویژگی‌های زودیافت خاک، از طریق توسعه توابع غیرپارامتریک می‌تواند رویکرد مناسبی باشد. در این پژوهش روش غیرپارامتریکی با عنوان K– نزدیک‌ترین همسایگی در تخمین CEC خاک استفاده شد و نتایج آن با یکی از پرکاربردترین روش‌های مرسوم مبتنی بر مدل‌های شبکه عصبی مصنوعی (ANN) مورد مقایسه قرار گرفت. ٦٨٣ نمونه خاک از مناطق مرکزی ایران انتخاب شدند که ۱۲۰ عدد از آن‌ها به‌عنوان داده‌های مورد آزمون (هدف) و ٥٦٣ عدد به‌عنوان بانک داده مرجع (آموزش) قرار گرفتند. مقادیر پارامترهای رس، سیلت، شن و کربن آلی خاک به‌عنوان متغیر مستقل ورودی (زودیافت) و CEC به‌عنوان متغیر وابسته خروجی بودند. نتایج نشان داد که بیش‌ترین خطای برآورد (MaxE) در روش K-NN برابر ٤/٨١ $cmol^+/kg$ و این مقدار در روش ANN برابر ٥/٢٦ $cmol^+/kg$ بود. ریشه میانگین مربعات خطا در روش K-NN، ۱/۵۱ و در روش ANN، ۱/۵۳ بود، که نشان می‌دهد هر دو روش قادرند با دقت بالا و یکسانی CEC خاک‌های هدف را پیش‌بینی نمایند. مقادیر مثبت آماره میانگین خطا (ME) برای این دو روش نیز نشان داد که هر دوی آن‌ها متمایل به برآورد کم‌تر مقدار CEC می‌باشند. همچنین نتایج بررسی کارایی مدل‌ها نشان داد که هر دو روش از کارایی بالایی (EF=۰/۸۸) در برآورد ظرفیت تبادل کاتیونی خاک برخوردار هستند.

واژه‌های کلیدی: مدل غیرپارامتریک، ظرفیت تبادل کاتیونی، K– نزدیک‌ترین همسایگی، مدل شبکه عصبی مصنوعی

* مسئول مکاتبه: azolfaghari@ut.ac.ir

مقدمه

ظرفیت تبادل کاتیونی (CEC)[1] یکی از ویژگی‌های مهم خاک است که بسیاری از عملکردهای اساسی خاک در ارتباط با نوسانات اسیدیته (pH)، نگهداری آب، مواد غذایی و آلاینده‌ها را کنترل می‌نماید و به همین دلیل این پارامتر، یکی از شاخص‌های کلیدی در مدیریت حاصل‌خیزی خاک و بهره‌وری زمین محسوب می‌شود (کارو و همکاران، ۲۰۰۱). تعیین سطوح بحرانی آلاینده‌های شیمیایی در خاک و پالایش آن‌ها از خاک، بدون دسترسی به داده‌های CEC امکان‌پذیر نمی‌باشد. داده‌های CEC خاک‌های زراعی می‌تواند تفسیر روشن‌تر و کامل‌تری از فرایندهای تغذیه گیاه، کاربرد کودها و اصلاح شیمیایی خاک‌ها را در اختیار پژوهشگران قرار دهد. دقیق‌ترین داده‌های CEC خاک‌ها در آزمایشگاه و از طریق اندازه‌گیری مستقیم به‌دست می‌آید، ولی روش‌های اندازه‌گیری این پارامتر به‌نسبت پرزحمت، پرهزینه و البته بسیار وقت‌گیر می‌باشند (سیبلد و همکاران، ۲۰۰۵؛ مک‌براتنی و همکاران، ۲۰۰۲). به همین دلیل تاکنون بسیاری از پژوهشگران بررسی‌های خود را بر روش‌های غیرمستقیم برآورد CEC از روی ویژگی‌های زودیافت خاک متمرکز نموده‌اند (سیبلد و همکاران، ۲۰۰۵). رگرسیون‌گیری بین ویژگی‌های پایه‌ای خاک به‌عنوان داده‌های ورودی و مقادیر CEC به‌دست آمده در آزمایشگاه، منجر به ایجاد معادله‌های چندمتغیره‌ای می‌شود که می‌تواند تخمین قابل‌قبولی از CEC خاک مناطق همگن را در اختیار پژوهشگران قرار دهد. این معادله‌ها که CEC را تابعی خطی از ویژگی‌های فیزیکی و شیمیایی خاک به‌خصوص مقدار رس و مواد آلی خاک تخمین می‌زنند تحت عنوان کلی توابع انتقالی خاک (PTFs)[2] نام گرفته‌اند (مک‌براتنی و همکاران، ۲۰۰۲). کریمیان (۱۹۹۸) سهم رس و مواد آلی در CEC ۱۵۰ نمونه از خاک‌های آهکی مناطقی استان فارس را در غالب توابع انتقالی بیان نمود. کروگ و همکاران (۲۰۰۰) از توابع انتقالی برای پیش‌بینی ظرفیت تبادل کاتیونی ۱٦٤۳ نمونه از خاک‌های دانمارک استفاده کردند. آن‌ها با متوسط خطای پیش‌بینی ۱/۹۹ $cmol^{+}/kg$ نشان دادند که ۹۰ درصد تغییرات CEC در خاک‌های مورد آزمایش، به‌وسیله مقدار رس و مواد آلی قابل‌توجیه بود. بل و وان‌کولن (۱۹۹۵) در پژوهشی در مکزیک تابعی ارایه نمودند که بیش از ۹٦ درصد تغییرات CEC را با مقادیر رس، ماده آلی و pH توجیه می‌نمود. از عیب‌های PTF ها و به‌طور کلی توابع پارامتریک در تخمین ویژگی‌های دیریافت خاک تعیین مقدار عددی هر یک از پارامترها

1- Cation Exchange Capacity
2- Pedotransfer Function

است که از طریق بهترین برازش معادله‌های یاد شده بر داده‌ها به‌دست می‌آید. از آن‌جا که مقادیر این پارامترها برای هر سری داده‌ها تغییر می‌کند، عیب‌های اصلی کاربرد این توابع این است که این معادله‌ها برای هر سری از داده‌های جدید نیازمند برازش دوباره و تعیین مقادیر جدید برای پارامترها است. ضمن آن‌که تحقق شرایط اعتبار رگرسیون‌گیری (مانند نرمال بودن توزیع خطاها حول میانگین و...) از محدودیت‌های کاربرد این توابع می‌باشد. علاوه‌بر این تخمین نتایج در سری‌های کوچک داده‌ای دارای اریب است (نمس و همکاران، ۲۰۰٦). اخیراً پژوهشگران با استفاده از روش‌های غیرپارامتریک مانند مدل‌های شبکه عصبی مصنوعی (ANN)[1] تخمین‌های دقیق‌تری از مقدار CEC ارایه نموده‌اند (معماریان‌فرد و بیگی‌هرچگانی، ۲۰۰۹؛ امینی و همکاران، ۲۰۰۵). شبکه‌های عصبی مصنوعی یکی از روش‌های هوشمند پردازش داده‌ها هستند که در ابتدا به‌وسیله پاچپسکی و همکاران (۱۹۹٦)، شاپ و بوتن (۱۹۹٦) و تاماری و همکاران (۱۹۹٦) مورد استفاده قرار گرفته‌اند. فریت و همکاران (۲۰۰۳) در تخمین رطوبت خاک با استفاده از شبکه‌های عصبی مصنوعی گزارش کردند که شبکه‌های عصبی مصنوعی ابزار مناسبی برای تخمین رطوبت خاک می‌باشند. پاچپسکی و همکاران (۱۹۹٦) با مقایسه مدل‌های رگرسیونی و شبکه‌های عصبی مصنوعی در ۲۳۰ نمونه خاک گزارش کردند که شبکه عصبی با دقت بیش‌تری ظرفیت نگهداری آب در خاک را برآورد می‌کند. شاپ و لجی (۱۹۹۸) نشان دادند که کیفیت پیش‌بینی توابع انتقالی توسط شبکه‌های عصبی وابسته به مجموعه داده‌هایی است که برای آموزش استفاده می‌شود و دقت برآوردها به‌طور مستقیم متأثر از کیفیت و نوع داده‌ها می‌باشد. تاواراکاوی و همکاران (۲۰۰۹) در اشتقاق توابع هیدرولیکی خاک با استفاده از روش‌های غیرپارامتری، بهبود چشم‌گیری را در تخمین‌های صورت گرفته مشاهده نمودند. یکی دیگر از روش‌های غیرپارامتریک که اخیراً در بسیاری از علوم از جمله کشاورزی (بنایان و هوگنبوم، ۲۰۰۹)، جنگل (جرتسن، ۲۰۰۷؛ لوپز و همکاران، ۲۰۰۱) و هیدرولوژی (کلارک و همکاران، ۲۰۰٤؛ یتس و همکاران، ۲۰۰۳) کاربرد چشم‌گیری یافته است مبتنی بر الگوشناسی و استفاده از اصل تشابه و نزدیکی داده‌ها است. چنین رویکردی در تخمین ویژگی‌های موردنظر، هنگامی‌که نوع رابطه بین داده‌های ورودی و خروجی روشن نباشد از سودمندی بیش‌تری برخوردار است (یاکوایتز، ۱۹۹۳؛ لال و شارما، ۱۹۹٦). روش K نزدیک‌ترین همسایگی (K-NN)[2] یکی از تکنیک‌های مبتنی بر چنین رویکردی است که در تخمین

1- Artificial Neural Network
2- K-Nearest Neighbor

برخی ویژگی‌های فیزیکی و شیمیایی خاک مورد استفاده قرار گرفته است. جلالی و همایی (۲۰۱۱) این روش را برای تخمین هدایت هیدرولیکی خاک با استفاده از برخی خصوصیات خاک مانند توزیع اندازه ذرات، هدایت الکتریکی اشباع، رطوبت اشباع، کربن آلی، مقدار مواد خنثی‌شونده، جرم ویژه حقیقی و ظاهری به‌کار بردند. آن‌ها گزارش کردند که این تکنیک در بیش‌تر موارد به شکل قابل‌قبولی توانایی تخمین کمیت موردنظر را دارد و می‌تواند جایگزین مناسبی برای اشتقاق توابع انتقالی خاک محسوب گردد. نمس و همکاران (۲۰۰۹) نیز با مقایسه کارایی توابع انتقالی پارامتریک و تکنیک K-NN در خاک‌های ایالت متحده، توانایی روش K-NN در تخمین توابع هیدرولیکی خاک را در مقیاس کل خاک‌های ایالت متحده بهتر از توابع پارامتریک ارزیابی نمودند. نمس و همکاران (۲۰۰٦) در برآورد نقاط پتانسیلی خاک (۳۳- و ۱۵۰۰- کیلوپاسکال) از روی توزیع اندازه ذرات، جرم ویژه ظاهری و درصد مواد آلی ۲۱۲۵ خاک از پایگاه داده امریکا (U.S. NRCS-SCS) توانایی دو روش ANN و K-NN را مورد بررسی قرار دادند و گزارش کردند که هر دو روش یاد شده از دقت یکسانی در تخمین و اشتقاق توابع توابع هیدرولیکی برخوردارند. آن‌ها همچنین کاربرد روش K-NN را در مواردی که توسعه دوباره توابع انتقالی نیاز نباشد به‌عنوان روش جایگزین مناسبی در اشتقاق این توابع پیشنهاد نمودند. در پژوهشی مشابه، حق‌وردی و همکاران (۲۰۱۰) نیز برای تخمین نقاط پتانسیلی آب در خاک (نقطه پژمردگی دایم و ظرفیت مزرعه) دو روش غیرپارامتریک ANN و K-NN را مورد مقایسه قرار دادند و گزارش کردند که روش K-NN از دقت بالاتری نسبت به ANN در خاک‌های شمال و شمال شرق ایران برخوردار است. از آن‌جا که تاکنون در ایران هیچ پژوهشی با هدف مقایسه عملکرد دو روش غیرپارامتریک K نزدیک‌ترین همسایگی (K-NN) و شبکه عصبی مصنوعی (ANN) بر روی CEC خاک‌های کشور گزارش نشده است، بنابراین در این پژوهش ضمن بررسی سودمندی کاربرد تکنیک K-NN در برآورد CEC برخی از خاک‌های ایران، دقت تخمین این کمیت با دو روش K-NN و ANN نیز مورد ارزیابی قرار خواهد گرفت.

مواد و روش‌ها

تعداد ٦٨٣ نمونه خاک از برخی استان‌های ایران شامل مناطقی از استان سمنان ۲۷۸ خاک، دشت قزوین ١٦٧ خاک، ساوجبلاغ از استان البرز (١٢٦ خاک) و ورامین (۱۱۲ خاک) از عمق ۲۵ سانتی‌متری سطحی، برداشت شد. از این تعداد ۱۲۰ عدد خاک به‌صورت تصادفی انتخاب و به‌عنوان خاک‌های

مورد آزمون (هدف) برای تخمین CEC در نظر گرفته شدند. همه نمونه‌ها پس از هوا خشک شدن و کوبیده شدن از الک ۲ میلی‌متری عبور داده شدند و برای انجام آزمون‌های فیزیکی و شیمیایی آماده شدند. توزیع اندازه ذرات به روش هیدرومتری (گی و بادر، ۱۹۸٦)، کربن آلی با روش اکسیداسیون تر والکلی- بلک (نلسون و سامرز، ۱۹۸۲) و اندازه‌گیری CEC نیز با روش استاندارد (چاپمن، ۱۹٦۵) انجام شد.

مدل شبکه عصبی مصنوعی (ANN): شبکه‌ای از عناصر پردازش ساده (نورون‌ها)، که می‌توانند رفتار پیچیده معینی را از ارتباط بین عناصر پردازش و پارامترهای عنصر نمایش دهند، اساس پردازش در این شبکه محسوب می‌شود. واحدهای محاسباتی شبکه (پرسپترون‌ها) می‌توانند به دو شکل تک‌لایه (SLP)[1] و یا چندلایه (MLP)[2] عمل نمایند. در این پژوهش از الگوریتم پرسپترون MLP با سه‌لایه برای برآورد داده‌های هدف استفاده شد. این نوع الگوریتم از یک لایه ورودی به‌منظور اعمال ورودی‌های مسأله برای آموزش شبکه با داده‌های توزیع اندازه ذرات، درصد کربن آلی و مقادیر ظرفیت تبادل کاتیونی این خاک‌ها، یک لایه پنهان با ۷ نورون و تابع عمل‌گر سیگموئیدی و یک لایه خروجی با تابع عمل‌گر خطی تشکیل شده است و قادر خواهد بود که برآوردی از ظرفیت تبادل کاتیونی خاک‌های مورد آزمون را با توجه به بانک داده‌های ورودی (مرجع) ارایه نمایند. برای سنجش عملکرد ANN در اشتقاق توابع و پیش‌بینی ظرفیت تبادل کاتیونی خاک‌ها، مقادیر آماره‌های ضریب همبستگی (r)، ریشه میانگین مربعات خطا (RMSE)، میانگین قدرمطلق خطا (MAE)، خطای ماکزیمم (ME)، میانگین باقی‌مانده‌ها (MR) و در نهایت کارایی مدل (EF) محاسبه شدند.

روش K نزدیک‌ترین همسایگی (K-NN): به‌طورکلی K-NN روشی برای طبقه‌بندی یک عنصر در یک مجموعه است که این کار را براساس نزدیک‌ترین خصوصیات سایر اعضای موجود در مجموعه (نمونه‌های آموزش‌دهنده) انجام می‌دهد. K-NN دارای یکی از ساده‌ترین الگوریتم‌های آموزش برای پیش‌بینی داده‌ها است و بر خلاف PTF های دیگر از هیچ تابع ریاضیاتی از پیش مشخص‌شده‌ای برای تخمین عناصر هدف استفاده نمی‌کند. الگوریتم K-NN به گروه موسوم به الگوریتم‌های یادگیرنده تنبل (lazy learning algorithms) تعلق دارد و داده‌ها را به‌گونه‌ای ذخیره می‌نماید که تا زمانی‌که دستوری برای برآورد جدید نباشد، فرایند آموزش روی این داده‌ها انجام نخواهد شد (نمس و همکاران،

1- Single-Layer Perceptron
2- Multi-Layer Perceptron

۲۰۰۶؛ نمس و همکاران، ۲۰۰۸). در این تکنیک، تعداد بهینه‌ای از نمونه‌های موجود در یک مجموعه که دارای شبیه‌ترین ویژگی‌ها به نمونه هدف باشند (K) در نظر گرفته می‌شود و سپس جایابی و طبقه‌بندی عنصر هدف در این مجموعه از داده‌ها، با تعیین فاصله‌ها و سپس وزن‌دهی عناصر آموزش‌دهنده صورت می‌پذیرد. در یک بانک داده مرجع خاک، نزدیک‌ترین (مشابه‌ترین) خاک‌ها به خاک مورد آزمون انتخاب می‌شوند (K) و این خاک‌ها با توجه به‌میزان مشابهت به خاک مورد آزمون وزن‌دهی می‌شوند. به این ترتیب که در ابتدا فواصل اقلیدسی بین خاک مورد آزمون و خاک‌های مرجع محاسبه شده، سپس تعداد K عدد از نزدیک‌ترین خاک‌ها براساس فاصله به‌دست آمده، وزن‌دهی می‌شوند. در انتها با توجه وزن هر یک از خاک‌ها در مجموعه‌ای با K عدد خاک (نزدیک‌ترین همسایه)، برآوردی از داده‌های هدف که مورد آزمون واقع شده‌اند، صورت می‌پذیرد (نمس و همکاران، ۲۰۰۶).

با توجه به موارد بالا در این پژوهش به‌ترتیب زیر عمل گردید.

الف) محاسبه فاصله اقلیدسی بین نمونه هدف با هر یک از نمونه‌های خاک در بانک داده مرجع از رابطه زیر انجام شد:

$$d_i = \sqrt{\sum_{j=1}^{x} \Delta a_{ij}^{2}} \qquad (1)$$

که در آن، d_i: نشان‌دهنده فاصله i امین خاک از بانک داده مرجع تا خاک هدف می‌باشد، Δa_{ij}: اختلاف فاصله

$$a_{ij(temp)} = \frac{[(a_{ij}) - \overline{a}_i]}{\sigma(a_j)} \qquad (۲)$$

که در آن، a_{ij}: نماینده j امین متغیر از i امین خاک است، $\overline{a_j}$ و $\sigma(a_j)$ به‌ترتیب میانگین و انحراف معیار مقادیر مشاهده‌ای از j امین متغیر در بانک داده‌های مرجع است.

سپس مقادیر نهایی تبدیل شده j امین متغیر از i امین خاک ($a_{ij(trans)}$) با توجه به دامنه کم‌ترین تا بیش‌تر مقدار $a_{ij(temp)}$ از رابطه زیر به‌دست آمد و به‌جای مقادیر عددی واقعی به‌عنوان داده ورودی مورد استفاده قرار گرفت (نمس و همکاران، ۲۰۰۶).

$$a_{ij} = \frac{\{\max[range(a_{j=1(temp)}),...range(a_{j=x(temp)})]\}}{range(a_{j(temp)})} \qquad (۳)$$

پس از نرمال کردن متغیرهای مستقل ورودی و محاسبه فاصله‌ها بین نمونه هدف و سایر نمونه‌های موجود در بانک داده مرجع، نمونه خاک‌های مورد مطالعه بر حسب مقدار محاسبه شده برای فاصله، به‌ترتیب صعودی مرتب شدند.

ب) یافتن تعداد بهینه خاک‌هایی که نزدیک‌ترین فاصله را با نمونه هدف دارند (مقدار K). به این منظور از روش ارزیابی تقاطعی (Cross Validation) استفاده شد. لال و شارما (۱۹۹۶) پیشنهاد کردند که اگر تعداد کل نمونه‌ها در بانک داده مرجع بیش از ۱۰۰ عدد باشد مقدار بهینه K برای رسیدن به کم‌ترین خطای برآوردها، تقریباً برابر ریشه دوم تعداد نمونه‌ها خواهد بود. نمس و همکاران (۲۰۰۶) در پژوهشی با اعمال اندازه‌های مختلف بانک داده مرجع از خصوصیات هیدرولیکی خاک نتیجه گرفتند که مقدار بهینه K به‌صورت تابعی از تعداد کل خاک‌های بانک مرجع (N) است. آن‌ها رابطه زیر را برای تعیین بهترین مقدار برای خاک پیشنهاد نمودند.

$$K = ۰/۶۵۵N^{۰/۴۹۳},\ R^۲ = ۰/۹۸۹ \qquad (٤)$$

در این پژوهش با توجه به تعداد نمونه‌های بانک مرجع مقدار خطای برآورد در دامنه‌ای از مقادیر مختلف K محاسبه شد و براساس روش ارزیابی تقاطعی، بهترین تعداد برای K، زمانی‌که خطای تخمین‌ها (RMSE)[1] به کم‌ترین مقدار خود رسید تعیین گردید.

1- Root Mean Square Error

ج) وزن‌دهی نمونه خاک‌های آموزش‌دهنده (K) براساس فاصله‌های به‌دست آمده بین خاک هدف و هر یک از خاک‌های تعیین شده در K طبق رابطه‌های زیر صورت گرفت.

$$w_i = \frac{d_{i(rel)}}{\sum_{i=1}^{K} d_{i(rel)}} \quad (۵)$$

که در آن، w_i: وزن i امین نمونه از خاک‌های آموزش‌دهنده (K) است و $d_{i(rel)}$ فاصله نسبی هدف و نمونه i می‌باشد که از رابطه زیر به‌دست آمده است.

$$d_{i(rel)} = \left(\frac{\sum_{i=1}^{K} d_i}{d_i} \right)^p \quad (۶)$$

نمس و همکاران (۲۰۰۶) کم‌ترین مقدار خطای تخمین در توابع هیدرولیکی خاک را در p های نزدیک به ۱ به‌دست آوردند و گزارش نمودند که حساسیت روش K-NN به مقادیر مختلف p بسیار اندک می‌باشد. در این پژوهش علاوه‌بر بهینه‌سازی تعداد نمونه خاک‌های نزدیک‌ترین همسایه (K)، بهترین مقدار p نیز از تکنیک ارزیابی تقاطعی به‌دست آمد.

د) برآورد مقادیر CEC در نمونه‌های هدف از طریق محاسبه میانگین وزنی CEC در نمونه خاک‌های K با استفاده از رابطه زیر صورت گرفت.

$$P_i = \sum_{i=1}^{K} w_i o_i \quad (۷)$$

که در آن، P_i: مقدار CEC پیش‌بینی شده خاک هدف، w_i: وزن مربوط به i امین خاک در K و O_i مقدار CEC اندازه‌گیری شده i امین خاک در K است.

ه) در آخرین مرحله مقادیر آماره‌های لازم برای مقایسه و ارزیابی قابلیت دو روش ANN و K-NN در پیش‌بینی CEC خاک‌های مختلف از رابطه‌های زیر به‌دست آمد.

$$r = \frac{n\left(\sum_{i=1}^{n}(p_i O_i)\right) - \left(\sum_{i=1}^{n} P_i\right)\left(\sum_{i=1}^{n} O_i\right)}{\sqrt{\left[n\sum_{i=1}^{n}(P_i)^2 - \left(\sum_{i=1}^{n}(P_i)\right)^2\right]\left[n\sum_{i=1}^{n}(O_i)^2 - \left(\sum_{i=1}^{n}(O_i)\right)^2\right]}} \quad (۸)$$

که در آن، r: ضریب همبستگی پیرسون، P_i و O_i: به‌ترتیب مقادیر CEC پیش‌بینی شده و مشاهده‌ای و n: تعداد نمونه‌های مورد آزمون است. این ضریب که بین ۱- تا ۱ تغییر می‌کند بیان‌کننده میزان هم‌روندی داده‌هاست. هرچه این مقدار به ۱ و یا ۱- نزدیک‌تر باشد، نشانه آن است که روند تغییرات مقادیر پیش‌بینی شده و مشاهده‌ای یکنواخت‌تر می‌باشد. به‌منظور مقایسه عملکرد دو روش ANN و K-NN آماره‌های ریشه میانگین مربعات خطا (RMSE)، میانگین قدرمطلق خطا (MAE)[1]، بیش‌ترین خطا (MaxE)[2]، میانگین باقی‌مانده (ME)[3] و کارایی مدل (EF)[4] با استفاده از رابطه‌های زیر به‌دست آمدند.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}} \quad (9)$$

$$MAE = 1/n \sum_{i=1}^{n} |P_i - O_i| \quad (10)$$

$$ME = \max |P_i - O_i| \quad (11)$$

$$MR = 1/n \sum_{i=1}^{n} (P_i - O_i) \quad (12)$$

$$EF = \frac{\sum_{i=1}^{n}(o_i - \bar{o})^2 - \sum_{i=1}^{n}(p_i - o_i)^2}{\sum_{i=1}^{n}(o_i - \bar{o})^2} \times 100 \quad (13)$$

که در آن‌ها، $\bar{o}$: میانگین مقادیر مشاهده‌ای و CEC می‌باشد.

در این پژوهش برنامه کامپیوتری در محیط برنامه‌نویسی Matlab (7.10) تهیه شد (Matlab 7.1, The Math works Inc., Natick, MA) و تمامی مراحل (الف) تا (ه) با استفاده از این برنامه به انجام رسید.

1- Mean Absolute Error
2- Maximum Error
3- Mean Error
4- Efficiency of Model

نتایج و بحث

واکنش خاک (pH) در نمونه‌های مورد مطالعه بین ۸/۳-۷/٦ قرار داشت. محدوده هدایت الکتریکی (EC) نمونه‌ها ۱۱-٤ دسی‌زیمنس بر متر و نسبت جذب سدیم (SAR) آن‌ها نیز کم‌تر از ۱۳ بود که نشان‌دهنده شوری متوسط تا زیاد خاک‌های مورد مطالعه بود. جدول ۱ خلاصه‌ای از ویژگی‌های اندازه‌گیری شده خاک‌های هدف و بانک مرجع را نشان می‌دهد. همان‌گونه که مشاهده می‌شود، خاک‌های مورد مطالعه در خصوصیاتی مانند توزیع اندازه ذرات و کربن آلی که از مهم‌ترین عوامل موثر بر CEC خاک‌ها هستند، دامنه به‌نسبت وسیعی را نشان می‌دهند. این امر را می‌توان از دلایل اصلی تنوع مقادیر CEC اندازه‌گیری شده در این پژوهش نیز به حساب آورد. جدول ۱، میانگین و انحراف معیار تقریباً یکسانی را بین خصوصیات خاک‌های مورد آزمون و بانک داده مرجع نشان می‌دهد. این امر می‌تواند به بهبود تخمین‌های CEC با هر دو روش غیرپارامتریک مورد مقایسه در این پژوهش کمک نماید. از آن‌جا که الگوریتم روش ANN از تمام داده‌های بانک مرجع و ارتباط بین آن‌ها برای آموزش استفاده می‌نماید، همگنی داده‌ها می‌تواند کمک قابل‌توجهی به بهبود نتایج برآوردها با این روش نماید این شرایط برای روش K-NN که از تعداد محدودتری از داده‌ها که کم‌ترین فاصله را با داده‌های هدف دارند (K)، به مراتب منافع کم‌تری را به‌دنبال خواهد داشت. حق‌وردی و همکاران (۲۰۱۰) در تخمین توابع هیدرولیکی خاک با مقایسه اثر کیفیت و نوع داده‌ها بر عملکرد دو روش ANN و K-NN گزارش کردند که حساسیت روش ANN به ناهمگنی داده‌ها بیش‌تر از روش K-NN بوده به‌طوری‌که همگن نبودن داده‌ها در این پژوهش باعث افزایش ۱۰۰ درصدی خطای برآوردها در این روش شد.

بهینه‌سازی هم‌زمان K و P با استفاده از روش ارزیابی تقاطعی به گونه‌ای انجام شد که به‌ازای هر مقدار P از ۲-۰/۲ (با فاصله‌های ۰/۲) K های مختلف از ۱ تا ٤۰ در نظر گرفته شد و برای هر ترکیب P و K برای هر یک از ۱۲۰ عدد خاک‌های مورد آزمون، برنامه کامپیوتری تهیه شده، یک بار اجرا گردید و مقادیر RMSE به‌دست آمد. شکل ۱ نمودار سه‌بعدی تغییرات RMSE را با تغییرات هم‌زمان p و K نشان می‌دهد. نتایج این بررسی نشان داد که کم‌ترین خطای تخمین CEC در P=۱ و K=۱٤ وجود دارد. نمس و همکاران (۲۰۰٦) نتیجه گرفتند که P می‌تواند اثر منفی کوچک بودن بانک داده مرجع را در عملکرد الگوریتم بهبود بخشد. در حقیقت مقادیر غیر از ۱ برای P، با تأثیر بر وزن‌دهی نقاط سبب تعدیل اثرات منفی کوچک بودن بانک داده مرجع می‌شود. به‌طوری‌که در مقادیر کوچک P (P<۱)، اثر وزن نقاط نزدیک به داده هدف کاهش و با افزایش P به بزرگ‌تر از ۱، این اثر افزایش

می‌یابد. در این پژوهش با توجه به نوع و تعداد خاک‌ها در بانک داده مرجع، کم‌ترین RMSE در P=۱ به‌دست آمد (شکل ۳). بر این اساس رابطه‌های ۵ و ٦ در P=۱ به‌صورت ساده زیر بازنویسی شدند و بهترین وزن‌ها برای خاک‌های K از رابطه زیر به‌دست آمد.

$$w_i = \frac{1/d_i}{\sum_{i=1}^{K} 1/d_i} \qquad (14)$$

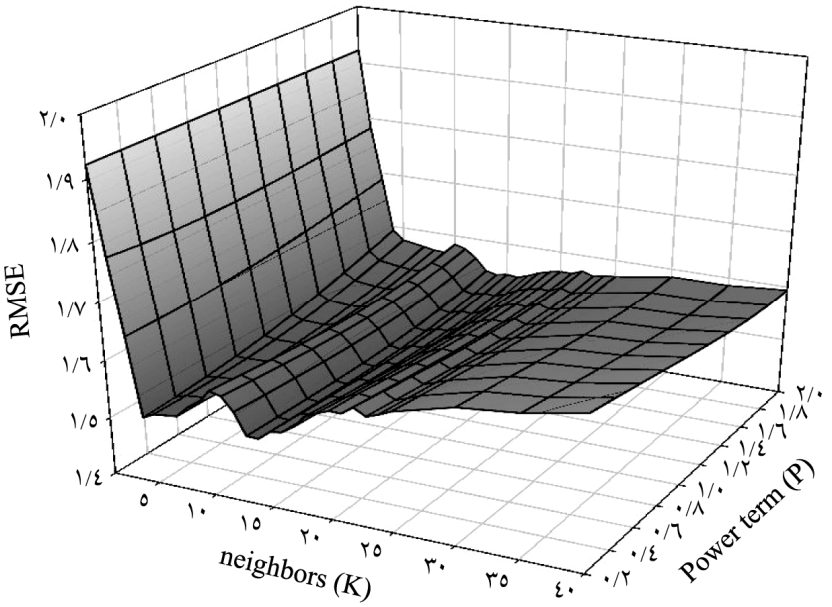
جدول ۱- خلاصه آماری برخی از ویژگی‌های خاک‌های بانک مرجع و هدف.

	خصوصیت	واحد	ماکزیمم	مینیمم	میانگین	انحراف معیار	میانه	ضریب تغییرات (درصد)
بانک داده مرجع (N=٥٦٣)	رس	(درصد)	٦٨/٠٠	٤/٧٢	٢٣/٩٢	١٢/٦٤	١٩/٠٠	٥٢/٨
	سیلت		٧٢/٠٠	٤/٠٠	٤٠/٤٥	١١/١٦	٤٢/٠٠	٢٧/٦
	شن		٩١/٢٨	٥/٣٨	٣٥/٨٦	١٣/٤٨	٣٦/٠٠	٣٧/٦
	کربن آلی		٣/٦٩	٠/٠١	٠/٨٢	٠/٥٨	٠/٥٩	٧٠/٧
	ظرفیت تبادل کاتیونی	$Cmol^{+} kg^{-1}$	٢٩/٤٠	٦/٥٠	١٢/٣٠	٤/٤٢	٩/٧٩	٣٥/٩
داده‌های هدف (N=١٢٠)	رس	(درصد)	٦٤/٢٥	٥/٩٢	٢٣/٢١	١٣/٨٩	١٨/٠٠	٥٩/٨
	سیلت		٧٣/٠٠	٦/٠٠	٤١/٦٣	١١/٠٢	٤٣/٥١	٢٦/٥
	شن		٧٦/٥٦	٤/٧٢	٣٦/٠٤	١٣/٥٦	٣٦/٠٠	٣٧/٦
	کربن آلی		٢/٦١	٠/٠٢	٠/٨٦	٠/٦٠	٠/٥٩	٦٩/٨
	ظرفیت تبادل کاتیونی	$Cmol^{+} kg^{-1}$	٢٣/٢٦	٦/٩٠	١٢/٠٦	٤/٤٥	٩/٥٢	٣٦/٩

N: تعداد داده‌ها است.

شکل ۲ مقادیر به‌دست آمده RMSE را در P=۱ در مقابل K های مختلف نشان می‌دهد، که بیش‌ترین مقدار خطا مطابق انتظار در حالتی که فقط یک خاک به‌عنوان نزدیک‌ترین همسایه برای تخمین‌ها استفاده شد (K=۱) به‌دست آمد و کم‌ترین آن در حالت بهینه در K=١٤ وجود داشت و با

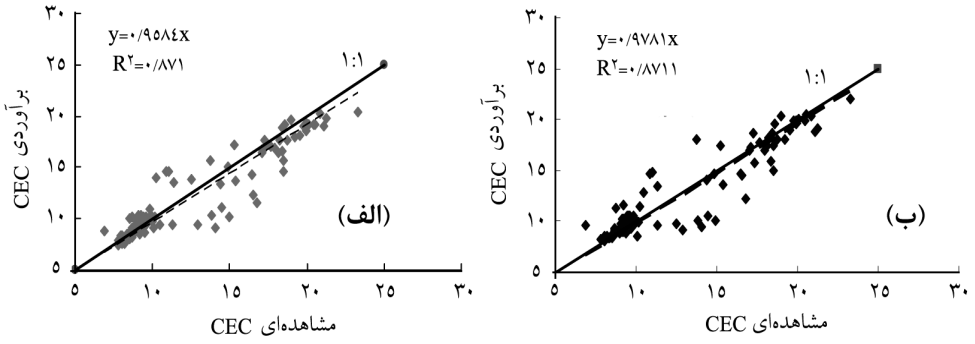
افزایش مقدار K به بیش از ١٤ خطای برآوردها دوباره افزایش یافت. در این پژوهش مقدار K به‌دست آمده از طریق ارزیابی تقاطعی هم‌خوانی بسیار خوبی با نتایج نمس و همکاران (٢٠٠٦) نشان می‌دهد (رابطه ٤) و با آن‌چه لال و شارما (١٩٩٦) توصیه نمودند، مغایرت داشت ($K = ٥٦٣^{٠/٥} \approx ٢٤$). شکل ٣ نیز مقادیر خطای برآورد را در مقابل P های مختلف ارایه می‌کند. با توجه به تغییرات بسیار اندک RMSE (در حدود ٠/٠٣)، این شکل بیان می‌کند که روش K-NN نسبت به تغییرات P حساسیت پایینی را در محدوده خاک‌های مورد مطالعه نشان می‌دهد و بنابراین انتخاب هر یک از مقادیر P در محدوده اعداد نزدیک به ١ تفاوت قابل‌توجهی را در نتایج تخمین‌ها ایجاد نمی‌کند. البته تعداد خاک‌های موجود در بانک داده مرجع عامل مهمی در تعیین P بهینه می‌باشد. به‌عبارت دیگر استفاده از رابطه ساده (١٤) به‌منظور وزن‌دهی خاک‌های K در بانک داده‌های مشابه می‌تواند نتایج قابل‌قبولی را ارایه نماید. نمس و همکاران (٢٠٠٦) در برآورد توابع هیدرولیکی با ٥ اندازه مختلف بانک داده مرجع مقادیر بین ١/١–٠/٩٥ را برای P بهینه به‌دست آوردند.



شکل ١- رابطه مجذور میانگین مربعات خطا (RMSE)، با تغییرات هم‌زمان K و P در یک نمودار سه‌بعدی.

عل

CEC پیش‌بینی شده در مقابل مقادیر مشاهده‌ای این کمیت در روش شبکه عصبی مصنوعی (الف) و K نزدیک‌ترین همسایگی (ب) ترسیم شده است. همان‌طورکه مشاهده می‌شود شیب معادله‌های خطوط برازش داده شده بر این نقاط تفاوت بسیار اندکی را با شیب خط ۱:۱ ترسیمی نشان می‌دهد. در این‌جا ضریب تبیین بالا (۰/۸۷) با شیب نزدیک به ۱، بی‌تردید بیانگر دقت بالا و یکسان هر دو روش در تخمین CEC خاک‌های مورد مطالعه می‌باشد. ضریب همبستگی بالا و بسیار معنی‌دار ۰/۹٤ (در سطح احتمال ۹۹ درصد) نشان‌دهنده روند یکسان تغییرات مقادیر برآوردی و مشاهده‌ای در روش K-NN بود که از این نظر هم تفاوتی با روش ANN نشان نداد. به‌منظور مقایسه عملکرد دو روش ANN و K-NN آماره‌های RMSE، MAE، ME، MR و EF را نیز مورد بررسی قرار گرفت. جدول ۲ این مقادیر را برای دو روش مورد آزمون نشان می‌دهد. نتایج نشان داد که در روش K-NN با K=۱٤ و P=۱ در بدترین حالت، CEC را با ٤/۸۱ $cmol^+/kg$ خطا (ME)، پیش‌بینی نمود. این شرایط در روش ANN با تخمین ضعیف‌تر و میزان خطای ۵/۲٦ $cmol^+/kg$ همراه بود. مقادیر کوچک آماره‌های RMSE و MAE در هر دو روش K-NN و ANN نشان داد که این دو روش در برآورد CEC خاک‌های مورد آزمون از دقت بالا و یکسانی برخوردارند. اگرچه این مقدار در روش K-NN از نظر عددی اندکی بهتر از روش ANN بود. مقادیر مثبت آماره MR برای این دو روش نیز نشان داد که هر دوی آن‌ها متمایل به برآورد کم‌تر مقدار پیش‌بینی شده نسبت به داده‌های مشاهده‌ای (Under estimation) می‌باشند. همچنین مقایسه مقادیر EF نشان داد که هر دو روش دارای کارایی بالایی (EF=۸۸ درصد) در برآورد ظرفیت تبادل کاتیونی از روی توزیع اندازه، درصد کربن آلی خاک می‌باشند و هیچ اختلافی از این نظر بین دو روش مشاهده نمی‌شود.



شکل ٤- رابطه بین مقادیر پیش‌بینی و مشاهده‌ای CEC با استفاده از دو روش ANN (الف) و K-NN (ب).

جدول ۲- آماره‌های مقایسه عملکرد دو روش K نزدیک‌ترین همسایگی (K-NN) و شبکه عصبی مصنوعی (ANN).

روش	r	RMSE	MAE	MaxE	ME	EF
K-NN	۰/۹٤	۱/۵۱	۰/۹۲	٤/۸۱	۰/۱	۸۸
ANN	۰/۹٤	۱/۵۳	۱/۰۰	۵/۲٦	۰/۱۸	۸۸

ریشه میانگین مربعات خطا (RMSE)، میانگین قدرمطلق خطا (MAE)، بیش‌ترین خطا (MaxE)، میانگین خطا (ME) بر حسب $cmol^+/kg$ و کارایی مدل (EF) بر حسب درصد می‌باشد.

نتیجه‌گیری

از آن‌جا که تاکنون در ایران هیچ پژوهشی با هدف مقایسه عملکرد دو روش غیرپارامتریک K نزدیک‌ترین همسایگی (K-NN) و شبکه عصبی مصنوعی (ANN) بر روی CEC خاک‌های کشور گزارش نشده است، ضرورت چنین بررسی مورد توجه نویسندگان قرار گرفت. بر این اساس در این پژوهش قابلیت دو روش نام برده در برآورد CEC خاک از روی ویژگی‌های زودیافتی مانند بافت و درصد کربن آلی خاک مورد ارزیابی و مقایسه قرار گرفت. بررسی ضرایب همبستگی در هر دو روش مورد مقایسه، هم‌روندی بالا و بسیار معنی‌داری ($P<۰/۰۱$) را بین مقادیر مشاهده‌ای و برآوردی CEC خاک‌ها نشان داد. در مدل K-NN که با استفاده از تشابه ۱٤ عدد از نزدیک‌ترین خاک‌های بانک مرجع به هدف، مقدار CEC را پیش‌بینی می‌کرد، بین مقادیر CEC برآوردی و اندازه‌گیری شده (نسبت به خط ۱:۱) ضریب تبیین (R^2) ۰/۸۷ مشاهده شد که به روشنی توانایی و دقت بسیار بالای این روش را در تخمین CEC خاک از روی ویژگی‌های یاد شده نشان می‌دهد. این دقت در تخمین CEC خاک‌ها در روش نام برده، با مدل ANN که از ٥٦۳ داده بانک مرجع برای آموزش استفاده می‌نمود، یکسان بود، به‌طوری‌که مقادیر RMSE محاسبه شده برای روش‌های K-NN و ANN به‌ترتیب ۱/۵۱ و ۱/۵۳ ($cmol^+/kg$) به‌دست آمد. مقدار مثبت میانگین خطاهای (ME) محاسبه شده نیز نشان داد که به‌طور کلی هر دو روش، مقدار CEC را کم‌تر از مقدار مشاهده‌ای این کمیت تخمین زده‌اند. کارایی (EF) ۸۸ درصدی برای هر دو روش، قابلیت بالا و یکسان آن‌ها را در تخمین CEC خاک‌ها نشان می‌دهد. بر این اساس و با توجه به کاربرد آسان روش K-NN در پیش‌بینی CEC خاک‌ها به‌خصوص زمانی‌که داده‌های موجود همگن نباشند و یا تعداد این داده‌ها برای آموزش نورون‌های شبکه عصبی کافی نباشد، استفاده از روش K-NN توصیه می‌شود.

منابع

1.Amini, M., Abbaspour, K.C., Khademi, H., Fathianpour, N., Afyuni, M., and Schulin, R. 2005. Neural Network Models to Predict cation exchange capacity in arid regions of Iran. Eur. J. Soil Sci. 56: 551-559.

2.Bannayan, M., and Hoogenboom, G. 2009. Using pattern recognition for estimating cultivar coefficients of a crop simulation model. Field Crop. Re. 111: 290-302.

3.Bell, M.A., and Van Kulen, H. 1995. Soil pedotransfer function for four Mexican soils. Soil. Sci. Soc. Am. J. 59: 865-871.

4.Carrow, R.N., Waddingtone, D.V., and Rieke, P.E. 2001. Turfgrass soil fertility and chemical problems. John Wiley and Sons, New York.

5.Chapman, H.D. 1965. Cationexchange capacity. Methods of Soil Analysis. Part 2. Chemical and microbiological properties. ASA-CSSA-SSSA Publisher, Madison. Wisconsin, USA.

6.Clark, M.P., Gangopadhyay, S., Brandon, D., Werner, K., Hay, L., Rajagopalan, B., and Yates, D. 2004. A resampling procedure for generating conditioned daily weather sequences. Water Resour. Res. 40. W04304 doi. 10.1029/2003 WR002747.

7.Frate, F.D., Ferrazoli, P., and Schiavon, G. 2003. Retrieving soil moisture and agricultural variables by microwave radiometry using neural network. Remote Sens. Environ. 84: 174-183.

8.Gee, G.W., and Bauder, J.W. 1986. Particle- size analysis. Methods of Soil Analysis Part 1. Physical and mineralogical methods, ASA-CSSA-SSSA Publisher, Madison. Wisconsin, USA.

9.Gjertsen, A.K. 2007. Accuracy of forest mapping based on Landsat TM data and a kNN-based method. Remote Sens. Environ. 110: 420-430.

10.Haghverdi, A., Ghahraman, B., Khoshnood Yazdi., A.A., and Arabi, Z. 2010. Estimating of water content in FC and PWP in north and north east of Iran's soil samples using k-nearest neighbor and artificial neural networks. J. Water Soil. 24: 4. 804-814. (In Persian)

11.Jalali, V.R., and Homaee, M. 2011. A nonparametric model by using k-nearest neighbor technique for predicting soil saturated hydraulic conductivity. J. Water Soil. 25: 2. 347-355. (In Persian)

12.Karimian, N.A. 1998. Clay and organic matter contribution in cation exchange capacity of calcareous soils from Fars province. Proceedings of 5th Iranian soil congress, Karaj, Iran. (In Persian)

13.Krogh, L., Madsen, H.B., and Greve, M.H. 2000. Cation exchange capacity pedotransfer functions for Danish soils. Acta Agric. Scand., Sect. B, Soil and Plant Sci. 50: 1-12.

14.Lall, U., and Sharma, A. 1996. A nearest-neighbor bootstrap for resampling hydrologic time series. Water Resour. Res. 32: 679-693.

15.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 Sens. Environ. 77: 251-274.
16.Mc Bratney, A.B., Minasny, B., Cattle, S.R., and Vervoort, R.W. 2002. From pedotransfer functions to soil inference systems. Geoderma. 109: 41-73.
17.Memarian Fard, M., and Beigi Harchagani, H. 2009. Comparison of artificial neural network and regressionpedotransfer functions models for prediction of network and regression pedotransfer functions models for prediction of J. Water Soil. 23: 4. 90-99. (In Persian)
18.Nelson, D.W., and Sommers, L.E. 1982. Total carbon, organic carbon, and organic matter. Methods of Soil Analysis. Part 2. Chemical and microbiological properties. ASA-CSSA-SSSA Publisher, Madison. Wisconsin, USA.
19.Nemes, A., Rawls, W.J., and Pachepsky, Ya.A. 2006. Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. Soil Sci. Soc. Am. J. 70: 327-336.
20.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. Environm Modell. Softw. 23: 254-255.
21.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. Sci. Soc. Am. J. 73: 1638-1645.
22.Pachepsky, Ya.A., Timlin, D., and Varallyay, G. 1996. Artificial neural networks to estimate soil water retention from easily measurable data. Soil Sci. Soc. Am. J. 6: 727-733.
23.Schaap, M.G., and Bouten, W. 1996. Modeling water retention curves of sandy soils using neural networks. Water Resour. Res. 32: 3033-3040.
24.Schaap, M.G., and Leij, F.L. 1998. Database-related accuracy and uncertainty of pedotransfer functions. Soil Science, 10: 765-779.
25.Seybold, C.A., Grossman, R.B., and Reinsch, T.G. 2005. Predicting cation exchange capacity for soil survey using linear models. Soil Sci. Soc. Am. J. 69: 856-863.
26.Tamari, S., Wosten, J.H.M., and Ruize, J.C. 1996. Testing an arifical neural networks for predicting soil hydraulic conductivity. Soil. Sci. Soc. Am. J. 60: 1732-1741.
27.Twarakavi, , J., and Schaap, M.G. 2009. Development of pedotransfer functions for estimation of soil hydraulic parameters using support vector machines. Soil. Sci. Soc. Am J. 73: 1443-1452.
28.Yakowitz, S. 1993. Nearest-neighbor estimation for null-recurrent Markov time series. Stoch. Proc. Appl. 48: 311-318.
29.Yates, D., Gangopadhyay, S., Rajagopalan, B., and Strzepek, K. 2003. A technique for generating regional climate scenarios using a nearest-neighbor algorithm. Water Resour. Res. 39:1199 doi 10.1029/2002WR001769.

J. of Soil Management and Sustainable Production, Vol. 3(1), 2013
http://ejsms.gau.ac.ir

Comparison of K-nearest neighbor and artificial neural network methods for predicting cation exchange capacity of soil

***A.A. Zolfaghari[1], M.T. Tirgar Soltani[2], T. Afshari Badrloo[3] and F. Sarmadian[4]**

[1]Ph.D. Student, Dept. of Soil Science Engineering, University of Tehran,
[2]M.Sc., Dept. of Soil Science Engineering, University of Tehran,
[3]M.Sc. Student, Dept. of Soil Science Engineering, Shahrood University of Technology,
[4]Associate Prof., Dept. of Soil Science Engineering, University of Tehran
Received: 05/29/2012; Accepted: 12/15/2012

Abstract

Cation exchange capacity (CEC) measurement is a very expensive and time-consuming method in large scale assessments. It can be an appropriate approach to predict CEC from readily available properties via developing nonparametric models. In the present study, a nonparametric technique has been used for estimating CEC and compared with the most common nonparametric models which is based on artificial neural networks (ANN). 683 soils were selected from central Iran that 120 of them were used as target data and the others (563) were the reference data set. The parameters, clay, silt, sand and organic carbon content were the input independent variables (readily available properties) and the CEC was as an output dependent variable in this work. The results showed that the maximum error (MaxE) in K-NN and ANN techniques were 4.81 and 5.26 $cmol^{+}/kg$, respectively. Root mean squared error (RMSE) for the K-NN and ANN were 1.51 and 1.53, respectively. This indicated that both methods are able to properly and equally predict CEC. The positive values of mean error (ME) showed that both models tended to underestimate CEC values in samples. The results analysis also showed that the efficiency of models (EF=0.88) were high by the estimation of CEC values in target soils.

Keywords: Nonparametric models, Cation exchange capacity, K-nearest neighbor, Artificial neural network model

* Corresponding Authors; Email: azolfaghari@ut.ac.ir