نشریه آب و خاک (علوم و صنایع کشاورزی)
جلد ۳۰، شماره ۳، مرداد – شهریور ۱۳۹۶، ص. ۹۲۸-۹۱۵

Journal of Water and Soil
Vol. 30, No. 3, Jul.-Agu. 2017, p. 915-928

تعیین روش بهینه پیش‌پردازش داده‌ها به منظور افزایش دقت شبیه‌سازی‌های شوری خاک سطحی (مطالعه موردی: منطقه مروست)

اعظم حبیبی پور[1*] – علی طالبی[2] – علی اکبر کریمیان[3] – فرهاد دهقانی[4] – محمد حسین مختاری[5]

تاریخ دریافت: ۱۳۹۵/۰۳/۲۲
تاریخ پذیرش: ۱۳۹۵/۱۱/۲۴

چکیده

در این تحقیق تأثیر روش‌های پیش‌پردازش داده‌های شوری درافزایش دقت شبیه‌سازی‌های صورت گرفته توسط الگوریتم درخت تصمیم در منطقه مروست مورد بررسی قرار گرفته است. به این منظور شبیه‌سازی‌ها در سه حالت (استفاده از داده‌های اصلی، استفاده از لگاریتم داده‌ها و استفاده از داده‌های استاندارد شده) صورت گرفت. نتایج نشان داد علیرغم معنی دار بودن ضریب همبستگی در هر سه حالت، میزان خطا در حالت استفاده از لگاریتم داده‌ها نسبت به دو حالت دیگر کمتر بوده و نتایج به واقعیت نزدیکتر می‌باشد. به طوری که این حالت (استفاده از لگاریتم داده ها)، درخت ایجاد شده قادر است با ترکیب "باند ۷، ارتفاع" و استفاده از ۵ قانون، میزان شوری خاک سطحی را برآورد نماید. با توجه به اینکه توزیع احتمالاتی حاکم بر داده‌های شوری منطقه، یکی از توزیع های خانواده لگاریتم(Log-Pearson 3) می‌باشد می‌توان اظهار داشت، کاهش خطا در حالت استفاده از لگاریتم داده‌ها در ارتباط نزدیک با توزیع احتمالاتی حاکم بر داده‌های شوری منطقه مورد بررسی در این تحقیق می‌باشد. لذا شبیه سازی با استفاده از لگاریتم داده‌ها به دلیل خطای کمتر و نیازمندی به داده‌های ورودی کمتر به عنوان مدل برتر شناخته شد.آماره های خطای R، Rmse، %Rmse، MAE وBias در این حالت ۰/۷۶، ۰/۴۹، ۳۸/۵۸، ۰/۳۷ و ۰/۱۴- بدست آمد.

واژه‌های کلیدی: بهنجارسازی داده‌ها، توزیع احتمالاتی، درخت تصمیم رگرسیونی، هدایت الکتریکی

مقدمه

شوری یکی از مشکلات خاک های مناطق خشک و نیمه خشک می‌باشد که شدت آن در اثر فرآیندهای طبیعی و فعالیت های انسانی رو به افزایش است. بیش از یک سوم خاک های شور دنیا و بخش اعظم خاک های ایران در مناطق خشک و نیمه خشک واقع شده است. از طرفی شناسایی و طبقه بندی خاک های شور به منظور مقابله با شرایط سخت و اعمال مدیریت صحیح امری ضروری است. روش های متعددی برای براورد میزان شوری خاک سطحی مورد استفاده محقیقن قرار گرفته است. مطالعه شوری خاک با استفاده از مطالعات میدانی بسیار پرهزینه است. لذا استفاده از تصاویر ماهواره ای و اطلاعات باندهای اصلی و مصنوعی آن‌ها رایج تر می‌باشد (۱، ۱۸، ۲۱). در همین ارتباط در مطالعات مختلف شاخص‌های متفاوتی به کار گرفته شده و نتایج متنوعی هم به دست آمده است. به طوری که خواجه‌الدین (۱۴) از داده‌های چندزمانه سنجنده MSS برای بررسی ارتباط بین داده‌های استخراج شده از تصاویر ماهواره ای با برخی خصوصیات خاک در منطقه جازموریان استفاده کرده است. وی رابطه مناسبی بین داده‌های سنجنده MSS و فاکتورهای مختلف اندازه گیری شده خاک نظیر درصد شن، سیلت، رس، درصد پوشش سنگ و سنگ ریزه، کلسیم و پتاسیم پیدا نکرد ولی رابطه مناسبی بین شوری و غلظت سدیم با داده‌های تصاویر ماهواره ای ماه جولای به دست آورد. دویودی (۱۰) به بررسی خصوصیات خاک های متأثر از شوری در دشت های آبرفتی منطقه اوتارپرادش هندوستان پرداخت. او با استفاده از داده‌های سنجنده MSS مربوط به سال‌های ۱۹۷۵ و ۱۹۹۲ و تکنیک آنالیز مؤلفه های اصلی PCA و نسبت گیری طیفی به این نتیجه رسید که مؤلفه سوم PCA حاصل از تمامی باندها و نسبت گیری طیفی بین MSS1 وMSS2 اطلاعات کاملی درباره خاک های متأثر از شوری ارائه خواهد داد. چیت ساز (۸) به تهیه نقشه شوری و قلیائیت خاک با استفاده از داده‌های

۱، ۲، ۳ و ۵- به ترتیب دانشجوی دکترای علوم و مهندسی آبخیزداری، دانشیار، استادیاران دانشکده منابع طبیعی و کویر شناسی، دانشگاه یزد
(*–نویسنده مسئول: Email: AzamHabibipoor@yahoo.com)
۴- استادیار مرکز ملی تحقیقات شوری
DOI: 10.22067/jsw.v31i3.55462

سنجنده TM در شمال شرقی اصفهان پرداخت. نتایج، مدل رگرسیونی مناسبی را ارائه داد که با استفاده از اطلاعات باندهای ۴، ۵ و ۶ ، تغییرات شوری نمونه‌های خاک سطحی منطقه را برآورد می‌نمود. نائیج نوری (۱۶) به تفکیک اراضی شور و گچی با داده‌های رقومی TM در دشت کاشان پرداخت و با استفاده از ضریب همبستگی پیرسون و آنالیز رگرسیونی نشان داد حداکثر همبستگی بین باندهای TM1, TM5،TM6 با مقادیر شوری سطحی و باندهای TM3، TM6،TM5 و TM7 با تغییرات گچ وجود دارد. جعفری گرزین (۱۳) برای برآورد شوری خاک سطحی از داده‌های رقومی ETM^+ ماهواره لندست ۷ استفاده کرد او باندهای ETM2، Brightness ،TM3 ،TMS135 ،ETM3P4 و SRVI1 را به عنوان مؤثرترین باندهای اصلی و مصنوعی در برآورد شوری معرفی کرده است. عبدی نام (۲) برای تهیه نقشه شوری خاک دشت قزوین از ایجاد همبستگی بین داده‌های ماهواره ای با مقادیر شوری خاک استفاده و به دلیل وجود همبستگی بالای ارقام رقومی باند ETM7 با مقادیر شوری، از داده‌های این باند استفاده کرد. الدیری و همکاران (۱۲)، به منظور تهیه نقشه شوری خاک و بررسی پایش شوری در بخشی از دره آرکانساز از داده‌های IKONOS استفاده کردند. نتایج نشان داد که باند سبز و مادون قرمز نزدیک برای مطالعات شوری خاک مناسب‌تر است. احمدیان و پاکپرور (۴)، به منظور بررسی روند تغییرات شوری خاک در دشت قهاوند (استان همدان) از دو سری اطلاعات داده‌های رقومی ماهواره لندست ۵ و۷ مربوط به سال‌های ۱۹۸۹ و ۲۰۰۰ میلادی کمک گرفتند. پژوهش، روی شاخص‌های مختلف PC57، PC1234، NDVI، Grenes وBrenes صورت گرفت. نتایج نشان دهنده افزایش ۳۱/۹ درصدی وسعت خاک‌های شور منطقه در طول ۱۱ سال مورد بررسی بوده است. بوسس و همکاران (۷) برای نقشه برداری شوری خاک در حوالی تکسکوکو مکزیک از داده‌های رقومی ETM^+ و عکس‌های هوایی استفاده کردند. در این میان با تعدیل کردن شاخص پوشش گیاهی (NDVI) شاخص طیفی جدیدی به نام Cosri تهیه نمودند. همبستگی بسیار بالای بین شوری و قلیائیت خاک با ارزش طیفی این باند ترکیبی (به ترتیب ۰/۸۸۵ و۰/۸۵۷) به صورت یک مدل رگرسیونی برای تهیه نقشه شوری خاک ارائه شد. ریورو و همکاران (۱۷) داده‌های طیفی ASTER۷ و ETM را در تهیه نقشه برخی خصوصیات خاک مورد استفاده قرار دادند. علیرغم وجود دامنه تغییرات زیاد و تفاوت زیاد مقادیر ماکزیمم و مینیمم در داده‌ها، نتایج رضایتبخش بود. دشتکیان و همکاران (۹)، نقشه شوری خاک را با استفاده از داده‌های ماهواره‌ای لندست و داده‌های میدانی در منطقه مروست تهیه و مورد بررسی قرار دادند. نتایج نشان داد روش میانگین رگرسیون‌ها با استفاده از اطلاعات استاندارد شده باندهای ۱، ۲ و ۳، مناسب‌ترین روش برای تهیه نقشه شوری خاک منطقه می‌باشد. تاج گردان و همکاران (۱۹) به تهیه نقشه شوری خاک سطحی با استفاده از داده‌های دورسنجی ETM^+ در منطقه آق قلای استان گلستان پرداختند و رابطه رگرسیونی ارائه دادند که می‌توانست با استفاده از باند اصلی ۴، مولفه‌های روشنایی و سبزینگی حاصل از تسلد کپ و مولفه حاصل از ادغام باند پانکروماتیک با باند اصلی به برآورد مقادیر شوری خاک سطحی در منطقه بپردازد.

همانطور که در مرور منابع هم به تفصیل بیان شده، تصاویر ماهواره ای مختلفی با قدرت تفکیک متفاوت وجود دارد. ولی ممکن است به دلایل مختلف (نظیر تحریم‌ها و بسته بودن سایت‌های دانلود تصویر، خراب بودن سنجنده ماهواره در تاریخ نمونه برداری صحرایی و غیرو) دستیابی به آن‌ها در تاریخ نمونه برداری صحرایی میسر نباشد. از طرفی روش‌های مختلفی هم برای به دست آوردن اطلاعات و خروج اطلاعات از هر پیکسل این تصاویر وجود دارد، به طوری که می‌توان ارزش پیکسلی باندهای اصلی را استخراج و از آن‌ها استفاده کرد و یا عملیات ریاضی (جمع، ضرب، تقسیم، آنالیز مولفه‌های اصلی و...) را بر روی تصویر اعمال نمود و سپس ارزش مصنوعی هر پیکسل را استخراج کرد. لذا محققین مختلف از تصاویر مختلف و پارامترهای مختلف استفاده کرده اند و نتایج هم متفاوت بوده است و تاکنون دستیابی به یک مدل جامع که بتواند شوری خاک سطحی را با استفاده از یک یا چند شاخص پیش بینی نماید میسر نشده است و در هر منطقه بسته به شرایط محیطی، پارامترهای مؤثر در شبیه سازی متغیر می‌باشد. ولی به هر صورت این مطالعات از این نظر ارزشمند است که می‌تواند میزان نمونه برداری های میدانی را کاهش دهد و در نتیجه در وقت و هزینه صرفه جویی خواهد شد. به عبارتی می‌توان با تعداد نمونه صحرایی کمتر و تلفیق آن با اطلاعات تصاویر ماهواره ای نقشه شوری خاک را با دقت قابل قبول ارائه داد.

از طرفی اکثر روش های هوش مصنوعی مثل شبکه عصبی مصنوعی جعبه سیاه هستند یعنی متغیرهای ورودی توسط کاربر تعیین و به مدل معرفی می‌گردند و مشخص نیست که رابطه بین پارامترهای مستقل و پارامتر وابسته چگونه حاصل می‌گردد. لذا باید در انتخاب نوع و تعداد ورودی‌ها دقت نمود. ولی الگوریتم درخت تصمیم، نیمه هوشمنداست و می‌تواند از بین پارامترهای ورودی مدل، موثرترین‌ها را انتخاب کند. از این رو استفاده از الگوریتم‌هایی نظیر درخت تصمیم که می‌تواند به صورت نیمه هوشمند، پارامترهای مؤثرتر در فرایند شبیه سازی را انتخاب نماید دقت شبیه سازی‌ها را بالا خواهد برد. به عبارتی کارکردن با داده‌های بزرگ و پیچیده یکی از مزایای مهم درخت تصمیم است. این الگوریتم در عین سادگی می‌تواند با داده‌های پیچیده و بزرگ به راحتی کار کند و از روی آنها قانون و رابطه بسازد و مشخص نماید که چگونه می‌توان از این متغیرهای مستقل به متغیر وابسته رسید. علاوه بر آن توجه به ماهیت داده‌های شوری و انتخاب روش های مناسب پیش‌پردازش بر روی

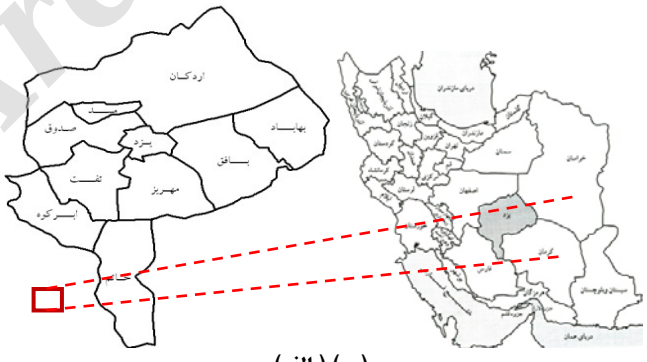
داده ها قبل از ورود به مدل‌های هوش مصنوعی از جمله مواردی است که می‌تواند نتایج حاصل از شبیه سازی ها را به واقعیت نزدیک سازد. روش های مختلفی برای پیش پردازش داده ها وجود دارد. این موضوع به ویژه زمانی اهمیت بیشتری می‌یابد که از چند پارامتر با ماهیت و بازه تغییرات مختلف برای شبیه سازی استفاده شود. هدف تمامی این روش ها اعمال عملیاتی بر روی داد ه های اصلی است تا داده‌ها را به داده هایی جدید با بازه تغییرات جدید یا توزیع مناسب تبدیل کند. به عبارتی این روش ها، مقیاس گذاری داده ها را تغییر می‌دهد. بعضی از این روش‌ها تبدیل خطی انجام می دهند. در این تبدیل تابع توزیع داده‌ها تغییر نمی‌کند. در مواردی که توزیع داده ها حول میانگین شان یک دست نباشد می توان از تبدیل‌هایی بر اساس توابع غیر خطی استفاده نمود. انتخاب روش مناسب بهنجار کردن (پیش پردازش) به نوع مساله برمی گردد و برای هر کابردی متفاوت است (۱۵). این پژوهش تلاش دارد تأثیر روش‌های پیش‌پردازش داده های شوری را در افزایش دقت شبیه سازی‌های صورت گرفته توسط الگوریتم درخت تصمیم را در منطقه مروست مورد بررسی قرار دهد. لازم به ذکر است با برازش داده‌های شوری در قالب یک تابع توزیع فراوانی می توان تا حدود زیادی رفتار پدیده شوری را شناسایی کرد. به این ترتیب بهترین تابع توزیعی که بر داده های شوری برازش می‌یابد رفتار پدیده شوری در منطقه را نشان می دهد. لذا فرض شده توزیع‌های احتمالاتی حاکم بر داده‌های شوری نقش مؤثری در انتخاب بهترین روش پیش‌پردازش داده ها قبل از شبیه سازی خواهد داشت. هدف این تحقیق، تعیین روش بهینه پیش‌پردازش داده ها یا ورودی های الگوریتم درخت تصمیم رگرسیونی است که می تواند نقش پیش پردازش داده‌ها را در افزایش کارایی درخت تصمیم در برآورد شوری خاک سطحی نشان دهد. پیش پردازش داده ها عملیاتی را بر روی داده‌های اصلی اعمال می کند که در آن داده‌ها، به داده‌هایی جدید با بازه تغییرات جدید یا توزیع مناسب تبدیل می شود. به این ترتیب با تغییر در روش پیش پردازش، مقیاس وتوزیع داده‌های ورودی تغییر می کند و این موضوع موجب تغییر در دقت شبیه سازی ها می شود. این موضوع زمانی اهمیت بیشتری می یابد که از چند پارامتر با ماهیت و بازه تغییرات مختلف برای شبیه سازی متغیر وابسته استفاده شود.

مواد و روش‌ها

- معرفی منطقه مورد مطالعه

منطقه مورد مطالعه، بخش اعظم دشت مروست واقع در استان یزد به مساحت ۸۸۹۸۲ هکتار می‌باشد که در طول جغرافیایی ۷۸۸۸۰۷ تا ۸۱۹۱۷۷ و عرض جغرافیایی ۳۳۵۰۸۶۵ تا ۳۳۷۷۴۳۲ (سیستم مختصات UTM زون ۳۹R) قرار دارد (شکل ۱). این محدوده، شهر مروست و مراتع مجاور آن را شامل می‌شود. موقعیت نقاط نمونه‌برداری خاک در شکل ۲ نمایش داده شده است.

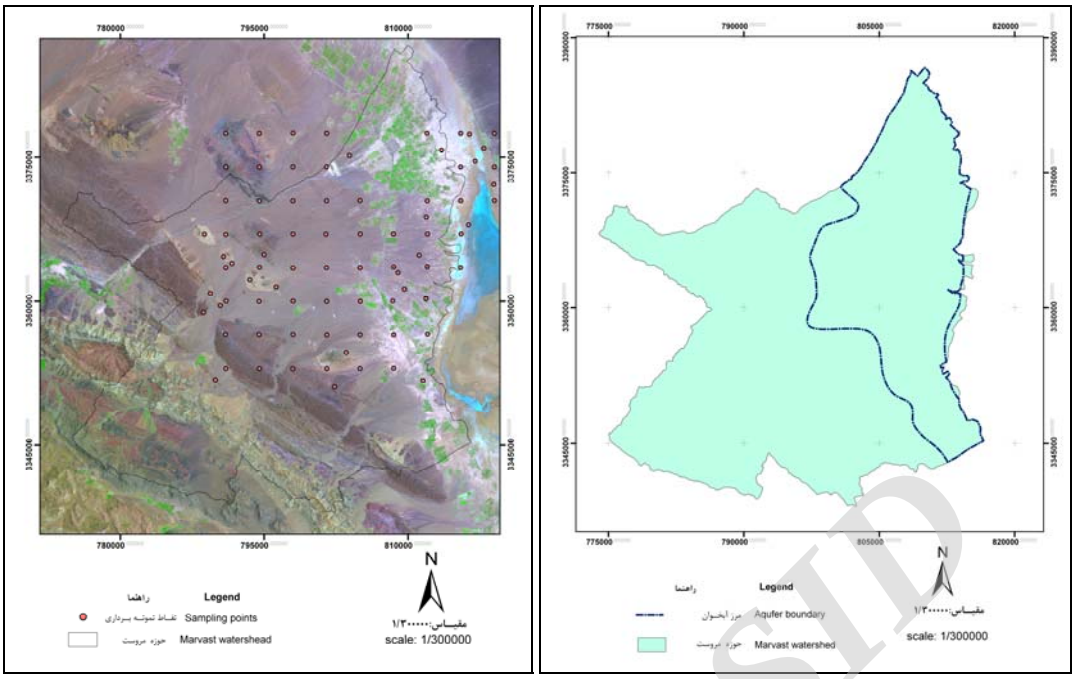
منطقه مورد مطالعه از غرب به ارتفاعات هم مرز با استان فارس و در شرق قسمتی از کویر مروست را در بر می‌گیرد. گودترین نقطه، واقع در کویر مروست با ۱۵۱۵ متر از سطح دریا و بلندترین نقطه آن مربوط به ارتفاعات جنوب غربی محدوده با ۲۰۶۰ متر از سطح دریا است. وجود ارتفاعات، انواع دشت سرها و کویر در این منطقه، موجب شده تا میزان شوری خاک منطقه بازه تغییرات وسیعی داشته باشد. میانگین دمای سالانه این منطقه ۱۸/۳ درجه سانتیگراد و میانگین بارندگی ۲۰ ساله در ایستگاه هواشناسی مروست ۷۷ میلی‌متر می‌باشد. اقلیم منطقه مورد پژوهش بر اساس روش دومارتن اصلاح شده خشک سرد تا فراخشک سرد می‌باشد.



(ب) (الف)

شکل۱- موقعیت منطقه مورد مطالعه در الف) ایران، ب) استان یزد.

Figure 1. Position of the studied area in a)Iran b)Yazd province.



(ج) (د)

شکل ۲- حوزه آبخیزمروست (ج) و موقعیت نقاط نمونه‌برداری در محدوده مورد مطالعه (د)

Figure 2. Marvast watershed and position of sampling point in the studied area.

- روش مطالعه

در این تحقیق از الگوریتم CART (Classification and regression tree)به عنوان یکی از انواع درختان تصمیم رگرسیونی جهت «برآورد شوری خاک» استفاده شد. درخت تصمیم یکی از پرکاربردترین روش هایی است که برای استنتاج استقرایی مورد استفاده قرار می گیرد. درخت تصمیم یک روش غیر پارامتریک با ساختار سلسله مراتبی داده و یادگیری نظارت شده است که با استفاده از استراتژی تقسیم و تفکیک داده، پیاده سازی می شود. در این روش، تقسیم بندی داده ها با استفاده از ویژگی های آنها به صورت یک درخت پیاده سازی می‌شود و گاهی برای خوانایی بیشتر به صورت قوانین اگر – آنگاه بیان می‌شود. این روش، بر اساس داده های آموزش در هر مرحله یکی از ویژگی های داده را انتخاب می کند و داده های آن مجموعه را به دو یا تعدادی بیشتر تقسیم بندی می کند و این کار را تا زمانی ادامه می دهد تا تمام داده های موجود در یک دسته دارای یک برچسب والد باشد. ساخت این درختان بر سه اصل استوار است:

۱-مجموعه‌ای از سؤالات به شکل ?x≤d که در آن x یک متغیر مستقل و d یک مقدار ثابت است وجواب هر سؤال بله/خیر است.

بهترین معیار شاخه زدن جهت انتخاب بهترین متغیرمستقل برای ایجاد شاخه.

ایجاد آمار خلاصه برای گره انتهایی(۶).

معیارهای مختلفی جهت ایجاد شاخه و تولید درخت تصمیم وجود دارد، ولی از آنجا که تحقیق حاضر به استفاده از درخت تصمیم رگرسیونی پرداخته، معیار مورد استفاده در این مدل که انحراف حداقل مربعات (LSD)[۳] نام دارد تشریح می‌گردد. این معیار به صورت زیر تعریف می‌شود:

$$SS(t) = \sum_{i=1}^{N_t} (y_i(t) - \bar{y}(t))^2 \quad (۱)$$

N_t : تعداد رکوردها (داده‌ها) در گره برگt .

$y_i(t)$: مقدار خروجی (متغیر هدف در گره برگ).

$\overline{y}_{(t)}$: میانگین مقادیر متغیر هدف برای همه گره‌ها.

حال متغیر ورودی S زمانی بهترین متغیر برای ایجاد شاخه در گره t می‌باشد که مقدار Q(s,t) را بیشینه نماید. به عبارتی وقتی کل داده ها در گره t قرار دارد، متغیر مستقلی برای تقسیم داده ها به دو دسته مجزا انتخاب می‌شود که مقدار خطای شاخه سمت راست و چپ را به کمترین مقدار خود برساند. این کار تا انتها به همین ترتیب ادامه می‌یابد تا داده ها در دسته های مجزا و همگن طبقه بندی شوند. آنگاه قوانین دخت تصمیم از روی همین طبقه بندی ایجاد می‌شود و درخت می‌تواند با دریافت داده های جدید، پیش بینی و برآورد خود را انجام دهد.

$$Q(s,t) = SS(t) - SS(t_R) - SS(t_L) \quad (۲)$$

که در آن $SS(t_R)$ و$SS(t_L)$ به ترتیب میزان$SS(t)$ یا خطای طبقه بندی در شاخه سمت راست و سمت چپ گره t می‌باشد. در این مقاله برای اجرای الگوریتم درخت تصمیم از نرم افزارMATLAB استفاده گردید.

- پارامترهای مورد استفاده درشبیه‌سازی

درشبیه‌سازی صورت گرفته در این تحقیق از ۴۶ پارامتر به عنوان متغیر مستقل و ورودی مدل استفاده شد. پارامترهای مزبور تقریبا تمامی پارامترهای مؤثر در شبیه سازی پدیده شوری را شامل می‌شود. در سایر مدل های هوش مصنوعی نظیر شبکه عصبی مصنوعی امکان به کار گیری تعداد زیاد متغیر مستقل وجود ندارد، چون تعداد ترکیب پارامترها بالا می‌رود. ولی الگوریتم درخت تصمیم این توانایی را دارد که از طیف وسیعی از پارامترهای مستقل، بهترین ها و موثرترین ها را جدا سازی و انتخاب کند. لذا برای هر نقطه نمونه برداری میدانی انجام شد. سپس ۱۴ عدد مربوط به انواع شاخص های شوری، ۷ عدد مربوط به باندهای اصلی تصویر ماهواره ای، ۱۵ عدد از طریق تجزیه مؤلفه های اصلی باندهای ماهواره، ۸ عدد مربوط به انواع شاخص پوشش گیاهی و ۳ عدد از متغیرهای فیزیوگرافی استخراج گردید. پارامترهای مذکور به مدل معرفی و شبیه‌سازی‌ها برای برآورد متغیرهدف (شوری خاک سطحی) صورت گرفت. متغیرهای مستقل مورد استفاده در این تحقیق وعلائم اختصاری آن ها درجدول (۳) نمایش داده شده است.

جدول ۱- متغیر های مستقل مورد استفاده در تحقیق.

Table 1. Independent variables used in study.

پارامتر Parameter	علامت اختصاری Symbol
شاخص های شوری Salinity indices	SI1-SI2-SI3-SI4-SI5-YSI-Ratio1-Ratio2-Ratio3-BI-NDSI-IB-SR-VSSI
باندهای اصلی ETM^+ Main bands of ETM^+	B1-B2-B3-B4-B4-B5-B7
تجزیه مولفه های اصلی باند ۱ تا ۷ Partial component analysis from band1 to band7	PCA1-7
تجزیه مولفه های اصلی باند ۲ تا ۴ Partial component analysis from band2 to band4	PCA2-4
تجزیه مولفه های اصلی باند ۲ تا ۵ Partial component analysis from band2 to band5	PCA2-5
شاخص های پوشش گیاهی Vegetation indices	NDVI-CVI-NRVI-TVI-CTVI-TTVI-AVI-MSAVI2
متغیرهای فیزیوگرافی Physiographic variables	ارتفاع- شیب- جهت Dem-Slop-Aspect
شاخص رطوبتی Moisture index	NDMI

شبیه‌سازی‌ها در سه حالت استفاده از داده های اصلی، استفاده از لگاریتم داده ها و استفاده از استفاده داده های نرمال شده برای برآورد پارامتر شوری صورتگرفت. برای نرمال سازی داد های از رابطه ۳ استفاده گردید.

$$X_n = \frac{X_o - X_{\min}}{X_{\max} - X_{\min}} \qquad (۳)$$

X_n و X_O: به ترتیب داده‌های نرمال شده و داده‌های اولیه رانشان می‌دهد.

X_{min} , X_{max}: به ترتیب مینیمم و ماکزیمم داده‌های اولیه را نشان می‌دهد.

- تقسیم‌بندی داده‌ها برای آموزش مدل

جهت اجرای الگوریتم درخت تصمیم رگرسیونی، داده‌ها به دو دسته تقسیم شدند: داده‌های آموزشی[۴] و داده‌های آزمایشی[۵]. در این تحقیق ۷۵ درصد ازکل داده‌ها به آموزش مدل تعلق گرفت و ۲۵ درصد باقیمانده به عنوان داده‌های آزمون به مدل معرفی گردید. برای انتخاب داده های آموزش و آزمون ابتدا داده ها با استفاده از نرم افزار Minitab تصادفی شدند. بدیهی است به این ترتیب دخالت کاربر در انتخاب ۷۵ و ۲۵ درصد داده ها کاهش می‌یابد. لازم به ذکر است داده‌های آزمون در مرحله آموزش مورد استفاده قرار نگرفته اند.

- ارزیابی کارآیی شبیه سازی ها

در نهایت جهت ارزیابی کارایی شبیه سازی های صورت گرفته

دراین تحقیق از چهار پارامتر آماری ذیل استفاده گردید:

$$RMSE = \left(\frac{\sum_{i=1}^{n} (p_i - o_i)^2}{n} \right)^{1/2} \quad (۴)$$

$$MAE = \frac{\sum |O_i - P_i|}{N} \quad (۵)$$

$$Bias = \overline{P} - \overline{O} \quad (۶)$$

$$R = \sqrt{1 - \frac{\sum_{i=1}^{n} (Oi - Pi)}{\sum_{i=1}^{n} (Oi - \overline{O})}} \quad (۷)$$

دراین روابط O_i و P_i به ترتیب مقادیر مشاهده شده و شبیه‌سازی شده، $\overline{O}$ و $\overline{P}$ میانگین مقادیر مشاهده شده و شبیه‌سازی شده و N تعداد نمونه‌ها می‌باشد.

نتایج و تحلیل نتایج

خلاصه آماره های توصیفی برای داده های مورد استفاده در این تحقیق در جدول شماره ۲ آمده است. همان طور که این جدول نشان می‌دهد مقدار EC در منطقه مورد مطالعه از ۰/۷ تا ۲۰۰ دسی زیمنس بر متر متغیر بوده و مقدار ضریب تغییرات آن ۱۲۲/۱۵ درصد می‌باشد که نشان دهنده تغییر پذیری زیاد متغیر می‌باشد. بر اساس طبقه بندی ویلدینگ (۲۰) خصوصیات خاک با ضریب تغییرات بیش از ۳۵ درصد دارای تغییر پذیری زیاد می‌باشد. تغییر پذیری زیاد متغیر شوری به دلیل توپوگرافی منطقه و وجود ارتفاعات، انواع دشت سرها و کویر در منطقه مطالعاتی می‌باشد.

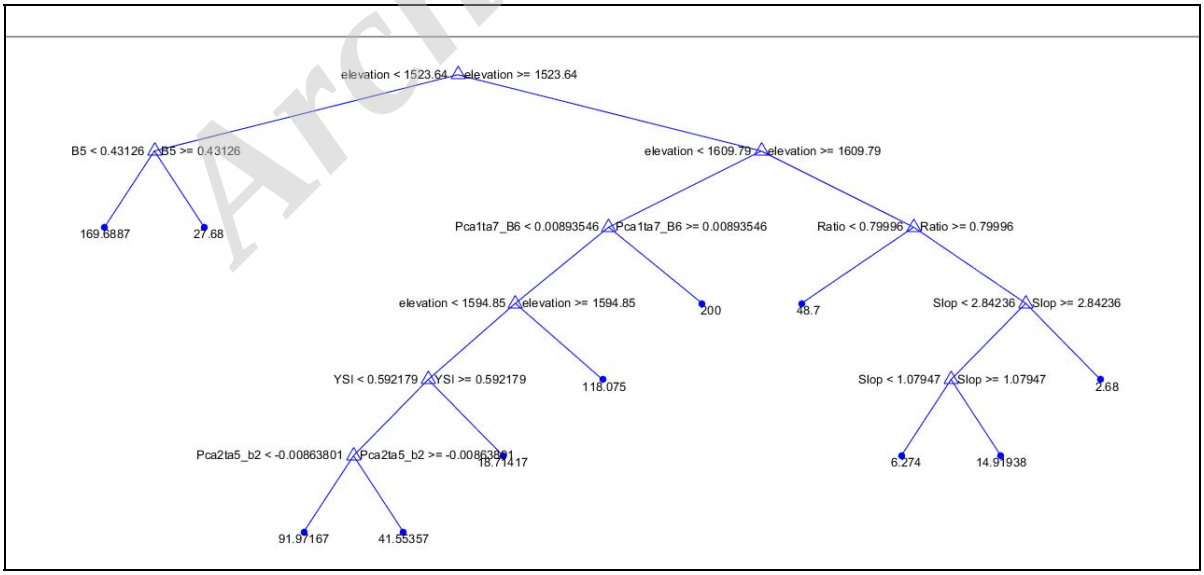
جدول۲- خصوصیات آماری داده های هدایت الکتریکی سطحی.

Table2.Statistical characteristic of data in this study.

پارامتر Parameter	حداقل minimum	حداکثر Maximum	میانگین Average	انحراف معیار Standard deviation	ضریب تغییرات Coefficient of variation
شوری EC(ds/m)	0.7	200	51.41	62.8	122.15

- **نتایج حاصل از اجرای الگوریتم درخت تصمیم درشرایط استفاده از داده های اصلی:**

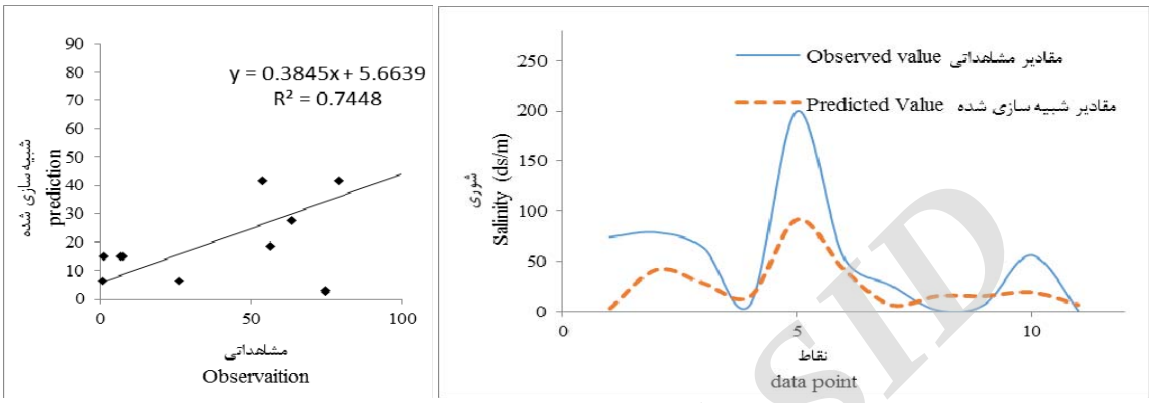
ساختار درخت ایجاد شده در شکل ۳ نشان می‌دهد در زمان استفاده از داده های اصلی، مدل از بین ۴۶ پارامتری که در اختیار داشته، از هفت پارامتر شامل (YSI, B5, PCA1_7, PCA2_5, Ratio-Dem, Slop) استفاده نموده است.



شکل ۳- ساختار درخت تصمیم ایجاد شده درحالت استفاده از داده های اصلی.

Figure 3. Structure of decision tree created using the original data.

شکل ۳ نشان می‌دهد که در شرایط استفاده از داده های اصلی، الگوریتم درخت رگرسیونی با استفاده از ۷ پارامتر و ۱۱ قانون به برآورد میزان شوری خاک سطحی پرداخته است. نمودار نقطه ای و نمودارخطی مربوط به نتایج مدل درخت تصمیم و مقادیر واقعی مربوط به حالت استفاده از داده های اصلی نیز در شکل (۴) نمایش داده شده است که نشان می‌دهد درخت تصمیم در این شرایط، کلیت روند داده ها را به خوبی شبیه سازی نموده است. ضریب همبستگی داده های مشاهداتی و پیش بینی شده ۰/۷۴ می‌باشد. بررسی آماره t نشان داد بین مقادیر مشاهداتی و پیش بینی شده در سطح ۵ درصد همبستگی معنی دار وجود دارد.

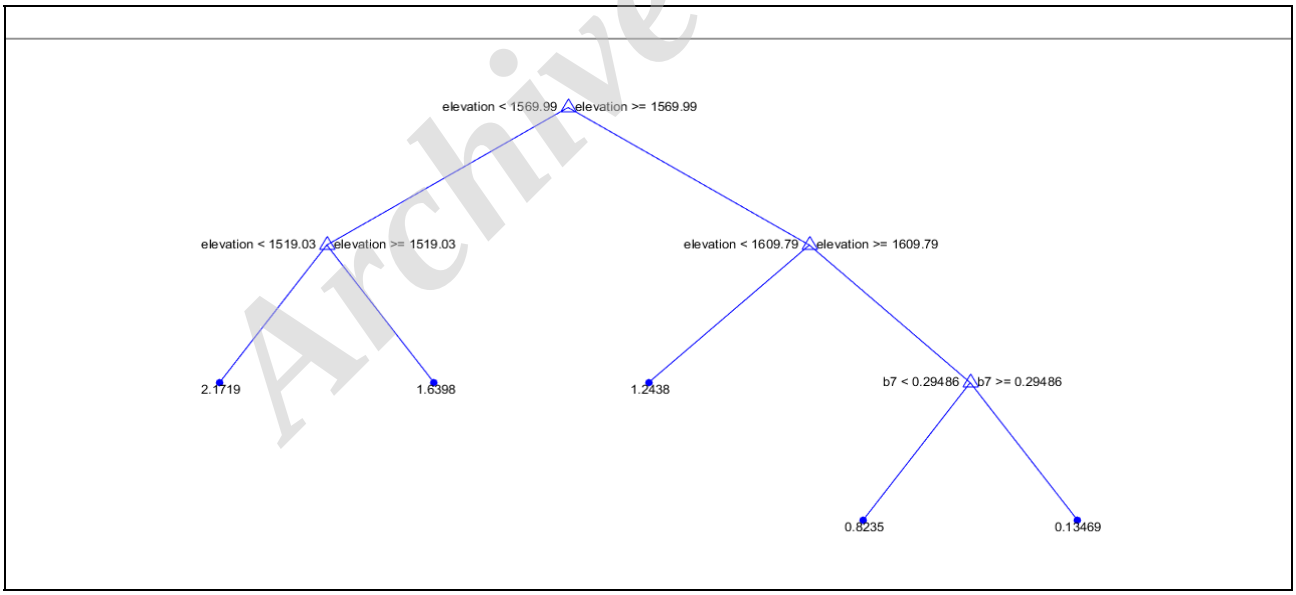


شکل ٤- نمودار تغییرات مقادیر پیش بینی شده درمقابل مقادیر واقعی در حالت استفاده از داده های اصلی.

Figure 4. Point and linear graph of simulated values vs observed values in condition using the original data.

- **نتایج مدل با استفاده از لگاریتم داده ها:**

ساختار درخت ایجاد شده در حالت استفاده از لگاریتم داده ها در شکل (۵) نمایش داده شده است.

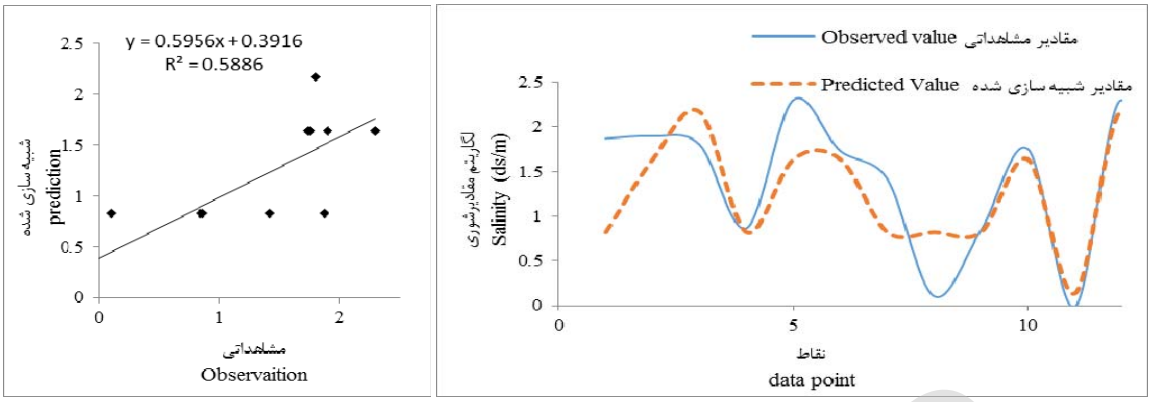


شکل ٥- ساختار درخت تصمیم در حالت استفاده از لگاریتم داده ها.

Figure 5. Structure of decision tree created using the logarithmic data.

شکل ۵ نشان می‌دهد که الگوریتم درخت تصمیم رگرسیونی در شرایط استفاده از لگاریتم داده ها، از میان ۴۶ پارامتری که در اختیار داشته، فقط از پارامتر "ارتفاع- ارزش رقومی باند هفت در پیکسل مورد نظر" استفاده نموده است. این درخت با استفاده از ۵ قانون به برآورد میزان شوری خاک سطحی منطقه پرداخته است. نمودارنقطه ای ونمودار خطی مربوط به نتایج مدل درختان تصمیم و مقادیر واقعی

مربوط به حالت استفاده از لگاریتم داده ها نیز درشکل (۶) نمایش داده شده است.



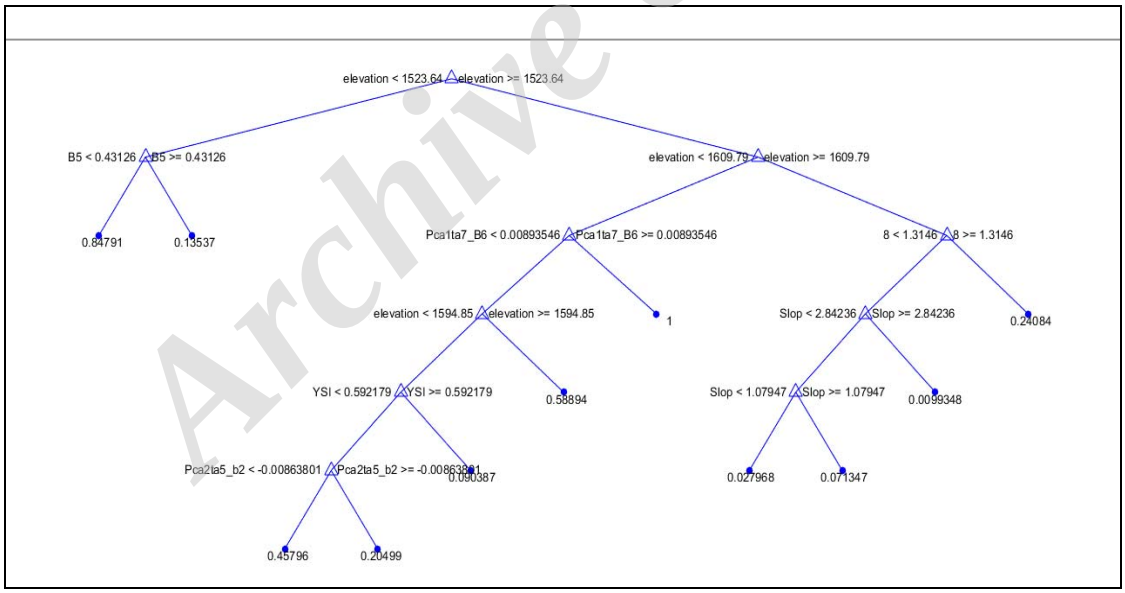
شکل٦- نمودار تغییرات مقادیر پیش بینی شده در مقابل مقادیر واقعی با استفاده از لگاریتم داده ها.

Figure 6. Point and linear graph of simulated values vs observed values in condition using the logarithmic data.

در حالت استفاده از لگاریتم داده ها نیز ضریب همبستگی داده های مشاهداتی و پیش بینی شده (۰/۵۹) از نظر آماری معنی دار می-باشد. علاوه بر آن نمودار خطی مقادیر مشاهداتی و پیش بینی شده نشان می‌دهد که عملکرد الگوریتم درخت تصمیم دراین حالت بهبود یافته است.

- **نتایج مدل با استفاده از داده های استاندارد شده**

در این بخش داده ها با استفاده از رابطه (۳) استاندارد شد و سپس به مدل معرفی و میزان شوری استاندارد شده، برآورد گردید. درخت تصمیم ایجاد شده در این حالت در شکل (۷) نمایش داده شده است.



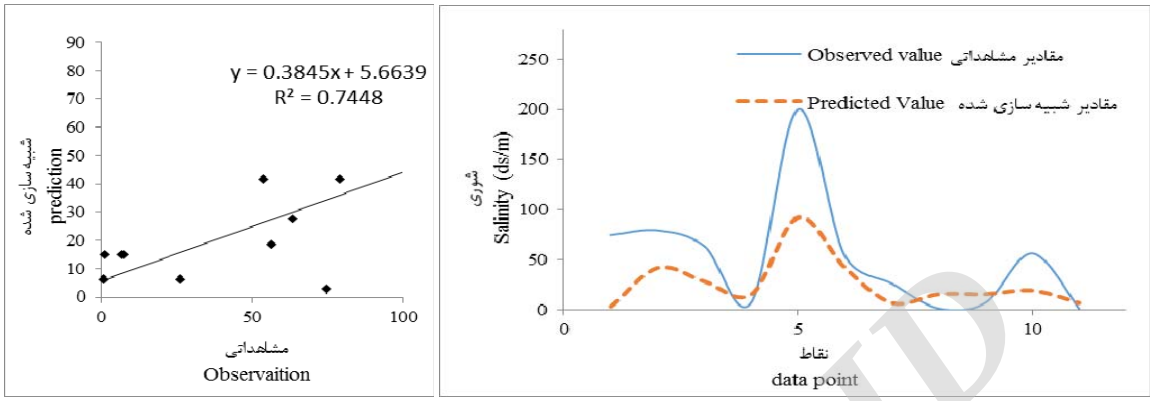
شکل۷ - ساختار درخت تصمیم در حالت سوم (استفاده از داده های نرمال شده).

Figure 7. Structure of decision tree created in the third mode (using the normalized data).

نتایج الگوریتم درخت تصمیم در شرایط استفاده از داده های استاندارد شده با شرایط استفاده از داده های اصلی مطابقت دارد (شکل ۳ و ۷). در این حالت نیز الگوریتم درخت رگرسیونی با استفاده از ۷ پارامتر و ۱۱ قانون به برآورد میزان شوری خاک سطحی پرداخته است و ساختار درخت ایجاد شده تغییری نکرده است. به نظر می‌رسد عدم تبعیت داده های شوری خاک منطقه از توزیع نرمال در این موضوع مؤثر است. لازم به توضیح است استاندارد کردن داده ها با استفاده از رابطه ۳ مقیاس داده ها را به خطی تبدیل می‌کند و آن ها

را در یک بازه (۱و۰) قرار می‌دهد. در مواردی که توزیع داده ها حول میانگین شان یک دست نباشد می‌توان از تبدیل هایی بر اساس توابع غیر خطی استفاده نمود (۱۵). نتایج این تحقیق نیز نشان داد تبدیل مقیاس داده ها با استفاده از توابع غیر خطی (لگاریتمی) برای بهنجار سازی داده ها مؤثرتر بوده است.

نمودارنقطه ای ونمودارخطی مربوط به نتایج مدل درخت تصمیم و مقادیر واقعی مربوط به حالت استفاده از میانگین متحرک سه ساله نرمال شده در شکل (۸) نمایش داده شده است.



شکل ۸- نمودار تغییرات مقادیر پیش بینی شده در مقابل مقادیر واقعی با استفاده از داده‌های نرمال شده.

Figure 8. Point and linear graph of simulated values vs observed values in condition using the normalized data.

مقایسه نتایج نهایی شبیه سازی حالت اول و سوم (انطباق شکل ۴ و ۸) نشان می‌دهد استفاده از داده‌های استاندارد شده در مقایسه با شرایط استفاده از داده‌ها اصلی، تأثیری در بهبود کارایی الگوریتم درخت تصمیم ندارد. به نظر می‌رسد این موضوع ناشی از نحوه کارکرد الگوریتم درخت تصمیم بوده و عملیات نرمال سازی که داده‌های موجود را در یک دامنه عددی صفر تا یک محدود می‌سازد، قدرت درخت را در انتخاب متغیرهای مهم تر در ایجاد شکاف (شاخه زنی) و تفکیک داده‌ها کاهش می‌دهد. نتایج نهایی حاصل از این تحقیق در جدول ۵ ارائه شده است.

.جدول ۵- نتایج شبیه سازی صورت گرفته در این تحقیق برای پیش بینی شوری خاک سطحی.

Table 5. Results of simulations for estimating soil salinity.

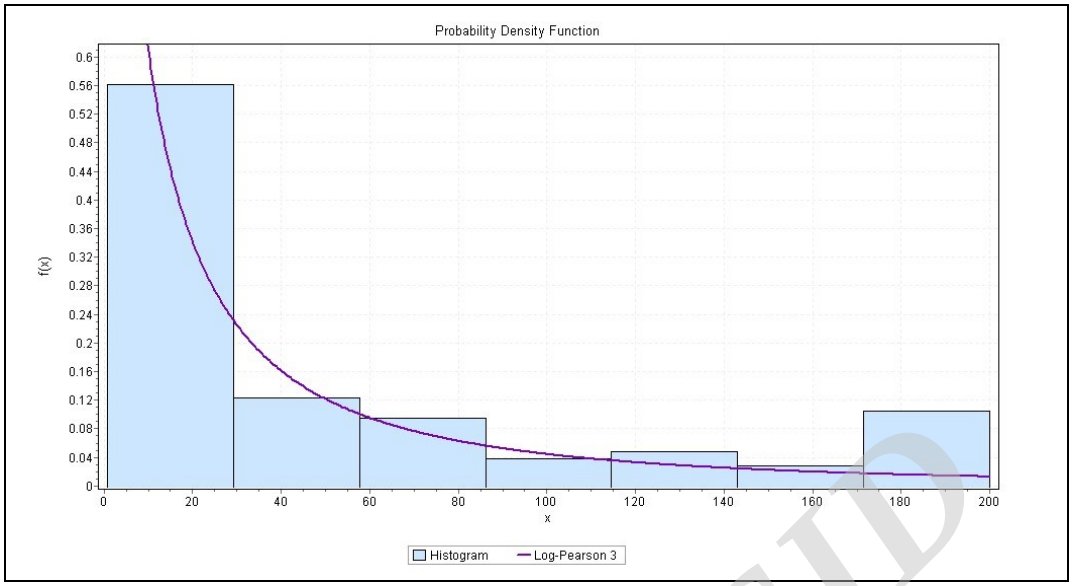
پارامتر Parameter	آماره های خطا statistical Error				
	R	Rmse	Rmse%	MAE	Bias
شوری EC	0.86	44.53	86.62	32.51	-26.23
لگاریتم مقادیر شوری Log EC	0.76	0.49	38.58	0.37	-0.14
مقادیر شوری نرمال شده Standardize	0.86	0.22	0.88	0.16	-0.13

بر اساس نتایج این تحقیق می‌توان گفت:

- پیش‌پردازش داده‌های موجود و استفاده از لگاریتم داده‌ها منجر به بهبود کارایی مدل درخت تصمیم رگرسیونی در منطقه مروست شده است. بررسی توزیع آماری داده‌های شوری منطقه با نرم افزار Easy fit نشان می‌دهد داده‌های مزبور از توزیع Log-Pearson 3 پیروی می کند که از خانواده توزیع های لگاریتمی می‌باشد (شکل ۹). از این رو استفاده از لگاریتم داده‌ها به عنوان ورودی منجر به بهبود نتایج شبیه سازی شده است. این موضوع با نتایج تحقیقات افخمی و همکاران (۳)، مبنی بر افزایش دقت شبیه سازی رسوب معلق با روش های هوش مصنوعی (شبکه های عصبی مصنوعی و سیستم استنتاج فازی-عصبی) در شرایط استفاده از لگاریتم داده‌ها تطابق دارد. لازم به ذکر است مطالعات بسیار محدودی در خصوص ارتباط توزیع احتمالاتی یک پدیده با روش بهنجار سازی و پیش پردازش داده انجام گرفته است. لذا به خاطر شباهت دو پدیده رسوب و شوری از نظر بازه تغییرات زیاد هر دوی آن ها (دامنه تغییرات زیاد) و موثر بودن تبدیل مقیاس غیر خطی نسبت به تبدیل مقیاس خطی در بررسی و شبیه سازی (لگاریتم داده‌ها نسبت به داده‌های استاندارد شده) می‌توان گفت برای شبیه سازی رفتار پدیده‌های با دامنه تغییرات زیاد بهتر است از تبدیل مقیاس غیر خطی روی داده‌ها استفاده نمود.

- در شبیه سازی های صورت گرفته در این تحقیق، زمانی که از لگاریتم داده‌ها برای اجرای مدل استفاده گردید، ترکیب "باند ۷، ارتفاع" به عنوان مناسب ترین حالت شناسایی شد. الگوریتم مزبور به صورت خودکار ۵ قانون زیر را ارائه داد. طبق نتایج این تحقیق، درخت ایجاد شده در این حالت قادر است با ۵ قانون، میزان شوری خاک سطحی را برآورد نماید (شکل ۵).

اگر ارتفاع منطقه، کمتر از ۱۵۱۹ متر باشد متوسط شوری خاک سطحی ۱۴۷/۹دسی زیمنس بر مترخواهد بود.



شکل۹ - تطابق داده‌های شوری منطقه مروست با توزیع احتمالاتی Log-Pearson 3

Figure 9. Matching of Salinity data in MARVAST region with Log-Pearson 3 Probabilistic distribution.

اگر ارتفاع منطقه، بین ۱۵۱۹ تا ۱۵۶۹/۹ متر باشد متوسط شوری خاک سطحی۴۳/۶دسی زیمنس بر مترخواهد بود.

اگر ارتفاع منطقه، بین ۱۵۶۹/۹ تا ۱۶۰۹/۸ متر باشد متوسط شوری خاک سطحی ۱۷/۵ دسی زیمنس بر مترخواهد بود.

اگر ارتفاع منطقه بیشتر یا مساوی ۱۶۰۹/۸ وارزش پیکسلی باند ۷ سنجنده ETM^+ در نقطه مورد نظرکمتر از ۰/۲۹۵ باشد شوری خاک سطحی ۴/۷ دسی زیمنس بر مترخواهد بود.

اگر ارتفاع منطقه بیشتر یا مساوی ۱۶۰۹/۸ و ارزش پیکسلی باند ۷ سنجنده ETM^+ در نقطه مورد نظر بیشتر یا مساوی ۰/۲۹۵ باشد شوری خاک سطحی ۱/۴ دسی زیمنس بر مترخواهد بود.

- در تمام شبیه سازی های صورت گرفته، پارامتر ارتفاع به عنوان ورودی توسط الگوریتم درخت تصمیم انتخاب شده است. این موضوع نشان دهنده اهمیت پارامتر ارتفاع در توزیع میزان شوری خاک سطحی منطقه می‌باشد. بررسی ضریب همبستگی پارامترهای مورد استفاده در شبیه سازی های تحقیق با میزان شوری خاک سطحی نشان داد پارامتر ارتفاع و مقادیر شوری خاک سطحی بیشترین همبستگی را با یگدیگر داشته و مقدار آن ۰/۵۳۲- می‌باشد.

- هرس، یکی از مراحل اصلی اجرای مدل درخت تصمیم است. در کد نویسی انجام شده در برنامه MATLAB می‌توان تعداد سطوح ایجاد شده در درخت تصمیم را بوسیله عملیات هرس کاهش داد. در این مقاله، نتایجی که از هر بار هرس به دست آمده، با نتایج اولیه مقایسه و بهترین نتیجه به عنوان نتیجه الگوریتم درخت تصمیم ارائه گردید. به عبارتی برای هر سه سناریو (داده‌های اصلی، لگاریتم داده‌ها و داده‌های استاندارد شده) سه درخت ایجاد شد که در هر سه مورد عملیات هرس اعمال شد که در غالب موارد منجر به افزایش کارایی درخت در برآورد شوری گردید. لازم به ذکر است بهترین نتیجه هرس در شرایط استفاده از لگاریتم داده‌ها، دارای سه سطح می‌باشد.

- به طور کلی، مدل درخت تصمیم به لحاظ سادگی و ایجاد قوانین برای شبیه سازی بر سایر روش های هوش مصنوعی برتری دارد. از طرفی، مدل درخت تصمیم قادر است بدون دخالت کاربر، ورودی‌های مهم تر را برای ایجاد قوانین استفاده و ورودی‌های ضعیف‌تر را حذف نماید. از این رو همان طور که شبیه سازی های صورت گرفته در این تحقیق نشان داد از میان ۴۶ متغیر مستقل معرفی شده به مدل، ۷ پارامتر در شرایط استفاده از داده‌های اصلی و داده‌های استاندارد شده (حالت اول و سوم) و دو پارامتر در شرایط استفاده از لگاریتم داده‌ها (حالت دوم) توسط الگوریتم درخت تصمیم مورد استفاده قرارگرفته است. بدیهی است از نظر کاربردی مدلی در اولویت قرار دارد که هم از نظر آماره های خطا قابل قبول بوده و هم به تعداد ورودی های کمتری نیاز داشته باشد (۱۱). توضیح اینکه؛ در یک شبیه سازی علاوه بر این که آماره های خطا بایستی قابل قبول باشند، نقش تعداد ورودی های مورد نیاز را هم نباید از نظر دور داشت، چراکه هر چه تعداد ورودی های مورد نیاز برای شبیه سازی کمتر باشد، در وقت و هزینه تحقیق صرفه جویی خواهد شد و کار با سهولت و دقت مناسب انجام خواهد شد.

نتیجه‌گیری کلی

منطقه مروست در استان یزد، یکی از مناطقی است که بـه لحـاظ نزدیکی به کویر با مشکل شوری خاک مواجه بوده و شناسایی و طبقه بندی خاک های منطقه برای اعمال مدیریت صحیح امـری ضـروری است. از طرفی نمونه برداری از خاک و مطالعـات آزمایشـگاهی بـرای برآورد شوری، مستلزم صرف زمان و هزینه بالایی است. از ایـن رو در این تحقیق، از تصاویر ماهواره ای استفاده و اطلاعات باندهای اصـلی و مصنوعی اسـتخراج گردیـد. تلفیـق ایـن اطلاعـات بـا پارامترهـای فیزیوگرافی به عنوان ابزاری برای برآورد میزان شوری مـورد اسـتفاده قرار گرفت تا هزینه و وقت مورد نیاز برای نمونه برداری های میـدانی کاهش یابد. داده‌هایی که به این ترتیب استخراج شدند حجیم بـوده و دارای مشکلاتی مانند نویز، بایاس، تغییرات شدید در بازه دینـامیکی و نمونه برداری می‌باشـند. بنـابراین داده‌کـاوی در دو مرحلـه ۱- انجـام عملیات پیش پردازش ۲- انتخاب مـدل مناسـب بـرای شـبیه سـازی صورت گرفت. سه روش برای پیش‌پردازش داده‌هـا و الگـوریتم نیمـه هوشمند درخت تصمیم برای استخراج دانش نهفتـه درون پایگـاه داده انتخاب گردید. از نتایج این پژوهش می‌توان دریافت درخـت تصـمیم رگرسیونی، دخالت کاربر را محدود می‌نماید. به طوری که در بهتـرین شبیه سازی صورت گرفته در این تحقیق، از میان ۴۶ پـارامتر مسـتقل معرفی شده به مدل، صرفا دو پـارامتر "ارتفـاع، ارزش رقـومی بانـد ۷ پیکسل در نقطه مورد نظر" انتخاب شد. علاوه بر آن نتایج نشـان داد هرچند الگوریتم درخت تصمیم محدودیتی از نظر تعداد ورودی نـدارد، اما کارایی این الگوریتم به عنوان یکی از مدل های هـوش مصـنوعی تا حدود زیادی تحت تأثیر ماهیت ورودی های معرفی شـده بـه مـدل می‌باشد. نتایج حاصل از شـبیه سـازی هـای ایـن تحقیـق نشـان داد علیرغم معنی دار بودن ضـریب همبسـتگی در سـه حالـت (داده‌هـای اصلی، لگاریتم داده‌ها، داده‌های استاندارد شده)، میزان خطـا در حالـت استفاده از لگاریتم داده‌ها نسبت به دو حالت دیگر کمتر بـوده و نتـایج به واقعیت نزدیکتر می‌باشد. با توجه بـه نتـایج بدسـت آمـده، بهتـرین توزیع احتمالاتی که بر داده‌های شوری منطقه برازش می‌یابـد، توزیـع احتمالاتی Log-Pearson 3 است که یکـی از توزیـع هـای خـانواده لگاریتم می‌باشد. با توجه به این موضوع، می‌توان اظهـار داشـت روش پیش پردازش داده‌ها در ارتباط نزدیک با توزیع احتمـالاتی حـاکم بـر داده می باشد و می‌تواند در تحقیقات بعدی مورد توجه محققـین قـرار گیرد.

منابع

1- Abdelfattah M., Shahid Sh., and Othman Y. 2009. Soil salinity mapping model developed using RS and GIS: A case study from Abu Dhabi, United Arab Emirates. European Journal of Scientific Research 26 (3): 342-351.
2- Abdinam A. 2005. Investigation of soil salinity mapping using the correlation between satellite data and numerical values of soil salinity in Ghazvin plain. Journal of research and building. (In Persian).
3- Afkhami H., Dastoorani M, T., and Fotouhi F. 2015. The impact probability distribution to increase accuracy of prediction of suspended sediment using artificial neural networks and neuro-fuzzy inference system (Case Study: Watershed Dez). Iranian Journal of Watershed Management Science and Engineering. (21), 21-35. (In Persian).
4- Ahmadian M., and Pakparvar V. 2006. Evaluation of Soil Salinity using RS & GIS in Ghahavand plain. Agriculture and Natural Resources Research Center of Hamadan. (In Persian).
5- Amini M. 1999. Investigation of geostatistics of soils salinity and alkalinity in selected soils in the Rodasht region. M.Sc Thesis of Soil Sciences. College of Agriculture. Isfahan University of Technology. 119p, (In Persian).
6- Breiman L., Friedman J., Olshen R., and Stone C. 1984. Classification and Regression Trees. Chapman & Hall/CRC Press, Boca Raton, FL.
7- Buces F.N., Siebe C., Cram S., and Palacio, J.L. 2006. Mapping soil salinity using a combined spectral response index for bare soil and vegetation: (A case study in the former lake Texcoco, Mexico). Journal of Arid Environments, 65:644-667.
8- Chitsaz V. 1999. Possibility investigation of mapping soil salinity and alkalinity in eastern region of Isfahan using TM Digital data. M.Sc Thesis. Isfahan University of Technology. 129p, (In Persian).
9- Dashtakian K., Pakparvar M., and Abdallai J. 2008. Investigation of mapping methods using Landsat data in Marvast region. Iranian Journal of Range and Desert Research.15 (2): 139-157. (In Persian).
10- Dwivedi R. S., and Sreenivas K. 1998. Image transforms as a tool for the study of soil salinity and alkalinity dynamics. International Journal of Remote Sensing, 19: 605-619.
11- Ebrahimian H., Liaghat A., and Bazrafshan M. 2011. Estimation of Some Climatic Parameters by Using Pedo-Transfer. Iranian Journal of Watershed Management Science and Engineering. (14), 77-85. (In Persian).
12- Eldiery A., Garcia L., and Reich R. 2005. Estimating soil salinity from remote sensing data in Corn fields. Hydrology days, 2005. Colorado State University fort Collins, co 80523-1372.
13- Jafari Gorzin B. 2002. Study of landsat ETM$^+$ capability in detecting salt affected lands (a case study in Gorgan Plain), a thesis of presented for M.Sc. Gorgan university of Agriculture and Natural Resource Science, college of Range and Watershed

Management, 127p.
14- Khajaldin S, J. 1996. Using data of Landsat MSS 5 for investigation of Plant communities and identify soil lands in Jazmoorian region. 02nd National Conference on desertification and desertification control methods. Kerman city. (In Persian).
15- Mohammadi Takami S, M. 2005. The methods of data processing and pattern recognition. K.N. Toosi University of Technology. (In Persian).
16- Naeijnoori R. 2001. Investigation on possibility of Separation salinity and gypsum land using TM data. M.Sc Thesis of desertification, collage of natural resource, Isfahan University of Technology. (In Persian).
17- Rivero R. G., Grunwald S., and Bruland G. L. 2007. Incorporation of spectral data into multivariate geostatistical models to map soil phosphorus variability in a Florida wetland, Geoderma, 140: 428-443.
18- Taghizadeh-Mehrjardi R., Minasny B., Sarmadian F., and Malone B, P. 2014. Digital mapping of soil salinity in Ardekan region, central Iran. Geoderma. 213: 15-28.
19- Tajgardan T., Ayoubei sh., Shetaei Sh., and Khormali F. 2009. Mapping of surface soil salinity using ETM^{+} data (case study: Northern Aq Qala , Gulistan Province. (In Persian).
20- Wilding L.P. 1985. Spatial variability: Its documentation, accommodation, and implication to soil survey. In: Nielsen, D.R., and J. Bouma, (eds.), Soil Spatial Variability, Pudoc, Wagenigen, the Netherlands. 166-194.
21- Wu J.,Vincent B., Yang, ., Bouarfa S., and Vidal A. 2008. Remote sensing monitoring of changes in soil salinity: A case study in Inner Mongolia, China. Journal of Sensors, 8: 7035-7049.

Journal of Water and Soil
Vol. 30, No. 3, Jul.-Agu. 2017, p. 915-928

نشریه آب و خاک (علوم و صنایع کشاورزی)
جلد ۳۰، شماره ۳، مرداد – شهریور ۱۳۹۶، ص. ۹۲۸-۹۱۵

Investigation of the Optimal Method of Data Processing to Increase Accuracy of Simulation of Surface Soil Salinity (Case study: *MARVAST*)

A. Habibipoor[1*] – A. Talebi[2] – A. A. Karimian[3] – F. Dehghani[4] – M. H. Mokhtari[5]

Received: 26-4-2016
Accepted: 31-5-2017

Introduction: Salinity is one of the problems of arid and semi-arid soils. Identification and classification of saline/alkaline soils is necessity for dealing with difficult situations and correct management. Considering the nature of salinity data and selection of befitting methods to process data before use artificial neural network, can result in better simulations. The aim of this study was to investigate the optimal method for data processing to enhance the accuracy of surface soil salinity simulation and improve the efficiency of decision tree algorithm.

Materials and Methods: The study area was 88940.4 hectares of *Marvast* plain located in central Iran (54° 5 ´ to 54° 18 ´ east longitude and 30° 10 ´ to 30° 35 ´ north latitude). This region faces with problems of soil and water resources salinity. In this study, the effect of data processing on increasing accuracy of simulation of soil surface salinity was assessed in *Marvast* region using decision tree algorithm. For this purpose, the decision tree algorithm was applied and simulation was performed using three approaches i.e. original data, logarithmic data and standardized data. Finally, five statistics including R ،Rmse ،%Rmse ،MAE and Bias were calculated to evaluate the performance of used simulation methods.

Results and Discussion: In this study, when the logarithmic data was used, the composition of band 7 – elevation was identified as the most appropriate condition. The created tree can estimate the soil salinity by five laws:

If elevation is less than 1519, then the average of surface soil salinity will be 147.9 ds/m.

If elevation is between 1519 to 1569.9, then the average of surface soil salinity will be 43.6 ds/m.

If elevation is between 1569.9 to 1609.8, then the average of surface soil salinity will be 17.5 ds/m.

If elevation is more or equal to 1609.8 and pixel value of band 7 (ETM+ sensor) in selected point is less than 0.295, then the average of surface soil salinity will be 4.7 ds/m.

If elevation is higher or equal to 1609.8 and pixel value of band 7 (ETM+ sensor) in selected point is more than or equal to 0.295, then the average of surface soil salinity will be 1.4 ds/m.

For the approach of using the logarithmic data, decision tree algorithm used two parameters out of 46 independent variables introduced into the model. R ،Rmse ،%Rmse ،MAE and Bias for this method was computed to be 0.76, 0.49, 38.57, 0.37 and -0.14, respectively. The application of logarithmic data was recognized as the best method considering the lower calculated error and its less input requirement. Using Easy fit software, the distribution of salinity data was found to be Log Pearson 3. Thus, the use of logarithmic data improved model performance. Our findings were in agreement with those of Afkhami et al (2015) who increased the simulation accuracy of suspended sediment with artificial intelligence methods (Artificial neural networks and ANFIS) using logarithmic data.

Conclusions: As effective factors for soil salinity simulation vary in different regions, application of a unique method and indicator to estimate soil salinity in deferent region may not be possible.. The application of semi intelligent algorithm which limits user intervention and selects effective parameters for simulation would increase the simulation accuracy. Furthermore, considering the nature of salinity data and selection of befitting methods to process before using decision tree algorithm can effectively improve model performance. The current study was conducted to select an appropriate approach to enhance the simulation accuracy of surface soil salinity. The results demonstrate that the performance of decision tree algorithm as one of the artificial intelligence models can be affected by input data. In this study, Log-Pearson3 distribution was defined as the distribution of salinity data. Moreover, despite existence of significant correlation coefficients for three simulation methods, the error was lower when logarithmic data was used. Since the probability distribution of salinity data in the studied area was logarithmic (Log-Pearson 3), the reduction in error rate can be attributed to

1, 2, 3, 5- PhD Student, Associate Professor and Assistant Professors Yazd University, Iran
(*-Corresponding Author Email: Email: AzamHabibipoor@yahoo.com)
4-Faculty of national salinity research center, Yazd, Iran

the probability distribution of salinity data.

Keywords: Normalization, Probability distribution, Regression decision tree, Electrical conductivity