19th Iranian Soil Science Congress
16-18 September, 2025

نوزدهمین کنگره علوم خاک ایران
۲۵ تا ۲۷ شهریور ۱۴۰۴

مدیریت جامع‌نگر و هوشمند خاک و آب
Holistic and Smart Soil and Water Management
دانشکدگان کشاورزی و منابع طبیعی دانشگاه تهران
College of Agriculture & Natural Resources, University of Tehran

۰۴۲۵۰-۳۲۰۳۱

بهبود مدل‌سازی داده‌های نامتوازن شوری خاک با استفاده از رویکرد پیش‌پردازش ترکیبی

زهرا رسائی[۱]، فریدون سرمدیان[۲*]، اعظم جعفری[۳]

۱- گروه علوم و مهندسی خاک، دانشکده کشاورزی، دانشکدگان کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران
۲- گروه علوم و مهندسی خاک، دانشکده کشاورزی، دانشکدگان کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران
fsarmad@ut.ac.ir
۳- گروه علوم خاک، دانشکده کشاورزی، دانشگاه شهید باهنر کرمان، کرمان، ایران

چکیده

نقشه‌برداری رقومی خاک (DSM) ابزاری مؤثر برای پایش و ارزیابی شوری خاک به شمار می‌رود. با این حال، داده‌های هدایت الکتریکی (EC) اغلب دارای توزیع‌های نامتوازن و چولگی شدید می‌باشند. این مسئله در منطقه نیمه‌خشک تا خشک آبیک قزوین، در ایران، به وضوح مشاهده می‌شود؛ به‌طوری که در مجموعه‌ای با ۲۸۰ نمونه خاک، تقریباً نیمی از داده‌ها غیرشور بوده، در حالی که داده‌های شوری، چولگی شدیدی نشان می‌دهند. برای رفع این عدم تعادل، سه رویکرد مورد بررسی قرار گرفت: تبدیل لگاریتمی داده‌ها، استفاده از روش نمونه‌گیری بیش از حد مصنوعی (SMOTE) و ترکیب این دو روش. پس از اعتبارسنجی عملکرد مدل‌ها، نقشه‌های پراکنش شوری و عدم قطعیت تهیه گردید. مدل‌سازی با داده‌های لگاریتمی جدید، بهترین نتایج را ارائه داد (R^2=۰/۹۳ و nRMSE=۰/۳۰) که نسبت به مدل مبتنی بر داده‌های اولیه، باعث افزایش ۴۴ درصدی R^2 و کاهش ۲/۸۱ واحدی در nRMSE شد (R^2=۰/۴۹ و nRMSE=۳/۱۱). همچنین، مدل ترکیبی به‌طور قابل توجهی عدم قطعیت را کاهش داد. یافته‌های این پژوهش، اهمیت پیش‌پردازش داده‌ها در مدل‌سازی توزیع شوری خاک با داده‌های نامتعادل را برجسته می‌سازد و گامی موثر در تهیه نقشه‌های شوری خاک برای مدیریت دقیق اراضی به شمار می‌رود.

واژگان کلیدی: توزیع نامتعادل؛ نقشه‌برداری رقومی خاک؛ شوری خاک؛ خشک و نیمه‌خشک؛ بوت‌استرپ

مقدمه

نقشه‌برداری رقومی (DSM) شوری خاک با استفاده از مدل‌های مختلف یادگیری ماشین (ML) مانند درخت رگرسیون (Taghizadeh-Mehrjardi et al., 2014)، جنگل تصادفی (Wang et al., 2024; Aksoy et al., 2024)، کوبیست (Omrani et al., 2021)، و همچنین ترکیب چندین روش (Naimi et al., 2021) انجام شده است. در این مطالعات، از متغیرهای محیطی متعددی بهره گرفته شده است؛ از جمله داده‌های سنجش از دور (Omrani et al., 2021; Wang et al., 2024)، مشتقات DEM مانند ارتفاع و شاخص رطوبت توپوگرافی (Taghizadeh-Mehrjardi et al., 2014)، و اطلاعات اقلیمی نظیر میانگین بارندگی سالانه (Wang et al., 2024).

یکی از چالش‌های اساسی در مدل‌سازی شوری خاک، عدم تعادل داده‌ها یا وجود چولگی شدید در توزیع آن‌ها است؛ مسئله‌ای که می‌تواند بر عملکرد الگوریتم‌های یادگیری ماشین تأثیر منفی بگذارد. برای اصلاح این مشکل، روش‌هایی مانند تبدیل‌های پیش‌پردازشی توسعه یافته‌اند. به‌عنوان مثال، Wang و همکاران (۲۰۲۰) با استفاده از تبدیل لگاریتمی، توزیع اجزای بافت خاک را اصلاح کردند و نشان دادند که مدل جنگل تصادفی با داده‌های تبدیل شده عملکرد بهتری نسبت به داده‌های خام دارد. Naimi و همکاران (۲۰۲۱) نیز از همین روش برای مدل‌سازی هدایت الکتریکی (EC) خاک استفاده کردند. در سال‌های اخیر، استفاده از روش نمونه‌برداری بیش از حد (SMOTE) به‌عنوان رویکردی نوین برای مقابله با عدم تعادل داده‌ها در مدل‌سازی شوری خاک گسترش یافته است. Noomen و Donyavi (۲۰۲۰) با ترکیب SMOTE و روش‌های زمین‌آماری، دقت مدل را در مناطق با تراکم بالای مشاهدات افزایش دادند. همچنین، Aksoy و همکاران (۲۰۲۴) از روش SMOTE برای اصلاح توزیع نامتوازن

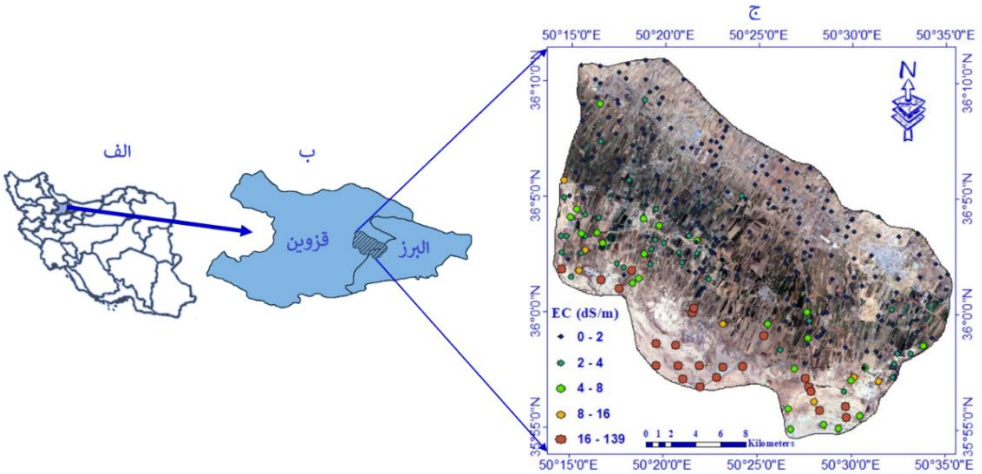
کلاس‌های شوری خاک استفاده کردند.

بنابراین، انتخاب یک روش مناسب برای اصلاح داده‌های نامتوازن در مدل‌سازی شوری خاک اهمیت زیادی دارد. در حالی که تبدیل لگاریتمی می‌تواند پیچیدگی‌هایی را به مدل تحمیل کند، الگوریتم‌های نمونه‌گیری جدید مانند SMOTE با اتکا به داده‌های اصلی، امکان تولید خروجی‌های دقیق‌تری را فراهم می‌آورند. بررسی‌های منابع بیانگر استفاده از توزیع‌های نرمال یا نمونه‌برداری بیش از حد داده به تنهایی برای رفع مشکل مدل‌سازی شوری خاک می‌باشند (Omrani et al., 2021; Aksoy et al., 2024). بنابراین، این پژوهش برای اولین بار به نقشه‌برداری شوری خاک سطحی (عمق ۰-۵۰ سانتی‌متر) در منطقه‌ای خشک و نیمه‌خشک با توزیع داده‌های شدیدا چوله و نامتوازن پرداخته است. در این راستا، اثر سه رویکرد شامل: ۱) نرمال‌سازی با تبدیل لگاریتمی، ۲) متعادل‌سازی داده‌های پیوسته با نمونه‌برداری جدید و ۳) ترکیب این دو روش بر بهبود دقت مدل بررسی شده است. یافته‌های این پژوهش می‌توانند به‌عنوان مبنایی در مطالعات آینده برای مدیریت داده‌های نامتوازن در مدل‌سازی ویژگی‌های پیوسته خاک در سایر مناطق مورد استفاده قرار گیرند.

مواد و روش‌ها

منطقه مورد مطالعه

این مطالعه در بخشی از دشت قزوین، با مساحت تقریبی ۶۰،۰۰۰ هکتار، انجام شده است (شکل ۱). منطقه مورد مطالعه در محدوده عرض‌های جغرافیایی ۳۵ درجه و ۵۵ دقیقه تا ۳۶ درجه و ۱۰ دقیقه شمالی و طول‌های جغرافیایی ۵۰ درجه و ۱۵ دقیقه تا ۵۰ درجه و ۳۵ دقیقه شرقی قرار دارد و ارتفاع آن از ۱،۱۴۳ متر تا ۱،۶۶۷ متر متغیر است. با حرکت از شمال به جنوب، میانگین دمای سالانه از ۱۳/۳ درجه سانتی‌گراد به ۱۵/۵ درجه سانتی‌گراد افزایش یافته، در حالی که میزان بارندگی از ۳۱۴/۲ میلی‌متر به ۲۵۳/۵ میلی‌متر کاهش می‌یابد. مواد مادری خاک‌های منطقه شامل رسوبات آبرفتی و کوهرفتی مربوط به دوره کواترنر در بخش‌های شمالی و مرکزی است، که در بخش‌های جنوبی به کفه‌های گلی و نمکی تغییر می‌یابد. از نظر توپوگرافی، منطقه دارای واحدهای فیزیوگرافی متنوعی شامل تپه‌ها (۷٪)، واریزه‌های سنگریزه‌ای (۹٪)، دشت دامنه‌ای (۳۲٪)، دشت آبرفتی (۴۳٪) و اراضی پست (۹٪) می‌باشد. کاربری‌های غالب شامل کشاورزی آبی و دیم، مراتع شور و غیرشور است. ریشه گیاهان عمدتاً در عمق ۵۰ سانتی‌متری از سطح خاک قرار دارند، بنابراین شوری در عمق ۰-۵۰ سانتی‌متری به‌عنوان متغیر هدف در نظر گرفته شده است.



شکل ۱- موقعیت منطقه مطالعاتی در ایران (الف) در مرز استان‌های البرز و قزوین (ب) و روی تصویر گوگل ارث به همراه تغییرات میزان شوری (EC) خاک سطحی مربوط به ۲۸۰ نقطه مطالعاتی (ج)

متغیرهای محیطی

متغیرهای محیطی با وضوح مکانی ۱۲/۵ متر تهیه شدند. داده‌های توپوگرافی از مدل رقومی ارتفاعی آلوس (ALOS) تهیه شدند. برای استخراج شاخص‌های گیاهی، شوری، کانی، رطوبت و نسبت‌های باندی، از تصاویر ماهواره‌ای سنتینل ۲ استفاده گردید. همچنین، میانگین بارندگی و دمای سالانه نیز به‌عنوان متغیرهای اقلیمی درنظر گرفته شدند. با استفاده از روش حذف بازگشتی ویژگی (RFE)، متغیرهای مهم شامل ارتفاع، شاخص همواری کف دره با درجه تفکیک بالا (MrVBF)، عمق دره (Valley

Depth)، شاخص رطوبت ساگا (Saga Wetness Index)، شاخص پوشش گیاهی عمودی (Perpendicular Vegetation Index)، شاخص شوری (Salinity Index)، میانگین بارندگی و دمای سالانه (Mean Annual Precipitation and Temperature) برای مدل‌سازی انتخاب شدند (Chen & Jeong, 2007).

داده‌های خاک و پیش‌پردازش آن‌ها

در این مطالعه، نمونه‌های خاک از عمق ۵۰-۰ سانتی‌متری در ۲۸۰ نقطه، که به صورت تصادفی انتخاب شده بودند، جمع‌آوری شدند (شکل ۱). پس از هوا خشک و خرد شدن، از الک ۲ میلی‌متری عبور داده شدند و هدایت الکتریکی آن‌ها در عصاره اشباع با استفاده از دستگاه رسانایی‌سنج اندازه‌گیری گردید.

برای اصلاح و پیش‌پردازش مقادیر هدایت الکتریکی، سه رویکرد زیر به‌کار گرفته شد:

- تبدیل لگاریتمی: مقادیر اولیه هدایت الکتریکی با استفاده از تابع log در نرم‌افزار R نرمال شدند.
- نمونه‌گیری بیش از حد: برای متعادل‌سازی مقادیر EC، از تابع SMOTE در نرم‌افزار R استفاده شد.
- ترکیب دو روش: مقادیر EC بدست آمده از روش SMOTE، نرمال‌سازی شدند.

مدل‌سازی و اعتبارسنجی مدل

برای مدل‌سازی و تهیه نقشه توزیع شوری خاک سطحی در منطقه مطالعاتی، از مدل جنگل تصادفی رگرسیونی (Random Forest Regression) استفاده شد. برای اعتبارسنجی مدل‌ها، از شاخص‌های آماری ضریب تعیین (R^2)، RMSE، و RMSE نرمال شده (nRMSE) استفاده شد (فرمول‌های ۱ تا ۳).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2} \quad (۱)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \quad (۲)$$

$$nRMSE = \frac{RMSE}{\bar{O}} \quad (۳)$$

در این معادلات، Oi و $\bar{O}$ به ترتیب مقادیر مشاهده شده و میانگین آن‌ها و Pi مقادیر پیش‌بینی شده است.

نقشه‌های توزیع شوری تهیه شدند و عدم قطعیت مدل‌ها با استفاده از تابع Bootstrapping در نرم‌افزار R بررسی شد.

نتایج و بحث

تحلیل توصیفی داده‌های خاک

مقادیر هدایت الکتریکی (EC) در این مطالعه در محدوده ۰/۱ تا ۱۳۹/۹ dS/m قرار داشتند که میانگین و انحراف معیار آن به ترتیب ۵/۷۳ و ۱۶/۵۸ dS/m می‌باشد (جدول ۱). توزیع مکانی EC نشان داد که مقادیر شوری از شمال به جنوب منطقه افزایش می‌یابد، به طوری که بیشترین میزان شوری در مناطق پست مشاهده می‌شود (شکل ۱). این تغییرات نشان دهنده تأثیر عوامل مختلف بر تشکیل و گسترش شوری خاک در منطقه مورد مطالعه است. در بخش‌های جنوبی و اراضی پست، عواملی نظیر دمای خاک و شرایط رطوبتی نقش مهمی در افزایش شوری خاک داشته‌اند (Taghizadeh-Mehrjardi et al., 2021). نرمال‌سازی لگاریتمی داده‌ها منجر به کاهش چولگی توزیع مقادیر EC شد. همچنین، چولگی داده‌ها در روش متعادل‌سازی نیز تا حد زیادی کاهش یافته است.

جدول ۱- خلاصه آمار توصیفی شوری خاک در عمق ۵۰-۰ سانتی‌متری

داده EC	حداقل	حداکثر	میانگین	واریانس	انحراف معیار	ضریب تغییرات	چولگی	کشیدگی
اصلی	۰/۱	۱۳۹/۹	۵/۷۳	۲۷۴/۹۶	۱۶/۵۸	۲۹۰	۵/۳۰	۳۴/۴۹
لگاریتمی	-۲/۳۰	۴/۹۴	۰/۴۴	۱/۷۳	۱/۳۱	۲۹۸	۱/۲۴	۴/۴۲
متعادل شده	۰/۱	۱۳۹/۹	۲۰/۴	۶۹۴/۶۵	۲۶/۳۶	۱۲۹	۲/۲۱	۷/۸۳
لگاریتمی متعادل شده	-۲/۳۰	۴/۹۴	۰/۴۴	۱/۷۳	۱/۳۱	۲۹۸	۱/۲۴	۴/۴۲

ارزیابی مدل‌ها

نتایج ارزیابی عملکرد مدل‌های جنگل تصادفی در جدول ۲ آورده شده است. همان‌طور که مشاهده می‌شود، مدل آموزش دیده با داده‌های اصلی شوری خاک نتوانسته است به خوبی تغییرات شوری خاک (۱۳۹/۹-۰/۱ dS/m) را نشان دهد. این مدل، مقادیر

پایین شوری را بیش‌برآورد و مقادیر بالا را کمتر از حد واقعی برآورد کرده است (۰/۵۳ dS/m-۹۹/۱۳). این عدم تطابق منجر به عملکرد ضعیف مدل با کم‌ترین دقت و بیشترین خطای پیش‌بینی (R^2=۰/۴۹ و RMSE=۳/۱۱) شده است. بنابراین، می‌توان دریافت که مدل جنگل تصادفی در مواجهه با توزیع نامتعادل داده‌ها دچار چالش شده است (Zeng et al., 2022; Zhang et al., 2022)، که بر ضرورت استفاده از داده‌های متعادل‌سازی شده برای بهبود عملکرد مدل تأکید دارد.

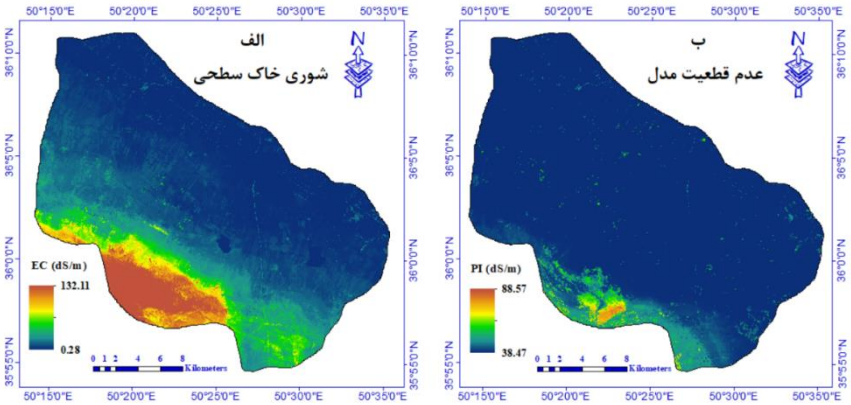
جدول ۲- نتایج ارزیابی مدل‌های جنگل تصادفی پیش‌بینی کننده تغییرات شوری خاک در عمق ۰-۵۰ سانتی‌متری

داده EC	R^2	RMSE	nRMSE	محدوده پیش‌بینی	محدوده عدم قطعیت
اصلی	۰/۴۹	۱۷/۸۰	۳/۱۱	۰/۵۳ تا ۹۹/۱۳	۵۸/۴۷ تا ۸۸/۵۷
لگاریتمی	۰/۵۶	۰/۹۷	۲/۲۰	۰/۳۰ تا ۸۳/۸۵	۲۴/۷۴ تا ۴۳/۹۲
متعادل شده	۰/۷۱	۱۰/۲۴	۰/۵۰	۰/۲۸ تا ۱۳۲/۱۱	۳۹/۵۳ تا ۸۰/۳۱
لگاریتمی متعادل شده	۰/۹۳	۶/۰۲	۰/۳۰	۰/۲۱ تا ۱۳۷/۲۹	۵۲/۲۵ تا ۸۴/۳۵

استفاده از داده‌های لگاریتمی باعث افزایش دقت مدل شده، اما مشکل کم‌برآورد مقادیر بالای شوری تشدید شده است. درواقع، تبدیل لگاریتمی نتوانست به‌درستی ارتباط بین متغیرهای کمکی و مقادیر EC را، به‌ویژه در مناطق پرخطر مانند اراضی پست با بیشترین میزان شوری (۱۳۹/۹ dS/m) ثبت کند. این مسئله در مناطق جنوبی که دارای مقادیر بحرانی شوری هستند، چالش‌های جدی در مدیریت اراضی ایجاد کرده و ممکن است موجب سردرگمی مسئولان و تصمیم‌گیرندگان در برنامه‌ریزی‌های مرتبط شود. در مقابل، مدل‌سازی با داده‌های متعادل شده نه تنها موجب افزایش دقت مدل (R^2) به میزان ۰/۲۲ و کاهش خطا (nRMSE) به مقدار ۲/۶۱ نسبت به داده‌های اصلی شده است، بلکه عملکرد بهتری نسبت به داده‌های لگاریتمی نیز نشان می‌دهد و دقت پیش‌بینی را ۱۵ درصد افزایش داده است. این نتایج بر اهمیت پیش‌پردازش متعادل‌سازی داده‌ها در بهبود دقت مدل تأکید دارند و با یافته‌های Elreedy و Atiya (۲۰۱۹) و Sharififar و همکاران (۲۰۱۹) هم‌راستا می‌باشد. اگرچه مدل همچنان دارای سطحی از خطا است، استفاده از داده‌های متعادل‌شده و لگاریتمی عملکرد آن را به‌طور قابل‌توجهی ارتقا داده است (R^2=۰/۹۳ و nRMSE=۰/۳۰). علاوه بر این، مدل توانسته است محدوده تغییرات شوری منطقه را با دقت بالاتری تخمین بزند (۰/۲۱ dS/m-۱۳۷/۲۹)، که بیشترین تطابق را با وضعیت واقعی منطقه در میان مدل‌های بررسی‌شده دارد. این یافته‌ها نشان می‌دهند که در مدل‌سازی ویژگی‌های خاک با توزیع نامتقارن و اریب، وجود تعداد کافی از نمونه‌های متعادل و توزیع نرمال نقش اساسی در افزایش دقت مدل دارد. شکل ۲ الف نقشه پراکنش مکانی شوری خاک سطحی برآورد شده بوسیله این مدل را نشان می‌دهد.

عدم قطعیت مدل‌ها

محدوده عدم قطعیت مدل‌ها در جدول ۲ ارائه شده است. یافته‌ها نشان می‌دهند که روش‌های مختلف پیش‌پردازش داده تأثیر قابل‌توجهی بر میزان عدم قطعیت مدل‌ها داشته‌اند. مدل مبتنی بر داده‌های اصلی، بیشترین میزان عدم قطعیت را نشان می‌دهد (۴۷/۵۸-۵۸/۸۸ dS/m)، که ناشی از توزیع نامتعادل داده‌ها و چولگی شدید داده‌ها است. این موضوع بیانگر حساسیت مدل به توزیع داده‌ها و درنتیجه عدم توانایی آن در تخمین دقیق شوری خاک می‌باشد (Zeng et al., 2022; Zhang et al., 2022). هرچند تبدیل لگاریتمی موجب کاهش نسبی پراکندگی داده‌ها شده، اما محدوده عدم قطعیت هنوز نسبتاً زیاد باقی مانده است (۲۸/۵۵ dS/m)، که بیانگر تأثیر محدود این روش در بهبود الگوی پیش‌بینی مدل است.



شکل ۲- پراکنش و عدم قطعیت شوری خاک سطحی منطقه مورد مطالعه بر اساس مدل پیش‌پردازش ترکیبی (متعادل-لگاریتم)

در مقابل، استفاده از داده‌های متعادل‌شده منجر به کاهش عدم قطعیت مدل به میزان ۱۰/۶۸ dS/m نسبت به مدل داده‌های اصلی شده، که نشان‌دهنده بهبود دقت مدل در تخمین مقادیر شوری است. ترکیب روش تعادل‌سازی و تبدیل لگاریتمی منجر به بهترین نتایج شده است، به‌گونه‌ای که محدوده عدم قطعیت به میزان قابل‌توجهی کاهش یافته و به ۵۲/۲۵ تا ۸۴/۳۵ dS/m رسیده است. این مدل دارای بالاترین مقدار R^2 (۰/۹۳) و کم‌ترین مقدار nRMSE (۰/۳۰) است، که نشان‌دهنده پیش‌بینی دقیق‌تر و کاهش عدم قطعیت مدل می‌باشد. این نتایج تأکید می‌کنند که بکارگیری روش‌های موثر پیش‌پردازش داده، به‌ویژه ترکیب نرمال‌سازی و متعادل‌سازی داده‌ها، نقش مهمی در کاهش عدم قطعیت و افزایش اعتماد به نتایج مدل دارد. همچنین، کوچک‌تر بودن بازه عدم قطعیت به معنای اعتمادپذیری بیشتر مدل در تخمین مقادیر شوری خاک است، که می‌تواند در مدیریت بهینه اراضی بسیار مفید واقع شود.

نقشه پراکنش مکانی عدم قطعیت مدل بدست آمده از داده‌های متعادل و نرمال‌سازی شده که بیشترین دقت را داشت، در شکل ۲ ب نمایش داده شده است. بالاترین عدم قطعیت در مناطق با توپوگرافی هموار و اراضی پست مشاهده شده، که با مقادیر بالای EC تطابق دارد. این مسئله احتمالا به دلیل تراکم پایین مشاهدات در این نواحی نسبت به بخش‌های مرتفع‌تر منطقه رخ داده است، که با گزارش Agaba و همکاران (۲۰۲۴) هم‌خوانی دارد. با این حال، این وضعیت می‌تواند ناشی از محدودیت متغیرهای کمکی در نمایش پویایی پیچیده شوری در این مناطق باشد. در این راستا، Aksoy و همکاران (۲۰۲۴) بیان کرده‌اند که نقشه‌برداری شوری خاک تحت تأثیر عوامل مختلفی مانند بافت خاک، رطوبت و خواص بازتاب طیفی قرار دارد و ممکن است در محیط‌های مختلف نتایج متفاوتی به همراه داشته باشد. علاوه بر این، آن‌ها تأکید کرده‌اند که تکیه بر مجموعه محدودی از متغیرهای کمکی ممکن است منجر به برور خطا در تخمین شوری خاک شود، به‌ویژه زمانی که این متغیرها نتوانند تغییرات مکانی و دینامیک پیچیده شوری منطقه را به‌طور دقیق نمایش دهند.

نتیجه‌گیری

نتایج این مطالعه نشان می‌دهد که پیش‌پردازش داده‌های پیوسته نقش مهمی در بهبود دقت مدل و کاهش عدم قطعیت پیش‌بینی دارد. اگرچه بسیاری از پژوهش‌های پیشین برای اصلاح توزیع نامتوازن داده‌های پیوسته از تبدیل‌های آماری مانند لگاریتمی استفاده کرده‌اند، یافته‌های ما نشان می‌دهد که اعمال روش‌های متعادل‌سازی میزان خطای مدل را به‌طور قابل‌توجهی کاهش داده است. در حالی که اکثر مطالعات روی متعادل‌سازی کلاس‌های خاک تمرکز داشته‌اند، تلاش‌های محدودی برای حل چالش عدم تعادل در داده‌های پیوسته انجام شده است. پژوهش حاضر با ارائه یک راهکار نوین برای متعادل‌سازی داده‌های پیوست، گامی مهم در افزایش دقت مدل‌های پیش‌بینی ویژگی‌های مدیریتی خاک، به‌ویژه شوری، برداشته است. علاوه بر این، نتایج به‌دست آمده از این مطالعه، در سطح جهانی جدید و نوآورانه تلقی می‌شود و می‌تواند به‌عنوان یک رویکرد موثر در کاهش عدم قطعیت پیش‌بینی در مدل‌سازی داده‌های خاک مورد استفاده قرار گیرد. رویکرد معرفی شده، نه‌تنها دقت پیش‌بینی مدل‌های شوری خاک را افزایش می‌دهد، بلکه می‌تواند به‌عنوان یک چارچوب کاربردی برای اصلاح داده‌های نامتوازن در سایر مطالعات محیطی نیز مورد استفاده قرار گیرد. این موضوع کاربردهای گسترده‌ای در زمینه مدیریت پایدار اراضی، برنامه‌ریزی منابع طبیعی، و توسعه کشاورزی پایدار خواهد داشت.

تشکر و قدردانی

این اثر تحت حمایت مادی صندوق حمایت از پژوهشگران و فناوران کشور (INSF) برگرفته شده از طرح شماره ۴۰۲۴۳۴۳، انجام شده است. بدین وسیله نویسنده اول مقاله مراتب قدردانی خود را از صندوق INSF بابت حمایت‌های مالی طرح و همچنین از دانشگاه تهران بابت فراهم آوردن امکانات اجرای آن اعلام می‌دارد.

فهرست منابع

Agaba, S., Ferré, C., Musetti, M., & Comolli, R. (2024). Mapping Soil Organic Carbon Stock and Uncertainties in an Alpine Valley (Northern Italy) Using Machine Learning Models. Land, 13(1), 78.

Aksoy, S., Sertel, E., Roscher, R., Tanik, A., & Hamzehpour, N. (2024). Assessment of soil salinity using explainable machine learning methods and Landsat 8 images. International Journal of Applied Earth Observation and Geoinformation, 130, 103879.

Chen, X. W., & Jeong, J. C. (2007). Enhanced recursive feature elimination. In Sixth international conference on machine learning and applications (ICMLA 2007) (pp. 429-435). IEEE.

Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. Information Sciences, 505, 32-64.

Naimi, S., Ayoubi, S., Zeraatpisheh, M., & Dematte, J.A. M. (2021). Ground observations and environmental covariates integration for mapping of soil salinity: a machine learning-based approach. Remote Sensing, 13(23), 4825.
Nauman, T. W., & Duniway, M. C. (2020). A hybrid approach for predictive soil property mapping using conventional soil survey data. Soil Science Society of America Journal, 84(4), 1170-1194.
Omrani, M., Shahbazi, F., Feizizadeh, B., Oustan, S., & Najafi, N. (2021). Application of remote sensing indices to digital soil salt composition and ionic strength mapping in the east shore of Urmia Lake, Iran. Remote Sensing Applications: Society and Environment, 22, 100498.
Sharififar, A., Sarmadian, F., Malone, B. P., & Minasny, B. (2019). Addressing the issue of digital mapping of soil classes with imbalanced class observations. Geoderma, 350, 84-92.
Taghizadeh-Mehrjardi, R., Minasny, B., Sarmadian, F., & Malone, B. P. (2014). Digital mapping of soil salinity in Ardakan region, central Iran. Geoderma, 213, 15-28.
Taghizadeh-Mehrjardi, R., Schmidt, K., Toomanian, N., Heung, B., Behrens, T., Mosavi, A., ..., & Scholten, T. (2021). Improving the spatial prediction of soil salinity in arid regions using wavelet transformation and support vector regression models. Geoderma, 383, 114793.
Torgo, L., Branco, P., Ribeiro, R. P., & Pfahringer, B. (2015). Resampling strategies for regression. Expert Systems, 32(3), 465-476. 4.
Wang, N., Chen, S., Huang, J., Frappart, F., Taghizadeh, R., Zhang, X., ..., & Shi, Z. (2024). Global Soil Salinity Estimation at 10 m Using Multi-Source Remote Sensing. Journal of Remote Sensing, 4, 0130.
Zeng, P., Song, X., Yang, H., Wei, N., Du, L. (2022). Digital Soil Mapping of Soil Organic Matter with Deep Learning Algorithms. ISPRS International Journal of Geo-Information, 11(5), 299.
Zhang, X., Xue, J., Chen, S., Wang, N., Shi, Z., Huang, Y., & Zhuo, Z. (2022). Digital mapping of soil organic carbon with machine learning in dryland of Northeast and North plain China. Remote Sensing, 14(10), 2504.

Improving Digital Mapping of Soil Salinity Through Hybrid Preprocessing of Imbalanced Data

Zahra Rasaei [1], Fereydoon Sarmadian [1], Azam Jafari [2]
[1]Soil Science Department, Faculty of Agricultural, University College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran
[2]Soil Science Department, Faculty of Agriculture, Shahid Bahonar University of Kerman, Kerman, Iran

Abstract
Digital Soil Mapping (DSM) is an effective tool for monitoring and assessing soil salinity. However, electrical conductivity (EC) data often exhibit highly imbalanced and skewed distributions. This issue is clearly observed in the semi-arid to arid region of Abyek, Qazvin, Iran, where nearly half of the 280 collected soil samples are non-saline, while the saline samples show severe skewness. To address this imbalance, three preprocessing approaches were evaluated: logarithmic transformation, synthetic minority oversampling (SMOTE), and a combination of both. After validating model performance, salinity distribution and uncertainty maps were generated. The model using the combined approach yielded the best results (R^2 = 0.93, nRMSE = 0.30), representing a 44% increase in R^2 and a 2.81-unit reduction in nRMSE compared to the model based on the original data (R^2 = 0.49, nRMSE = 3.11). Moreover, the combined method significantly reduced predictive uncertainty. These findings highlight the importance of data preprocessing in modeling soil salinity with imbalanced datasets and represent a valuable step toward producing more accurate salinity maps for precision land management.
Keywords: Imbalanced distribution, Soil salinity, Arid and semiarid, Bootstrap