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

تحلیل اثر داده‌های نامتعادل بر دقت و خودهمبستگی مکانی مدل‌سازی کربن آلی خاک

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

۱- گروه علوم خاک، دانشکده کشاورزی، دانشگاه شهید باهنر کرمان، کرمان، 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 و مهندسی خاک، دانشکده کشاورزی، دانشکدگان کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران

چکیده

توزیع نامتعادل داده‌ها یکی از چالش‌های مهم در مدل‌سازی مکانی متغیرهای خاکی، به‌ویژه کربن آلی خاک (SOC)، محسوب می‌شود. این پژوهش با هدف بررسی تأثیر متعادل‌سازی داده‌ها بر دقت مدل و ساختار مکانی باقیمانده‌ها، در بخشی از دشت قزوین با مساحت حدود ۶۰۰۰۰ هکتار انجام شد. در این راستا، ابتدا میزان SOC در ۲۸۰ نمونه سطحی (عمق ۰–۳۰ سانتی‌متر) اندازه‌گیری و مجموعه‌ای از متغیرهای محیطی مانند پارامترهای توپوگرافی، شاخص‌های پوشش گیاهی، رطوبت و اقلیم جمع‌آوری شد. پس از انتخاب متغیرهای مهم با استفاده از روشRFE، مدل جنگل تصادفی (RF) در دو سناریوی استفاده از داده‌های نامتعادل و متعادل‌شده با روش SMOTE اجرا گردید. عملکرد مدل‌ها با شاخص‌های RMSE، R^2 و Moran's I مورد ارزیابی قرار گرفت. نتایج نشان داد که مدل RF با داده‌های متعادل‌شده (RMSE=0.3، R^2=0.60) در مقایسه با داده‌های نامتعادل (RMSE=0.4، R^2=0.40) عملکرد بهتری از خود نشان داد. همچنین شاخص خودهمبستگی مکانی موران برای باقیمانده‌ها در حالت متعادل‌شده از ۰/۳۰ به ۰/۰۳ کاهش یافت. بررسی کورلوگرام نیز حاکی از کاهش ساختار مکانی در خطاهای مدل پس از متعادل‌سازی داده‌ها بود. این یافته‌ها نشان می‌دهد که متعادل‌سازی آماری داده‌ها، علاوه بر بهبود دقت عددی مدل، به کاهش خودهمبستگی مکانی در خطاها و افزایش کیفیت فضایی مدل منجر می‌شود.

واژگان کلیدی: شاخص موران، SMOTE، نقشه‌برداری رقومی خاک

مقدمه

کربن آلی خاک (Soil Organic Carbon – SOC) یکی از مهم‌ترین شاخص‌های کیفیت خاک و عاملی کلیدی در فرآیندهای زیست‌محیطی از جمله چرخه جهانی کربن، تولید محصولات کشاورزی، پایداری اکوسیستم‌ها و مقابله با تغییر اقلیم به شمار می‌آید (Lal, 2004; Minasny et al., 2017). برآورد دقیق و مکانی SOC نه‌تنها در مدیریت پایدار منابع خاک و آب ضروری است، بلکه به عنوان ابزاری مؤثر در برنامه‌ریزی‌های اقلیمی، سیاست‌گذاری‌های کشاورزی و پایش سلامت خاک در مقیاس‌های محلی تا جهانی مطرح شده است (Adhikari & Hartemink, 2016; Hengl et al., 2017). با توجه به هزینه و زمان‌بر بودن نمونه‌برداری میدانی و آزمایشگاهی، بهره‌گیری از روش‌های مدل‌سازی رقومی خاک (Digital Soil Mapping – DSM)و الگوریتم‌های یادگیری ماشین به‌منظور پیش‌بینی مکانی SOC در سال‌های اخیر به‌طور گسترده مورد توجه قرار گرفته است (Heung et al., 2016; Wadoux et al., 2021).

یکی از چالش‌های اساسی در فرآیند مدل‌سازی، توزیع نامتوازن داده‌ها (Data Imbalance) است که به‌ویژه در متغیرهایی مانندSOC که دارای توزیع چوله و غیریکنواخت در سطح منطقه هستند، به‌وضوح مشاهده می‌شود. در بسیاری از موارد،

درصد زیادی از نمونه‌ها مقادیر SOC پایینی دارند و تنها تعداد معدودی دارای مقادیر بالا هستند. این نابرابری در توزیع داده می‌تواند موجب تعصب (bias) در فرآیند یادگیری مدل شده، دقت پیش‌بینی در مقادیر اقلیت را کاهش داده و ساختارهای مکانی واقعی متغیر را مخدوش نماید (Zhao et al., 2019).

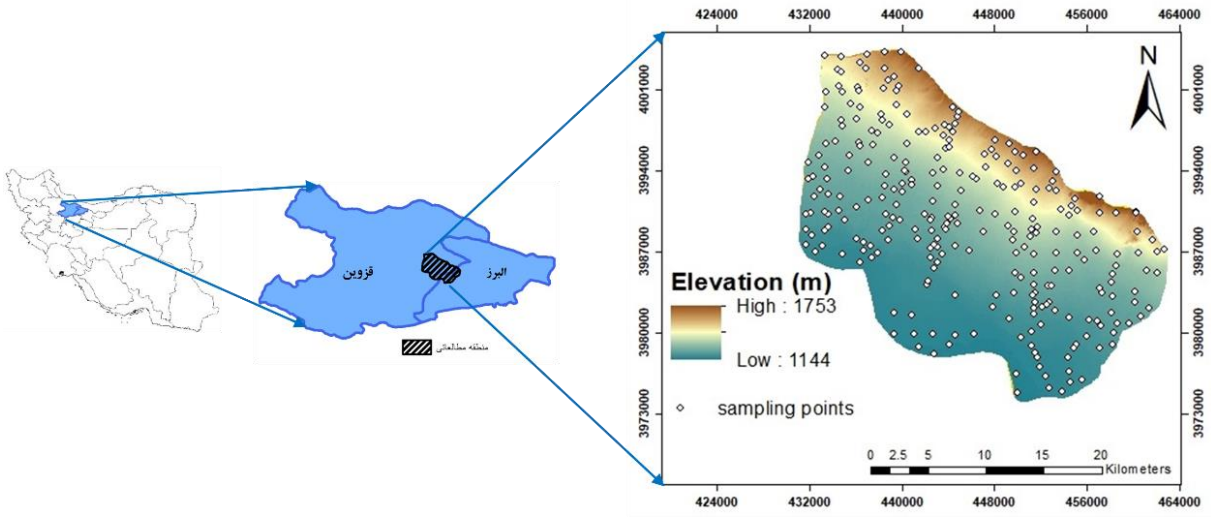
از سوی دیگر، یکی از مفاهیم کلیدی در تحلیل داده‌های مکانی، خودهمبستگی مکانی (Spatial Autocorrelation) است که نشان‌دهنده میزان شباهت بین مقادیر یک متغیر در موقعیت‌های مکانی مجاور می‌باشد. شاخص موران (Moran's I) به‌طور گسترده برای سنجش خودهمبستگی مکانی باقیمانده‌های مدل‌ها استفاده می‌شود. اگر باقیمانده‌های مدل دارای خودهمبستگی مکانی بالا باشند، نشان‌دهنده‌ی آن است که مدل قادر به بازنمایی صحیح روابط فضایی متغیر نبوده است. از این رو، تحلیل خودهمبستگی مکانی نه‌تنها ابزاری آماری، بلکه ابزاری تشخیصی برای ارزیابی و بهبود مدل‌سازی مکانی محسوب می‌شود (Hengl et al., 2007; Brus et al., 2011). با وجود پیشرفت‌های صورت گرفته، پژوهش‌های اندکی به بررسی مستقیم تأثیر عدم تعادل آماری داده‌ها بر ساختار مکانی باقیمانده‌ها در مدل‌سازی متغیرهای خاکی پرداخته‌اند. بنابراین، ضروری است رابطه میان عدم تعادل توزیع داده‌ها و خودهمبستگی مکانی در فرآیند مدل‌سازی، به‌ویژه برای متغیرهای محیطی مهمی مانند SOC، مورد بررسی دقیق قرار گیرد.

از این رو، پرسش اساسی این پژوهش آن است که چگونه توزیع نامتوازن داده‌ها می‌تواند باعث افزایش خودهمبستگی مکانی در باقیمانده‌های مدل گردد؟ همچنین آیا اصلاح عدم تعادل داده‌ها می‌تواند منجر به کاهش خودهمبستگی مکانی و افزایش دقت مکانی پیش‌بینی شود؟ بنابراین، هدف این مطالعه بررسی تأثیر توزیع داده‌ها (متعادل یا نامتعادل) بر میزان دقت و خودهمبستگی مکانی باقیمانده‌های مدل در مدل‌سازی کربن آلی خاک با استفاده از مدل جنگل تصادفی است. برای این منظور، ابتدا توزیع مقدار کربن آلی خاک در مجموعه‌ای از ۲۸۰ نمونه سطحی متعادل‌سازی و اصلاح شده و سپس عملکرد مکانی مدل در دو حالت با و بدون تعادل توزیع داده‌ها مقایسه شده است.

مواد و روش‌ها

تشریح منطقه مطالعاتی و داده‌های خاک

این پژوهش در بخشی از دشت قزوین با مساحت تقریبی ۶۰۰۰۰ هکتار انجام شد (شکل ۱). منطقه مورد مطالعه بین 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 واقع شده و ارتفاع آن از ۱۱۴۴ متر در پایین‌ترین نقطه تا ۱۷۵۳ متر در بلندترین نقطه متغیر است. مواد مادری خاک در بخش‌های شمالی و مرکزی منطقه عمدتاً شامل رسوبات آبرفتی و کوهرفتی مربوط به دوره کواترنر و در نواحی جنوبی کفه‌های گلی و نمکی می‌باشند. از نظر توپوگرافی، منطقه دارای تنوع فیزیوگرافی قابل توجهی است که شامل تپه‌ها، واریزه‌های سنگریزه‌ای، دشت‌های دامنه‌ای، دشت‌های آبرفتی و اراضی پست می‌شود. کاربری اراضی عمدتاً شامل اراضی زراعی آبی و دیم، مراتع شور و غیرشور است.



شکل ۱- منطقه مورد مطالعه و پراکنش مکانی نقاط نمونه‌برداری شده

به منظور انجام این مطالعه، ۲۸۰ نمونه سطحی (عمق ۰-۳۰ سانتی‌متری) طبق الگوی نمونه‌برداری تصادفی طبقه‌بندی شده از قسمت‌های مختلف منطقه برداشت شد. بعد از کوبیدن و عبور نمونه‌های خاک از الک ۲ میلی‌متری، میزان درصد کربن آلی نمونه‌ها به روش سوزاندن تر (Walkely and Black, 1934) اندازه‌گیری شد.

مدل‌سازی با الگوریتم جنگل تصادفی

در این پژوهش، از الگوریتم جنگل تصادفی (Random Forest) به‌منظور مدل‌سازی مکانی کربن آلی خاک (SOC) استفاده شد. این الگوریتم به‌عنوان یکی از روش‌های قدرتمند یادگیری ماشین، با بهره‌گیری از مجموعه‌ای از درخت‌های تصمیم‌گیری و فرآیند نمونه‌گیری بوت‌استرپ، قادر است روابط غیرخطی پیچیده بین متغیر هدف و مجموعه‌ای از متغیرهای مستقل را به خوبی شناسایی نماید. پیاده‌سازی این مدل با استفاده از بسته‌ی randomForest در محیط نرم‌افزار R انجام شد.

متغیرهای محیطی مورد استفاده در مدل‌سازی، دارای وضوح مکانی ۱۲/۵ متر بودند. داده‌های توپوگرافی از مدل ارتفاعی دیجیتال ALOS استخراج شده و برای محاسبه شاخص‌های مرتبط با پوشش گیاهی، شوری، کانی‌ها، رطوبت و نسبت‌های طیفی، تصاویر ماهواره‌ای Sentinel-2 به‌کار گرفته شد. علاوه‌براین، میانگین دمای سالانه و میانگین بارندگی سالانه به‌عنوان شاخص‌های اقلیمی لحاظ شدند. به‌منظور انتخاب مؤثرترین متغیرها، از روش حذف بازگشتی ویژگی‌ها (Recursive Feature Elimination - RFE) استفاده شد (Chen & Jeong, 2007). بر این اساس، متغیرهای مهم شامل ارتفاع، عمق دره، شاخص سبزینگی، شاخص رطوبت ساگا و میانگین بارندگی سالانه برای مدل‌سازی نهایی برگزیده شدند.

متعادل‌سازی داده‌ها

بررسی توزیع آماری کربن آلی خاک نشان داد که داده‌ها به‌شدت نامتوازن هستند، به‌طوری‌که بیش از ۷۰ درصد نمونه‌ها دارای مقادیر کمتر از ۱ درصد بودند. برای کاهش اثر این عدم تعادل در فرآیند یادگیری مدل، از روش SMOTE (Synthetic Minority Oversampling Technique) در محیط R استفاده شد. این روش با استفاده از نمونه‌گیری مصنوعی از مقادیر حداقل، موجب متعادل‌سازی توزیع SOC گردید.

اعتبارسنجی مدل

به‌منظور ارزیابی دقت مدل‌ها، از روش اعتبارسنجی متقابل ۱۰-تایی (10-fold cross-validation) استفاده شد. برای اعتبارسنجی مدل‌ها، از شاخص‌های آماری ضریب تعیین (R^2) و RMSE، استفاده شد.

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

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

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

مقدار شاخص موران با استفاده از ماتریس وزنی همسایگی ایجادشده بر مبنای فاصله جغرافیایی بین نمونه‌ها و تابع moran.test از بسته‌ی spdep در محیط R محاسبه گردید. کاهش مقدار Moran's I در باقیمانده‌ها نشانه‌ای از بهبود دقت مکانی مدل است.

نتایج و بحث

بررسی مشخصات آماری کربن آلی خاک (SOC) در عمق ۰ تا ۳۰ سانتی‌متری نشان می‌دهد (جدول ۱) که این متغیر از تنوع قابل‌توجهی در منطقه برخوردار است. مقدار SOC بین ۰ تا ۲/۸ درصد متغیر بوده و میانگین آن برابر با ۰/۸۷ درصد محاسبه شد. انحراف معیار نسبتاً بالا (۰/۴۸) و ضریب تغییرات (CV) برابر با ۰/۵۶ نشان‌دهنده پراکندگی متوسط تا زیاد داده‌ها در سطح منطقه است. توزیع داده‌ها چوله‌ی مثبت (Skewness =۱/۱۴) و کشیدگی بالاتر از توزیع نرمال (Kurtosis =۱/۴۳) دارد که بیانگر توزیع نامتقارن با غلبه مقادیر پایین SOC در نمونه‌ها می‌باشد. این ویژگی آماری، احتمال ایجاد عدم تعادل (Imbalance) در داده‌ها را افزایش می‌دهد و بر عملکرد مدل‌های پیش‌بینی اثرگذار است، لذا در ادامه تحلیل، تأثیر متعادل‌سازی داده‌ها بر دقت و ساختار مکانی مدل بررسی می‌گردد.

جدول ۱- شاخص‌های آماری کربن آلی خاک (SOC) در عمق ۰-۳۰ سانتی‌متر

حداقل	حداکثر	میانگین	انحراف معیار	واریانس	چولگی	کشیدگی	ضریب تغییرات

۰/۵۶	۱/۴۳	۱/۱۴	۰/۲۳	۰/۴۸	۰/۸۷	۲/۸۰	۰/۰۰

متعادل‌سازی داده‌ها

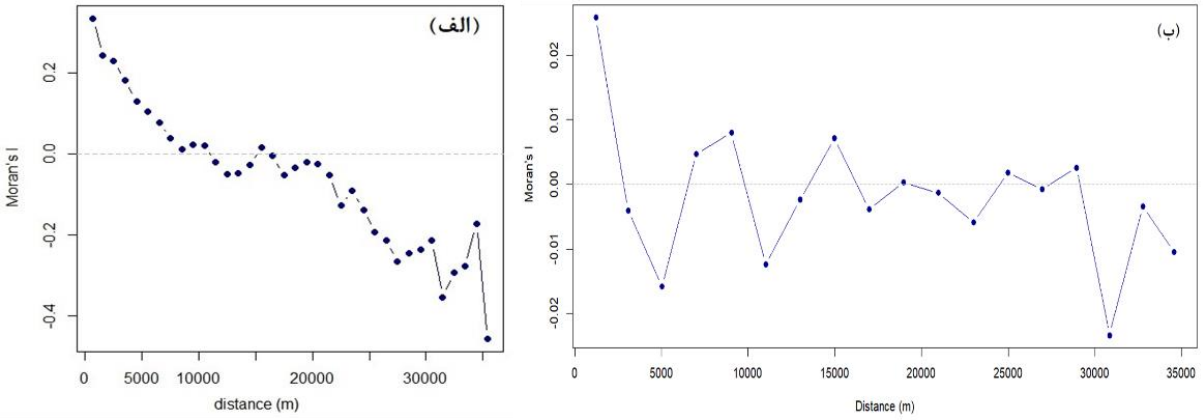
مقایسه عملکرد مدل جنگل تصادفی در دو سناریوی استفاده از داده‌های نامتعادل و متعادل‌شده، نشان می‌دهد که متعادل‌سازی آماری داده‌ها نه‌تنها موجب بهبود دقت پیش‌بینی مدل می‌شود، بلکه می‌تواند به‌طور مؤثری ساختار مکانی خطاها را کاهش دهد. در مطالعات اخیر نیز تأکید شده است که عدم تعادل در توزیع داده‌ها، به‌ویژه در متغیرهای محیطی با توزیع چوله مانند کربن آلی خاک، می‌تواند منجر به سوگیری مدل نسبت به مقادیر غالب شده و موجب بروز الگوهای فضایی در باقیمانده‌ها گردد (Santos et al., 2020; Li et al., 2021; Ebrahimi et al., 2023). در مقابل، به‌کارگیری روش‌هایی مانند SMOTE ، ضمن بهبود عملکرد عددی مدل، موجب کاهش خودهمبستگی مکانی در باقیمانده‌ها و افزایش قابلیت تعمیم‌پذیری مدل می‌شود (Chen et al., 2022; Wang et al., 2024).

در حالت استفاده از داده‌های نامتعادل، مقدار ریشه میانگین مربعات خطا (RMSE) برابر با ۰/۴ و ضریب تعیین (R^2) معادل ۰/۳۹ به‌دست آمد (جدول ۲). همچنین مقدار شاخص خودهمبستگی مکانی Moran's I برای باقیمانده‌های مدل ۰/۳۰ محاسبه شد، که نشان‌دهنده‌ی وجود ساختار مکانی نسبتاً قوی در خطاهای مدل است. تحلیل کورلوگرام (correlogram) نیز بیانگر آن بود که مقدار شاخص موران با افزایش فاصله بین نقاط کاهش می‌یابد (شکل ۲)، که دلالت بر خودهمبستگی مکانی مثبت در فواصل کوتاه دارد. چنین رفتاری در باقیمانده‌ها نشان می‌دهد که مدل در بازنمایی صحیح الگوهای فضایی SOC ناکارآمد عمل کرده و اطلاعات مکانی به‌طور کامل از طریق مدل‌سازی دریافت نشده‌اند (Li et al., 2021; Nussbaum et al., 2018).

در مقابل، پس از متعادل‌سازی داده‌ها با استفاده از روش SMOTE و اصلاح نسبت نمونه‌ها بین بیشترین و کمترین مقادیر ، عملکرد مدل به‌طور محسوسی بهبود یافت. در این شرایط، مقدار RMSE به ۰/۳ کاهش یافته و R^2 به ۰/۶ افزایش پیدا کرد. همچنین مقدار Moran's I به ۰/۰۳ رسید که بیانگر کاهش شدید خودهمبستگی مکانی در باقیمانده‌ها است. افزون بر این، روند تغییرات Moran's I در کورلوگرام برای داده‌های متعادل‌شده، نوساناتی اندک و حول صفر نشان داد که دلالت بر عدم وجود ساختار مکانی معنادار در خطاها دارد (Ebrahimi et al., 2023; Chen et al., 2022).

این نتایج با یافته‌های مطالعات مشابه هم‌راستا است که نشان داده‌اند عدم تعادل در داده‌ها منجر به تعصب مدل نسبت به مقادیر غالب شده و دقت پیش‌بینی در مقادیر با کمترین فراوانی را به‌شدت کاهش می‌دهد (Scholten, et al. 2023). علاوه‌بر این، این تعصب می‌تواند ساختارهای فضایی پنهان را در باقیمانده‌ها حفظ کند. در حالی‌که متعادل‌سازی داده‌ها به کمک روش‌هایی مانند SMOTE باعث توزیع متوازن‌تر خطاها در فضای جغرافیایی شده و کیفیت فضایی مدل را افزایش می‌دهد (Santos et al., 2020; Wang et al., 2024).

در مجموع، یافته‌های این پژوهش نشان می‌دهد که متعادل‌سازی آماری داده‌ها نه‌تنها باعث بهبود شاخص‌های عددی عملکرد مدل می‌شود، بلکه به‌طور معناداری خودهمبستگی مکانی باقیمانده‌ها را کاهش می‌دهد و مدل را از نظر مکانی پایاتر و قابل اطمینان‌تر می‌سازد.

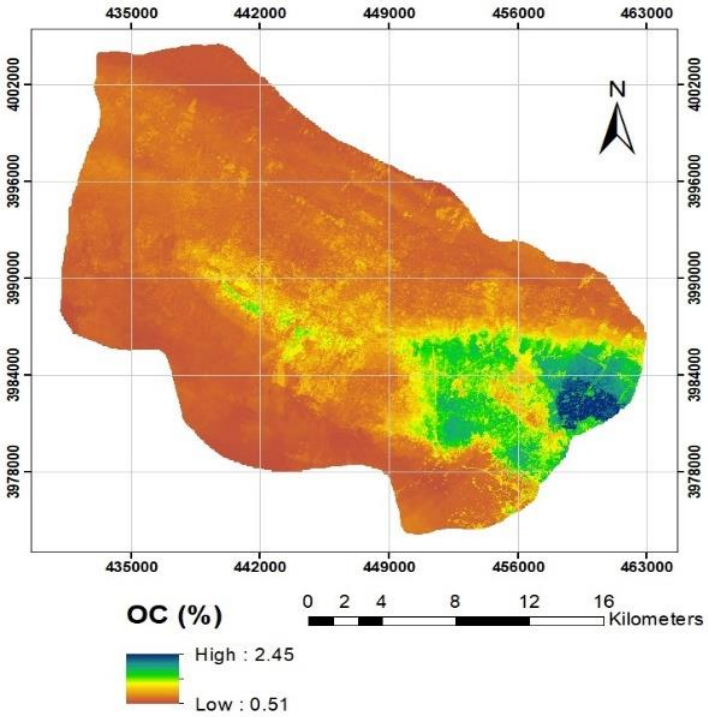


شکل ۲- روند تغییرات شاخص موران در داده‌های نامتعادل (الف) و متعادل شده (ب)

جدول ۲- مقایسه شاخص‌های ارزیابی مدل جنگل تصادفی در دو حالت داده‌های نامتعادل و متعادل‌شده

شاخص ارزیابی	داده‌های نامتعادل	داده‌های متعادل‌شده
ریشه میانگین مربعات خطا (RMSE)	۰/۴۰	۰/۳۰
ضریب تعیین (R^2)	۰/۳۹	۰/۶۰
شاخص خودهمبستگی مکانی (Moran's I)	۰/۳۰	۰/۰۳
روند شاخص موران با فاصله	کاهش تدریجی	ثابت و نزدیک به صفر
توزیع فضایی باقیمانده‌ها	خودهمبسته	تصادفی

درنهایت، نقشه توزیع مکانی کربن آلی خاک براساس داده‌های متعادل شده تهیه شد (شکل ۳). در قسمت شمال شرق، شرق و مرکز منطقه مورد مطالعه بیشترین مقدار کربن آلی مشاهده شد که به علت وجود کاربری کشت آبی، دیم و همچنین وجود مراتع غیر شور می‌باشد در حالیکه در قسمت جنوب و جنوب شرق، به علت وجود دشت‌های شور و عدم وجود پوشش گیاهی، مقدار کربن آلی کاهش یافته است.



شکل ۳- توزیع مکانی کربن آلی خاک براساس داده‌های متعادل شده

نتیجه‌گیری

نتایج این مطالعه نشان داد که عدم تعادل در توزیع داده‌ها می‌تواند منجر به کاهش دقت پیش‌بینی و حفظ ساختارهای مکانی در خطاهای مدل گردد. استفاده از روش SMOTE برای متعادل‌سازی آماری داده‌ها، نه‌تنها عملکرد مدل جنگل تصادفی را از نظر شاخص‌های RMSE و R^2 بهبود بخشید، بلکه به‌طور معناداری خودهمبستگی مکانی باقیمانده‌ها را کاهش داد. کاهش مقدار شاخص موران و پایداری کورلوگرام حول صفر در مدل متعادل‌شده، حاکی از کارایی بیشتر مدل در بازنمایی الگوهای فضایی SOC بود. بنابراین، در مطالعات آینده توصیه می‌شود پیش از مدل‌سازی، وضعیت تعادل داده‌ها بررسی و در صورت نیاز، اصلاح گردد تا نتایج دقیق‌تر و معتبرتری حاصل شود. ترکیب روش‌های یادگیری ماشین با تکنیک‌های آماری متعادل‌سازی، می‌تواند به عنوان رویکردی مؤثر در مدل‌سازی مکانی متغیرهای خاکی مورد استفاده قرار گیرد.

فهرست منابع

Adhikari, K., & Hartemink, A. E. (2016). Linking soils to ecosystem services — A global review. Geoderma, 262, 101–111.
Brus, D. J., Kempen, B., & Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. European Journal of Soil Science, 62(3), 394–407.
Chen, Z., et al. (2022). Mitigating spatial bias in random forest models with balanced sampling. Remote Sensing, 14(12), 2809.
Ebrahimi, A., et al. (2023). Improving prediction of soil properties by handling imbalanced data in machine learning models. Environmental Modelling & Software, 161, 105594.
Hengl, T., et al. (2017). SoilGrids250m: Global gridded soil information based on machine learning. PLOS ONE, 12(2), e0169748.
Hengl, T., Heuvelink, G. B. M., & Stein, A. (2007). Comparison of kriging with external drift and regression-kriging. Geoderma, 140(1–2), 480–496.
Heung, B., Bulmer, C. E., & Schmidt, M. G. (2016). Predictive soil parent material mapping at a regional-scale: A Random Forest approach. Geoderma, 279, 68–83.
Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. Science, 304(5677), 1623–1627.
Li, W., et al. (2021). Effect of sample imbalance on digital soil mapping accuracy. Catena, 203, 105309.
Minasny, B., et al. (2017). Soil carbon 4 per mille. Geoderma, 292, 59–86.
Nussbaum, M., Papritz, A., Baltensweiler, A., & Walthert, L. (2018). Evaluating digital soil mapping models for soil organic carbon at the national scale. Geoderma, 318, 45–58.
Santos, H. G., et al. (2020). Addressing data imbalance in soil mapping using synthetic oversampling. Geoderma, 373, 114399.
Scholten, T., et al. (2023). Coping with imbalanced data problem in digital mapping of soil functional classes using SMOTE. European Journal of Soil Science, 74(2), e13368.
Wadoux, A. M., et al. (2021). Machine learning for digital soil mapping: Applications, challenges, and perspectives. Earth-Science Reviews, 223, 103824.
Walkley, A., & Black, I. A. (1934). An examination of the Degtjareff method for determining soil organic matter and a proposed modification of the chromic acid titration method. Soil Science, 37(1), 29–38.
Wang, Y., et al. (2024). Spatial implications of data distribution in machine learning-based soil mapping. Soil Systems, 8(1), 21.

Evaluating the Influence of Data Imbalance on Accuracy and Spatial Autocorrelation in Soil Organic Carbon Modeling

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

Abstract
Data imbalance in environmental datasets presents a significant challenge in the spatial modeling of soil properties, particularly soil organic carbon (SOC). This study explores the effect of data balancing on model accuracy and the spatial structure of residuals across a 60,000-hectare area in the Qazvin Plain, Iran. A total of 280 surface samples (0–30 cm depth) were analyzed for SOC content, alongside a suite of environmental covariates including topographic metrics, vegetation indices, soil moisture parameters, and climatic variables. Key predictors were identified using recursive feature elimination (RFE), and random forest (RF) models were implemented under two scenarios: imbalanced data and balanced data via the synthetic minority oversampling technique (SMOTE). Model performance was evaluated using root mean square error (RMSE), coefficient of determination (R^2), and Moran's I index of spatial autocorrelation. The SMOTE-adjusted RF model exhibited improved predictive accuracy (RMSE = 0.3, R^2 = 0.60) relative to the imbalanced case (RMSE = 0.4, R^2 = 0.40). Moreover, Moran's I for model residuals dropped substantially from 0.30 to 0.03 following data balancing. Correlogram analysis corroborated this reduction, indicating weakened spatial structure in residuals. These results highlight the value of statistical data balancing in enhancing both numerical precision and spatial consistency in SOC prediction models.
Keywords: Moran's I, SMOTE, Spatial modeling