

پیش‌بینی و نقشه‌برداری کربن آلی خاک با استفاده از سنجش از دور و الگوریتم‌های یادگیری ماشین

ثریا بندک^{۱*}، سید علیرضا موحدی نائینی^۱، چوقی بایرام کمکی^۲، محمد کاکوئی^۳ و جوشم ورلست^۴

(تاریخ دریافت: ۱۴۰۱/۸/۲۹؛ تاریخ پذیرش: ۱۴۰۱/۱۲/۲۴)

چکیده

کربن آلی خاک (SOC) که یکی از حیاتی‌ترین خصوصیات فیزیکی و شیمیایی خاک است، نقش مهمی در تولید پایدار کشاورزی دارد و از تخریب و فرسایش خاک جلوگیری می‌کند. داده‌کاوی و مدل‌سازی مکانی همراه با تکنیک‌های یادگیری ماشینی به منظور بررسی میزان کربن آلی خاک مبتنی بر داده‌های سنجش از دور به صورت گسترده مورد توجه قرار گرفته است. هدف از این مطالعه بررسی تغییرات مقدار SOC با استفاده از تکنیک سنجش از دور در مقایسه با روش‌های نمونه‌برداری مزرعه‌ای در مناطقی از گنبد کاووس و جنگل نیلی آزادشهر است. نمونه‌های خاک به منظور اندازه‌گیری مقدار SOC در آزمایشگاه جمع‌آوری شد. داده‌ها به دو دسته ۷۰ درصد برای واسنجی (آموزشی) و ۳۰ درصد اعتبارسنجی (آزمون) دسته‌بندی شدند و از الگوریتم‌های یادگیری ماشین شامل جنگل تصادفی، ماشین بردار پشتیبان، درخت تصمیم پیشرفته، و افزایش گرادیان اکستریم برای برآورد و تهیه نقشه کربن آلی خاک استفاده شد. متغیرهای کمکی پیش‌بینی کربن آلی خاک شامل باندها، شاخص‌های منتج از ماهواره لندست ۸ و سنتینل ۲، اقلیم و توپوگرافی است. نتایج نشان داد که استخراج مؤلفه‌های مربوط به باندها همراه با محاسبه شاخص‌هایی مانند تفاوت پوشش گیاهی نرمال شده، شاخص خیسسی، شاخص همواری دره به‌عنوان متغیرهای کمکی نقش مهمی در برآورد صحیح‌تر مقدار ماده آلی خاک دارند. همچنین مقایسه رگرسیون‌های مختلف تخمین نشان داد، مدل جنگل تصادفی در ماهواره سنتینل ۲ و مدل درخت تصمیم پیشرفته در لندست ۸ به ترتیب با مقادیر ضریب تبیین، خطای جذر میانگین مربعات و میانگین خطای مطلق ۰/۶۴، ۰/۰۵ و ۰/۱۷ بهترین کارایی را نسبت به سایر رویکردهای استفاده شده در برآورد کربن آلی خاک سطحی در منطقه مطالعاتی ارائه کردند. نتایج این مطالعه بر قابلیت داده‌های سنجش از دور و مدل یادگیری در تخمین مکانی کربن آلی خاک دلالت دارد و می‌تواند به‌عنوان روشی جایگزین برای روش‌های آزمایشگاهی در تعیین برخی ویژگی‌های خاک از جمله کربن آلی خاک مورد توجه قرار گیرد.

واژه‌های کلیدی: کربن آلی خاک، یادگیری ماشین، رگرسیون، سنجش از دور، اعتبارسنجی

۱. گروه علوم خاک، دانشکده آب و خاک، دانشگاه علوم کشاورزی و منابع طبیعی گرگان، گلستان، ایران.
۲. گروه بیابان، دانشکده مرتع و آبخیزداری، دانشگاه علوم کشاورزی و منابع طبیعی گرگان، گلستان، ایران.
۳. گروه مهندسی الکترونیک، دانشگاه صنعتی چالمرز، گوتنبرگ، سوئد.
۴. گروه سنجش از دور، دانشکده علوم زمین، دانشگاه والنسیا، والنسیا، اسپانیا.

*: مسئول مکاتبات: پست الکترونیکی: soraya.bandak@gmail.com

مقدمه

مقدار ماده آلی در خاک، یکی از منابع اصلی کربن خاک، تغذیه گیاه و تعیین کننده میزان حاصلخیزی خاک است. همچنین، مقدار ماده آلی در خاک نقش مهمی در چرخه آب و کیفیت آن ایفا می‌کند (۲۴). کشور ایران، با میانگین بارش سالانه ۲۵۰ میلی‌متر، یکی از کم آب‌ترین کشورهای جهان با شرایط اقلیمی خشک و نیمه‌خشک است. مقدار و کیفیت ماده آلی خاک در حاصلخیزی، مدیریت کوددهی و آبیاری محصولات گوناگون کشاورزی اهمیت فراوانی دارد از این رو، در دشت‌های وسیع کشور، لزوم تهیه نقشه ماده آلی خاک کاملاً احساس می‌شود (۱). کیفیت نمونه‌برداری و دقت در آنالیزهای آزمایشگاهی، در فرایند تهیه نقشه ماده آلی خاک، اهمیت فراوانی دارد. به‌منظور تعیین توزیع کربن آلی در خاک، روش‌های گوناگونی مانند نمونه‌برداری شبکه‌ای و یا منطقه‌ای به کار می‌رود. از آنجا که نمونه‌برداری خاک از نظر فنی دشوار، پرهزینه و آماده‌سازی نمونه‌ها نیز برای تجزیه و تحلیل زمان‌بر است، همواره تعداد محدودی نمونه از اراضی تهیه و آنالیز می‌شود. بیان مقدار کربن آلی خاک در اراضی و در قالب کمیته عددی همچون میانگین کلی مقدار کربن آلی خاک، به‌ویژه در اراضی پهناور، کافی نیست و نیاز مبرم به روش‌های کم هزینه و دقیق‌تر احساس می‌شود (۳).

در دهه‌های اخیر پژوهش‌های بی‌شماری در خصوص امکان استفاده از فناوری‌های نوین در فعالیت‌های کشاورزی به‌منظور اجرای راهکارهای مدیریتی دقیق‌تر صورت گرفته است. تکنیک‌هایی چون تحلیل تغییر مکانی تولید محصول و استفاده از تکنیک‌های سنجش از دور (Remote Sensing) و سیستم‌های اطلاعات جغرافیایی (GPS)، بر مبنای اطلاعات و داده‌های کشاورزی منجر به بهبود قابل توجهی در فعالیت‌های کشاورزی شده است (۵). استفاده از فناوری‌های نوین، امکان دسترسی به اطلاعات با حجم و دقت بیشتر در زمان کمتر را فراهم می‌کند. این ویژگی‌های منحصربه‌فرد، کاربرد آن را در علوم مختلف روز

به روز گسترش می‌دهد. یکی از علومی که توانسته است کمک شایانی به پیشرفت آن نماید علوم خاک است (۵).

بررسی مقدار کربن آلی خاک با استفاده از روش‌های معمول و قدیمی، علاوه بر زمان‌بر بودن، هزینه‌های فراوانی را در پی خواهد داشت. در این زمینه، سنجش از دور روش بسیار مفیدی است که می‌توان از آن، به‌منظور به‌دست آوردن لایه‌های اطلاعاتی از خاک و پوشش گیاهی دشت‌ها و مناطق گوناگون، استفاده کرد. ویژگی‌هایی مانند فراهم آوردن دید وسیع و یکپارچه از منطقه، تکرارپذیری، در دسترس بودن اطلاعات و دقت بالای آنها، صرفه‌جویی در زمان و هزینه از جمله دلایلی است که استفاده از این‌گونه اطلاعات را به‌منظور بررسی مقدار ماده آلی خاک و نیز کنترل تغییرات آن، بر دیگر روش‌های رایج برتری می‌دهد. کاربرد روش‌های نوین سنجش از دور تصاویر ماهواره‌ای و استفاده از شاخص‌های طیفی مبتنی بر بازتاب طیفی پدیده‌ها می‌تواند نتایج رضایت بخش و مطلوبی برای پژوهشگران داشته باشد (۷).

پژوهش‌های انجام شده نشان داده‌اند که انعکاس حاصل از باندهای طیفی با ویژگی‌های خاک همبستگی دارد و قادر است خصوصیات فیزیکی، شیمیایی و بیولوژیکی را با کاهش هزینه و افزایش سرعت نسبت به روش‌های تجزیه آزمایشگاهی تخمین بزند (۸). تصاویر ماهواره‌ای توسط پژوهشگران فراوانی در زمینه علوم خاک به روش‌های مختلفی استفاده شده است (۹). افزایش رطوبت خاک و میزان ماده آلی منجر به کاهش در مقادیر بازتابی از سطح زمین می‌شود و بر اساس یافته‌های بدست آمده باندهای مادون قرمز نزدیک (۲/۱۶ تا ۲/۹ میکرومتر) بیشترین همبستگی را با خصوصیات طیفی ماده آلی نشان می‌دهند (۱۰). مقدار ماده آلی سطحی خاک در طبیعت دارای تغییرات زمانی و مکانی بوده که برآورد میزان و مقدار آن توسط روش‌های سنجش از دور قابل بررسی است. از آنجایی که تصاویر ماهواره‌ای دارای قابلیت‌های متمایزی نظیر پوشش تکراری، چند طیفی بودن، دید وسیع و تنوع فرم

مورد مطالعه باشد. در مطالعات شناسایی خاک‌ها، به دلیل وسعت زیاد، بالا بودن هزینه‌های اجرا، غیرقابل کنترل بودن تغییرات و عوامل تغییر دهنده، نمونه‌برداری باید هماهنگ طراحی شود. در ابتدا، نقشه‌ها و تصاویر ماهواره‌ای جمع‌آوری می‌شود. در این مرحله نیاز به محاسبه تبدیل‌های رادیومتری، انعکاس و تصحیح‌های اتمسفری نبود چون تصویر دریافتی از Google Earth Engine پردازش شده است. از سری نقشه‌های پایه مثل شیب، ارتفاع محلی، شاخص‌های سنجش از دوری قطعه‌بندی در Google Earth Engine بر اساس لایه‌های اطلاعاتی و نقشه‌های پایه به روش قطعه‌بندی قطعات مشابه استفاده می‌شود تا به صورت لایه‌برداری تهیه شود. سپس با روش نمونه‌برداری تصادفی ۳۹۴ نمونه در سطحی حدود ۳۷۳۳۰ هکتار برداشت می‌شود. نمونه‌برداری به صورت نقطه‌ای است که دایره‌ای به شعاع ۵ متر ترسیم شده و هشت نمونه از عمق صفر تا ۱۰ سانتی‌متری گرفته و با هم ترکیب می‌شود. پس از هوا خشک شدن و عبور تمامی نمونه‌ها از الک دو میلی‌متر، مقدار کربن آلی خاک به روش واکلی بلاک (۲۹) برای تمامی نمونه‌ها اندازه‌گیری و در نهایت SOC محاسبه شد.

متغیرهای کمی محیطی

در مطالعه حاضر، از باندهای مربوط به تصاویر سنجنده OLI لندست ۸ (۲۴ ژوئن) به ترتیب در مسیر و ردیف ۱۶۴ و ۳۶ دریافت شده از Google Earth Engine استفاده شد. با توجه به هدف پژوهش، باندهای مرئی (باندهای ۲، ۳ و ۴)، مادون قرمز نزدیک (باند ۵)، مادون قرمز طول موج کوتاه (باندهای ۶ و ۷) با اندازه پیکسل‌های ۳۰ متر و باندهای حرارتی (باندهای ۱۰ و ۱۱) با اندازه پیکسل‌های ۱۰۰ متر برای استخراج ارزش‌های طیفی متناظر با نقاط زمینی و تجزیه و تحلیل استفاده شد.

ماهواره سنتینل ۲

این ماهواره ۱۳ باند طیفی در محدوده‌های مرئی، مادون قرمز نزدیک و مادون قرمز طول موج کوتاه را تصویربرداری

اطلاعاتی هستند این تکنیک‌ها می‌تواند به عنوان یک منبع جایگزین در تهیه اطلاعات در زمینه‌های مختلف مطرح شود. در کشور ایران هنوز هم بیشتر نقشه‌های موجود مربوط به ویژگی‌های کلیدی خاک (کربن آلی، شوری، بافت خاک و ...) در مناطق با کاربری‌های مختلف گنبد کاووس، با قدرت تفکیک مکانی بالا برای هرگونه مدیریت و بهره‌برداری از اراضی موجود نیست. بنابراین با توجه به اهمیت کربن خاک به عنوان یک ویژگی پویا و تأثیرگذار بر روی قابلیت و تناسب اراضی، پژوهش حاضر با اهداف زیر طراحی شده است: (۱) تهیه نقشه تغییرات کربن خاک بر اساس پروژه جهانی نقشه‌برداری رقومی (۲) تعیین مهم‌ترین فاکتورهای محیطی پیش‌بینی کننده تغییرات کربن خاک و (۳) مقایسه عملکرد الگوریتم‌های یادگیری ماشین جنگل تصادفی، افزایش گرادیان اکسترمم، درخت تصمیم پیشرفته و ماشین بردار پشتیبان در مدل‌سازی مکانی کربن خاک.

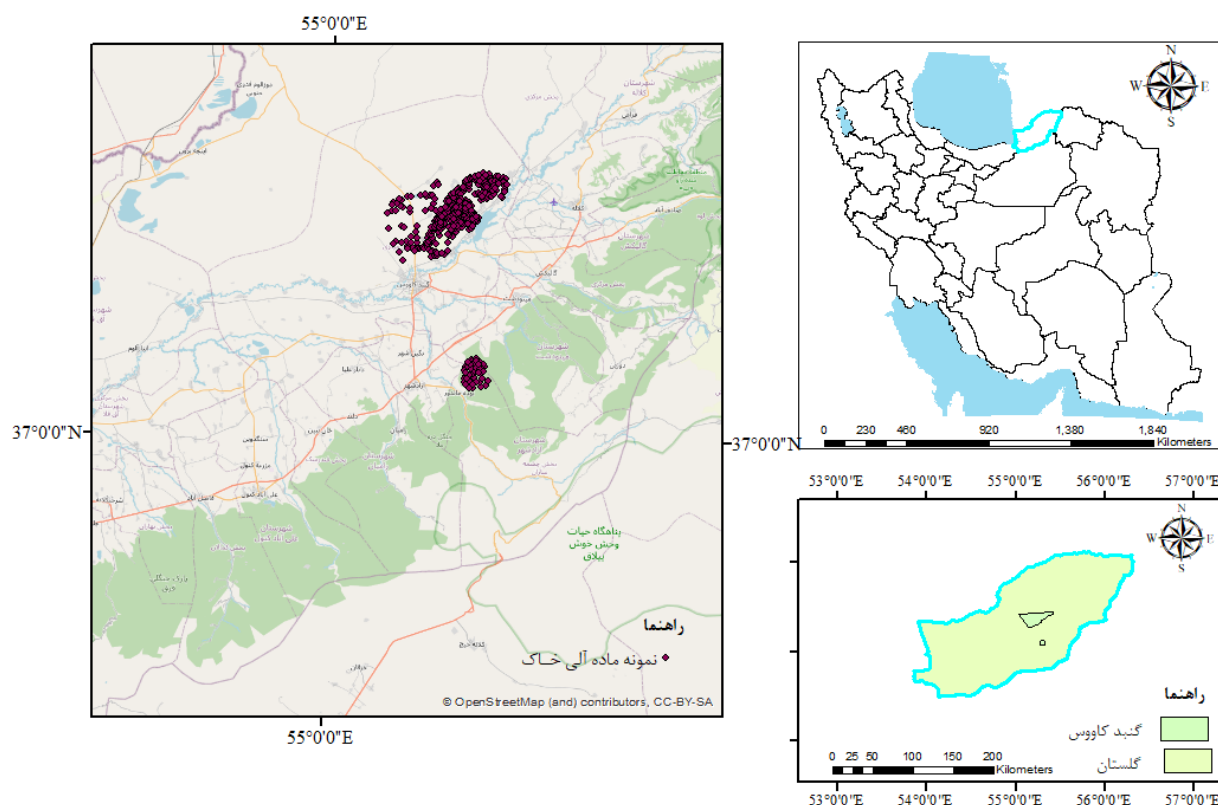
مواد و روش‌ها

منطقه مطالعاتی

منطقه مورد مطالعه ۳۷۳۳۰ هکتار است که در شمال شرق شهرستان گنبد کاووس در امتداد ساحل راست رودخانه گرگانرود بین ۳۷ درجه و ۱۵ دقیقه تا ۳۷ درجه و ۲۵ دقیقه عرض شمالی و ۵۵ درجه و ۱۰ دقیقه تا ۵۵ درجه و ۲۲ دقیقه طول شرقی در دامنه‌های جنوبی تپه ماهورهای شمالی گرگانرود واقع شده است (شکل ۱). متوسط بارندگی سالیانه ۶۲۶ میلی‌متر و متوسط درجه حرارت سالیانه ۱۸/۶ درجه سانتی‌گراد است. این منطقه دارای رژیم‌های رطوبتی اریدیک، یودیک و زیریک و حرارتی ترمیک و مزیک است. واحدهای فیزیوگرافی منطقه شامل دشت‌های دامنه‌ای، کوه و تپه و کاربری‌های اصلی منطقه شامل زراعت، مرتع و جنگل است.

طراحی نقاط نمونه‌برداری

طراحی نقاط نمونه‌برداری باید مبتنی بر نیازهای مطالعه، اهداف مقیاس، فرضیه‌های اولیه و بیان گر کل دامنه و تغییرات جامعه



شکل ۱. منطقه مطالعاتی

مدل‌های یادگیری ماشین جنگل تصادفی، ماشین بردار پشتیبان، درخت تصمیم و افزایش گرادیان اکسترمم اقدام به مدل‌سازی در زبان برنامه نویسی کولب (پایتون) می‌شود. نقشه‌های نهایی به کمک نرم‌افزار Google Earth Engine تهیه و استفاده شد.

برآورد کربن آلی خاک

در این پژوهش از رویکردهای مدل جنگل تصادفی، ماشین بردار پشتیبان، درخت تصمیم و XGBoost برای مدل‌سازی مکانی کربن آلی خاک استفاده شد. از یادگیری ماشین برای پیش‌بینی توزیع فضایی ماده آلی خاک به‌عنوان متغیر وابسته بر اساس شاخص‌های سنجش از دور، توپوگرافی و سایر عوامل محیطی (به‌عنوان متغیر مستقل) استفاده می‌شود. ماشین‌های بردار پشتیبان که برای اولین بار توسط لویز

می‌کند. توانایی تفکیک مکانی ماهواره Sentinel 2 در باندهای مختلف از ۱۰ تا ۶ متر متغیر است که در عرض برداشت بالا (۲۹۰ کیلومتر) برداشت می‌شود. توان تفکیک زمانی Sentinel 2 در شرایط جوی مناسب و بدون ابر، ۵ روز در استوا است. در عرض‌های میانی توان تفکیک زمانی این ماهواره (در حالت استفاده از دو ماهواره آن) ۲ تا ۳ روز خواهد بود که در مقایسه با سایر ماهواره‌ها به‌مراتب کمتر است، این ماهواره از توان تفکیک زمانی بالایی برخوردار است. تصاویر Sentinel 2 با داشتن این مزیت، می‌توانند برای نظارت بر تغییرات پوشش گیاهی در فصل رشد بسیار کاربردی باشند.

برای محاسبه شاخص‌های طیفی در این پژوهش، از بین شاخص‌های گوناگون شاخص پوشش گیاهی، رطوبت، اقلیم و توپوگرافی محاسبه شد (جدول ۱). سپس با روش داده‌کاوی بر اساس داده‌های میدانی با استفاده از

جدول ۱. متغیرهای کمکی محیطی استفاده شده در پیش‌بینی مکانی کربن آلی خاک

منبع	رابطه	مخفف	شاخص
(۲)	$\frac{(1+i)(NIR - Red)}{(NIR + Red + L)}$	SAVI	شاخص گیاهی تعدیل کننده اثر خاک
(۲۶)	$\frac{(NIR - Red)}{(NIR + Red)}$	NDVI	شاخص گیاهی اختلاف نرمال
(۷)	$\frac{2.5(NIR - Red)}{(NIR + 6R - 7.5B) + 1}$	EVI	شاخص گیاهی بهبود یافته
(۳۱)	$(NIR2 - Red2) / (NIR2 + Red2)$	GDVI	شاخص گیاهی تعمیم یافته
(۱۷)	$\frac{(NIR - SWIR)}{(NIR + SWIR)}$	NDWI	شاخص نرمال شده تفاوت آب
(۳۳)	$\frac{(GREEN - MIR)}{(GREEN + MIR)}$	MNDWI	شاخص اصلاحی نرمال شده تفاوت آب
(۱۸)	$NDSI = (Red - NIR) / (Red + NIR)$	NDSI	شاخص شوری
(۱۵)	Word Clim	MAP	میانگین بارندگی سالانه
(۱۵)	Word Clim	MAT	میانگین دمای سالانه
(۱۶)	DEM	MrVBF	شاخص همواری دره با درجه تفکیک بالا
(۴)	DEM	TWI	شاخص خیزی

شاید به گونه‌ای بتوان محبوبیت کنونی روش ماشین بردار پشتیبان را با محبوبیت شبکه‌های عصبی در دهه گذشته مقایسه کرد. علت این قضیه نیز قابلیت استفاده این روش در حل مسائل گوناگون است، در حالی که روش‌هایی مانند درخت تصمیم‌گیری را نمی‌توان به راحتی در مسائل مختلف به کار برد. به هر حال مشابه با هر مدل ریاضی و آماری، مدل‌های ماشین بردار پشتیبان نیز معایبی دارند. وجود تعداد زیاد متغیر ورودی، اصلی‌ترین مشکل در توسعه این مدل‌ها است. تعداد زیاد متغیرهای ورودی ممکن است مانع یافتن مدل بهینه توسط ماشین بردار پشتیبان شود. بر این اساس برای انجام تحلیل رگرسیون داده‌های ورودی از فضای اولیه به فضایی با ابعاد بالاتر انتقال داده می‌شوند که این کار با توابع کرنل انجام می‌گیرد. همچنین کرنل‌های مختلفی به منظور ارزیابی مدل در روش‌های مبتنی بر هسته در نظر

گرانادوس و همکاران (۱۹) در سال ۱۹۹۶ معرفی شد، یک سیستم یادگیری است که ایده اصلی آن بر مبنای محاسبه تابع رگرسیون خطی در فضای بالاتری از ویژگی بنا شده است (۲۰). از الگوریتم ماشین بردار پشتیبان، در هر جایی که نیاز به تشخیص الگو یا دسته‌بندی اشیاء در کلاس‌های خاص باشد می‌توان استفاده کرد. ماشین بردار پشتیبان از یک روش که trick kernel نامیده می‌شود، برای تبدیل داده‌ها استفاده و سپس بر اساس این تبدیل، مرز بهینه بین خروجی‌های ممکن را پیدا می‌کند. به عبارت ساده تبدیلات بسیار پیچیده را انجام می‌دهد، سپس مشخص می‌کند چگونه داده‌ها را بر اساس برجسب‌ها یا خروجی‌های تعریف شده، جدا کرد. یکی از روش‌هایی که در حال حاضر به صورت گسترده برای مسئله دسته‌بندی استفاده می‌شود، روش ماشین بردار پشتیبان است.

جدول ۲. کرنل‌های روش هسته مینا و پارامترهای انتخابی هریک از آنها

پارامتر انتخابی	فرمول	تابع
C, ϵ	$X^T Y$	خطی
C, d, r, Y, ϵ	$(YX^T y + r)^d$	پلی نومیال
C, Y, ϵ	$\text{Exp}(-\gamma \ X - Y\ ^2)$	گوسین (RBF)

رگرسیون درختی توسعه یافته

رگرسیون درختی توسعه یافته از مجموعه روش‌های یادگیری ماشین و ترکیبی از دو تکنیک آماری بوستینگ و رگرسیون درختی است (۲۴، ۱۴). بوستینگ روشی است که در آن مدل‌های درختی به صورت تکرارپذیر با زیرمجموعه‌ای از داده‌های آموزشی برازش داده می‌شوند. در برازش رگرسیون درختی توسعه یافته باید دو پارامتر، میزان یادگیری و پیچیدگی درختی مشخص شود. میزان یادگیری سهم هر درخت متوالی را در مدل نهایی تعیین می‌کند و پیچیدگی درخت، اثرهای اصلی یا اثرهای متقابل بین متغیرها را نشان می‌دهد (۲۵).

در ادامه در این پژوهش مقدار کربن خاک با چهار الگوریتم یادگیری ماشین شامل رگرسیون درخت تصمیم پیشرفته، افزایش گرادیان اکستریم، ماشین بردار پشتیبان و جنگل تصادفی با استفاده از زبان برنامه نویسی پایتون (گوگل کولب) و بسته Scikit-learn برنامه ریزی و برآورد شد. بسته Scikit-learn از بسته‌های متن باز و قدرتمند در زبان برنامه نویسی پایتون است که برای اهداف یادگیری ماشین به کار می‌رود. در مدل جنگل تصادفی از سه متغیر قابل تنظیم برای بهبود مدل شامل تعداد متغیرهای استفاده شده در ساخت هر درخت، تعداد درختان جنگل تصادفی و حداقل گره‌های انتهایی استفاده می‌شود (۱۲). متغیرهای قابل تنظیم برای بهبود مدل ماشین بردار پشتیبان، بستگی به نوع مجموعه‌ای توابع ریاضی (Kernel) استفاده شده دارد. این مجموعه توابع می‌تواند انواع متفاوتی داشته باشند که به عنوان مثال می‌توان به توابع غیرخطی، چندجمله‌ای، تابع پایه شعاعی (RBF به اختصار یا Radial Basis Function) اشاره

گرفته شده است که شامل کرنل خطی، کرنل چندجمله‌ای و کرنل گوسین (RBF) هستند. جدول ۲ کرنل‌های استفاده شده در روش هسته مینا و پارامترهای مربوط به هریک از این کرنل‌ها را نشان می‌دهد. که در آن C پارامتر جریمه مربوط به میزان خطای مشاهداتی، ϵ میزان اپسیلون و بیانگر مقدار انعطاف تابع، r مقدار ضریب تابع در کرنل پلی‌نومیال و γ مقدار گاما است که نقش مؤثر و تعیین کننده‌ای در تعیین دقت مدل دارد.

جنگل تصادفی

مدل جنگل تصادفی یک فن یادگیری محبوب است که در تحلیل سنجش از دور به کار می‌رود. این مدل پایدارتر از مدل‌های سنتی تک درخت است و می‌تواند عملکرد را با کاهش پدیده‌های بیش از حد دانشی و بیش از حد مناسب تقویت کند (۲۱).

رگرسیون افزایش گرادیان اکستریم

الگوریتم XGBoost که توسط چن و گوسترین (۶) پیشنهاد شد، نوعی روش اجرایی کارآمد برای افزایش صحت و عملکرد ماشین افزایش گرادیان و به خصوص افزایش دقت طبقه‌بندی درخت‌های رگرسیون است. به خاطر پردازش منظم و موازی، XGBoost در مقایسه با GBM عملکرد بهتری دارد. این الگوریتم برای آموزش بهتر، تمام فاکتورهای پیش‌بینی کننده را همزمان ترکیب می‌کند (۲۳).

جدول ۳. فرآپارامترهای مورد نیاز برای بهینه کردن مدل رگرسیون استفاده شده

پارامتر بهینه	فرا پارامترها	الگوریتم
RBF	kernel type	
1	c	ماشین بردار پشتیبان
'scale'	σ	
3	Max depth	
'log2'	max_features	جنگل تصادفی
100	n_estimators	
'log2'	max_features	
1000	n_estimators	درخت تصمیم پیشرفته
3	Max_depth	
42	random_state	
3	max_features	افزایش گرادیان اکسترمم
0.25	gamma	

یادگیری مدل) و داده‌های اعتبارسنجی (۳۰ درصد داده‌ها برای آزمون و صحت‌سنجی) تقسیم شدند. در انتها برای ارزیابی دقت مدل‌های تخمین میزان کربن آلی خاک، از آماره‌های ضریب تبیین (R^2)، جذرمیانگین مربعات خطا (RMSE)، میانگین خطای مطلق (MAE)، معیارهای اطلاعات آکائیک (Akaike (AIC)) و معیار اطلاعات بی‌زین (BIC) استفاده شد که معادلات این روش‌ها در روابط (۱ تا ۵) آورده شده است.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - p_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - p_i)^2} \quad (2)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - p_i| \quad (3)$$

در این روابط n تعداد نمونه‌ها، y_i مقدار اندازه‌گیری شده کربن آلی خاک، p_i مقدار پیش‌بینی شده کربن آلی خاک و \bar{y} میانگین مقادیر اندازه‌گیری شده است.

$$AIC = 2K - 2\log(L) \quad (4)$$

$$BIC = 2\log(n) k - 2\ln(L) \quad (5)$$

کرد. در این پژوهش از تابع RBF استفاده شد که دو متغیر مهم قابل تنظیم آن شامل جریمه و گاما است متغیرهای قابل تنظیم برای مدل درخت تصمیم پیشرفته و افزایش گرادیان اکسترمم در جدول ۳ آورده شده است. پس از برنامه‌ریزی مدل‌ها، بهترین متغیرهای مدل‌ها با استفاده از تابع Grid Search CV در بسته Scikit-learn تعیین شد.

اهمیت متغیرهای مستقل

اهمیت نسبی متغیرها با استفاده از شاخص درصد افزایش میانگین خطای پیش‌بینی در زبان برنامه‌نویسی پایتون با تابع Importance Feature در بسته Learn-Scikit محاسبه شد. در این روش یک متغیر پیش‌بینی‌کننده، زمانی با اهمیت شناخته می‌شود که به هم‌زدن و جابه‌جایی مقادیر آن منجر به افزایش خطای پیش‌بینی شود (۱).

ارزیابی اعتبار مدل

برای مدل‌سازی کربن آلی خاک، داده‌ها به دو گروه داده‌های واسنجی (۷۰ درصد داده‌ها برای آموزش و

تصادفی، ماشین بردار پشتیبان، درخت تصمیم و افزایش گرادیان اکستریم آنالیز شده نتایج حاصل از آنها به منظور بررسی کارایی روش در برآورد کربن آلی خاک در منطقه مورد مطالعه با ماهواره‌های لندست ۸ و سنتینل ۲ به صورت مجزا مورد بررسی قرار گرفت. در این پژوهش، قابلیت تخمین مقادیر مواد آلی خاک با استفاده از باندهای سنجنده OLI ماهواره لندست ۸ و سنتینل ۲ شاخص‌های به دست آمده از آنها، در منطقه زهکشی شده گنبد کاووس، بررسی شد. سه نوع کاربری موجود در منطقه شامل اراضی کشاورزی، مرتع جنگلی مورد بررسی قرار گرفت.

نمودار پراکندگی ماده آلی خاک با استفاده از لندست ۸

نمودار پراکندگی ماده آلی خاک در شکل ۲ نشان داده شده که با توجه به این نمودار بیشترین مقدار رگرسیون مربوط به مدل درخت تصمیم پیشرفته است، پس این مدل با دقت بیشتری توانست میزان ماده آلی را پیش‌بینی کند. میزان $RMSE=0/05$ ، $RME=0/17$ ، $R=0/55-0/64$ همچنین مدل ماشین بردار پشتیبان بعد از درخت تصمیم بیشترین تخمین ماده آلی خاک داشت. الگوریتم جنگل تصادفی (۵۷٪) و XGBoost (۵۳٪) نیز کمترین ضریب تبیین را برای تخمین ماده آلی خاک داشتند.

عملکرد مدل با لندست ۸ و سنتینل ۲

با وجود اینکه در سال‌های اخیر، مدل‌های یادگیری ماشین به طور گسترده برای تخمین کربن خاک در مقیاس منطقه‌ای و جهانی استفاده شده است با این حال، تاکنون پژوهشی در مورد عملکرد مدل‌های مختلف یادگیری ماشین در تخمین کربن خاک در منطقه گنبد ارزیابی نشده است. در این پژوهش، عملکرد ۴ مدل رگرسیون جنگل تصادفی، درخت تصمیم پیشرفته، ماشین بردار پشتیبان و افزایش گرادیان اکستریم مورد ارزیابی قرار گرفت.

لندست ۸: در جدول ۴، نتایج اعتبارسنجی توابع رگرسیونی پیشنهادی در برآورد ماده آلی خاک برای سری داده‌های آزمون، بر اساس آماره‌های مختلف ارائه شده است. همان‌گونه که در

جایی که K تعداد متغیرهای مستقل است L ارزش احتمال و n تعداد اندازه‌گیری‌های ثابت شده است.

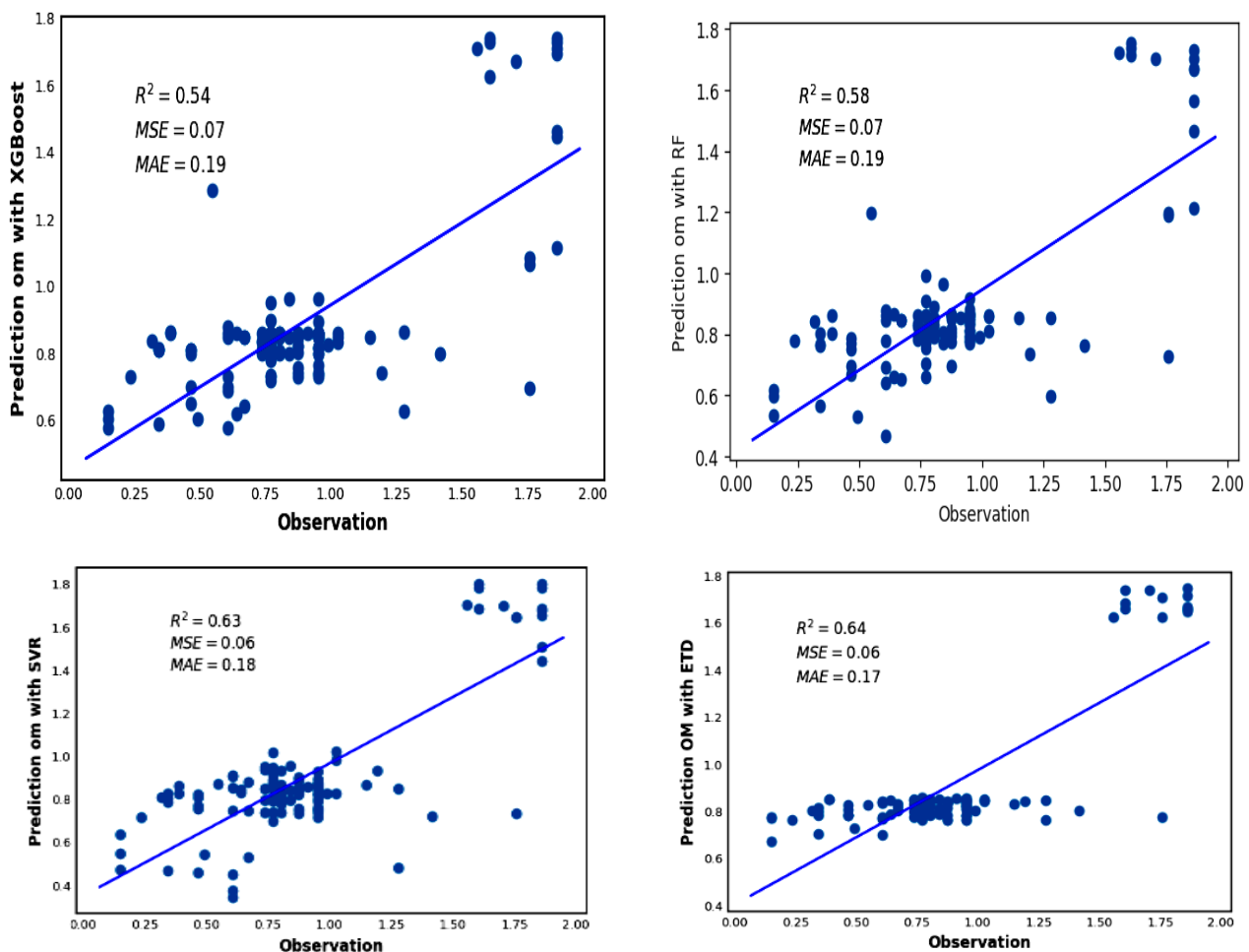
ارزیابی عدم قطعیت

عدم قطعیت پیش‌بینی کربن خاک در این مطالعه با استفاده اعتبارسنجی متقابل (Cross Validation) انجام شد. اعتبارسنجی متقابل یا متقاطع بر مبنای داده‌های مشاهداتی است که در هنگام ساختن مدل استفاده نمی‌شوند و برای بررسی و سنجش کارایی مدل برای پیش‌بینی داده‌های جدید کاربرد دارند. در مدل‌سازی یادگیری ماشینی، لازم است پارامترهای مدل برآورد شوند. تعداد پارامترها با پیچیدگی مدل ارتباط مستقیم دارد و با افزایش پارامترها یا متغیرهای مستقل مدل، کارایی مدل نیز بدلیل کاهش میانگین مربعات نمونه جمع‌آوری شده با افزایش متغیرها افزایش می‌یابد. بنابراین برای سهولت انجام فرایند مدل‌سازی باید تعداد پارامتر (متغیر) های مدل را بصورت بهینه تعیین کرد که در نتیجه مناسب‌ترین مدل برای نمونه، با بیشترین تعداد پارامتر حاصل خواهد شد.

به عبارت دیگر، هدف در اعتبارسنجی متقابل، دستیابی به مدلی است که تعداد پارامترهای آن بهینه باشد. یعنی پیدا کردن مدلی است که دچار بیش برازش نباشد. این در حالی است که این مدل، برای نمونه دیگری از همان جامعه استفاده نشود، چون در این صورت افزایش متغیرها، ممکن است به کارایی مدل آسیب برساند (بیش برازش مدل). برای حل این مسئله می‌توان از اعتبارسنجی متقابل یا متقاطع استفاده کرد که هدف آن تعیین تعداد پارامترها یا متغیرهای مناسب مدل است.

نتایج و بحث

داده‌های کربن آلی بدست آمده از نمونه برداری زمینی در سال ۲۰۲۰ به همراه داده‌های انعکاسی حاصل از باندهای مختلف ماهواره لندست ۸ و سنتینل ۲ نیز شاخص‌های مختلف گیاهی و رطوبتی محاسبه شده، در زبان برنامه‌نویسی و با استفاده از روش‌های رگرسیون یادگیری ماشین جنگل



شکل ۲. رابطه پیش‌بینی مدل و مقادیر مشاهده‌ای در چهار مدل رگرسیون جنگل تصادفی، درخت تصمیم پیشرفته، ماشین بردار پشتیبان و افزایش گرادیان اکستریم با ماهواره لندست ۸

الگوریتم جنگل تصادفی (۰/۰۷)، برای رگرسیون ماشین بردار پشتیبان، درخت تصمیم پیشرفته و افزایش گرادیان اکستریم به ترتیب (۰/۰۶، ۰/۰۶، ۰/۰۷) به دست آمد، که در مجموع نشان‌دهنده دقت متوسط مدل‌های پیشنهادی در برآورد کربن آلی خاک است.

سنتینل ۲: نتایج بررسی و آزمون روش‌های مختلف رگرسیون نشان‌دهنده عملکرد مناسب روش رگرسیون جنگل تصادفی است. مدل جنگل تصادفی بالاترین عملکرد (RMSE = 0.05, $R^2 = 0.64$) را برای برآورد کربن خاک به همراه داشت. پس از مدل جنگل تصادفی، مدل درخت

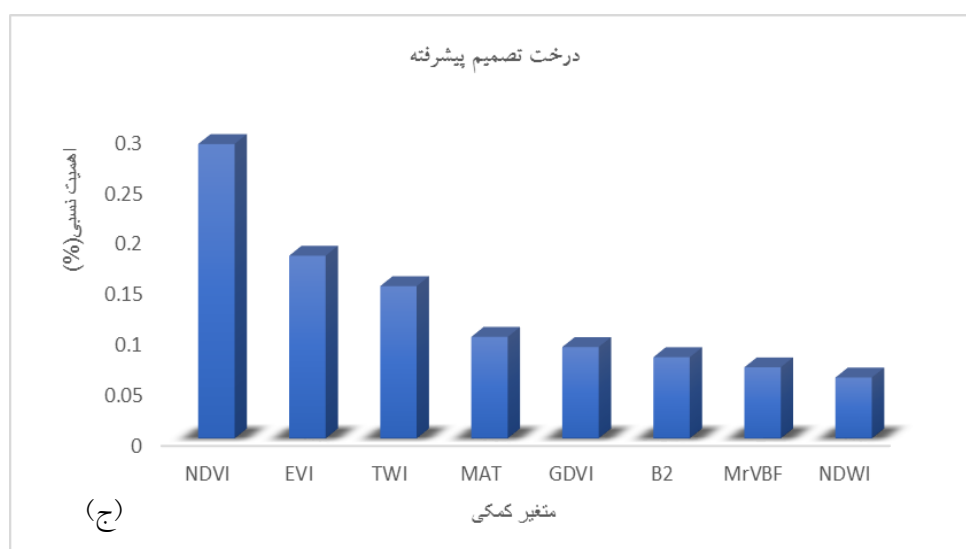
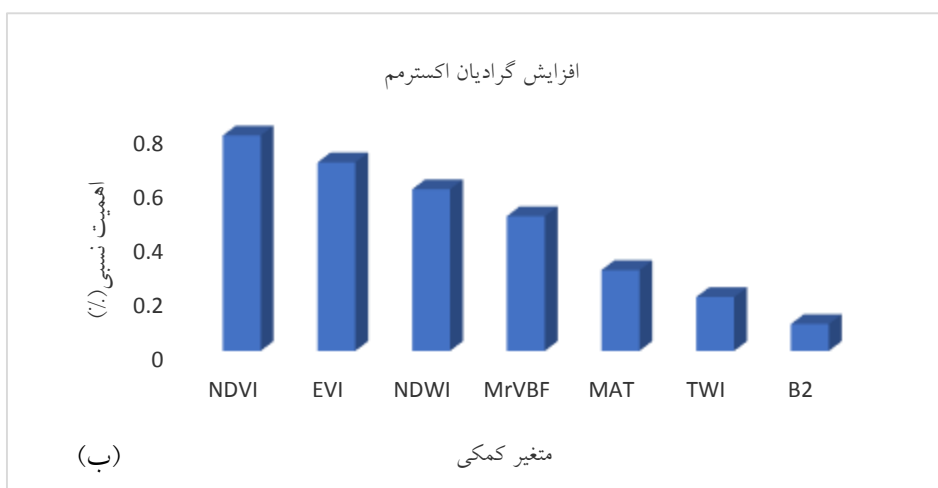
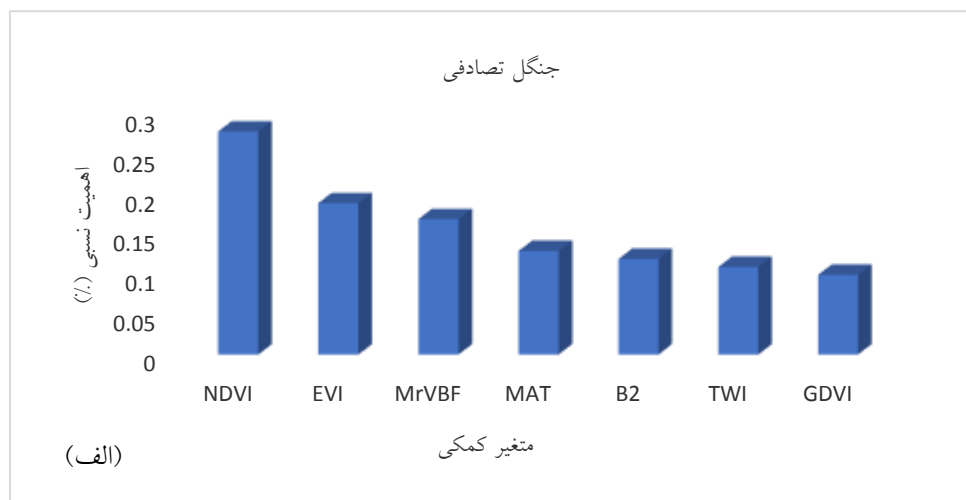
این جدول دیده می‌شود ضریب تبیین همه الگوریتم‌ها برای منطقه مورد مطالعه، بیشتر از ۵۰٪ است. این بدین معنی است که این توابع در برآورد کربن آلی خاک در مرحله ارزیابی، از دقت پیش‌بینی خوبی برخوردار هستند و این را می‌توان بدین ترتیب توضیح داد که کربن خاک وابستگی قوی به متغیرهای مستقل انتخاب شده دارد. هر ۴ مدل یادگیری ماشین دارای قابلیت‌های کافی برای یادگیری این وابستگی‌های اساسی برای تخمین کربن خاک هستند، اگرچه اصول یادگیری هر مدل تفاوت‌هایی دارند. با توجه به نتایج حاصل از آماره RMSE، مقدار این آماره برای

جدول ۴. معیارهای ارزیابی مدل‌های یادگیری ماشین برای پیش‌بینی کربن آلی خاک

معیار اطلاعاتی بیزی (BIC)	معیار اطلاعاتی اکائیک (AIC)	درصد میانگین قدر مطلق خطا (PMAE)	میانگین قدر مطلق خطا (MAE)	جذر میانگین مربعات خطا (RMSE)	ضریب تبیین (R ²)	رگرسیون
-۸۹۳	-۹۱۹	۵/۷۸	۰/۱۳	۰/۰۳	۰/۷۵	واسنجی
-۲۸۴	-۳۰۳	۶/۰۳	۰/۱۸	۰/۰۶	۰/۵۷	اعتبارسنجی
-۸۴۷	-۸۷۲	۹/۶۸	۰/۱۴	۰/۰۳	۰/۷۱	واسنجی
-۲۴۹	-۲۸۲	۷/۲۵	۰/۱۹	۰/۰۷	۰/۵۳	اعتبارسنجی
-۷۰۰	-۷۴۴	۸/۰۶	۰/۱۶	۰/۰۶	۰/۵۵	واسنجی
-۲۷۹	-۳۱۲	۸/۰۲	۰/۱۷	۰/۰۵	۰/۶۴	اعتبارسنجی
-۷۴۱	-۷۸۴	۵/۸۹	۰/۱۵	۰/۰۵	۰/۶۱	واسنجی
-۲۷۴	-۳۰۸	۶/۷۳	۰/۱۷	۰/۰۵	۰/۶۲	اعتبارسنجی
-۸۷۰	-۹۰۶	۵/۲۴	۰/۱۳	۰/۰۳	۰/۷۵	واسنجی
-۲۸۸	-۳۱۶	۶/۳۰	۰/۱۷	۰/۰۵	۰/۶۴	اعتبارسنجی
-۶۸۵	-۷۲۲	۷/۸۰	۰/۱۷	۰/۰۶	۰/۵۱	واسنجی
-۲۷۱	-۲۹۶	۹/۲۳	۰/۱۶	۰/۰۶	۰/۵۶	اعتبارسنجی
-۷۲۰	-۷۵۲	۶/۲۳	۰/۱۶	۰/۰۶	۰/۵۴	واسنجی
-۲۹۱	-۳۱۵	۷/۸۴	۰/۱۷	۰/۰۵	۰/۶۳	اعتبارسنجی
-۶۵۸	-۷۰۵	۶/۹۸	۰/۱۷	۰/۰۷	۰/۴۹	واسنجی
-۲۶۳	-۲۹۹	۷/۵۶	۰/۱۸	۰/۰۶	۰/۶۰	اعتبارسنجی

قابل چشم‌پوشی است. رگرسیون افزایش گرادیان اکستریم پایین‌ترین عملکرد را برای پیش‌بینی ماده آلی داشت. این الگوریتم‌ها بیشترین تخمین را با ماده آلی داشتند. در این زمینه، نتایج مطالعات (۲۸) نیز نتایج این پژوهش را تأیید می‌کند. درصد میانگین مطلق خطا برای هر ۴ الگوریتم کمتر از ده درصد شد و این نشان‌دهنده عملکرد مناسب مدل‌های به کار برده شده است.

تصمیم پیشرفته ($R^2 = 0.63, RMSE = 0.05$) عملکرد مناسب‌تری را برای کربن آلی خاک به همراه داشت (جدول ۴). همان‌طور که جدول نشان می‌دهد مقدار R^2 در الگوریتم جنگل تصادفی بیشتر از روش درخت تصمیم بود و برای این مدل برابر با ۰/۶۴ بدست آمد. بین الگوریتم جنگل تصادفی، درخت تصمیم و ماشین بردار پشتیبان اختلاف اندکی برای تخمین ماده آلی خاک وجود داشت که



شکل ۳. اهمیت نسبی متغیرها مستقل در برآورد کربن در سه مدل رگرسیون جنگل تصادفی (الف)، افزایش گرادیان اکسترمم (ب) و درخت تصمیم پیشرفته (ج) با ماهواره لندست ۸

اهمیت ویژگی متغیر وابسته با لندست ۸

اهمیت نسبی ویژگی‌های مستقل برای تخمین ماده آلی در شکل ۳ آورده شده است. نتایج نشان دهنده این است که مدل جنگل تصادفی همانطور که در شکل ۳ دیده می‌شود، شاخص نرمال شده اختلاف پوشش گیاهی مهم‌ترین ویژگی و به دنبال آن EVI (۰/۱۸)، $MrVBF$ (۰/۱۵)، MAT (۰/۱۱)، $B2$ (۰/۱۱) و TWI (۰/۱۰) و شاخص پوشش گیاهی تفاوت تعمیم یافته برای پیش‌بینی کربن خاک مهم‌ترین متغیر بودند.

در مدل افزایش گرادیان اکسترمم اهمیت نسبی متغیرهای مختلف برای پیش‌بینی کربن بر اساس انتخاب فرآپارمترهای بهینه در شکل (شکل ۳-ب) نشان داده شده است. به دنبال آن شاخص بهبود یافته پوشش گیاهی، شاخص اختلاف نرمال شده آب، شاخص توپوگرافی همواری دره، میانگین دمای سالانه، شاخص خیسی و باند ۲ از مهم‌ترین متغیرهای مستقل برای برآورد کربن آلی خاک بودند.

عامل مثبت کنترل کننده تغییرپذیری کربن خاک در منطقه مورد مطالعه، مطابق (شکل ۳-ج)، با درخت تصمیم، شاخص نرمال شده پوشش گیاهی و شاخص بهبود یافته گیاهی از مهم‌ترین متغیرهای کمکی برای پیش‌بینی کربن خاک بودند. اهمیت نسبی سایر شاخص‌ها به شرح زیر شاخص خیسی، میانگین دمای سالانه، شاخص پوشش گیاهی تعمیم یافته، باند، شاخص همواری دره و شاخص نرمال شده اختلاف آب است. مدل‌های درخت تصمیم، جنگل تصادفی و افزایش گرادیان اکسترمم، این مزیت را دارد که می‌تواند اهمیت نسبی متغیر پیش‌بینی کننده را درجه‌بندی کند.

سه مدل (جنگل تصادفی، افزایش گرادیان اکسترمم و درخت تصمیم) نشان داد که شاخص‌های پوشش گیاهی اهمیت بالایی در پیش‌بینی ماده آلی خاک دارند. این مطالعه مطابق با مطالعه فلاحت‌کار و همکاران (۱۳) که نشان داد شاخص اختلاف نرمال شده آب یکی از مهم‌ترین متغیرهای کمکی برای پیش‌بینی کربن خاک است، نشان داد که استخراج مؤلفه‌های

مربوط به باندها همراه با محاسبه شاخص‌هایی مانند تفاوت پوشش گیاهی نرمال شده، به‌عنوان متغیرهای کمکی، نقش مهمی در برآورد صحیح‌تر مقدار ماده آلی خاک دارند. محتوای کربن آلی خاک به شدت وابسته به پوشش گیاهی طبیعی و بقایای گیاهی باقی مانده پس از برداشت گیاه بود که با نتایج پژوهش (۲۷) مطابقت دارد.

علاوه بر این، $NDVI$ اهمیت بیشتری برای محتوای میزان ماده آلی خاک در مقایسه با سایر شاخص‌های گیاهی سنجش از دور مانند شاخص بهبود یافته EVI دارد. اگرچه در بسیاری از پژوهش‌ها EVI بهتر از $NDVI$ است، نتایج ما نشان داد که $NDVI$ برای توزیع ماده آلی خاک در مقایسه با EVI اهمیت بیشتری دارد. که ممکن است اهمیت بیشتر شاخص بهبود یافته گیاهی نسبت به $NDVI$ به شرایط توپوگرافی مرتبط باشد (۱۱). گزارش شده است که شاخص پوشش گیاهی بهبود یافته نسبت به $NDVI$ به توپوگرافی حساس‌تر است (۲۷). در همین حال، باند سبز Landsat-8 ($B2$) سهم بیشتری در میزان ماده آلی خاک نسبت به $B3$ و $B4$ نشان داد. سهم بالای $MrVBF$ در تغییرپذیری ماده آلی خاک در این مطالعه را می‌توان به رسوب ذرات و رسوبات ریز غنی شده با ماده آلی از ارتفاعات در دره‌های پایین با مناطق مسطح و کم ارتفاع نسبت داد. TWI تنها ۳ درصد محتوای ماده آلی خاک را به همراه داشت که نشان می‌دهد ماده آلی خاک تمایل دارد در مناطق مرطوب‌تر تجمع پیدا کند. تقی‌زاده مهرجردی و همکاران (۲۷) نشان داد که شاخص رطوبت مهم‌ترین متغیر زمین برای پیش‌بینی ماده آلی خاک است.

نمودار پراکنندگی ماده آلی خاک با استفاده از سنتینل ۲

مقایسه مقادیر برآورد شده کربن آلی خاک برای سال ۲۰۲۰ با ماهواره سنتینل ۲، با مقادیر نمونه‌برداری شده آن در این سال، به صورت نمودار ۳ نشان داده شده است. همان‌طور که دیده می‌شود مقدار R^2 در شکل ۳ بیشتر از مقدار آن در برآوردهای

آن، وضعیت پوشش گیاهی، مواد معدنی، خصوصیات فیزیکی و شیمیایی خاک و همچنین به مواد و روش‌های بکار گرفته شده نسبت داد که این مطلب تأکید می‌کند که با انجام پژوهش‌های متعدد و در مناطق مختلف بایستی به بررسی قابلیت داده‌های طیفی ماهواره‌ای در برآورد مشخصه‌های خاک پرداخت. با توجه به نتایج ارائه شده می‌توان گفت که متغیرهای کمکی حاصل از هر دو سنجنده به خوبی توانستند از طریق طول موج‌های منعکس شده در منطقه مطالعاتی ایجاد تفاوت کرده و در نتیجه، نقش مهمی در تخمین مقادیر مربوط به ویژگی‌های مقدار ماده آلی خاک ایفا کنند. افزون بر آن، رابطه بین باندهای انتخاب شده و متغیرهای کمکی استخراج شده نشان داد که باندهای ۲، ۳، ۴ نشان‌دهنده باندهای (RGB) و شاخص پوشش گیاهی به دست آمده توسط یک سنجنده تصویربرداری از زمین، بیشترین تأثیر را در پیش‌بینی محتوی ماده آلی در منطقه مورد مطالعه داشتند و سایر متغیرهای کمکی، NDSI، MNDWI نقش کمتری در پیش‌بینی مقدار تجمع ماده آلی خاک در منطقه مورد مطالعه داشتند.

عدم قطعیت پیش‌بینی کربن آلی خاک با ارزیابی متقابل

در جدول ۵ نتایج عدم قطعیت برای ۴ مدل آورده شده است که بر اساس نتایج بدست آمده، در هر دو ماهواره لندست ۸ و سنتینل ۲ جنگل تصادفی نسبت به رگرسیون ماشین بردار پشتیبان، افزایش گرادیان اکسترمم و درخت تصمیم پیشرفته کارایی بهتری در تعیین عدم قطعیت پیش‌بینی کربن آلی در منطقه مورد مطالعه دارد. عملکرد بهتر RF می‌تواند به دلیل توانایی آن در مدل سازی پایگاه داده‌های بزرگ و توانایی ادغام متغیرهای ورودی زیاد بدون تغییر متغیر باشد. پس از مدل جنگل تصادفی، درخت تصمیم پیشرفته عملکرد مناسب‌تری را برای کربن آلی خاک داشت و ماشین بردار پشتیبان نیز عملکرد مناسبی برای عدم قطعیت داشت. در حالی که افزایش گرادیان اکسترمم کمترین عملکرد را داشت.

ماهواره لندست ۸ است که نشان‌دهنده تناسب بیشتر کربن آلی برآورد شده با مقادیر واقعی آن در این سنجنده است. گرچه انتظار می‌رود که این پایین بودن مقدار برآورد نامناسبی از مقادیر کربن آلی خاک را به دنبال داشته باشد، اما دیده می‌شود که کربن آلی برآورد شده با ماهواره سنتینل ۲ با مقادیر واقعی آن تفاوت‌های بسیار کمی دارد. در بررسی وجود تفاوت‌های کوچک و بزرگ در برآورد پارامترهای مربوط به خاک باید بیان کرد که عوامل متعددی وجود دارد که می‌تواند انعکاس‌های خاک را در اکوسیستم‌های مختلف به‌ویژه در اکوسیستم پیچیده کشاورزی تحت تأثیر قرار دهد و باعث شود سنجنده‌ها در برآورد این پارامترهای خاکی دچار مشکل شوند. این مشکل می‌تواند تحت تأثیر مسائل زیادی مثلاً دقت در زمان و مکان نمونه‌برداری، انتخاب سنجنده مناسب و وجود یا عدم وجود تصاویر با کیفیت مطلوب در زمان مد نظر تغییر کند. با توجه به شکل مقادیر واقعی و پیش‌بینی شده نزدیک به خط یک به یک پراکنده شده‌اند. در نتیجه، می‌توان استدلال کرد که همبستگی مثبت و کاملی بین داده‌ها برقرار است و مدل از دقت زیادی برخوردار است. نتایج مشابه توسط ویکاکسل و همکاران (۳۰) ارائه شده است. برطبق یافته‌های این محققان، مدل جنگل تصادفی می‌تواند نتایج خوبی برای پیش‌بینی مواد مغذی خاک ارائه دهد.

اهمیت ویژگی متغیر وابسته با سنتینل ۲

سهم هر کدام از متغیرهای مستقل در برآورد متغیر وابسته نیز قابل مشاهده است و آن گونه که مشخص است شاخص پوشش گیاهی بیشترین تأثیر مثبت را در برآورد کربن آلی داراست.

این تفاوت می‌تواند به دلیل اختلاف چند روزه زمان عکس‌برداری ماهواره لندست ۸ و سنتینل ۲ و نیز حساسیت‌های متفاوت سنجنده‌های این دو ماهواره نسبت به پدیده‌ها باشد. دلیل وجود نتایج ضد و نقیض در پژوهش‌ها را می‌توان به عوامل زیادی از جمله شرایط منطقه مورد مطالعه از نظر میزان مواد آلی و توزیع مکانی

جدول ۵. عدم قطعیت رگرسیون‌های استفاده شده با روش ارزیابی مقاطع

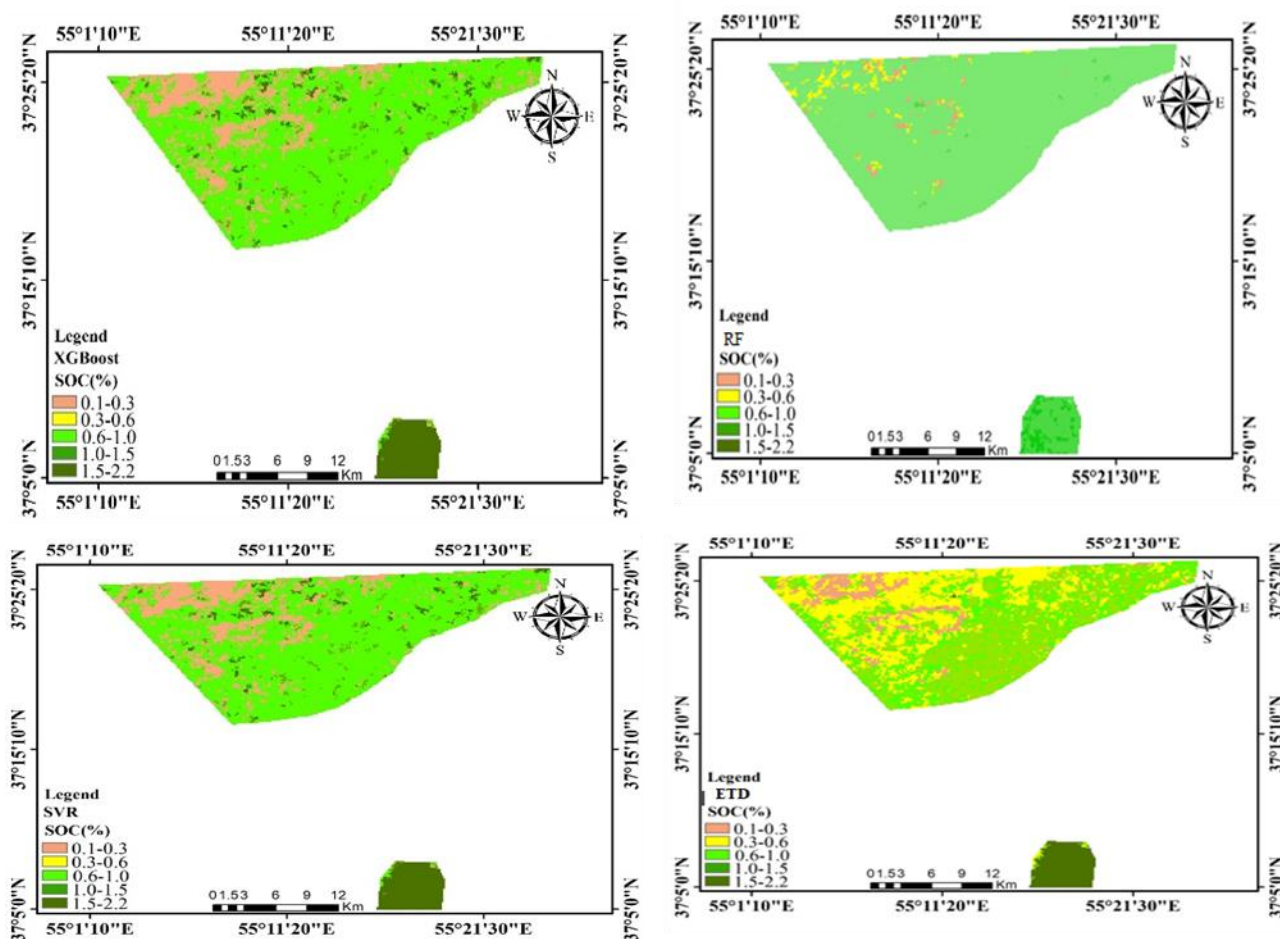
اعتبار سنجی				واسنجی				نوع نقطه نوع نقطه
MEA ± میانگین اشتباه معیار	PMAE ± میانگین اشتباه معیار	RMSE ± میانگین اشتباه معیار	R ² ± میانگین اشتباه معیار	MEA ± میانگین اشتباه معیار	PMAE ± میانگین اشتباه معیار	RMSE ± میانگین اشتباه معیار	R ² ± میانگین اشتباه معیار	
۰/۰۳ ± ۰/۰۶	۵/۸۴ ± ۰/۷	۰/۳۳ ± ۰/۱۵	۰/۶۸ ± ۰/۲۱	۰/۰۳ ± ۰/۰۶	۶/۰۶ ± ۰/۰۵	۰/۱۷ ± ۰/۰۱	۰/۷۲ ± ۰/۰۵	جنگل تصادفی
۰/۰۵ ± ۰/۰۴	۶/۰۴ ± ۰/۳	۰/۳۳ ± ۰/۱۵	۰/۰۶ ± ۰/۰۸	۰/۰۵ ± ۰/۰۲	۹/۷۰ ± ۰/۰۵	۰/۱۷ ± ۰/۰۱	۰/۶۹ ± ۰/۲۰	ماشین بردار پشتیبان
۰/۰۴ ± ۰/۰۶	۹/۷۲ ± ۱/۶	۰/۳۶ ± ۰/۱۹	۰/۵۷ ± ۰/۰۷	۰/۰۲ ± ۰/۰۸	۹/۴۷ ± ۸/۲۱	۰/۲۴ ± ۰/۱۷	۰/۶۶ ± ۰/۰۵	افزایش گرادیان اکسترمم
۰/۰۶ ± ۰/۰	۹/۰۸ ± ۰/۷	۰/۳۸ ± ۰/۱۷	۰/۶۲ ± ۰/۲۴	۰/۰۴ ± ۰/۰۱	۹/۷۸ ± ۱/۸	۰/۱۷ ± ۰/۰۱	۰/۶۲ ± ۰/۱۸	درخت تصمیم پیشرفته

سننیل ۲								
۰/۱۴ ± ۰/۱۴	۶/۷۸ ± ۱/۷۱	۰/۳۵ ± ۰/۱۸	۰/۶۶ ± ۰/۱۹	۰/۰۴ ± ۰/۰۱	۹/۹۲ ± ۵/۹۸	۰/۲۱ ± ۰/۰۲	۰/۷۰ ± ۰/۲۴	جنگل تصادفی
۰/۱۳ ± ۰/۱۲	۱۰/۲ ± ۱/۷۴	۰/۳۵ ± ۰/۱۸	۰/۵۷ ± ۰/۱۶	۰/۰۵ ± ۰/۰۱	۹/۱۴ ± ۷/۹۲	۰/۲۱ ± ۰/۰۲	۰/۶۸ ± ۰/۱۴	ماشین بردار پشتیبان
۰/۱۵ ± ۰/۱۶	۹/۷۳ ± ۱/۸۶	۰/۳۶ ± ۰/۱۹	۰/۶۰ ± ۰/۲۷	۰/۰۶ ± ۰/۰۲	۹/۴۶ ± ۸/۲۰	۰/۲۴ ± ۰/۰۱	۰/۶۸ ± ۰/۳۳	افزایش گرادیان اکسترمم
۰/۱۴ ± ۰/۱۳	۹/۸۲ ± ۱/۷	۰/۳۴ ± ۰/۱۷	۰/۵۶ ± ۰/۲۵	۰/۰۵ ± ۰/۰۱	۷/۱۵ ± ۶/۲۷	۰/۲۲ ± ۰/۰۲	۰/۶۰ ± ۰/۰۲	درخت تصمیم پیشرفته

تهیه نقشه ماده آلی خاک با استفاده از الگوریتم‌های یادگیری ماشین

از آنجایی که گیاهان سبز و سالم باعث افزایش جذب در طیف مرئی می‌شوند و به این دلیل که باند ۳ یا باند قرمز اصولاً تحت تأثیر بازتاب از گیاهان سبز قرار می‌گیرد (در محدوده جذب کلروفیل قرار دارد)، با کاهش بازتاب و یا افزایش جذب در محدوده باند ۳ (محدوده مرئی) برآورد

کربن آلی خاک افزایش یافته و ارتباط باند ۳ با مقدار کربن آلی برآورد شده عکس یکدیگر خواهد بود. همان‌طور که در جدول ۵ دیده می‌شود در اراضی زهکشی شده مقادیر کربن آلی برآورد شده با سنجنده و نمونه‌برداری شده در منطقه با یکدیگر متناسب بوده و یا در بازه‌های نزدیک به یکدیگر قرار گرفته‌اند. این موضوع از طرفی نشان‌دهنده دقت بالای ماهواره لندست ۸ در برآورد این پارامتر خاکی



شکل ۶. نقشه ماده آلی خاک با چهار رگرسیون جنگل تصادفی، ماشین بردار پشتیبان، افزایش گرادیان اکسترم و درخت تصمیم پیشرفته

به‌دست آمده از پژوهش آنها، روش جنگل تصادفی شاخص‌های خطای بسیار کم و ضریب تبیین بیشتری (۹۷٪) نسبت به مدل رگرسیون خطی چندگانه داشته است ($RMSE = ۰/۶۶۲$ ، سایر پژوهشگران مانند (۳۲، ۲۷، ۲۲، ۵) و نیز پژوهش (۳۴) که مدل جنگل تصادفی را در استفاده از سنجش از دور و تصاویر ماهواره‌ای می‌تواند بر محدودیت‌های ناشی از روش‌های سنتی غلبه کرده و به‌عنوان یک جایگزین مناسب پایش کیفیت خاک با امکان نمایش نتایج در مقیاس‌های زمانی و مکانی مختلف، به خصوص برای مناطق پیش‌بینی کلاس‌ها و ویژگی‌های خاک با درصد خطای کم مطلوب دانسته‌اند، مطابقت داشت.

است. از طرف دیگر می‌توان گفت که تصویر ماهواره‌ای مربوط به سال ۲۰۲۰ که از مجموع تصاویر این ماهواره بارگیری شده در نزدیک‌ترین زمان به زمان نمونه‌برداری زمینی برداشت شده، یا اینکه شرایط آب و هوایی در هنگام برداشت این تصویر به شرایط آب و هوایی زمان نمونه‌برداری بسیار نزدیک بوده و بدین ترتیب این سنجنده توانسته است تخمین تقریباً مناسبی را با در نظر گرفتن پدیده‌های موجود مانند رویش پوشش گیاهی تازه و غیره ارائه دهد (شکل ۶). در تعدادی از مطالعات به دو استفاده از متغیرهای محیطی و سنجش از دور جهت پیش‌بینی ماده آلی خاک اشاره شده است (۳۲). مدل‌های جنگل تصادفی و رگرسیون خطی چندگانه را برای پیش‌بینی توزیع ماده آلی خاک به کار برده‌اند. بر اساس نتایج

نتیجه گیری

ماده آلی در خاک سطحی ضریب همبستگی بین شاخص‌ها، باندهای مورد مطالعه و ماده آلی خاک افزایش می‌یابد. این امر، در پژوهش‌های دیگری که در مناطق خشک و نیمه خشک صورت گرفته است نیز دیده می‌شود. این علت به‌گونه‌ای است که فرایند افزایش بازتابش سطحی خاک بر اثر کاهش تراکم پوشش گیاهی، کاهش مقدار ماده آلی خاک، کاهش رطوبت خاک و همچنین ایجاد و گسترش لکه‌ها و پوسته‌های نمکی در سطح خاک تأثیرگذار است.

وسیع‌تر استفاده شود. دوره بازگشت مناسب، قدرت تفکیک مکانی بالا، نظارت بدون وقفه و گردآوری اطلاعات در مقیاس وسیع از مزایایی است که استفاده از تصاویر ماهواره‌ای دارد که در این مطالعه به آن اشاره شد. با توجه به افزایش روزافزون قابلیت‌های مختلف رادیومتری، طیفی، زمانی و مکانی تصاویر ماهواره‌ای و توسعه روش‌های پردازش تصاویر می‌توان سنجش از دور را به‌عنوان یک روش جدید و مؤثر از نظارت بر کیفیت خاک یاد کرد. نتایج این پژوهش نشان داد که با افزایش مقدار

منابع مورد استفاده

1. Adjuik, T. A. and S. C. Davis. 2022. Machine Learning Approach to Simulate Soil CO₂ Fluxes under Cropping Systems. *Agronomy* 12: 197.
2. Alhammadi, M. S. and E. P. Glenn. 2008. Detecting date palm trees health and vegetation greenness change on the eastern coast of the United Arab Emirates using SAVI. *International Journal of Remote Sensing* 29: 1745-1765.
3. Basak, D., S. Pal and D. C. Patranabis. 2007. Support Vector Regression Neural Information Processing—Letters and Reviews.
4. Böhner J, and T. Selige. 2006. Spatial prediction of soil attributes using terrain analysis and climate regionalisation. *Gottinger Geographische Abhandlungen* 115:13-28.
5. Camera, C., Z. Zomeni, J. S. Noller, A. M. Zissimos, I. C. Christoforou and A. Bruggeman. 2017. A high resolution map of soil types and physical properties for Cyprus: A digital soil mapping optimization. *Geoderma* 285: 35-49.
6. Chen, T. and C. Guestrin. Xgboost: A scalable tree boosting system. *In: Proceedings of 2016 the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, Washington.*
7. Chen, X., D. Yang, J. Chen and X Cao. 2015. An improved automated land cover updating approach by integrating with downscaled NDVI time series data. *Remote Sensing Letters* 6: 29-38.
8. Cheng, Y., P. Li, G. Xu, Z. Li, Yu, S. Cheng, B. Zhao and F. Wang. 2018. Factors that influence soil total phosphorus sources on dam fields that are part of ecological construction programs on the Loess Plateau, China. *Catena* 171: 107-114.
9. Danesh, M., H. Bahrami, S. K. Alavipanah and A. A. Nourouzi. 2010. A synchronous investigation of soil geometric mean particle diameter and lime, using remote sensing technology (Case Study: Pol-e-Dokhtar, the Southwest of Lorestan Province, Iran). *Agriculture Science Technology* 12: 479-494
10. Daniel, K.W., N. K. Tripathi and K. Honda. 2002. An approach for estimating soil organic matter content using synthetic IRS satellite data in tropical soils of Lopburi, Thailand. *Journal of Soil Research* 40: 240-246.
11. Emadi, M., R. Taghizadeh-Mehrjardi, A. Cherati, M. Danesh, A. Mosavi and T. Scholten. 2020. Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran. *Remote Sensing* 12: 2234.
12. Estévez, V., A. Beucher, S. Mattbäck, A. Boman, J. Auri, K.M. Björk and P. Österholm. 2022. Machine learning techniques for acid sulfate soil mapping in southeastern Finland. *Geoderma* 406: 115446.
13. Falahatkar, S., S.M. Hosseini, S. Ayoubi and A. Salmanmahiny. 2016. Predicting soil organic carbon density using auxiliary environmental variables in northern Iran. *Archives of Agronomy and Soil Science* 62: 375-393.
14. Fan, J., X. Wang, L. Wu, H. Zhou, F. Zhang, X. Yu, ... and Y. Xiang. 2018. Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. *Energy Conversion and Management* 164: 102-111.
15. Fick, S. E. and R. J. Hijmans. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* 37: 4302-4315.
16. Gallant, J. C. and J. M. Austin. 2015. Derivation of terrain covariates for digital soil mapping in Australia. *Soil Research* 53: 895-906.

17. Gao, B. C. 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment* 58: 257-266.
18. Khan, M.T.H., A. Ather, K.D. Thompson and R. Gambari. 2005. Extracts and molecules from medicinal plants against herpes simplex viruses. *Antiviral research* 67: 107-119.
19. López-Granados, F., M. Jurado-Expósito, J. M. Peña-Barragán and L. García-Torres. 2005. Using geostatistical and remote sensing approaches for mapping soil properties. *European Journal of Agronomy* 23: 279-289.
20. Mahmoudzadeh, H., H. R. Matinfar and R. Taghizadeh-Mehrjardi. 2021. Soil organic carbon digital mapping (case study: kamyaran county, kurdistan province). *Journal of Soil Management and Sustainable Production* 10: 77-98.
21. Mehrabi Gohari, E., H. R. Matinfar, A. Jafari, R. Taghizadeh-Mehrjardi and F. Khayamim. 2020. Comparing different statistical models and pre-processing techniques for estimation several chemical properties of the soil using vnr/swir spectrum. *Iranian Journal of Remote Sensing & GIS* 11: 47-60.
22. Pahlavan-Rad, M.R., F. Khormali, N. Toomanian, C. W. Brungard, F. Kiani, C. B. Komaki and P. Bogaert. 2016. Legacy soil maps as a covariate in digital soil mapping: a case study from Northern Iran. *Geoderma* 279: 141-148.
23. Pölsterl, S., 2020. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *Journal of Machine Learning Research* 21: 1-6.
24. Post, W.M., R.C. Izaurralde, L. K. Mann and N. Bliss. 2001. Monitoring and verifying changes of organic carbon in soil. *Climate Change* 51: 73-99.
25. Raeesi, M., A. Zolfaghari M. Yazdani and M. Sabetizade. 2020. Investigating the Ability of Landsat 8 and Sentinel 2A Satellite Images for Estimating Soil Organic Matter and Available Phosphorus in Semnan Plain. *Iranian Journal of Soil Research* 34: 111-125.
26. Rouse Jr, J.W., R. H. Haas, D. W. Deering, J. A. Schell and J. C. Harlan. 1974. Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation remote sensing center Texas A&M University College Station, Texas 77843.
27. Taghizadeh-Mehrjardi, R., K. Nabiollahi and R. Kerry. 2016. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. *Geoderma* 266: 98-110.
28. Vapnik, V., S. Golowich and A. Smola. 1996. Support vector method for function approximation, regression estimation and signal processing. *Advances in Neural Information Processing Systems* 9.
29. Walkley, A. and I. A. Black. 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: 29-38.
30. Wilcox, C. H., B.E. Frazier and S.T. Ball. 1994. Relationship between soil organic carbon and Landsat TM data in Eastern Washington. *Photogrammetric Engineering and Remote Sensing* 60: 777-781.
31. Wu, W., A. S. Mhaimed, W. M. Al-Shafie, F. Ziadat, B. Dhehibi, V. Nangia and E. De Pauw. 2014. Mapping soil salinity changes using remote sensing in Central Iraq. *Geoderma Regional* 2: 21-31.
32. Wu, Z., C. Shen and A. Van Den Hengel. 2019. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition* 90: 119-133.
33. Xu, H., 2006. Modification of normalized difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing* 27: 3025-3033.
34. Zhang, Y., B. Sui, H. Shen and L. Ouyang. 2019. Mapping stocks of soil total nitrogen using remote sensing data: A comparison of random forest models with different predictors. *Computers and Electronics in Agriculture* 160: 23-30.

Predicting and Mapping Soil Organic Carbon Using Remote Sensing and Machine Learning Algorithms

S. Bandak^{1*}, S. A. R. Movahedi Naeini¹, C. Bairam Komaki², M. Kakooei³ and J. Verlst⁴

(Received: November 20-2022 ; Accepted: March 15-2023)

Abstract

Soil organic carbon (SOC) is one of the most important components of soil physical and chemical properties that have an important role in sustainable production in agriculture and preventing soil degradation and erosion. Data mining approaches and spatial modeling besides machine learning techniques to investigate the amount of soil organic carbon using remote sensing data have been widely considered. The objective of the present study was the evaluation of SOC using the remote sensing technique compared with field methods in some areas of the Gonbad Kavous and Neli forests of Azadshar. The soil samples were collected from the soil surface (0-10 cm depth) to estimate the SOC. Data were categorized into two categories: 70% for training and 30% for validation. Three machine learning algorithms including Random forest (RF), support vector machine, extra tree decision, and XGBoost were used to prepare the organic soil carbon map. In the present study, auxiliary variables for predicting SOC included bands related to Lands 8 OLI and sentinel 2 measurement images, topography, and climate. The results showed that the extraction of the components related to the bands along with the calculation of indicators such as normalized vegetation difference, wetness index, and the MrVBF index as auxiliary variables play an important role in more correct estimation of the amount of soil organic matter. Comparison of different estimation regressions showed that the Sentinel 2 random forest model and in Landsat8 with the values of coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MEA) of 0.64, 0.05, and 0.17, respectively, was the best performance ratio compared to other approaches used in the study to estimate the organic carbon content of surface soil in the study area. In general, the results of this study indicated the ability of remote sensing techniques and learning models in the spatial estimation of soil organic carbon. So, this method can be used as an alternative to laboratory methods in determining soil organic carbon.

Keywords: Soil organic carbon, Machine learning, Regression, Remote sensing, Validation

1. Department of Soil Science, College of Agriculture, Gorgan University of Agricultural Sciences and Natural Resources, Golestan, Iran.

2. Department of Arid Zone Management, Faculty of Rangeland and Watershed Management, Gorgan University of Agricultural Sciences and Natural Resources, Golestan, Iran.

3. Department of Computer Science, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden.

4. Image Processing Laboratory (IPL)—Laboratory for Earth Observation (LEO), University of Valencia, 46003 Valencia, Spain

*: Corresponding author, Email: soraya.bandak@gmail.com