سنجش از دور و سامانه اطلاعات جغرافیایی در منابع طبیعی (سال سیزدهم / شماره سوم) پائیز ۱۴۰۱
نمایه شده در سایت: پایگاه استنادی علوم جهان اسلام، جهاد دانشگاهی، مگ ایران، نورمگز، سیویلیکا، گوگل اسکولار
آدرس وب سایت : http://girs.iaubushehr.ac.ir

مقاله پژوهشی

ارائه روش تلفیقی کاهش نویز- داده‌کاوی برای تخمین ماده آلی خاک با طیف‌سنجی VNIR

الهه اکبری، سهام میرزایی، آرا تومانیان، علی درویشی بلورانی، حسینعلی بهرامی

دریافت: ۱ شهریور ۱۴۰۰/ بازنگری: ۱۲ شهریور ۱۴۰۰/ پذیرش:۱۳ شهریور ۱۴۰۰
دسترسی اینترنتی: ۱۳ شهریور ۱۴۰۰/ دسترسی چاپی: ۱ مهر ۱۴۰۱

چکیده

پیشینه و هدف خاک به‌عنوان منبع طبیعی ناهمگن و بزرگ‌ترین مخزن کربن آلی در اکوسیستم زمینی، از فرآیندها و مکانیسم‌های پیچیده‌ای تشکیل‌شده است. ضرورت برآورد اطلاعات دقیق خاک در مقیاس ملی و منطقه‌ای به‌منظور بهبود مدیریت خاک و درک خصوصیات خاک و چگونگی تأثیرگذاری آن در کشاورزی، منجر به علاقه‌مند شدن محققین به این حوزه شده است. محتوای (SOM) به‌عنوان شاخص کیفیت خاک در حاصلخیزی آن و تولید مواد غذایی تأثیرگذار است و نیز به‌عنوان یک متغیر کلیدی در مباحث محیطی و کشاورزی محسوب می‌شود. جمع‌آوری تعداد زیادی داده خاک دقیق باهدف مدیریت منابع غذایی برای جمعیت آینده ضروری است. بنابراین استفاده از روش‌های تخمین سریع و ارزان و البته افزایش دقت برآورد محتوای SOM در ارزیابی و مدیریت منابع خاک می‌تواند کمک‌کننده باشد. در کشاورزی دقیق، مقیاس اطلاعات خاک موردنیاز برای مدیریت اراضی و محصول بسیار کوچک‌تر بوده و به‌طورمعمول مقیاس جمع‌آوری داده‌های میدانی جوابگوی این نیاز نیست. نمونه‌برداری و آنالیز تعداد زیاد نمونه خاک و تهیه نقشه توزیع SOM، برای مناطق وسیع و بزرگ، بسیار دشوار است. علاوه بر این، روش‌های سنتی آزمایشگاهی تجزیه‌وتحلیل خاک برای نمونه‌برداری زیاد نیاز به نیروی کار بیشتر بوده و علاوه براین زمان‌بر و هزینه‌بر است و نیاز به اپراتور آزمایشگاه متخصص دارد. هدف از تحقیق حاضر، مقایسه عملکرد دو روش PLSR و روش یادگیری ماشین درخت رگرسیون ارتقا یافته (BRT) برای پیش‌بینی مواد آلی خاک با استفاده از طیف VNIR، است.

الهه اکبری(✉) [۱]، سهام میرزایی [۲]، آرا تومانیان [۳]، علی درویشی بلورانی [۳]، حسینعلی بهرامی [۴]

۱. استادیار گروه سنجش‌ازدور و **GIS**، دانشکده جغرافیا و علوم محیطی، دانشگاه حکیم سبزواری، سبزوار، ایران
۲. دانشجوی دکتری سنجش‌ازدور و سیستم اطلاعات جغرافیایی، دانشکده جغرافیا، دانشگاه تهران، ایران
۳. دانشیار گروه سنجش‌ازدور و سیستم اطلاعات جغرافیایی، دانشکده جغرافیا، دانشگاه تهران، تهران، ایران
۴. استاد، گروه خاکشناسی، دانشکده کشاورزی، دانشگاه تربیت مدرس، تهران، ایران

پست الکترونیکی مسئول مکاتبات : e.akbari@hsu.ac.ir

https://doi.org/10.30495/GIRS.2022.1938557.1935
http://dorl.net/dor/20.1001.1.26767082.1401.13.3.1.3

با استفاده از ترکیب تبدیل موجک و تشخیص باندهای مستقل، نویزهای موجود در داده‌های طیف‌سنجی خاک کاهش‌یافته است. علاوه بر این، طیف‌ها یا باندهای مستقل و مؤثر در طیف‌سنجی مواد آلی خاک انتخاب گردیدند. براین اساس، در این تحقیق، روش‌های Wavelet-PCA-PLSR و Wavelet-PCA- BRT توسعه داده شده است و کارایی هر یک از آن‌ها ارزیابی می‌گردد.

مواد و روش‌ها ۴۲ نمونه خاک از منطقه ناهمگن کشاورزی شهری در تهران در ۳۰-۰ سانتی‌متر خاک جمع‌آوری گردید. ماده آلی خاک با استفاده از روش والکی بلک و بازتاب طیفی خاک با استفاده از طیف‌سنج FieldSpec3 اندازه‌گیری شد. مشتق اول و دوم بازتاب، جذب طیفی و مشتق اول و دوم آن محاسبه گردید. به‌منظور کاهش نویز و هموارسازی طیف، از روش تبدیل موجک تابع ماتریس Sym8 استفاده‌شده است. همچنین، تبدیل موجک به‌منظور نشان دادن و بارزسازی ویژگی‌ها در طیف انجام می‌شود. از تجزیه‌وتحلیل مؤلفه‌های اصلی و آزمون هادلینگز با فاصله اطمینان ٪۹۵ به‌منظور تشخیص داده‌های پرت استفاده شد. پس از حذف داده پرت از هر مجموعه، روش PLSR و درخت رگرسیون ارتقا یافته بر روی بازتاب، جذب و مشتق اول و دوم آن‌ها در ۵ سطح از تبدیل موجک اجراشده است. سپس، با مقایسه نتایج، مدل مناسب از طریق اعتبارسنجی انتخاب شد. در هنگام استفاده از نمونه عددی، به‌جای درخت تصمیم‌گیری از درخت رگرسیون استفاده می‌شود، اما روند آن‌ها یکسان است. در درخت رگرسیون از جستجو حریصانه استفاده می‌شود. بنابراین، با پاسخ دادن به سؤال باینری که حداکثر اطلاعات در مورد متغیر پاسخ از طریق کدام نود به دست می‌آید، گره ریشه و دو فرزند آن تعیین می‌گردد. این فرایند در هر گره فرزند تکرار می‌شود. تولید ساختمان درخت به‌صورت بازگشتی تکرار شده است و یک معیار توقف معمولی در نظر گرفته می‌شود. معیار توقف می‌تواند نظیر رسیدن به انشعابی که قابل‌تقسیم نیست و اطلاعات کمتری می‌دهد و یا زمانی که اطلاعات در گره حاوی کمتر از، پنج درصد از کل داده‌ها است، باشد. همچنین، سعی در به حداقل رساندن اندازه درخت است. برای تقسیم گره، عامل جینی، عامل آنتروپی و غیره به‌منظور به حداقل رساندن این عوامل استفاده‌شده است. علاوه بر این، در هر شاخه، مجموع مربع خطاها محاسبه‌شده و آن‌هایی که مقادیر حداقل دارند، انتخاب می‌شود.

روش درخت رگرسیون ارتقا یافته، دو روش درخت رگرسیون و تکنیک ارتقا را به‌منظور بهبود توان پیش‌بینی هرکدام از آن‌ها ترکیب می‌کند. به‌منظور کالیبراسیون و اعتبارسنجی مدل، به‌طور تصادفی به ترتیب ۳۰ و ۱۲ نمونه خاک انتخاب و برای بیان صحت مدل‌ها از آماره‌های R^2 و RMSE استفاده‌شده است. برای انتخاب بهترین فاکتور تولید مدل PLSR برای هر طیف، واریانس و باقی‌مانده مقادیر برآوردی و RMSE اعتبارسنجی استفاده شد. درنهایت، برای ایجاد سطح پیوسته و آگاهی از نحوه تغییر مواد آلی خاک در منطقه، نقشه مواد آلی خاک با استفاده از تصویر ماهواره‌ای لندست OLI و روش با دقت بیشتر تولید شد.

نتایج و بحث برآورد رضایت‌بخش میزان SOM، ایجاد سطوح پیوسته با دقت بیشتر براساس کاهش نویز و حفظ داده‌های مفید، همواره موردتوجه محققین بوده است. در این تحقیق نیز با استفاده از داده‌های طیف‌سنجی خاک و اندازه‌گیری آزمایشگاهی میزان مواد آلی، سعی در برآورد چنین سطح پیوسته‌ای به‌منظور تخمین SOM بوده است. با استفاده از تبدیل موجک و حذف داده‌های پرت براساس هادلینگز در روش PCA، داده‌های مفید برای تولید سطح پیوسته استخراج شدند. در این روش، باندها یا طیف‌های مستقل و مؤثر در مدل باقی می‌مانند. درحالی‌که، لین و همکاران به‌منظور انتخاب باندهای مناسب در تخمین مواد آلی خاک از روش تبدیل موجک و همبستگی استفاده نموده‌اند. با استفاده از روش همبستگی در مناطق ناهمگن همانند منطقه موردمطالعه در این تحقیق، نتایج رضایت بخشی به دست نمی‌آید. روش PCA به‌طور غیر نظارت‌شده، با در نظر گرفتن مقادیر داده، اجزای اصلی و مقادیر و بردارهای ویژه را محاسبه نموده و سعی در ماکزیمم نمودن ماتریس کوواریانس براساس تجزیه مقادیر منفرد دارد. مدل‌های تخمین مواد آلی خاک به دو روش PLSR و BRT برای طیف بازتابی، جذبی و مشتق اول و دوم آن‌ها، اجرا شد. بررسی نتایج به‌دست‌آمده از توسعه این دو مدل حاکی از این است که مدل BRT، با مقادیر RMSE و R^2، به ترتیب ۰/۵۸ و ۰/۹۴، در داده مشتق دوم طیف اصلی، نتایج بهتری را به دست آورده است. از طرفی، مقادیر RMSE و R^2 در مدل PLSR برای داده مشتق اول طیف اصلی، به ترتیب ۱/۲۰۳۳۸ و ۰/۹۳۸ به‌دست‌آمده است. به‌طورکلی مقایسه RMSE مدل BRT و مدل PLSR، دلالت بر نتایج بهتر مدل BRT در این منطقه دارد.

نتیجه‌گیری نتایج این تحقیق مؤید این مطلب است که در مناطق ناهمگن کشاورزی - شهری، می‌توان از پتانسیل مدل‌های توسعه داده‌شده Wavelet-PCA-PLSR و Wavelet-PCA-BRT برای تخمین مواد آلی خاک استفاده نمود. چرا که اندازه‌گیری میدانی ویژگی‌های شیمیایی خاک نظیر مواد آلی بسیار زمان و هزینه‌بر است. علاوه بر این، امکان اندازه‌گیری این ویژگی‌ها در پوشش وسیع وجود ندارد. با استفاده از این توابع پیوسته و تصویر ماهواره‌ای، می‌توان نقشه مقادیر مواد آلی خاک را در پوشش وسیع تولید نمود تا از آن بتوان در مطالعاتی نظیر پتانسیل کشت، حاصلخیزی خاک و توسعه پایدار آن بهره‌برداری نمود.

واژه‌های کلیدی: طیف‌سنجی، ماده آلی خاک، رگرسیون کمترین مربعات جزئی، درخت رگرسیون ارتقا یافته، جنوب غربی تهران

لطفاً به این مقاله استناد کنید: اکبری، ا.، میرزایی، س.، تومانیان، آ.، درویشی بلورانی، ع.، بهرامی، ح. ع. ۱۴۰۱. ارائه روش تلفیقی کاهش نویز- داده‌کاوی برای تخمین ماده آلی خاک با طیف‌سنجی VNIR، نشریه سنجش‌ازدور و سامانه اطلاعات جغرافیایی در منابع طبیعی، ۱۳(۳): ۲۱-۱. https://doi.org/10.30495/GIRS.2022.1938557.1935

مقدمه

خاک به‌عنوان منبع طبیعی ناهمگن و بزرگ‌ترین مخزن کربن آلی در اکوسیستم زمینی، از فرآیندها و مکانیسم‌های پیچیده‌ای تشکیل‌شده است (۳۰). ضرورت برآورد اطلاعات دقیق خاک در مقیاس ملی و منطقه‌ای به‌منظور بهبود مدیریت خاک و درک خصوصیات خاک و چگونگی تأثیرگذاری آن در کشاورزی (۲۶)، منجر به علاقه‌مند شدن محققین به این حوزه شده است.

محتوای مواد آلی خاک (Soil Organic Matter, SOM) به‌عنوان شاخص کیفیت خاک در حاصلخیزی آن و تولید مواد غذایی تأثیرگذار است (۲، ۳ و ۳۰) و نیز به‌عنوان یک متغیر کلیدی در مباحث محیطی و کشاورزی محسوب می‌شود (۱۱). جمع‌آوری تعداد زیادی داده خاک دقیق با هدف مدیریت منابع غذایی برای جمعیت آینده ضروری است (۱). بنابراین استفاده از روش‌های تخمین سریع و ارزان و البته افزایش دقت برآورد محتوای SOM در ارزیابی و مدیریت منابع خاک می‌تواند کمک‌کننده باشد (۹).

در کشاورزی دقیق، مقیاس اطلاعات خاک موردنیاز برای مدیریت اراضی و محصول بسیار کوچک‌تر بوده و به‌طورمعمول مقیاس جمع‌آوری داده‌های میدانی جوابگوی این نیاز نیست (۱۷). نمونه‌برداری و آنالیز تعداد زیاد نمونه خاک و تهیه نقشه توزیع SOM، برای مناطق وسیع و بزرگ، بسیار دشوار است (۳۰). علاوه بر این، روش‌های سنتی آزمایشگاهی تجزیه‌وتحلیل خاک برای نمونه‌برداری زیاد نیاز به نیروی کار بیشتر بوده و علاوه براین زمان‌بر و هزینه‌بر است و نیاز به اپراتور آزمایشگاه متخصص دارد.

یکی از رایج‌ترین روش‌های تخمین پارامترهای خاک، استفاده از طیف‌سنجی مرئی و مادون‌قرمز نزدیک (VNIR) در محدوده طیفی ۳۵۰-۲۵۰۰ نانومتر است (۱۹). طیف‌سنجی آزمایشگاهی و میدانی و سنجش‌ازدور ابرطیفی، پتانسیل ارزیابی محتوای SOM را به دلیل فراوانی اطلاعات طیفی داراست (٤). این روش به‌عنوان مکمل روش‌های تجزیه‌وتحلیل آزمایشگاهی در برآورد خصوصیات خاک استفاده می‌گردد. به‌طوری‌که تمایل پژوهشگران به استفاده از روش‌های تجزیه‌وتحلیل VNIR با توجه به‌سرعت، مقرون‌به‌صرفه بودن و مهم‌تر از آن مجاز بودن تعداد نمونه-برداری بالا رو به افزایش است (۱۳، ۱۸، ۲۳ و ۲۵). پیک‌های مختلف بازتاب‌های طیفی برای مواد آلی مختلف ارائه‌شده است که مهم‌ترین آن‌ها ترکیبات آروماتیک در طول‌موج‌های ۱۶۱۱-۱۱۰۰-۸۲۵ نانومتر، آمین در طول‌موج‌های ۲۰۶۰-۱۵۰۰-۱۰۰۰-۷۵۱ نانومتر، کربوکسیلیک اسید در طول‌موج‌های ۱۹۳۰-۱٤٤۹، آمیدها در طول‌موج‌های ۲۰۳۳-۱۵۲٤ نانومتر، ترکیبات آلیفاتیک در طول‌موج‌های ۲۲۷۵-۱۷۰٦، فنول در طول‌موج ۱۹٦۱ نانومتر، پلی ساکارید در طول‌موج ۲۱۳۷ نانومتر، کربوهیدرات در طول‌موج ۲۳۸۱ نانومتر است (٦ و ۱۵).

تاکنون نقشه‌برداری و طبقه‌بندی خاک از طریق روش‌های مختلف، نظیر روش‌های آماری مانند رگرسیون مؤلفه‌های اصلی (Principal Component Regression, PCR) (۱۸)، رگرسیون کمترین مربعات جزئی (Partial Least Squares Regression, PLSR) (۱٦ و ۱۸)، و نیز استفاده از روش‌های یادگیری ماشین نظیر انواع مختلف شبکه‌های عصبی مصنوعی، درخت‌های تصمیم‌گیری، جنگل تصادفی و ماشین بردار پشتیبان (۱۰، ۱۱، ۱۷، ۱۸ و ۲٤) انجام‌شده است. در زمینه مواد آلی خاک نیز تاکنون مطالعات متعددی نظیر استفنس و همکاران (۲۲) با استفاده از فن‌آوری ابرطیفی و روش‌های یادگیری ماشین، صورت گرفته است. نامبردگان با استفاده از طیف‌سنجی VNIR به تهیه نقشه SOM در مناطق همگن دارای مواد آلی پرداخته‌اند. یانگ و لی (۲۹) محتوای SOM را از طریق ترکیب طیف‌سنجی خاک و رگرسیون خطی چند متغیره گام به گام کمی‌سازی نموده‌اند. در میان روش‌های کارآمد در ایجاد مدل‌های قابل‌اعتماد در زمینه سنجش‌ازدور ابرطیفی، برای تخمین محتوای SOM، روش رگرسیون کمترین مربعات جزئی بیشتر استفاده‌شده است. نوتیکا و همکاران (۲۰) و وهلند و همکاران (۲۷) از روش طیف‌سنجی و تخمین میزان مواد آلی خاک با استفاده از روش PLSR استفاده نموده‌اند و

قابلیت این روش را در تخمین مواد آلی خاک اذعان نموده‌اند. بااین‌حال، تجزیه‌وتحلیل PLSR و پردازش مواد آلی خاک بسیار تحت تأثیر نویزهای موجود در داده‌های طیف‌سنجی است (۷). ازاین‌رو، لین و همکاران (۱۳) با استفاده روش تبدیل موجک- همبستگی- PLSR به تخمین دقیق‌تر مواد آلی خاک از طریق طیف‌سنجی و کاهش نویز پرداخته‌اند. در مورد روش یادگیری ماشین درخت رگرسیون ارتقا یافته (Boosted Regression Tree, BRT) نیز تحقیقات اندکی در زمینه بررسی مواد آلی خاک از طریق طیف‌سنجی انجام‌شده است. در اکثریت مقالات به بررسی ارتباط مواد آلی خاک و عوامل محیطی به روش BRT پرداخته‌شده است. ویسکارا راسل و بهرنس (۲٤)، به مقایسه PLSR و روش‌های داده‌کاوی نظیر BRT، بردار پشتیبان تصمیم و جنگل تصادفی پرداخته‌اند و به کارایی بالاتر روش‌های داده‌کاوی اشاره‌کرده‌اند. لئو و همکاران (۱٤) با استفاده از داده طیف‌سنجی و نمونه مواد آلی، به روش BRT، میزان مواد آلی خاک را با دقت قابل‌قبول تخمین زده‌اند.

هدف از این تحقیق، توسعه روش تلفیقی برمبنای تبدیل موجک- تحلیل مؤلفه‌های اصلی- یادگیری ماشین براساس مقایسه عملکرد دو روش PLSR و BRT برای پیش‌بینی مواد آلی خاک با استفاده از طیف VNIR، است. به‌طوری‌که در این روش توسعه داده‌شده، با استفاده از ترکیب تبدیل موجک و تشخیص باندهای مستقل، نویزهای موجود در داده‌های طیف-سنجی خاک کاهش‌یافته است. علاوه بر این، طیف‌ها یا باندهای مستقل و مؤثر در طیف‌سنجی مواد آلی خاک انتخاب گردیدند. براین اساس، در این تحقیق، روش‌های Wavelet-PCA-PLSR و Wavelet-PCA-BRT توسعه داده‌شده است و کارایی هر یک از آن‌ها ارزیابی می‌گردد.

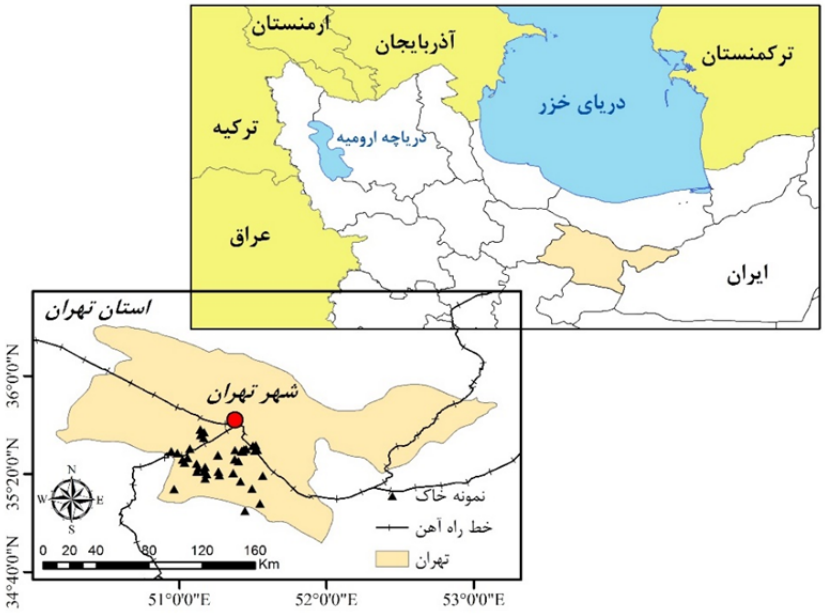
روش تحقیق

منطقه موردمطالعه

استان تهران به مرکزیت شهر تهران، با وسعتی حدود ۱۲۹۸۱ کیلومتر مربّع، بین ۳٤/۵ تا ۳٦/۰۵ درجه عرض شمالی و ۵۰ تا ۵۳/۳ درجه طول شرقی واقع‌شده است. این استان از شمال به استان مازندران، از جنوب به استان قم، از جنوب‌غربی به استان مرکزی، از غرب به استان البرز و از شرق به استان سمنان محدود است. ایستگاه‌های نمونه‌برداری خاک اندازه-گیری شده در این تحقیق، در جنوب غربی استان تهران در نواحی کشاورزی روستایی شهر ری و اسلامشهر بین ۳۵/۰۸ تا ۳۵/٦۳ درجه عرض جغرافیایی و ۵۱ تا ۵۱/٦٦ درجه طول جغرافیایی واقع‌شده است (شکل ۱).

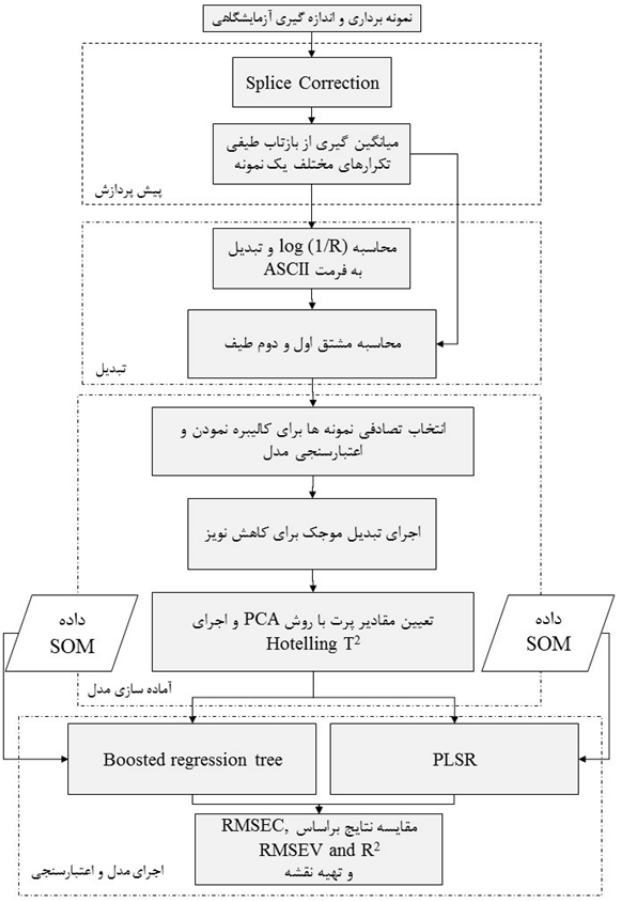
مواد و روش‌ها

شکل ۲، مراحل اجرایی تحقیق را نشان می‌دهد که از چهار مرحله تشکیل‌شده است. بخش اول شامل نمونه‌برداری، آماده‌سازی نمونه‌ها و اندازه‌گیری آزمایشگاهی و طیف‌سنجی می‌شود، بخش دوم شامل پیش‌پردازش داده‌های طیفی و محاسبات مشتق بازتاب و جذب طیفی می‌شود، بخش سوم آماده‌سازی مدل و بخش چهارم مدل‌سازی و ارزیابی صحت نتایج را شامل می‌شود.



شکل۱. موقعیت جغرافیایی نقاط نمونه‌برداری خاک در استان تهران

Fig. 1. Geographical location of soil samples in Tehran province



شکل۲. فلوچارت روش تحقیق

Fig. 2. Flow chart of research methodology

نمونه‌برداری و اندازه‌گیری آزمایشگاهی

۴۲ نمونه خاک از منطقه ناهمگن کشاورزی شهری در تهران در ۳۰-۰ سانتی‌متر خاک جمع‌آوری گردید. اسیدیته خاک با استفاده از pH متر و کربنات کلسیم خاک با استفاده از کلسیمتر و دو روش اسکیبلر اندازه‌گیری شد. بافت خاک و ماده آلی خاک به ترتیب با استفاده از روش‌های هیدرومتری و والکی-بلک اندازه‌گیری شد. در این تحقیق، اندازه‌گیری بازتاب طیفی خاک با استفاده از طیف‌سنج FieldSpec 3 ساخت شرکت ASD در محدوده ۲۵۰۰-۳۵۰ نانومتر با زاویه دید سنجنده ۲٤ درجه انجام شد. برای کالیبره کردن دستگاه، از صفحه مرجع سفید استفاده شد. این اندازه‌گیری در یک اتاق تاریک برای جلوگیری از اثر نورهای زائد انجام‌شده است. سنجنده به‌صورت عمود بر نمونه خاک تنظیم و زاویه نور ۷۵ درجه، تنظیم شد. همچنین، فاصله بین سنجنده و نمونه ۷ سانتی‌متر در نظر گرفته‌شده است. برای هر نمونه ۷ بار اندازه‌گیری طیفی تکرار شد. در شکل ۳ نحوهٔ نمونه‌برداری، اندازه‌گیری آزمایشگاهی و طیف‌سنجی نشان داده‌شده است.

شکل۳. نمونه‌برداری، اندازه‌گیری‌های آزمایشگاهی و طیفی نمونه‌های خاک

Fig.3. Sampling, laboratory measurement and spectrometry of soil samples

پیش‌پردازش و تبدیلات طیفی

به دلیل اینکه دستگاه طیف‌سنج از سه مجموعه آشکارساز با جنس‌های مختلف ساخته‌شده است، نتایج این آشکارسازها ممکن است در اثر عوامل مختلف کاملاً منطبق نباشد. ابتدا تصحیح بایاس میان طیف‌ها و پر کردن شکاف‌ها با تابع Splice correction در نرم‌افزار ViewSpec انجام شد. با استفاده از فیلتر Savitzky-Golay چندجمله‌ای درجه دو و اندازه پنجره ۱۱ طیف‌ها نرم شدند. سپس میانگین حسابی ۷ تکرار طیفی برای هر نمونه در نرم‌افزار MATLAB محاسبه شد. ازآنجایی‌که مشتق داده‌های طیفی، اثر نویزهای با فرکانس پایین را کاهش می‌دهد (۱۲)، مشتق اول و دوم متوسط طیف محاسبه شد. همچنین، تفاوت طیفی در ناحیه مرئی را می‌توان با لگاریتم معکوس گرفتن از طیف بارز نمود، و علاوه بر این، تأثیر تنوع نوردهی را به حداقل رساند (۲۸). در تحقیق حاضر، علاوه بر محاسبه مشتق اول (FDR) و مشتق دوم (SDR) بازتاب طیفی اصلی (REF)، جذب (1/R) و مشتق اول (log(1/R))' و دوم "(log(1/R)) جذب محاسبه گردید. به‌منظور توسعه روش‌های Wavelet-PCA-PLSR و Wavelet-PCA-

BRT، به ترتیب تبدیل موجک، تحلیل مؤلفه‌های اصلی و الگوریتم PLSR یا BRT اجرا گردید که در ادامه تشریح می‌شود.

روش رگرسیون کمترین مربعات جزئی - تبدیل موجک گسسته

تعداد زیادی نویز در منحنی انعکاس طیفی به‌ویژه در طیف خاک به چشم می‌خورد و طیف تبدیل یافته نیز دندانه دندانه است (۱۳). بنابراین، در این تحقیق، به‌منظور کاهش نویز و هموارسازی طیف، از روش تبدیل موجک تابع ماتریس Sym8 بر اساس لین و همکاران (۱۳) استفاده‌شده است. همچنین، تبدیل موجک به‌منظور نشان دادن و بارزسازی ویژگی‌ها در طیف انجام می‌شود. تبدیل موجک گسسته (Discrete Wavelet Transform, DWT) همانند تبدیل فوریه، برای تبدیل فضای طیف به فضای دیگری با ویژگی معنی‌دار، استفاده می‌شود. DWT با استفاده از رابطهٔ ۱ محاسبه شد.

$$x(t) = \sum_{j=1}^{l} \sum_{k=0}^{2^l} c_{j,k}\, \Psi_{j,k} \qquad [۱]$$

در این رابطه؛ $\Psi 0,0$ موجک والد است که سایر موجک‌ها ($\Psi_{j,k}$) از آن مشتق شده‌اند، $x(t)$ طیف، l سطح تجزیه DWT و $c_{j,k}$ ضریب موجک محاسبه‌شده به‌وسیله تولیدات بین $x(t)$ و $\Psi_{j,k}$ است (رابطهٔ ۲).

$$C_{j,k} = \langle x(t) | \Psi_{j,k} \rangle \qquad [۲]$$

پس از انجام روش کاهش نویز تبدیل موجک، تجزیه‌وتحلیل مؤلفه‌های اصلی (PCA) به‌منظور تشخیص داده‌های پرت استفاده شد. برای تشخیص داده‌های پرت، مقدار حد آستانه از طریق آزمون هادلینگز (T^2 Hotelling's) (رابطهٔ ۱) محاسبه شد (۸). در این پژوهش نیز از آزمون هادلینگز با در نظر گرفتن فاصله اطمینان ۹۵٪ استفاده‌شده است (رابطهٔ ۳).

$$T^2_{l,n,a} = \frac{l(n-1)}{n-1} F_{l,n-1,a} \qquad [۳]$$

در این رابطه؛ n تعداد نمونه‌ها، L تعداد مؤلفه‌ها، α سطح معنی‌داری (معمولاً بین ۱٪ و ۵٪) و $F_{L, n-l}$، توزیع فیشر F با L و n _ L درجه آزادی است. زمانی که مقدار T^2 بیش از مقدار حد آستانه باشد، آن داده، پرت تلقی می‌گردد (۸).

به‌منظور کالیبراسیون و اعتبارسنجی مدل، به‌طور تصادفی به ترتیب ۳۰ و ۱۲ نمونه خاک انتخاب‌شده است. پس از حذف داده پرت از هر مجموعه، روش PLSR و درخت رگرسیون ارتقا یافته بر روی ٦ ویژگی ازجمله بازتاب طیفی اصلی (REF)، مشتق اول (FDR) و مشتق دوم (SDR) بازتاب هر طیفی اصلی، لگاریتم معکوس، مشتق اول و دوم آن اجراشده است. علاوه بر این، این روند در ۵ سطح از تبدیل موجک محاسبه شد. سپس، با مقایسه نتایج، مدل مناسب از طریق اعتبارسنجی انتخاب شد. مدل کلی روش PLSR از طریق رابطه‌های ٤ و ۵ محاسبه گردید.

$$X = TP^T + E \qquad [٤]$$

$$Y = UQ^T + F \qquad [۵]$$

در این رابطه؛ X ماتریس مقادیر پیش‌بینی کننده ورودی، Y ماتریس پاسخ، T و U به ترتیب ماتریس مؤلفه X و Y است.P و Q، به ترتیب، ماتریس لودینگ (Loading) در راستای X و Y، و ماتریس E وF ماتریس خطا هستند. این تجزیه با هدف به حداکثر رساندن کوواریانس بین T و U انجام می‌شود. برای انجام PLSR در داده‌های غیرخطی، از تابع کرنل استفاده می‌شود.

روش درخت رگرسیون ارتقا یافته

در هنگام استفاده از نمونه عددی، به‌جای درخت تصمیم-گیری از درخت رگرسیون استفاده می‌شود، اما روند آن‌ها یکسان است. در درخت رگرسیون از جستجو حریصانه استفاده می‌شود. بنابراین، با پاسخ دادن به سؤال باینری که حداکثر اطلاعات در مورد Y از طریق کدام نود به دست می‌آید، گره ریشه و دو فرزند آن تعیین می‌گردد. این فرایند در هر گره

فرزند تکرار می‌شود. تولید ساختمان درخت به‌صورت بازگشتی تکرار شده است و یک معیار توقف معمولی در نظر گرفته می‌شود. معیار توقف می‌تواند نظیر رسیدن به انشعابی که قابل‌تقسیم نیست و اطلاعات کمتری می‌دهد و یا زمانی که اطلاعات در گره حاوی کمتر از، پنج درصد از کل داده‌ها است، باشد. همچنین، سعی در به حداقل رساندن اندازه درخت است. برای تقسیم گره، عامل جینی، عامل آنتروپی و غیره به‌منظور به حداقل رساندن این عوامل استفاده‌شده است. علاوه بر این، در هر شاخه، مجموع مربع خطاها محاسبه‌شده و آن‌هایی که مقادیر حداقل دارند، انتخاب می‌شود. همچنین، در درخت رگرسیون، فرایند هرس برای کاهش بیش برازش (Over-fitting) استفاده می‌گردد (۵). خروجی چندبعدی درخت رگرسیون غیرخطی به‌صورت رابطهٔ ٦ است.

$$H(x; A, R) = \sum_{k=1}^{K} a_k \mathbb{J}(X \in r_k) \qquad [٦]$$

در این رابطه؛ $P = f(A_m;R_m)g^M \ m=0$ ، مجموعه پارامترهای درخت رگرسیون را ارائه می‌دهد. $R = f(r_1; : : : ; r_K)$ مجموعه تقسیمات مجزای داده ورودی و $A = f(a_1; : : : ; a_K)$، مجموعه بردارها، M تعداد درخت و k نود برگ است. در درخت رگرسیون، مهم‌ترین موضوع انتخاب بهترین مقادیر پارامترهای A و R است. مجموع وزنی مربعات خطا برای هر گره برگ K به‌صورت رابطه ۷ محاسبه می‌شود.

$$S_k = \sum_{i \in r_k} w \|t_i - a_k\|_2^2 \qquad [٧]$$

تابع درخت رگرسیون به‌عنوان یک گروه از رگرسیون‌های درختی H بر اساس رابطه ۸ تعیین می‌شود.

$$F(x; P) = \sum_{m=0}^{M} H(x; A_m, R_m) \qquad [٨]$$

روش درخت رگرسیون ارتقا یافته، دو روش درخت رگرسیون و تکنیک ارتقا را به‌منظور بهبود توان پیش‌بینی هرکدام از آن‌ها ترکیب می‌کند (۲۸). این روش، توسط الگوریتم افزایش گرادیان، سعی در به حداقل رساندن رابطهٔ ۹ را دارد.

$$F^*(x) = \underset{F(x)}{\operatorname{argmin}} \sum_{i=1}^{N} w_i \psi(t_i, F(x_i)) \qquad [٩]$$

$a(t; F(x)) = \|t - F(x)\|_2^2$، تابع کاهش مربع خطا و N تعداد نمونه‌هاست. شبه کد درخت رگرسیون ارتقا یافته به شرح زیر است (جدول ۱): که در آن ^{o} یک پارامتر انقباض به‌منظور جلوگیری از بیش برازش است.

جدول۱. شبه کد درخت رگرسیون ارتقا یافته

Table 1. Pseudo-code of the BRT

1: $F_0(x) = \bar{t}$ weighted mean
2: for $m = 1$ to M do
3: $\tilde{t}_i = t_i - F_{m-1}(x_i), i = 1, \dots, N$
4: $(A_m, R_m) = \operatorname{argmin}_{A,R} \sum_{i=1}^{N} w_i \|\tilde{t}_i - H(x_i; A, R)\|_2^2$
5: $F_m(x) = F_{m-1}(x) + \nu H(x; A_m, R_m)$
6: end for

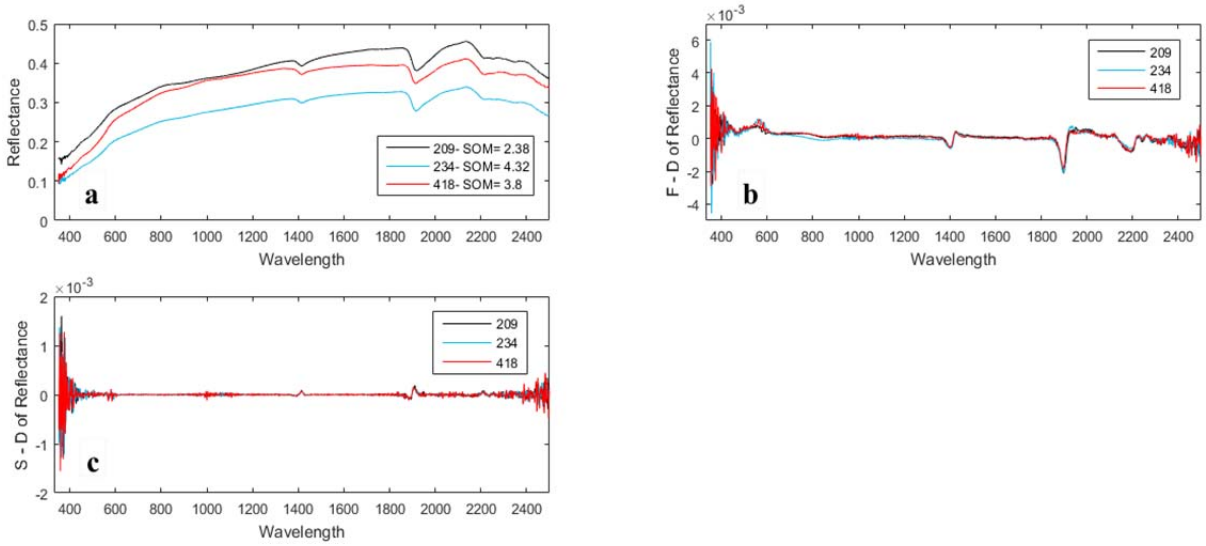
اعتبارسنجی مدل و تهیه نقشه

برای کالیبراسیون و اعتبارسنجی مدل، ۳۰ و ۱۲ نمونه خاک، به ترتیب به‌صورت تصادفی انتخاب شدند. به‌منظور بررسی دقت از ضریب تعیین (R^2) و ریشه میانگین مربعات خطای اعتبارسنجی به روش (RMSE) موردبررسی قرار گرفت. علاوه بر این، برای انتخاب بهترین فاکتور تولید مدل PLSR برای هر طیف، واریانس و باقی‌مانده مقادیر برآوردی و RMSE در هر دو مدل کالیبراسیون و اعتبارسنجی در نظر گرفته شد. درنهایت، با استفاده از تصویر ماهواره‌ای لندست OLI مربوط به تاریخ نمونه‌برداری خاک و طیف‌سنجی (آبان ۱۳۹۳)، و به روشی که با دقت بیشتر ارزیابی شد، نقشه مقادیر مواد آلی خاک منطقه تولید گردید.

نتایج

تفسیر منحنی بازتابندگی طیفی خاک و مواد آلی

به علت رعایت حجم مقاله، تنها منحنی رفتار طیفی سه نمونه خاک اندازه‌گیری شده، در شکل ٤ ارائه می‌گردد. اختلاف بین بازتابندگی خاک در نمونه‌های مختلف با مقادیر متفاوت SOM (به ترتیب در نمونه‌های ۲۰۹، ۲۳٤ و ٤۱۸، مقادیر ۲/۳۸، ٤/۳۲ و ۳/۸ درصد با جنس خاک لوم) در شکل a-٤ به چشم می‌خورد. به‌طوری‌که با افزایش محتوای SOM، بازتابندگی خاک به‌طورکلی کاهش‌یافته است. در این طیف‌های خاک به‌طور نمونه، سه باند جذبی قابل‌توجه در طول‌موج‌های ۱٤۲٦، ۱۹۲۰ و ۲۲۲۸ دیده می‌شود. به‌طوری‌که این باندهای جذبی به‌وضوح در شکل b-٤ و c-٤ مشهود است. این باندهای جذبی معروف به باندهای جذبی آب می‌باشند و از آن‌ها در شناسایی مواد موجود در خاک می‌توان استفاده نمود.



شکل۴. (a) بردار طیفی بازتابندگی طیف، (b) مشتق اول و (c) مشتق دوم بازتابندگی در سه نمونه خاک و میزان SOM آن‌ها (٪)

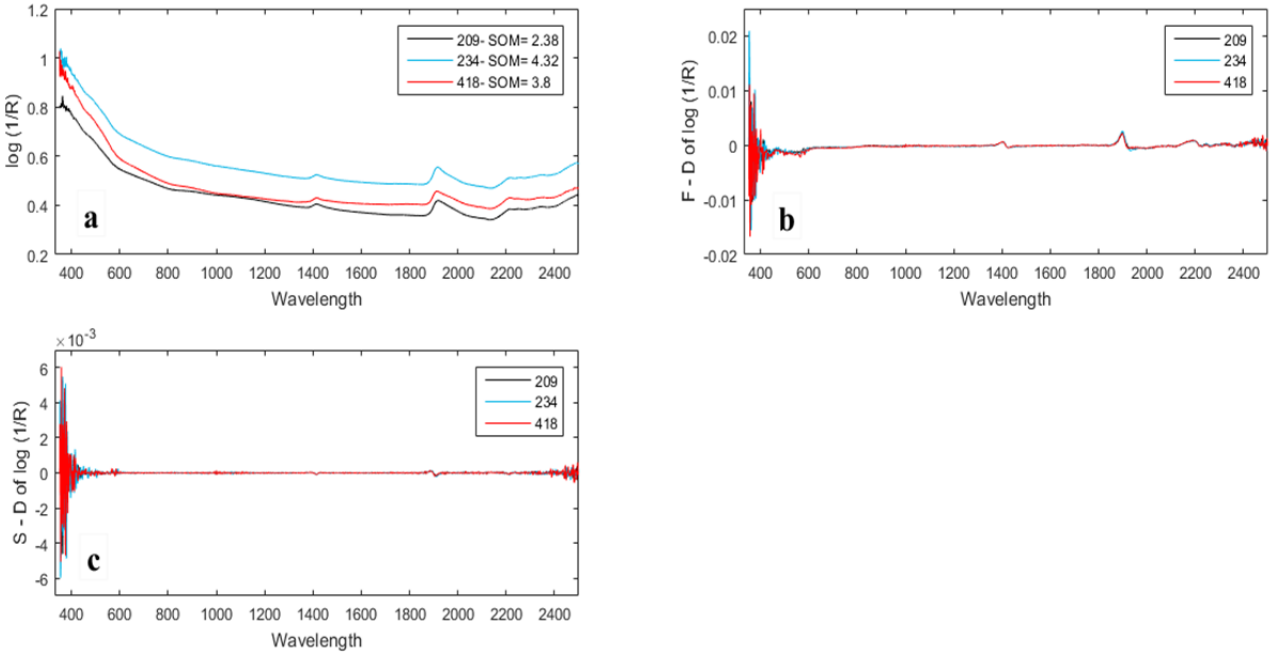
Fig. 4. (a) spectral reflectance vectors of the original spectrum, (b) is the first derivative and (c) is the second derivatives in the three measured soil samples and their levels of SOM

در جدول ۲، اطلاعات آماری مواد آلی خاک در نمونه‌های اندازه‌گیری شده در آزمایشگاه ارائه‌شده است. همان‌طور که در این اطلاعات آماری مشهود است، اختلاف زیادی بین میانگین و میانه وجود دارد که این دلیل بر ناهمگنی منطقه و وجود داده‌های پرت است. بنابراین روش PCA و تبدیل موجک، به‌منظور حذف و تعدیل نویزها و داده‌های پرت استفاده‌شده است.

جدول۲. اطلاعات آماری مواد آلی خاک در مجموعه داده‌های کالیبراسیون و اعتبارسنجی

Table 2. Summary statistics of the SOM in the calibration and validation dataset

SOM (%)	حداقل	میانه	میانگین	حداکثر
نمونه کالیبراسیون	۰/۱٤	۰/۸٤	۱/۲٦۱۳۱۳	٤/۳۲
نمونه اعتبارسنجی	۰/۱٤	۰/٦۱	۰/۷۳٤۱٦۷	۱/۳۳



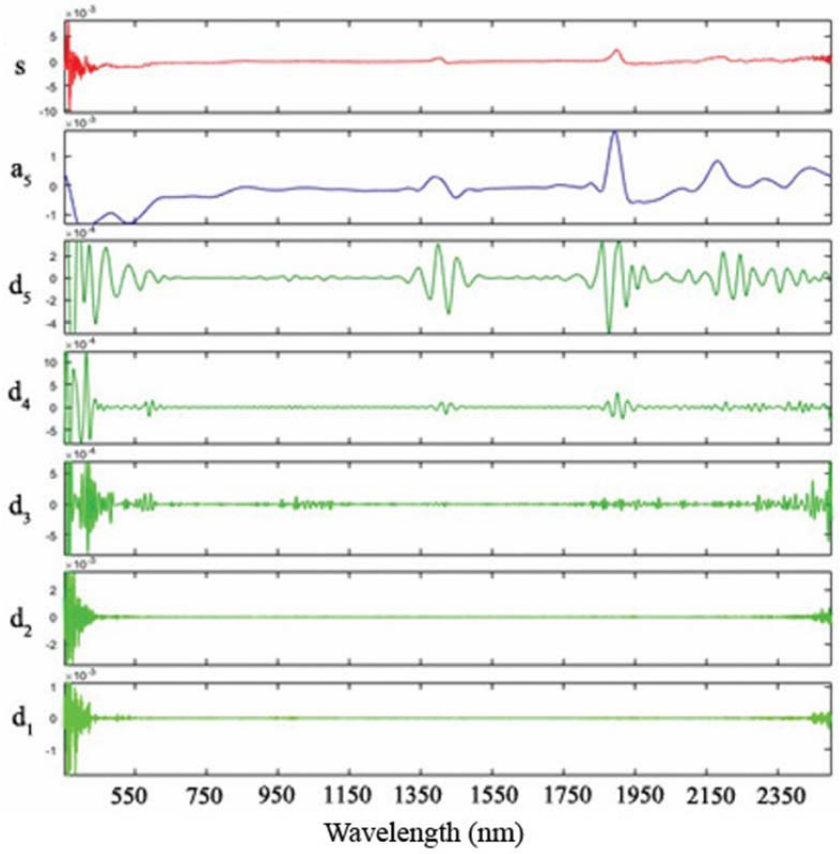
شکل۵. (a) بردار طیفی log (1/R)، (b) مشتق اول و (c) مشتق دوم آن در سه نمونه خاک اندازه‌گیری شده و میزان SOM آن‌ها

Fig. 5. (a) Spectral reflectance vector log (1/R), (b) the first derivative, (c) the second derivative in the three soil samples and their levels of SOM

تبدیل موجک گسسته و حذف داده‌های پرت

طیف‌های نمونه‌های خاک به FDR، SDR، log (1/R)، (log (1/R))' و (log (1/R))" تبدیل‌شده‌اند. در شکل ۵-a، ۵-b و ۵-c به ترتیب مشتق اول طیف اصلی، مشتق دوم آن، log (1/R)، (log (1/R))' و (log (1/R))" ارائه‌شده است.

همان‌طور که در شکل ٤ و ۵ نیز مشهود است، طیف‌های خاک با نویزهای دندانه‌دندانه همراه است. به‌منظور حذف نویزها و باقی‌مانده اطلاعات مفیدتر، از تبدیل موجک گسسته سطح ۵ به‌منظور تجزیه طیف و حذف نویز استفاده شد. به استناد تحقیقات مختلف که به‌منظور حذف نویز از تبدیل sym8 استفاده کرده‌اند، در این تحقیق نیز از این روش برای حذف نویز بهره‌برداری شد. به علت رعایت اختصار، تنها یک نمونه از نتایج تبدیل موجک بر روی مشتق اول جذب طیفی ارائه می‌گردد. در شکل ٦، به ترتیب از بالا، مشتق اول جذب طیفی، تقریب سطح ۵، جزییات سطح ۵، الی ۱، با استفاده از تبدیل موجک گسسته نمایش داده می‌شود.



شکل۶. منحنی طیف به ترتیب از بالا، '(log (1/R))، تقریب سطح ۵، جزییات سطح ۵، الی ۱. با استفاده از تبدیل موجک گسسته.

Fig. 6. Spectral curves of (log (1/R))', level 5 approximation and details of level 5 to 1 from top to down.

بر اساس اطلاعات آماری به‌دست‌آمده از هر سطح تبدیل موجک، سطح مناسب‌تر برای حذف نویز در داده موردنظر انتخاب‌شده است. در جدول ۳، اطلاعات آماری در مورد هر سطح موجک در طیف مشتق اول جذب طیفی ارائه می‌گردد و با در نظر گرفتن تشابه میانگین و میانه، سطح مناسب‌تر انتخاب گردید.

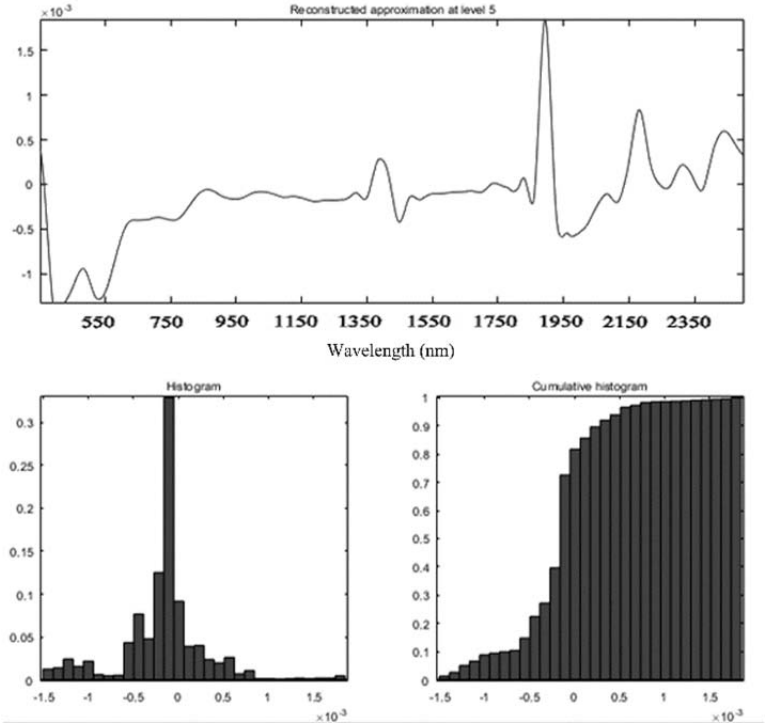
جدول۳. اطلاعات آماری در مورد هر سطح تبدیل موجک در مشتق اول جذب طیفی.

Table 3. statistical data related to each level of wavelet transform in the (log (1/R))' spectrum.

سطح تبدیل	حداقل	حداکثر	میانگین	میانه	انحراف معیار
(log (1/R))'	- ۰/۰۱۰٤	۰/۰۰۸۰۵	- ۰/۰۰۰۲۵۲۱	-۰/۰۰۱۳	۰/۰۰۰۷۱۹
تقریب سطح ۱	-۰/۰۰۹٤۵۵	۰/۰۰۷۹٦	۰/۰۰۰۱۵۹۲	-۰/۰۰۰۱۳۰۲	۰/۰۰۰٦۹۸۸
تقریب سطح ۲	-۰/۰۰٦۳٤۱	۰/۰۰٦٦۷۱	-۵/۲ e-05	-۰/۰۰۰۱۳۳٤	۰/۰۰۰۵۹۵٦
تقریب سطح ۳	-۰/۰۰۲٦۳٦	۰/۰۰۳٤٤۵	-۰/۰۰۰۱۰۲۳	- ۰/۰۰۰۱۳۵٤	۰/۰۰۰۵۲٦۸
تقریب سطح ٤	- ۰/۰۰۱۷۹٦	۰/۰۰۲۰۳	-۷/٤ e-05	-۰/۰۰۰۱۳٤۱	۰/۰۰۰٤۸۸٦
تقریب سطح ۵	-۰/۰۰۱۵۰۵	۰/۰۰۱۸۵۳	-۰/۰۰۰۱۱۵٦	-۰/۰۰۰۱۳۲	۰/۰۰۰٤٦٦٦

به‌طوری‌که نزدیکی مقادیر میانگین و میانه و مقدار کمتر انحراف معیار در داده تقریب به‌دست‌آمده از تبدیل موجک، دلالت بر نرمال بودن هیستوگرام و کاهش مقادیر نویز است.

بر اساس اطلاعات به‌دست‌آمده از ۵ سطح تبدیل موجک، سطح پنجم به دلیل نزدیکی مقادیر میانه و میانگین به سطح نرمال نزدیک‌تر است (شکل ۷).



شکل۷. هیستوگرام و منحنی سطح ۵ تبدیل موجک برای داده مشتق اول جذب طیفی

Fig. 7. Histogram and curves of level 5 wavelet transform for the data of the (log (1/R))' spectrum.

بعدازاین مرحله، به‌منظور حذف مقادیر پرت از داده‌ها، از روش PCA با محاسبه مقدار حد آستانه هادلینگز، استفاده‌شده است. در این داده‌ها با مقدار نمونه (حاصل‌ضرب تعداد طیف در تعداد نمونه کالیبراسیون)، مؤلفه و سطح اطمینان به ترتیب، ٦٤٥٠٠، ۳۰ و ۹۵ درصد (میزان آلفا ۰/۰۵)، میزان توزیع فیشر و حد آستانه هادلینگز، ۱/٦۲۲۳ و ٤٨/٦٩، به ترتیب به‌دست‌آمده است.

برآورد ماده آلی خاک با PLSR و BRT و تهیه نقشه

بعد از حذف مقادیر پرت، مدل‌های PLSR و BRT برای پیش‌بینی مقادیر SOM استفاده‌شده است. نتایج به‌دست‌آمده برای مدل PLSR در هر تبدیل طیف، بر اساس ۲۰ عامل (مؤلفه یا فاکتور) و انتخاب بهترین فاکتور تولید مدل PLSR برای واریانس مقادیر برآوردی و باقی‌مانده و RMSE در هر دو مدل کالیبراسیون و اعتبارسنجی به‌دست‌آمده است. به علت رعایت اختصار، از ارائه این مقادیر برای واریانس مقادیر برآوردی و باقی‌مانده و RMSE در هر دو مدل کالیبراسیون و اعتبارسنجی برای کلیه ۲۰ فاکتور خودداری شده است. جدول ٤، مقادیر باقی‌مانده Y برای بازتاب و جذب طیفی و تبدیلات آن‌ها را در بهترین مؤلفه‌های یافت شده در دو مدل کالیبراسیون و اعتبارسنجی در تبدیلات مختلف ارائه می‌دهد. برای انتخاب فاکتور مناسب‌تر در پیش‌بینی مقادیر SOM، مجموعه این مقادیر و نیز مقادیر RMSE و R^2 در نظر گرفته‌شده است.

جدول۴. مقادیر باقی‌مانده Y، واریانس Y و RMSE در بهترین مؤلفه‌های یافت شده در دو مدل کالیبراسیون و اعتبارسنجی در تبدیلات مختلف

Table 4. Residual values of Y, explained variance of Y and RMSE in the best achieved components in the two calibration and validation models for different transforms

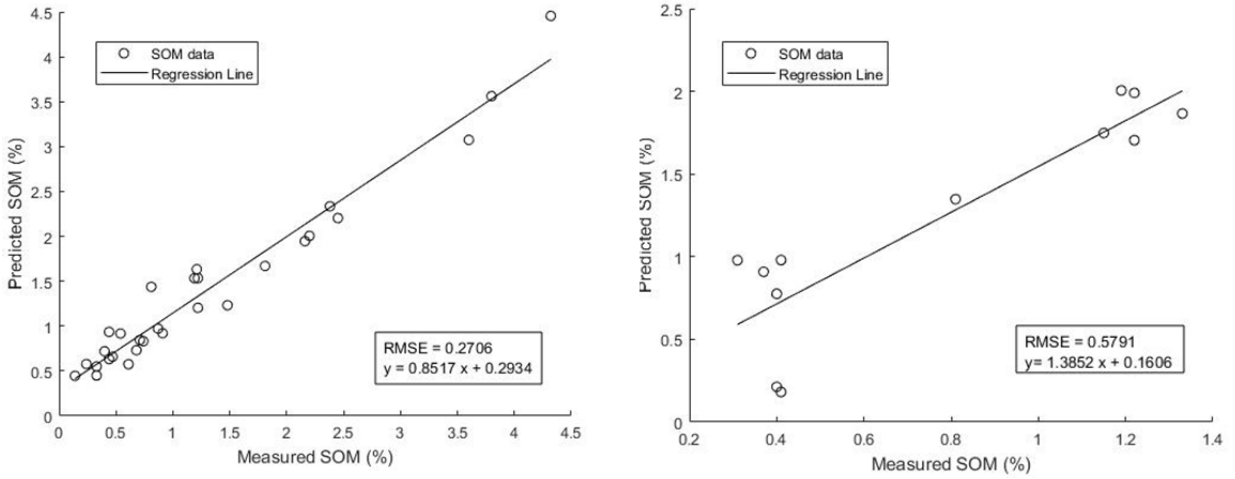
تبدیل	مؤلفه	باقیمانده Y		واریانس Y		RMSE	
		کالیبراسیون	اعتبارسنجی	کالیبراسیون	اعتبارسنجی	کالیبراسیون	اعتبارسنجی
بازتاب	۵	۰/۶۸۳	۰/۹۸۳	۴۳/۶۲۶	۲۲/۱۲۶	۰/۸۲۶	۰/۹۹۴
مشتق اول	۹	۰/۰۲۶	۱/۰۹۹	۹۷/۸۱۹	۱۹/۱۰۹	۰/۱۶۳	۱/۰۳۴
مشتق دوم	۳	۰/۲۹۱	۱/۳۸۸	۷۵/۹۹۴	۴/۴۲۵	۰/۵۳۹	۱/۰۸۵
جذب	۹	۰/۲۲۶	۱/۱۵۲	۸۱/۲۹۸	۱۵/۸۳۳	۰/۴۷۶	۰/۹۹۶
مشتق اول	۶	۰/۱۵۷	۰/۹۶۷	۸۷/۰۱۴	۱۴/۵۲۵	۰/۳۹۷	۱/۰۲۰
مشتق دوم	۷	۰/۹۹۸	۱/۴۶۳	۹۱/۷۶۱	۴/۵۰۷	۰/۳۱۶	۱/۲۰۷

جدول ۵، نتایج رگرسیون دو مدل PLSR و BRT را در تبدیلات مختلف طیف نشان می‌دهد. با مقایسه نتایج در مدل PLSR، این‌طور استنباط می‌شود که مشتق اول طیف اصلی برای تبدیل موجک سطح ۵، نتایج بهتری را برآورد نموده است. به‌طوری‌که مقادیر RMSE و R^2 به ترتیب ۱/۰۳۳۸ و ۰/۹۳۸ به‌دست‌آمده است. در مدل BRT، با مقایسه نتایج کالیبراسیون و اعتبارسنجی، این‌طور استنباط می‌شود که این مدل در داده مشتق دوم طیف اصلی با مقادیر RMSE و R^2 به ترتیب ۰/۵۸ و ۰/۹۴، نتایج بهتری را به دست آورده است. در مرتبه دوم و سوم اهمیت، مدل BRT در دو مشتق دوم جذب طیفی و مشتق اول بازتاب طیفی با مقادیر RMSE و R^2 به ترتیب ۰/۷۳۳۳ و ۰/۹۳۸۷ و ۰/۷۶۱۳ و ۰/۹۳۷۹ می‌باشند. با مقایسه نتایج دو مدل مشخص می‌شود که مدل BRT، نتایج بهتری در هر دو مدل کالیبراسیون و اعتبارسنجی به دست آورده است. درنتیجه از این مدل، می‌توان برای تولید سطح پیوسته از مقادیر SOM با استفاده داده‌های طیف‌سنجی با دقت به‌مراتب بالاتر از روش‌های موجود استفاده کرد. شکل‌های ۸ و ۹، نمودارهای مقادیر اندازه‌گیری شده در مقابل پیش‌بینی‌شده با استفاده از داده مشتق دوم طیف اصلی به روش BRT و داده مشتق اول طیف اصلی به روش PLSR را به ترتیب نشان می‌دهد. با استفاده از داده مشتق دوم طیف اصلی به روش BRT و داده تصویر ماهواره‌ای لندست OLI برای تاریخ نمونه-برداری، نقشه مقادیر ماده آلی خاک تولید گردید (شکل ۱۰).

جدول۵. نتایج آزمون دو مدل PLSR و BRT.

Table 5. The results of the two PLSR and BRT models.

مدل	طیف	مؤلفه	$RMSE_C$	$RMSE_V$	R^2
PLSR	جذب	۹	۰/٤٧٦	۰/۹۹٥	۰/۸۱
	مشتق اول جذب	٦	۰/۳۹٦	۱/۰۱۹	۰/۸۷
	مشتق دوم جذب	۷	۰/۳۲	۱/۲۱	۰/۹۲
	بازتاب	٥	۰/۸۳	۰/۹۹۳	۰/٤۳٦
	مشتق اول بازتاب	**۹**	**۰/۲٦۳**	**۱/۰۳٤**	**۰/۹۳۸**
	مشتق دوم بازتاب	۳	۰/٥۳۹	۱/۰۸٥	۰/۷٦
BRT	Log (1/R)	-	۰/۳۰٦	۱/۱٥۰	۰/۹۲۳
	مشتق اول log (1/R)	-	۰/۳٦۰	۱/۰۳۹	۰/۸۹۳
	مشتق دوم log (1/R)	-	۰/۲۷۲	۰/۷۳۳	۰/۹۳۹
	طیف اصلی	-	۰/۳۰۲	۰/۹٤۸	۰/۹۲٥
	مشتق اول طیف اصلی	-	۰/۲۷٤	۰/۷٦۱	۰/۹۳۸
	مشتق دوم طیف اصلی	-	**۰/۲۷۱**	**۰/٥۷۹**	**۰/۹۳۹**

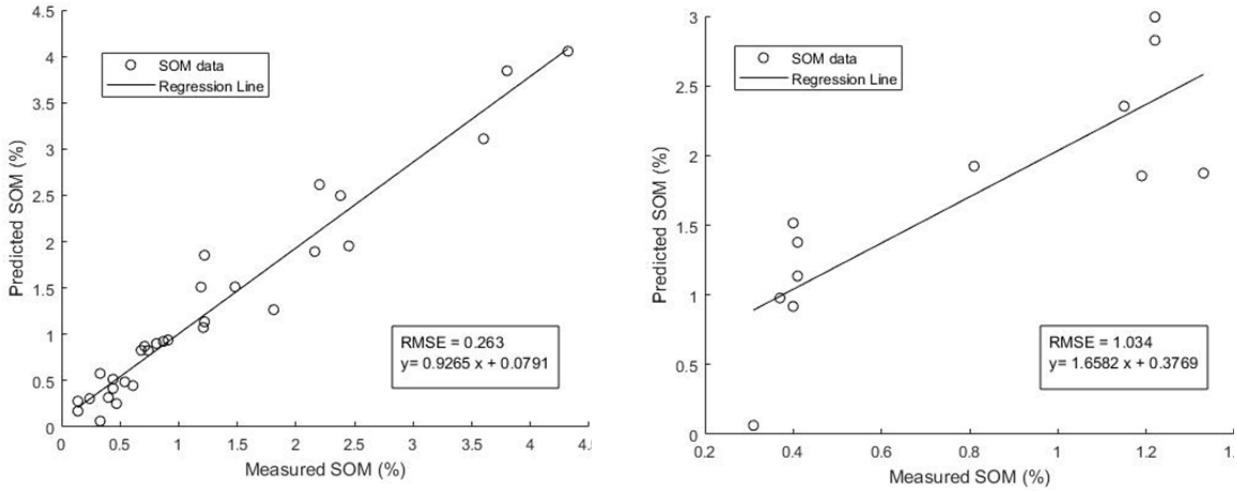


الف ب

شکل۸. نمودار مقادیر اندازه‌گیری شده در مقابل پیش‌بینی شده با استفاده از داده مشتق دوم بازتاب طیفی به روش BRT

الف) داده کالیبراسیون ب) داده اعتبارسنجی

Fig. 8. Histogram of measured vs. predicted values using the data related to the SDR via the BRT model
a) Calibration data b) validation data

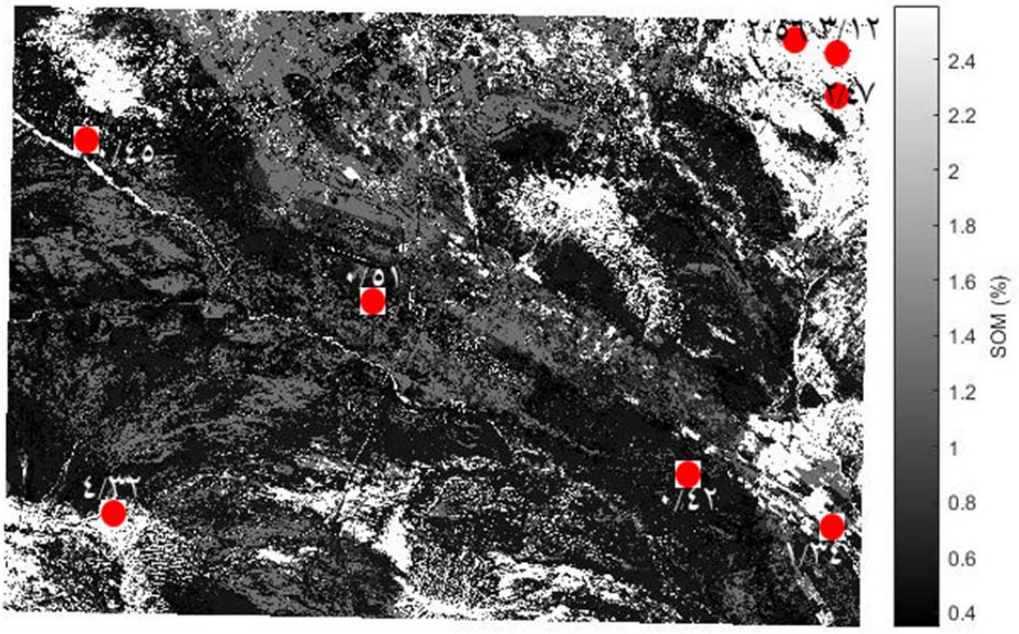


الف ب

شکل۹. نمودار مقادیر اندازه‌گیری شده در مقابل پیش‌بینی شده با استفاده از داده مشتق اول طیف اصلی به روش PLSR

الف) داده کالیبراسیون ب) داده اعتبارسنجی

Fig. 9. Histogram measured vs. predicted values using the data related to the SDR via the PLSR model.
a) Calibration data b) validation data



شکل۱۰. نقشه مقادیر برآورد شده ماده آلی خاک با استفاده از داده مشتق دوم بازتاب طیفی به روش BRT به همراه تعدادی نمونه خاک اندازه‌گیری شده

Fig. 10. Map of predicted values of soil organic matter using the data related to the SDR via the BRT model with some soil sampling.

بحث و نتیجه‌گیری

نتایج اندازه‌گیری آزمایشگاهی نشان داد نمونه‌های خاک در چهار کلاس بافتی رس، لوم ماسه‌دار، لوم رس‌دار و لوم رس و لای‌دار قرار می‌گیرند. کمترین و بیشترین میزان کربنات کلسیم نمونه‌های خاک به ترتیب ۱۳ و ۲۳ درصد است. کمترین و بیشترین میزان اسیدیته نمونه‌های خاک به ترتیب ۷/۲٤ و ۷/۹٤ است. کمترین و بیشترین میزان ماده آلی خاک به ترتیب ۰/۱٤ و ٤/۳۲ درصد است. اما این داده‌ها به‌صورت پراکنده است، برآورد رضایت‌بخش میزان SOM، ایجاد سطوح پیوسته با دقت بیشتر بر اساس کاهش نویز و حفظ داده‌های مفید، همواره موردتوجه محققین بوده است. در این تحقیق نیز با استفاده از داده‌های طیف‌سنجی خاک و اندازه‌گیری آزمایشگاهی میزان مواد آلی، سعی در برآورد چنین سطح پیوسته‌ای به‌منظور تخمین SOM بوده است. به‌طوری‌که با استفاده از تبدیل موجک و حذف داده‌های پرت براساس هادلینگز در روش PCA، داده‌های مفید برای تولید سطح پیوسته استخراج‌شده‌اند. در این روش، باندها یا طیف‌های مستقل و مؤثر در مدل باقی می‌مانند. درحالی‌که، لین و همکاران (۱۳) به‌منظور انتخاب باندهای مناسب در تخمین مواد آلی خاک از روش تبدیل موجک و همبستگی استفاده نموده‌اند. این در حالی است که با استفاده از روش همبستگی در مناطق ناهمگن همانند منطقه موردمطالعه در این تحقیق، نتایج رضایت بخشی به دست نمی‌آید. روش PCA به‌طور غیر نظارت‌شده، با در نظر گرفتن مقادیر داده، اجزای اصلی و مقادیر و بردارهای ویژه را محاسبه نموده و سعی در بیشینه نمودن ماتریس کوواریانس بر اساس تجزیه مقادیر منفرد (Singular Value Decomposition, SVD) دارد.

بعد از حذف داده‌های پرت و باقی‌مانده داده‌های مفید، مدل‌های تخمین مواد آلی خاک به دو روش PLSR و BRT، بر روی‌داده‌های اندازه‌گیری شده در جنوب غربی تهران توسعه داده‌شده است. به‌منظور بارز نمودن اطلاعات و ویژگی‌ها در طیف‌سنجی خاک از تبدیلات مختلف طیف نظیر جذب، مشتق اول و مشتق دوم برای بازتابندگی و جذب طیفی نیز قبل از اجرای تبدیل موجک و حذف داده‌های پرت استفاده‌شده است. درنتیجه این دو مدل تولید سطوح پیوسته، برای این شش دسته ویژگی، اجراشده است تا مدل با دقت بهتر برآورد شود. بررسی نتایج به‌دست‌آمده از توسعه این دو مدل حاکی از این است که مدل BRT، با مقادیر RMSE و R^2، به ترتیب ۰/۵۸ و ۰/۹٤، در داده مشتق دوم طیف اصلی، نتایج بهتری را به دست آورده است. از طرفی، مقادیر RMSE و R^2 در مدل PLSR برای داده مشتق اول طیف اصلی، به ترتیب ۱/۰۳۳۸ و ۰/۹۳۸ به‌دست‌آمده است. این در حالی است که به‌طورکلی مقایسه نتایج کالیبراسیون و اعتبارسنجی مدل BRT و PLSR، دلالت بر نتایج بهتر مدل BRT در این منطقه دارد. مقایسه با تحقیقات مشابه دلالت بر این دارد که مرولوس و همکاران (۱۷)، با استفاده از دو روش PLSR و رگرسیون درختی، در تخمین داده SOM، به مقادیر R^2 به ترتیب ۰/۷۱۱ و ۰/۷۸۵۸ دست‌یافته‌اند. علاوه بر این، ویسکارا راسل و بهرنس (۲٤)، به مقایسه PLSR و روش‌های داده‌کاوی BRT، و سایر روش‌ها پرداخته‌اند و به کارایی بالاتر روش‌های داده‌کاوی اشاره‌کرده‌اند. آن‌ها میزان مواد آلی خاک را با R^2، ۰/۸۱ و ۰/۸۳ به ترتیب در PLSR و BRT برآورد نموده‌اند. لئو و همکاران (۱٤) نیز میزان مواد آلی خاک را با استفاده از داده طیف‌سنجی و نمونه مواد آلی، به روش BRT، با R^2، ۰/۸۵ تخمین زده‌اند. ناوار و همکاران (۱۹) در برآورد ماده آلی خاک با استفاده از روش یادگیری ماشین Cubist در سطوح اندازه‌گیری آزمایشگاهی و مزرعه در نمونه‌های مرطوب خاک، به مقادیر R^2 به ترتیب ۰/۸۹ و ۰/۷٦ دست یافتند. میرزایی و همکاران (۱٦) با استفاده از روش PLSR در نمونه‌های مرطوب خاک برداشت‌شده از مناطق کشاورزی استان‌های تهران و لرستان، ماده آلی خاک را با R^2 معادل ۰/۵۹ برآورد نمودند.

درنهایت، برای ایجاد سطح پیوسته و آگاهی از نحوه تغییر مواد آلی خاک در منطقه، نقشه مواد آلی خاک با استفاده از تصویر ماهواره لندست OLI و روش BRT تولید شد. نتایج این تحقیق مؤید این مطلب است که در مناطق ناهمگن کشاورزی – شهری، می‌توان از پتانسیل مدل‌های توسعه

داده‌شده تحقیق حاضر تحت عنوان Wavelet-PCA-PLSR و Wavelet-PCA-BRT برای تخمین مواد آلی خاک استفاده نمود. چراکه اندازه‌گیری میدانی ویژگی‌های شیمیایی خاک نظیر مواد آلی بسیار زمان و هزینه‌بر است. علاوه بر این، امکان اندازه‌گیری این ویژگی‌ها در پوشش وسیع وجود ندارد. با استفاده از این توابع پیوسته و تصویر ماهواره‌ای می‌توان، نقشه مقادیر مواد آلی خاک را در پوشش وسیع تولید نمود تا از آن بتوان در مطالعاتی نظیر پتانسیل کشت، حاصلخیزی خاک و توسعه پایدار آن بهره‌برداری نمود.

تقدیر و تشکر

نویسندگان بدین‌وسیله مراتب سپاس و قدردانی خود را از گروه خاکشناسی دانشگاه تربیت مدرس، به دلیل در اختیار قرار دادن طیف‌سنج برای انجام این پژوهش، ابراز می‌دارند.

References

1. Alavipanah S.K, Damavandi A.A, Mirzaie S, Rezaie A, Matinfar H.R, Hamzeh S, Teymori H, Javad Zarrin I. 2016. Remote sensing application in evaluation of soil characteristics in desert areas. Natural Environment Change, 2(1): 1-24.
2. Attaeian B, Shojaeefar S, Zandieh V, Hashemi S.S. 2018. Study of soil organic carbon changes in two critical and vulnerable areas of Qahavand plain rangelands using remote sensing and GIS. RS & GIS for Natural Resources, 8(4): 76-90. (In Persian).
3. Dai F, Zhou Q, Lv Z, Wang X, Liu G. 2014. Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau. Ecological Indicators, 45: 184-194. doi: https://doi.org/10.1016/j.ecolind.2014.04.003.
4. Doetterl S, Stevens A, Van Oost K, Quine T.A, Van Wesemael B. 2013. Spatially-explicit regional-scale prediction of soil organic carbon stocks in cropland using environmental variables and mixed model approaches. Geoderma, 204: 31-42. doi:https://doi.org/10.1016/j.geoderma.2013.04.007
5. Friedman J.H. 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics: 1189-1232. doi:https://doi.org/10.1214/aos/1013203451.
6. Castaldi F, Palombo A, Pascucci S, Pignatti S, Santini F, Casa R. 2015. Reducing the Influence of Soil Moisture on the Estimation of Clay from Hyperspectral Data: A Case Study Using Simulated PRISMA Data. Remote Sensing, 7(11): 15561-15582. https://doi.org/10.3390/rs71115561.
7. Groenigen J.W, Mutters C.S, Horwath W.R, Van Kessel C. 2003. NIR and DRIFT-MIR spectrometry of soils for predicting soil and crop parameters in a flooded field. Plant and Soil, 250(1): 155-165. doi:https://doi.org/10.1023/A:1022893520315.
8. Hotelling H. 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24(6): 417-441. doi: 10.1037/h0071325.
9. Khanamani A, Jafari R, Sangoony H, Shahbazi A. 2011. Evaluation of soil status using RS and GIS technology (Case study: Segzi plain). Journal of Applied RS & GIS Techniques in Natural Resource Science, 2(3): 25-37. https://www.sid.ir/en/journal/ViewPaper.aspx?id=250690. (In Persian).
10. Kuang B, Tekin Y, Mouazen A.M. 2015. Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon, pH and clay content. Soil and Tillage Research, 146: 243-252. doi:https://doi.org/10.1016/j.still.2014.11.00.
11. Lacoste M, Minasny B, McBratney A, Michot D, Viaud V, Walter C. 2014. High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. Geoderma, 213: 296-311. doi:https://doi.org/10.1016/j.geoderma.2013.07.002.
12. Liaghat S, Ehsani R, Mansor S, Shafri H.Z, Meon S, Sankaran S, Azam S.H. 2014. Early detection of basal stem rot disease (Ganoderma) in oil palms based on hyperspectral reflectance data using pattern recognition algorithms. International Journal of Remote Sensing, 35(10): 3427-3439. doi:https://doi.org/10.1080/01431161.2014.903353.
13. Lin L, Wang Y, Teng J, Wang X. 2016. Hyperspectral analysis of soil organic matter in coal mining regions using wavelets, correlations, and partial least squares regression. Environmental Monitoring and Assessment, 188(2): 1-11. doi:https://doi.org/10.1007/s10661-016-5107-8.
14. Liu L, Ji, M, Dong Y, Zhang R, Buchroithner M. 2016. Quantitative retrieval of organic soil properties from Visible Near-Infrared Shortwave Infrared (Vis-NIR-SWIR) spectroscopy using

fractal-based feature extraction. Remote Sensing, 8(12): 1035. doi:https://doi.org/10.3390/rs8121035.
15. McCarty G.W, Reeves J.B, Reeves V.B, Follett R.F, Kimble J.M. 2002. Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement. Soil Science Society of America Journal, 66(2): 640-646. doi:https://doi.org/10.1016/j.geoderma.2009.04.005.
16. Mirzaei S, Darvishi Boloorani A, Bahrami H.A, Alavipanah, S.K, Mousivand A. 2021. Moisture influence reducing on soil reflectance using EPO for organic carbon prediction. 7th International Conference on Agriculture, Environment, Urban and Rural. Tbilisi, Georgia. 16 June. https://civilica.com/doc/1256685. (In Persian).
17. Morellos A, Pantazi X.E, Moshou, D, Alexandridis T, Whetton R, Tziotzios G, Wiebensohn J, Bill R, Mouazen A.M. 2016. Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. Biosystems Engineering. doi:https://doi.org/10.1016/j.biosystemseng.2016.04.018.
18. Mouazen A.M, Kuang B, De Baerdemaeker J, Ramon H. 2010. Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. Geoderma, 158(1): 23-31. doi:https://doi.org/10.1016/j.geoderma.2010.03.001.
19. Nawar S, Abdul Munnaf M, Mouazen A.M. 2020. Machine learning based on-line prediction of soil organic carbon after removal of soil moisture effect. Remote Sensing, 12(8): 1308. https://doi.org/10.3390/rs12081308.
20. Nocita M, Kooistra L, Bachmann M, Müller A, Powell M, Weel S. 2011. Predictions of soil surface and topsoil organic carbon content through the use of laboratory and field spectroscopy in the Albany Thicket Biome of Eastern Cape Province of South Africa. Geoderma, 167: 295-302. doi:https://doi.org/10.1016/j.geoderma.2011.09.018.
21. Ghazi M, Bahrami H.A, Darvishi Boloorani A, Mirzaei S. 2018. Estimating the measure of the soil's lime in dust's centers of Tehran province by using of VINR spectroscopy and satellite images of OLI. RS & GIS for Natural Resources, 8(4): 1-16, https://www.sid.ir/en/journal/ViewPaper.aspx?id=597225 (In Persian).
22. Steffens M, Kohlpaintner M, Buddenbaum H. 2014. Fine spatial resolution mapping of soil organic matter quality in a Histosol profile. European Journal of Soil Science, 65(6): 827-839. doi: https://doi.org/10.1111/ejss.12182.
23. Tekin Y, Kuang B, Mouazen A.M. 2013. Potential of on-line visible and near infrared spectroscopy for measurement of pH for deriving variable rate lime recommendations. Sensors, 13(8): 10177-10190. doi:https://doi.org/10.3390/s130810177.
24. Viscarra Rossel R.A, Behrens, T. 2010. Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, 158(1): 46-54. doi:https://doi.org/10.1016/j.geoderma.2009.12.025.
25. Viscarra Rossel R.A, Hicks W.S. 2015. Soil organic carbon and its fractions estimated by visible–near infrared transfer functions. European Journal of Soil Science, 66(3): 438-450. doi:https://doi.org/10.1111/ejss.12237.
26. Viscarra Rossel R.A, Cattle S.R, Ortega A, Fouad Y. 2009. In situ measurements of soil colour, mineral composition and clay content by vis–NIR spectroscopy. Geoderma, 150(3): 253-266. doi:https://doi.org/10.1016/j.geoderma.2009.01.025.
27. Vohland M, Besold J, Hill J, Fründ H.C. 2011. Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. Geoderma, 166(1): 198-205. doi:https://doi.org/10.1016/j.geoderma.2011.08.001.
28. Wang Y, Wang F, Huang J, Wang X, Liu Z. 2009. Validation of artificial neural network techniques in the estimation of nitrogen concentration in rape using canopy hyperspectral reflectance data. International Journal of Remote Sensing, 30(17): 4493-4505. doi:https://doi.org/10.1080/01431160802577998.
29. Yang H, Li J. 2013. Predictions of soil organic carbon using laboratory-based hyperspectral data in the northern Tianshan Mountains, China. Environmental Monitoring and Assessment, 185(5): 3897-3908. doi:https://doi.org/10.1007/s10661-012-2838-z.
30. Yang R.M, Zhang G.L, Liu F, Lu Y.Y, Yang F, Yang F, Yang M, Zhao Y.G, Li D.C. 2016. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. Ecological Indicators, 60: 870-878. doi:https://doi.org/10.1016/j.ecolind.2015.08.036.

RS & GIS for Natural Resources
(Vol. 13/ Issue 3) Autumn 2022
Indexed by ISC, SID, Magiran, Noormags, Civilica, Google Scholar
journal homepage : www.girs.iaubushehr.ac.ir

Original paper

Integrated noise reduction-data mining method for soil organic matter prediction by VNIR spectrometry

Elahe Akbari, Saham Mirzaei, Ara Toomanian, Ali Darvishi Boloorani, Hossein Ali Bahrami

Received: 23 August 2021 / Received in revised form 3 September 2021 / Accepted: 4 September 2021
Available online 4 September 2021 / Available print 23 September 2022

Abstract

Background and Objective Soil as a heterogeneous natural resource and the largest organic carbon storage in terrestrial ecosystems is composed of complicated processes and mechanisms. The necessity of accurately estimating soil properties on the national and regional scales for improving soil management, and understanding their influence on agriculture have resulted in attracting researchers' attentions to this field. Soil Organic Matter (SOM) is considered as an indicator of soil quality in fertility and food production. It is also considered as a key variable in environmental and agricultural issues. Thus, using rapid and cost effective and more accuracy estimation of the SOM content in soil resources assessment and management can be helpful.

E. Akbari(✉)[1], S. Mirzaei [2], A. Toomanian [3], A. Darvishi Boloorani [3], H. A. Bahrami [4]

1. Assistant Professor, Department of Remote Sensing and GIS, Faculty of Geography and Environmental Sciences, Hakim Sabzevari University, Sabzevar, Iran
2. PhD. Student of Remote Sensing and Geographical Information System, Faculty of Geography, University of Tehran, Iran
3. Associate Professor, Department of Remote Sensing and Geographical Information System, Faculty of Geography, University of Tehran, Iran
4. Professor, Department of Soil Science, Faculty of Agriculture, Tarbiat Modares University, Iran

e-mail: e.akbari@hsu.ac.ir

https://doi.org/10.30495/GIRS.2022.1938557.1935
http://dorl.net/dor/20.1001.1.26767082.1401.13.3.1.3

In precision agriculture, the scale of soil data required for management of lands and products is very large. The scale of collecting filed data usually cannot fulfil those needs. Sampling, preparing and analyzing the large number of soil samples as well as producing the distribution map for large areas are very difficult. In addition, traditional laboratory methods of soil analysis are boring, time-consuming, and costly. In fact, they need specialized laboratory operators. The aim of the present study is to compare the performance of the two Partial Least Squares Regression (PLSR) and Boosted Regression Tree (BRT) for predicting SOM using VNIR spectrometry data. With the use of combining Wavelet transform and diagnosis of independent bands, noises existing in soil spectroscopic data has reduced. In addition, independent and effective spectra and bands in spectroscopy of SOM were selected. Consequently, in the present research, Wavelet-PCA-PLSR and Wavelet-PCA- BRT models were developed and performance were assessed.

Materials and Methods 42 surface (0-30cm) soil samples in the heterogeneous areas of urban-agricultural regions in Tehran province were collected. Soil Organic Carbon (OC) measured using Walki Black method and the samples' spectrums were measured by ASD FieldSpec-3 spectrometer. First and second derivitation of spectral reflectance and absorbance were calculated.

To reduce noises and smooth the spectrum, Sym8 matrix function of wavelet transform was used, wavelet transform is conducted to show and reconstruct characteristics in the spectrum. Principal component analysis and Hotelling's T^2 test with 95% confidence level were used for outlier detection. PLSR and BRT was conducted onreflectance, absorbance and their first and second derivatives, at five levels of wavelet transform. Then, by comparing the results, the appropriate model was selected via validation. For doing the PLSR in nonlinear data, Kernel functions were used. When using numerical samples, regression trees are used instead of decision trees. But their processes are the same. In regression trees, the greedy algorithm was used. Therefore, by answering the binary question through which node the maximum data about respons variable is obtained, the root node and its two children are obtained. Producing the structure of trees is recursively repeated and a typical stopping criterion is considered. The stopping criterion can be as achievement to a split which cannot be divided and provides fewer data, or when data in the node contain 5% of the total data. Moreover, the tree size should be minimized. For splitting the node, the Ginny factor, entropy factor, etc. were used for minimizing those factors. In addition, the total square error is calculated in each branches and those with minimized values are selected. In addition, in the regression tree, the pruning process is employed for over-fitting. The BRT consists of the two regression tree and boosting techniques for improving the predictability of each of them. For calibration and validation of the model, 30 and 12 soil samples were randomly selected, respectively and R^2 and RMSE were used for quantify the accuracy of models. Moreover, to select the best production factor of the PLSR mode, explained variance residual values and RMSE of validation were considered. Finally, soil organic matter map was produced using Landsat OLI satellite imagery and the proofed method for the study area.

Results and Discussion The SOM value acceptably, the creation of continuous mappings with more accuracy based on noise reduction and retention of suitable data have always received researchers' attentions. The present study tried to find the better method such a more accurate quantization of SOM using soil spectroscopic data. Using wavelet transform and outlier removal based on Hotelling's T^2 via the PCA, the suitable data were extracted for producing the more accurate quantization. In this method, independent and effective bands or spectra remain in the model, while Lin et al. used wavelet transform and correlation techniques for selecting appropriate bands in estimating SOM. Since the soil reflectance is more complex and affected by several factors, using correlation method in these heterogeneous areas such as the area studied in the present study does not lead to acceptable results. Considering the data values, the unsupervised PCA method calculates principle components and eigenvalues and eigenvectors. It also tries to maximize the covariance matrix based on Singular Value Decomposition (SVD). SOM estimation models were developed using the PLSR and BRT for reflectance and absurbance spectra and their first and second derivation. Based on the results, the BRT method with RMSE and R^2 values as 0.58 and 0.94, respectively leads in the better results for the data of the second derivation of reflectance. Moreover, values of RMSE and R^2 in the PLSR were obtained as 1.0338 and 0.938, respectively for the data related to the second derivation of reflectance. However, comparing RMSE of the BRT and PLSR shows better results of the BRT model.

Conclusion In that field measurements of chemical properties of soil such as organic matters are critically time-consuming and costly. Furthermore, measuring those properties is not possible in the large samples. So, the results of the present study indicate that in heterogeneous agricultural-urban areas, potential of the developed models such as wavelet-PCA-PLSR and wavelet-PCA-BRT can be used for estimating SOM. Meanwhile, these two algorithms do not make distributional assumptions and therefore, there are no strong assumptions about normality. Using continuous functions and satellite imagery, the map of the level of SOM in large scales can be prepared in order that it can be utilized in studies such as cultivation potential, soil fertility, and sustainable development of soil.

Keywords: Spectroscopy, Soil organic matter, Partial Least Squares Regression (PLSR), Boosted Regression Tree (BRT), Southwest of Tehran

Please cite this article as: Please cite this article as: Akbari E, Mirzaei S, Toomanian A, Darvishi Boloorani, A, Bahrami, H. A. 2022. Integrated noise reduction-data mining method for soil organic matter prediction by VNIR spectrometry. Journal of RS and GIS for Natural Resources, 13(3): 1-21. https://doi.org/10.30495/GIRS.2022.1938557.1935