سنجش از دور و سامانه اطلاعات جغرافیایی در منابع طبیعی (سال دوازدهم / شماره اول) بهار ۱۴۰۰
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

شناسایی خاستگاه‌های هواویزهای اتمسفری با استفاده از سنجش از دور و داده‌کاوی (مطالعه موردی: استان یزد)

محمد کاظمی، علیرضا نفرزادگان، فریبرز محمدی، علی رضایی لطیفی

دریافت: ۲۳ دی ۱۳۹۸/ پذیرش: ۲۹ شهریور ۱۳۹۹
دسترسی اینترنتی: ۱ فروردین ۱۴۰۰

چکیده

پیشینه و هدف کشور ایران به دلیل قرار گرفتن در کمربند خشک و نیمه‌خشک جهان، در معرض پدیده‌های محلی و منطقه‌ای گردوغبار قرار دارد. میانگین روزهای توأم با گردوغبار در استان یزد بالغبر ۴۳ روز در سال است و این مهم به نحوی بر سلامت و کیفیت زندگی مردم اثرات مخربی وارد آورده است. میزان غلظت ذرات معلق و شاخص عمق اُپتیکی هواویز (AOD) در پی وقایع گردوغبار یکی از شاخص‌های کیفیت هوا است. بنابراین بررسی و تهیه نقشه‌های پهنه‌بندی حساسیت باهدف شناسایی مناطق دارای قابلیت بالای تولید گردوغبار، در محدودهٔ فعالیت‌های بشری دارای اهمیت است و جهت کاهش خسارات احتمالی و مدیریت خطر، اقداماتی مانند پهنه‌بندی عرصه‌های مختلف تولید گردوغبار می‌تواند مؤثر واقع شود. هدف از پژوهش حاضر پهنه‌بندی پتانسیل عرصه‌های مختلف مستعد گردوغبار با استفاده از مدل‌های داده‌کاوی و شناسایی مهم‌ترین متغیرها بر این پدیده و بهره‌مندی از سنجش‌ازدور در این راستا در استان یزد است.

مواد و روش‌ها در این تحقیق ابتدا متغیرهای اقلیمی مختلف (از تصاویر ماهواره‌ای مختلف) ازجمله سرعت باد در ارتفاع ده متری سطح زمین (Vs)، رطوبت خاک (Soil)، بارش تجمعی (Pr)، شاخص خشک‌سالی پالمر (Pdsi)، شاخص پوشش گیاهی نرمال شده (NDVI)، خشکی خاک یا کمبود آب خاک (Def)، تبخیر و تعرق مرجع (Pet) و واقعی (Aet)، بعد توپوگرافی (TD)، رادیانس طول‌موج کوتاه رسیده به زمین (Srad)، حداقل دمای هوا (Tmmn)، حداکثر دمای هوا (Tmmx)، فشار بخار (Vap)، کمبود فشار بخار(Vpd) و درصد رس (Clay) با استفاده از کدنویسی در سامانه آنلاین گوگل ارث انجین (GEE) استخراج شدند. سپس نمونه‌ها از مناطق بحرانی و مستعد گردوغبار در سیستم اطلاعات جغرافیایی و به کمک تصاویر AOD مودیس استخراج شدند و این ویژگی و همچنین سایر ویژگی‌ها در متغیرهای اقلیمی وارد سه مدل داده‌کاوی الگوریتم درختان رگرسیون و طبقه‌بندی (CART)، رگرسیون انطباقی چندمتغیره اسپیلاین (MARS) و درختان رگرسیون چندگانه

محمدکاظمی [۱]، علیرضا نفرزادگان(✉) [۲]، فریبرز محمدی [۱،۳]، علی‌رضایی لطیفی [۱،۴]

۱. استادیار مرکز مطالعات و تحقیقات هرمز، دانشگاه هرمزگان، بندرعباس، ایران
۲. استادیار گروه مهندسی منابع طبیعی، دانشکده کشاورزی و منابع طبیعی، دانشگاه هرمزگان، بندرعباس، ایران
۳. استادیار گروه علوم و مهندسی آب، مجتمع آموزش عالی میناب، دانشگاه هرمزگان، میناب، ایران
۴. استادیار گروه فیزیک، دانشکده علوم، دانشگاه هرمزگان، بندرعباس، ایران

پست الکترونیکی مسئول مکاتبات: a.r.nafarzadegan@gmail.com

http://dorl.net/dor/20.1001.1.26767082.1400.12.1.4.5

جلد ۱۲، شماره ۱، بهار (۱٤۰۰): ۸۵-٦۳.

جمع‌شدنی (TreeNet) شدند. درنهایت نتایج پیش‌بینی این مدل‌های داده‌کاوی در سیستم اطلاعات جغرافیایی تبدیل به نقشه و پهنه‌های مختلف پتانسیل خطر خیزش گردوغبار شدند.

نتایج و بحث درروش CART متغیرهایی همچون شاخص پوشش گیاهی نرمال‌شده، تبخیر و تعرق واقعی، مدل رقومی ارتفاع، طول‌موج کوتاه رسیده به سطح زمین، شاخص خشک‌سالی پالمر، سرعت باد و درصد رس، گره‌های انتهایی جهت شناسایی مناطق با میانگین بالای عمق اپتیکی هواویزها است. در این روش رطوبت خاک، مدل رقومی ارتفاعی و تبخیر تعرق رفرنس بیشترین اهمیت نسبی را در شناسایی مناطق بحرانی خیزش گردوغبار نشان دادند. ضریب همبستگی مدل مقدار ۰/۸۵ را نشان داد. نتایج داده‌کاوی به روش MARS نشان داد متغیرهای تبخیر و تعرق واقعی، رطوبت خاک و شاخص خشک‌سالی پالمر بیشترین اهمیت نسبی را در شناسایی مناطق بحرانی خیزش گردوغبار داشته‌اند. ضریب همبستگی مدل مقدار ۰/۷۲ را نشان داد. همچنین درروش TreeNet متغیرهای رطوبت خاک، شاخص خشک‌سالی پالمر و تبخیر و تعرق واقعی بیشترین اهمیت نسبی را نشان دادند. ضریب همبستگی مدل ۰/۷۵ بود. همچنین مناطق با حساسیت بسیار زیاد ، زیاد ، متوسط ، کم و بسیار کم به ترتیب حدود ۱٦٪ ، ۱۹٪ ، ۲٦٪ ، ۲۰٪ و ۲۰٪، استان یزد را اشغال کردند.

نتیجه‌گیری با توجه به نتایج یادشده در مورد شناسایی تأثیرگذارترین متغیرها بر گردوغبار در مناطق مختلف، نمی‌توان یک یا چند متغیر را در پدیده خیزش گردوغبار برای همه مناطق، مشترک در نظر گرفت و این مهم از منطقه به منطقه‌ای دیگر تغییر می‌کند. کما اینکه متغیرهای زمین‌شناسی و کاربری اراضی در پژوهش حاضر جزء متغیرهایی بودند که هیچ‌گونه اثری بر متغیر وابسته یعنی حساسیت به گردوغبار نداشتند. در پژوهش حاضر، اشتراکات متغیرهای مستقل مهم و چرخه تصمیم‌گیری شامل تبخیر و تعرق واقعی، رطوبت خاک، شاخص خشک‌سالی پالمر، سرعت باد، ارتفاع، شاخص پوشش گیاهی و حداقل دمای روزانه بودند. هیچ‌کدام از پژوهش‌های مرتبط در مورد موضوع پژوهش، در انتخاب بهترین مدل داده‌کاوی، همپوشانی نداشتند و مدل داده‌کاوی واحدی برای بررسی حساسیت مناطق مختلف به پدیده گردوغبار در ایران یافت نشد. شایان‌ذکر است، در این پژوهش مدل الگوریتم درختان رگرسیون و طبقه‌بندی انتخاب شد. پژوهش حاضر در نوع مدل‌های داده‌کاوی استفاده‌شده و متغیرهای مستقل با پژوهش‌های یادشده متفاوت بوده و با توجه به عدم همپوشانی نتایج انتخاب مدل برتر، نمی‌توان نسخه واحدی برای انتخاب بهترین مدل داده‌کاوی برای ایران در بحث گردوغبار ارائه نمود. لذا پیشنهاد می‌شود از بهترین مدل‌های منتخب در پژوهش‌های یادشده برای داده‌کاوی پدیده گردوغبار در پژوهش‌های آتی استفاده و موردقیاس قرار گیرند.

واژه‌های کلیدی: عمق اُپتیکی هواویز، متغیرهای مکانی، یادگیری ماشین، پهنه‌بندی

لطفاً به این مقاله استناد کنید: کاظمی، م.، نفر زادگان، ع. ر.، محمدی، ف.، رضایی لطفی، ع. ۱٤۰۰. شناسایی خاستگاه‌های هواویزهای اتمسفری با استفاده از سنجش از دور و داده‌کاوی (مطالعه موردی: استان یزد). نشریه سنجش‌ازدور و سامانه اطلاعات جغرافیایی در منابع طبیعی، ۱۲(۱): ۸۵-٦۳.

مقدمه

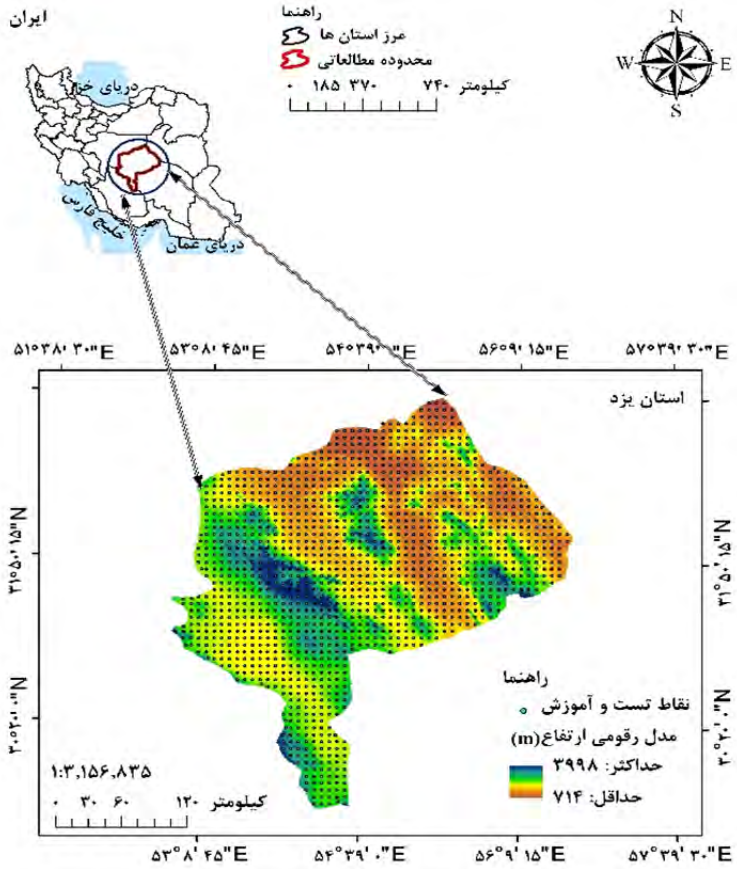
خاورمیانه یکی از پنج منطقه مهم جهان در تولید گردوغبار است (۲٤). ایران به‌دلیل قرار گرفتن در مناطق خشک و نیمه‌خشک جهان در معرض سیستم‌های متعدد گردوغبار محلی و فرامنطقه‌ای است (۲۳). طوفان‌های گردوغبار علاوه بر پوشانده شدن اراضی زراعی و گیاهان به‌وسیله مواد بادآورنده، باعث نابودی اراضی حاصل‌خیز و کاهش تولید بیولوژیک و تنوع زیستی، می‌شود و ماندگاری ساکنان را به‌شدت تحت تأثیر قرار می‌دهد. طوفان‌های گردوغبار در انتقال عوامل بیماری‌زای خطرناک به انسان، آلودگی آب‌وهوا و آسیب رساندن به عملکرد دستگاه تنفسی نقش دارد (۱۵). طوفان‌های گردوغبار در استان یزد، امری عادی است و میانگین روزهای توأم با طوفان گردوخاک در استان به ٤۳ روز در سال می‌رسد (۱٦). این پدیده سبب مشکلات بسیاری برای مردم استان شده است، به‌عنوان نمونه در خردادماه ۱۳۸۲ طوفان گردوغبار با سرعت ۲٦/٤ متر بر ثانیه شهر یزد را در نوردید و دید افقی به صفر رسید. در ساعات بعد از طوفان دمای هوا ۱٦ درجه سانتی‌گراد اُفت داشت و بیش از ۱۷٦٦۳ میلیون ریال در شهر یزد و بیش از ۱٦۰۹۱۱ میلیون ریال خسارت در کل استان واردشده است (٤). بنابراین بررسی و تهیه نقشه‌های پهنه‌بندی حساسیت باهدف شناسایی مناطق دارای قابلیت بالای تولید گردوغبار، در محدودهٔ فعالیت‌های بشری دارای اهمیت است و جهت کاهش خسارات احتمالی و مدیریت خطر، اقداماتی مانند پهنه‌بندی عرصه‌های مختلف تولید گردوغبار می‌تواند مؤثر واقع شود. تا به امروز استخراج متغیرهای مختلف از تصاویر ماهواره‌ای، با استفاده از نرم‌افزارهای مربوطه محاسبه می‌شد، که محاسبات هر یک از این پارامترها ماه‌ها به طول می‌انجامید. زیرا برای محاسبات تصاویر ماهواره‌ای، ابتدا باید تصحیح هندسی و اتمسفری انجام شود تا از درصد خطاها کاسته شود، سپس باید برای هر پارامتر، تعداد زیادی محاسبه صورت گیرد و تعداد زیادی تصویر با حجم زیاد تولید و طبقه‌بندی شود تا بتوان به یک تصویر نهایی دست پیدا کرد. این عملیات هم به زمان زیاد و هم به میزان زیاد حافظه ذخیره‌سازی بزرگی برای تصاویر محاسبه‌شده نیاز دارد (۳). اخیراً سامانه آنلاین موتور گوگل ارث این مشکل را برطرف نموده و حجم انبوهی از متغیرها را می‌توان با کد نویسی استخراج نمود. در همین راستا یکی از روش‌های پهنه‌بندی حساسیت، استفاده از روش‌های داده‌کاوی است. داده‌کاوی، استخراج دانش در پایگاه داده‌ها نامیده می‌شود و روشی برای کشف اطلاعات سودمند جدید و بالقوه از بین حجم انبوهی از اطلاعات است (۲۱). هم‌چنین از دیگر روش‌های مؤثر در شناسایی کانون‌های گردوغبار، استفاده از روش‌های سنجش‌ازدور است. در این خصوص با استفاده از تصاویر ماهواره مودیس مناطق برداشت گردوغبار و خصوصیات کاربری اراضی، پوشش گیاهی و خاک‌شناسی این کانون‌ها مشخص‌شده است. نتایج این تحقیقات نشان داده است که، پراکنش کانون‌های برداشت گردوغبار در خاک‌های حساس به فرسایش، اراضی دیم و مناطق با پوشش گیاهی ضعیف بوده است (۲۱ و ۲۲). دانش‌شهرکی و همکاران (٦) تغییرات فصلی و مکانی نرخ گردوغبار حمل شده از روی شهرهای دشت سیستان و ارتباط آن با برخی پارامترهای اقلیمی، بررسی‌شده است. نتایج پژوهش مذکور نشان داده است که مقدار میانگین نرخ گردوغبار حمل شده در دشت سیستان با سرعت باد، دمای هوا، دمای خاک در عمق ۵ سانتی‌متری و تبخیر و تعرق همبستگی مثبت و معنی‌داری دارد و با رطوبت نسبی همبستگی منفی و معنی‌داری را نشان داده است. هم‌چنین همبستگی بین بارندگی و میانگین نرخ گردوغبار حمل شده در سطح احتمال ۰/۰۵ معنی‌دار نبوده است. بروغنی و پورهاشمی (۵) که در گستره استان خراسان رضوی انجام‌شده است، با استفاده از تصاویر ماهواره‌ای مودیس در بازه زمانی ۲۰۰۵ تا ۲۰۱٦، پهنه‌بندی طوفان‌های گردوغبار انجام‌شده است. در پژوهش نامبردگان، از دو مدل وزن واقعه و نسبت فراوانی برای بررسی تأثیر متغیرهای خاک، شیب، شاخص پوشش گیاهی نرمال‌شده، لیتولوژی، فاصله از رودخانه و ژئومورفولوژی بر حساسیت منطقه به گردوغبار استفاده‌شده است. نتایج پژوهش مذکور نشان داده است که

متغیرهای ژئومورفولوژی، کاربری اراضی و شیب بیش‌ترین نقش را در وقوع گردوغبار در استان خراسان رضوی داشته است. مساحت مناطق با خطر زیاد و خیلی زیاد گردوغبار به‌ترتیب ٥٤/٩٥ ٪ و ٥٨/٢٣ ٪ از مساحت منطقه را به خود اختصاص داده‌اند. هم‌چنین با استفاده از دو روش داده‌کاوی جنگل تصادفی و رگرسیون لجستیک و متغیرهای نوع خاک، سنگ‌شناسی، شیب، اختلاف پوشش گیاهی نرمال شده، فاصله از رودخانه، واحدهای ژئومورفولوژی و کاربری اراضی به پهنه‌بندی خطر مناطق برداشت گردوغبار در استان خراسان رضوی بررسی‌شده است. با توجه به نتایج، دو متغیر شیب و کاربری اراضی را دارای بیش‌ترین اهمیت نسبی در ایجاد کانون‌های برداشت گردوغبار داشته است. غلامی و همکاران (۱۰) با استفاده از ۱۲ فاکتور اقلیمی، مشخصات خاک و سطح زمین و نیز هشت روش داده‌کاوی، نقشه‌های پتانسیل مکانی منشأ ریزگردهای استان خوزستان تهیه‌شده است. در این پژوهش از روش‌های داده‌کاوی و سیستم اطلاعات جغرافیایی به پهنه‌بندی پتانسیل خیزش گردوغبار استفاده‌شده است. نتایج نشان داده است که مدل داده‌کاوی ترکیبی EM، بیشترین دقت پیش‌بینی در شناسایی منشأ ریزگردها داشته است و از بین متغیرهای مستقل ورودی، سرعت باد دارای بیش‌ترین اهمیت نسبی در بروز پدیده گردوغبار استان خوزستان است. سبحانی و همکاران (۲۷) در تحقیقی با عنوان مدل‌سازی و پیش‌بینی گردوغبار در ایران از دو مدل شبکه عصبی ANFIS و RFB استفاده کردند. بازه زمانی تحقیق ایشان ۲۹سال و از داده‌های گردوغبار، دما و رطوبت نسبی ۲۸ ایستگاه زمینی درگیر گردوغبار شدید در ایران استفاده کردند. نتایج ارزیابی مدل‌ها با شاخص RMSE حکایت از برتری مدل RFB در مورد نتایج پیش‌بینی داشت. محمدخان (۱۹) به بررسی روند تغییرات طوفان‌های گردوغبار ایران از سال‌های ١٣٦٤ الی ١٣٨٤ پرداخت و بیان کرد که گردوغبار در ایران با پارامترهای تبخیر، دما و بارش همبستگی دارد و با ژئومورفولوژی و پارامتر ارتفاع دارای همبستگی نیست. ابراهیمی‌خوسفی و همکاران (۷) در تحقیقی با عنوان ارزیابی مدل‌های یادگیری ماشین برای پیش‌بینی تغییرات زمانی شاخص توفان گردوغبار در مناطق خشک ایران، در بازه زمانی ۲۰۰۰ تا ۲۰۱۸ به ارزیابی کاربرد نه مدل یادگیری ماشینی شامل رگرسیون چندمتغیره تطبیقی، انتخاب عملگر حداقل جمع‌شدگی مطلق، نزدیک‌ترین همسایه، الگوریتم ژنتیک، ماشین بردار پشتیبان، کوبیست، شبکه عصبی مصنوعی، افزایش گرادیان شدید و جنگل تصادفی پرداختند. مدل رگرسیون چندمتغیره اسپیلاین و پارامترهای اقلیمی شاخص پوشش گیاهی بارزسازی شده برای فصل بهار، حداکثر سرعت باد برای فصل تابستان، پاییز و فصول گردوغباری به‌ترتیب، به‌عنوان بهترین مدل پیش‌بینی داده‌کاوی و شاخص‌های اقلیمی مؤثر بر شاخص طوفان گردوغبار مناطق خشک ایران شناسایی شدند. در زمینه شناسایی کانون‌های برداشت گردوغبار و منشأ آن‌ها مطالعات متعددی در دنیا انجام‌شده است، اما پژوهش مستندی در خصوص پهنه‌بندی گردوغبار مناطق با استفاده از مدل‌های CART، MARS و TreeNet و مقایسه این سه مدل در ایران مشاهده نشده است. هدف از پژوهش حاضر پهنه‌بندی پتانسیل عرصه‌های مختلف مستعد گردوغبار با استفاده از مدل‌های داده‌کاوی، شناسایی مهم‌ترین متغیرها بر این پدیده در استان یزد و بهره‌مندی از سنجش‌ازدور و سیستم اطلاعات جغرافیایی در این خصوص است.

منطقه مورد مطالعه

ازنظر موقعیت جغرافیایی منطقه موردمطالعه در محدوده جغرافیایی بین ´١٥ °٣٠ تا ´٢٠ °٣٢ عرض شمالی و ´٣٨ °٥١ تا ´٣٩ °٥٧ طول شرقی واقع‌شده است (شکل ۱). اغلب مناطق استان یزد دارای اقلیم خشک و بیابانی تا فراخشک است. ازجمله عوامل خشکی آن، غالب بودن سیستم پرفشار جنب حاره، تبخیر و تعرق بالا است. به علت ناهنجاری‌های اقلیمی نظیر کاهش میزان بارش، افزایش دما و موقعیت جغرافیایی استان، تقریباً نیمی از مساحت آن را اراضی بیابانی پوشانده که همواره در معرض فرسایش بادی و توفان‌های گردوغبار قرار

می‌گیرند. آمار و اطلاعات هواشناسی نشان می‌دهد فراوانی پدیده گردوغبار ازجمله توفان‌های گردوغبار در منطقه مطالعاتی از مقدار بالایی برخوردار است (۱۸).

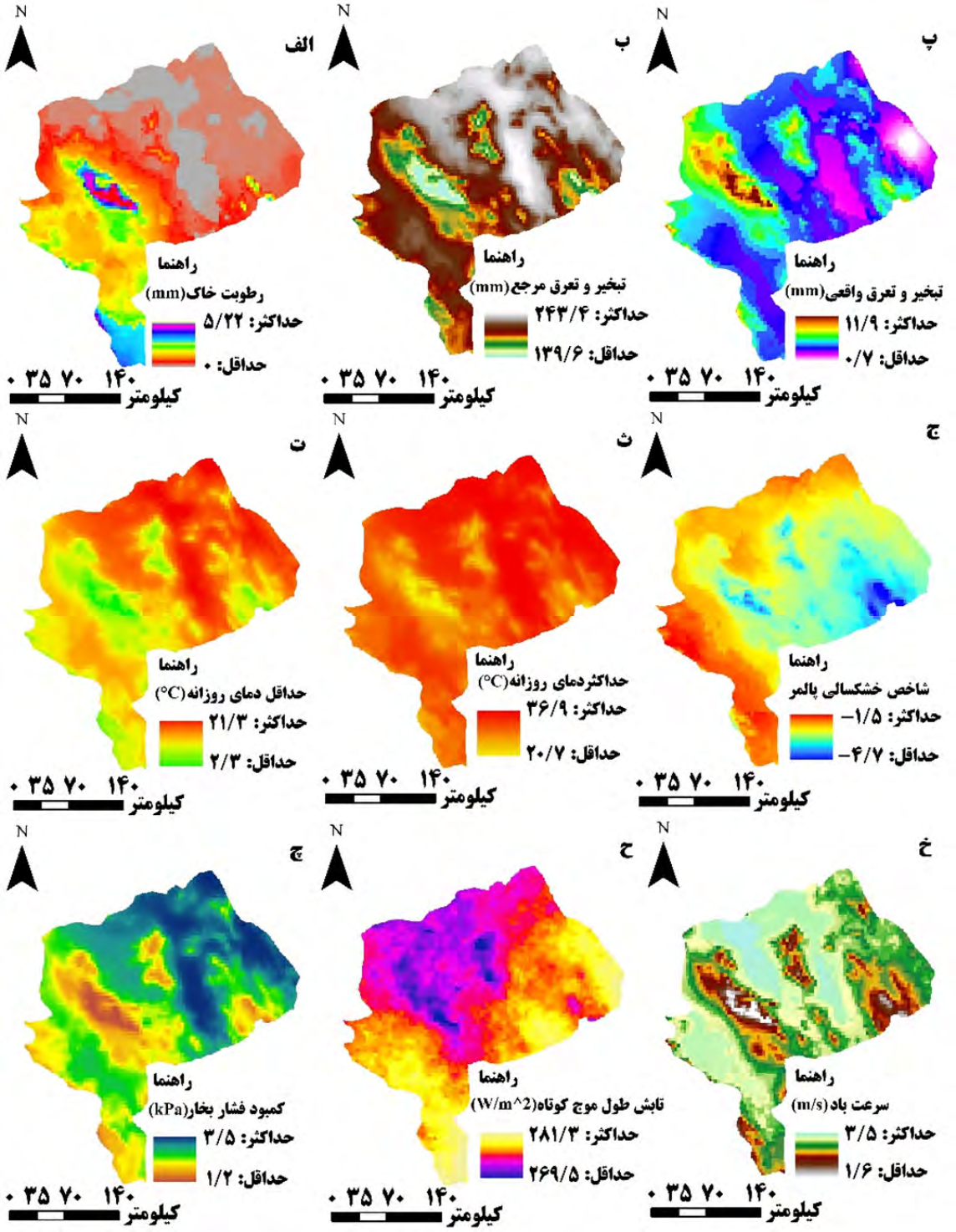


شکل ۱. نقشه محدوده مطالعاتی و نقاط نمونه‌برداری استفاده‌شده برای فرآیند یادگیری و آزمون

Fig. 1. Map of study area and sampling points used for learning and test procedure

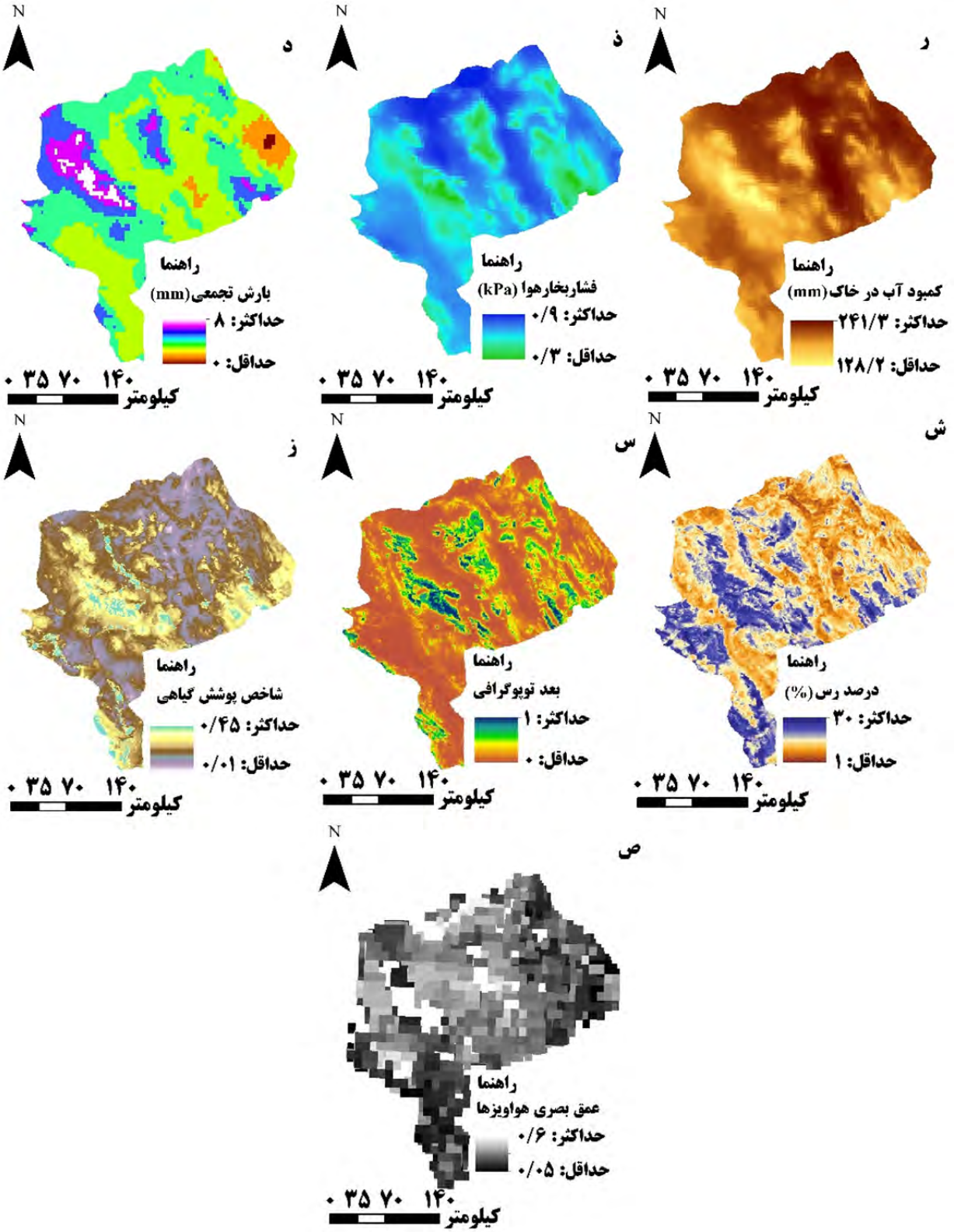
روش تحقیق

در پژوهش حاضر با استفاده از کد نویسی با زبان جاوا (Java Script) در محیط موتور گوگل ارث و برمبنای شیپ-فایل مرز استان یزد اقدام به اخذ داده‌های ماهواره‌ای براساس پروداکت‌های اقلیم و بیلان آب دانشگاه آیداهو، گردوغبار سنجنده ترا مودیس با دقت ٤ کیلومتر و درصد رس اُپن‌لند شد. تمامی این تصاویر برای ماه می هرسال در بازه سال‌های ۲۰۰۰ تا ۲۰۱۷ و برای یک بازه زمانی ۱۷ ساله اخذ شد و سپس متغیرها در این تصاویر در سیستم اطلاعات جغرافیایی برای بازه زمانی مذکور میانگین‌گیری شدند. در پژوهش حاضر، ۱۵ متغیر مستقل شامل سرعت باد در ارتفاع ده متری سطح زمین (vs)، رطوبت خاک (soil)، بارش تجمعی (pr)، شاخص خشک‌سالی پالمر (pdsi)، شاخص پوشش گیاهی نرمال شده (NDVI)، خشکی خاک یا کمبود آب خاک (def)، تبخیر و تعرق مرجع (pet) و واقعی (aet)، بعد توپوگرافی (TD)، رادیانس طول‌موج کوتاه رسیده به زمین (srad)، حداقل دمای هوا (tmmn)، حداکثر دمای هوا (tmmx)، فشار بخار (vap)، کمبود فشار بخار(vpd) و درصد رس (clay) موردبررسی قرار گرفت که در شکل ۲ ارائه‌شده است.



شکل۲. رطوبت خاک (الف)، تبخیر و تعرق مرجع (ب)، تبخیر و تعرق واقعی (پ)، حداقل دمای روزانه (ت)، حداکثر دمای روزانه (ث)، شاخص خشک‌سالی پالمر (ج)، کمبود فشار بخار (چ)، تابش طول‌موج کوتاه (ح)، سرعت باد (خ)

Fig. 2. Map of Soil moisture (a), Reference evapotranspiration (b), Actual evapotranspiration (c), Minimum daily temperature (d), Maximum daily temperature (e), Palmer drought index (f), Vapor pressure deficit (g), Shortwave radiation (h), Wind speed (i)



(ادامه) شکل۲. نقشه بارش تجمعی (د)، فشار بخار هوا (ذ)، کمبود آب در خاک (ر)، شاخص پوشش گیاهی (ز)، بُعد توپوگرافی (س)، درصد رس (ش) و عمق بصری هواویزها (ص)

(Continued). Fig. 2. Map of Cumulative precipitation (j), Vapor pressure (k), Soil water deficit (l), Vegetation index (m), Topographic dimension (n), Percentage of clay (o), and Aerosol optical depth (p)

هم‌چنین برای ماه آوریل تا می (با بررسی نمودار روند تغییرات عمق اپتیکی در سامانه گوگل ارث انجین، در بازه زمانی ۲۰۰۰ تا ۲۰۱۷ و شناسایی فراوانی ماه‌های دارای بیشترین میانگین عمق اپتیکی آئروسل‌ها، این ماه انتخاب شد. قابل‌ذکر است؛ بزرگ‌ترین واقعه گردوغبار طبق نمودار سری زمانی تغییرات عمق اپتیکی آئروسل برای ماه‌های اکتبر و سپتامبر سال‌های ۲۰۰۱ و ۲۰۰۸ با مقادیر ۰/۹۰۷ و ۰/۹۰۶ رخ داده‌است) هرسال بیش‌ترین مقادیر شاخص عمق اپتیکی آئروسل با ابزار تبدیلی سنجنده مودیس براساس پروداکت گردوغبار این سنجنده، میانگین‌گیری و سپس برای بازه زمانی یادشده محاسبه شد. قابل‌ذکر است که این داده‌ها از پروداکت MOD04 باهدف الگوریتم تیره (۱۳) در محصولات مودیس مخصوص هواویز در مقیاس روزانه از وبگاه ناسا دانلود شدند.

الگوریتم درختان رگرسیون و طبقه‌بندی (CART)

ایده اصلی این روش، درخت‌های دسته‌بندی و رگرسیون تقسیم داده‌ها به بخش‌های کوچک‌تر است که به‌طوری‌که این بخش‌ها حاوی اطلاعات تاحد ممکن تفکیک‌شده باشند (۱۱). یک درخت تصمیم‌گیری، مدلی غیر پارامتری است که به پراکنش خاص در مورد داده‌های متغیرهای مستقل وابسته نیست و روشی قابل‌اتکا در داده‌کاوی است. این روش انعطاف‌پذیر، دارای رویکردی مستحکم در تقسیم دوتایی و تصمیم در مورد بهترین اندازه درخت است. این روش برای داده‌کاوی و ایجاد مدل‌های پیش‌بینی‌کننده استفاده می‌شود (۱۷) و با داده‌های ناهمگون و ساختارهای غیرخطی سازگاری بالایی دارد. درخت با تقسیم مکرر مقادیر داده‌های متغیر وابسته توسط متغیرهای مستقل ساخته می‌شود. در هر تقسیم به دو گروه تقسیم می‌شوند که تا حد ممکن یکنواخت هستند و با بقیه فرق دارند. تقسیم شدن تا جایی ادامه پیدا می‌کند که به چندین گره نهایی با کم‌ترین میزان خطا برسد. به عبارتی منظور از همگن بودن گره این است که همه رکوردهای موجود در آن متعلق به یک دسته خاص باشند؛ چون در این صورت گره تبدیل به برگ می‌شود و الگوریتم موجود به دنبال ویژگی‌هایی از مجموعه داده‌هاست که خاصیت جداکنندگی بیش‌تری دارند (۲۶). اندازه درخت به تعداد گره‌های نهایی بستگی دارد. در این روش درخت تا حد ممکن رشد کرده و سپس عملیات هرس انجام می‌شود تا به یک سایز بهینه برسد. هرس‌شدن درخت بر اساس شاخص Cost complexity اتفاق می‌افتد. این روش نسبت به تغییرات یکنواخت متغیر مستقل تغییر نمی‌کند. در بین روش‌های یادگیری ماشینی، این روش بیش‌تر در بین پژوهشگران استفاده می‌شود و تفسیر ساده و راحتی دارد. در این روش از داده‌های یادگیری که ۷۰٪ داده‌ها و داده‌های آزمون که ۳۰٪ داده‌ها هستند جهت مدل‌سازی استفاده شد (۲۸). در داده‌های پژوهش حاضر داده مفقودشده وجود نداشت. هرچقدر یک بعد یا ویژگی، شاخص جینی کوچک‌تری داشت، آن ویژگی اطلاعات بیش‌تری ارائه کرد و توانست در درخت ساخته‌شده، بالاتر و نزدیک به ریشه قرار گیرد. هم‌چنین این الگوریتم از جداکننده جانشین بهره گرفت تا بهترین استفاده از داده با مقادیر گمشده را داشته باشد (۲۰). معیار مورداستفاده در این درختان انحراف حداقل مربعات بود و مطابق با رابطه ۱ محاسبه گردید.

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

در این رابطه؛ N_t تعداد داده‌ها در گره برگ t، $y_i(t)$ متغیر هدف در گره برگ، $\bar{y}(t)$ میانگین مقادیر متغیر هدف برای همه گره‌ها. متغیر ورودی $SS(t)$ زمانی بهترین متغیر برای ایجاد شاخه در گره t است که مقدار $Q(s,t)$ در رابطه ۲ را بیشینه کند.

$$Q(s,t) = SS(t) - SS(t_R) - SS(t_L) \qquad [۲]$$

در این رابطه؛ $SS(t_L)$ و $SS(t_R)$ به ترتیب میزان $SS(t)$ در شاخه سمت چپ و راست گره t می‌باشند. درختان ایجادشده ممکن نتایج بسیار پیچیده‌ای را نشان دهند و شامل صدها سطح شوند، لذا قبل از امتیازبندی آن‌ها برای داده‌های جدید بهینه‌سازی شدند (۱۲).

رگرسیون انطباقی چندمتغیره اسپیلاین (MARS)

این الگوریتم داده‌کاوی برای داده‌های متغیر وابسته پیوسته و یا دوتایی کاربرد دارد. این روش به شکلی مؤثر الگوی داده‌ای رابین متغیرها پیدا می‌کند که برای دیگر روش‌ها و مدل‌های رگرسیونی سخت و یا حتی غیرممکن است. مدل رگرسیونی که خوب توسعه‌یافته است می‌تواند برای پیش‌بینی‌ها و داده‌کاوی مناسب باشد. این مدل دارای خروجی پیوسته است که در یک قاعده و روند آرام مطابق با تغییرات داده‌های ورودی تغییر می‌کند که برخلاف روش‌های درخت تصمیم‌گیری است. در این روش اجرای مدل به شکل خودکار انجام می‌شود. به عبارتی انتخاب متغیرهای مربوط از متغیرهای بی‌ارتباط باهدف، تعیین روابط متقابل بین متغیرهای پیش-بینی‌کننده و اجرای آزمون‌های خودکار اضافی جهت جلوگیری از برازش بیش‌ازحد انجام می‌شود. این روش روابط متقابل بین متغیرها را مدل‌سازی می‌کند و تأثیر آن‌ها را روی متغیر وابسته بررسی می‌کند. هم‌چنین تأثیرات تک متغیر، روی متغیر وابسته نیز در نظر گرفته می‌شود. این روش روابط غیرخطی (گره‌ها) و روابط خطی رگرسیونی بین متغیرهای مستقل و وابسته را باهم و به شکل پیوسته در نظر می‌گیرد و از این روش برای کشف روابط غیرخطی بین متغیرها استفاده می‌شود (۹). این مدل قابلیت تفسیر بالایی دارد و نیازمند در نظر گرفتن فرض آماری بین متغیرها نیست (۱). در این الگوریتم بین متغیر وابسته و مستقل ابتدا گره‌ها پیدا می‌شوند و سپس رسم خط بین آن‌ها صورت می‌گیرد و در مرحله بعد تضمین پاسخ‌های پیوسته انجام می‌شود. در قیاس با سایر مدل‌های غیرخطی که فقط یک مجموعه از ضرایب را به داده‌ها اعمال می‌کنند؛ این روش با برازش دادن توابع چندجمله‌ای منطقه‌ای جداگانه برای هر یک از زیرمجموعه از داده‌ها اقدام به تشخیص الگوهای پیچیده می‌نماید (۳۰). مدل مارس بهینه طی یک فرایند دومرحله‌ای انجام می‌شود. در مرحله اول توسعه و در مرحله دوم عملیات هرس اتفاق می‌افتد. در مرحله نخست شاهد توابع پایه که مدلی بزرگ است ساخته می‌شود. این فرآیند بزرگ و منعطف شدن با اضافه شدن توابع پایه، تا جایی ادامه پیدا می‌کند که به یک مقدار بیشینه در توابع پایه برسد. در مرحله بعد عملیات حذف توابع پایه از مدل انجام می‌گیرد تا توابع ضعیف از چرخه خارج شوند و به مدل بهینه با کم‌ترین خطا خواهد رسید. در پژوهش حاضر به‌جای درصدی از داده‌ها به‌عنوان داده‌های تست و یادگیری، بر اساس درجه آزادی و بدون داده تست انجام شد و همه داده‌ها یادگیری محسوب شدند. هم‌چنین جهت انتخاب مدل بهینه از شاخص GCV استفاده شد. رگرسیون انطباقی چندمتغیره اسپیلاین مطابق با رابطه‌های ۳، ٤ و ٥ انجام شد (۲۹).

$$\hat{Y} = \hat{f}(x) = a_0 + \sum_{m=1}^{M} a_m B_m(X) \qquad [۳]$$

در این رابطه؛ a_0 یک مقدار ثابت است، M تعداد تِرم‌های غیرصفر همان گره‌ها که توابع پایه در آن‌ها تقسیم می‌شوند، a_m ضرایب مربوط به mاُمین تابع پایه و $B_m(X)$ مربوط به mاُمین تابع پایه برای مدل است که مطابق با رابطه ٤ محاسبه می‌شود.

$$B_m(X) = \prod_{i=1}^{K_m} \left[S_{i,m} (X_{v(i,m)} - t_{i,m}) \right]_+^q \qquad [٤]$$

در این رابطه؛ K_m درجه تعامل بین متغیرها در mاُمین تابع پایه، $S_{i,m=\pm1}$ مقداری بین ۱- تا ۱+ و $X_{v(i,m)}$ متغیر vاُم است که در آن $1 \le X_{v(i,m)} \le k$ که k تعداد کل متغیرهای ورودی و m شماره تابع پایه است. $t_{i,m}$ مکان گره در هر یک از متغیرهای پیش‌بینی متغیر وابسته است. q توان تابع پایه است و اندیس + به معنی بخش مثبت عبارت داخل براکت است. بدین معنی که اگر عبارت داخل براکت بزرگ‌تر از صفر بود، نتیجه خود عبارت داخل براکت می‌شود و در غیر این صورت نتیجه برابر صفر خواهد بود. جهت انتخاب بهترین مدل رگرسیونی از اعتبارسنجی متقاطع تعمیم‌یافته استفاده می‌شود (رابطه ٥).

$$V = \frac{(1/n)\sum_{i=1}^{n}[y_i - \hat{f}(x)_i]^2}{[1-(C(M)/n)]^2} \qquad [٥]$$

در این رابطه؛ y_i مقادیر واقعی کلاس موردنظر، $\hat{f}(x)_i$ مقدار تخمین زده‌شده برای مقادیر واقعی کلاس موردنظر، n تعداد کل مشاهدات، C(M) معیار هزینه–جریمه یک مدل است که شامل M تابع پایه است. به‌عبارت‌دیگر C(M) برابر تعداد مؤثر درجه آزادی است که به‌موجب آن، یک جریمه را برای اضافه کردن متغیرهای ورودی بیش‌تر به مدل اضافه می‌کند (۳۰).

درختان رگرسیون چندگانه جمع‌شدنی (TreeNet)

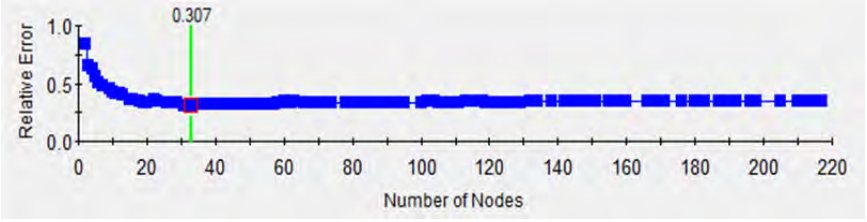
این روش از صدها و هزاران درخت تشکیل‌شده که هرکدام دارای گره‌های متعددی هستند. هر درخت نقش کوچکی در تشکیل درخت اصلی مدل‌دارند. این روش برای اکثر مسائل مدل‌سازی قابل‌استفاده است. این روش بر روی دقت پیش‌بینی تمرکز دارد و ترکیب مدل‌هایی با بالاترین کیفیت را انجام می‌دهد تا در یک مدل منفرد قرار گیرند. این روش داده‌های مشکوک را نادیده می‌گیرد. هم‌چنین این روش از رویکرد Gradient boosting بهره می‌برد (۸ و ۹). درخت رگرسیون چندگانه جمع‌شدنی یک الگوریتم دارای تکرار هست که بر اساس رابطه ٦ در هر تکرار m، یک درخت رگرسیون مانند: $T_m(x;\{R_{jm}\}, j=1,\dots,j)$ ساخته شد، که در هرکدام یک متغیر پیش‌بینی‌کننده x، مناطق جداکننده j و یک مقدار ثابت جداگانه پیش‌بینی شد.

$$T_m(x;\{R_{jm}\}_1^j = \sum_{i=1}^{j} \bar{y}_{jm}\, I(x \in R_{jm}) \qquad [٦]$$

در این رابطه؛ $\bar{y}_{jm}$ میانگین مقدار شبه پیش‌بینی در هر موقعیت R_{jm} در تکرار mاُمین است (۸ و ۹). در پژوهش حاضر از داده‌های یادگیری و آزمون به نسبت ۷۰٪ به ۳۰٪ و به‌صورت کاملاً تصادفی استفاده شد. شایان‌ذکر است که تعداد ۲۰۰ درخت حداقل با ٦ گره برای مدل‌سازی تنظیم شد.

نتایج و بحث

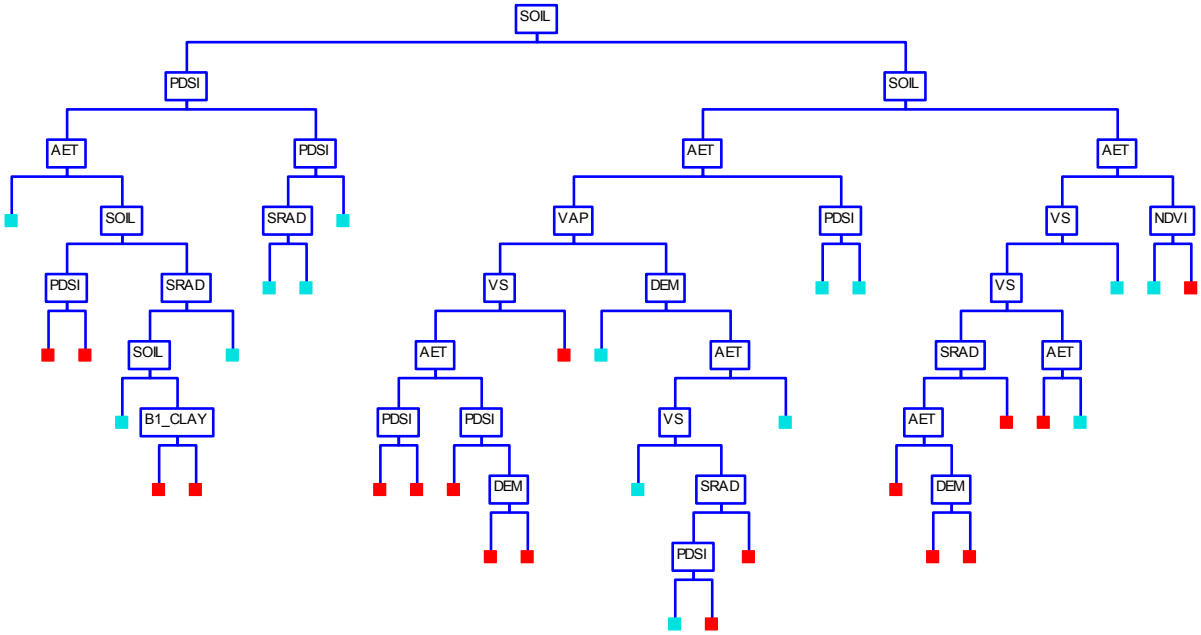
نتایج داده‌کاوی بین متغیر وابسته (عمق اپتیکی آئروسل) و متغیرهای مستقل (شاخص‌های اقلیمی) در پژوهش حاضر با استفاده از روش CART به شرح ذیل بود. در این روش از ۲۲۰ گره استفاده‌شده است که در گره ۳۳، بهترین اندازه درخت بر اساس مقدار بهینه بین خطای نسبی و تعداد گره با مقدار ۰/۳۰۷ حاصل شد و درخت هرس شد که در شکل ۳ ارائه‌شده است.



شکل ۳. تشخیص بهترین درخت تصمیم با تعداد ۳۳ گره

Fig. 3. Identifying the best decision tree with 33 nodes

مهم‌ترین گره‌ها در ریشه‌زنی درخت تصمیم شامل رطوبت خاک، شاخص خشک‌سالی پالمر و تبخیر و تعرق مرجع می‌باشند. میانگین‌های بالای شاخص عمق اپتیکی هواویزها به رنگ قرمز در شکل ٤ نشان داده‌شده است و متغیرهایی همچون شاخص پوشش گیاهی نرمال‌شده، تبخیر و تعرق واقعی، مدل رقومی ارتفاع، طول‌موج کوتاه رسیده به سطح زمین، شاخص خشک‌سالی پالمر، سرعت باد و درصد رس، گره‌های انتهایی جهت شناسایی مناطق با میانگین بالای عمق اپتیکی هواویزها می‌باشند. اهمیت نسبی متغیرها در فرآیند مدل‌سازی در روش CART محاسبه و در جدول ۱ ارائه‌شده است.



شکل۴. درخت تصمیم و گره‌های مادر (رنگ قرمز گره‌های مهم را نشان می‌دهد)

Fig. 4. Decision tree and mother nodes (red indicates important nodes)

جدول۱. اهمیت نسبی متغیرها با استفاده از روش CART

Table 1. Relative importance of variables using CART method

نام متغیر	اهمیت نسبی	شمای میزان اهمیت نسبی
SOIL	۱۰۰	
DEM	۸۳/۹٤	
PET	۷۷/۱٤	
AET	٦۷/۳۱	
TMMX	٦۷/٦۳	
TMMN	٦۷/٤۸	
PDSI	٦٦/۰۹	
VPD	٦۲/۱۳	
SRAD	٤۸/۹٤	
VS	۳٥/۸٥	
PR	۳٥/۱۷	
VAP	۳۰/۳۳	
DEF	۲۱/۰۰	
NDVI	۸/۱۳	
TD	٥/۱۷	
CLAY	۲/۳۹	
BULK	۰/۷۱	
SOILPH	۰/۳٤	

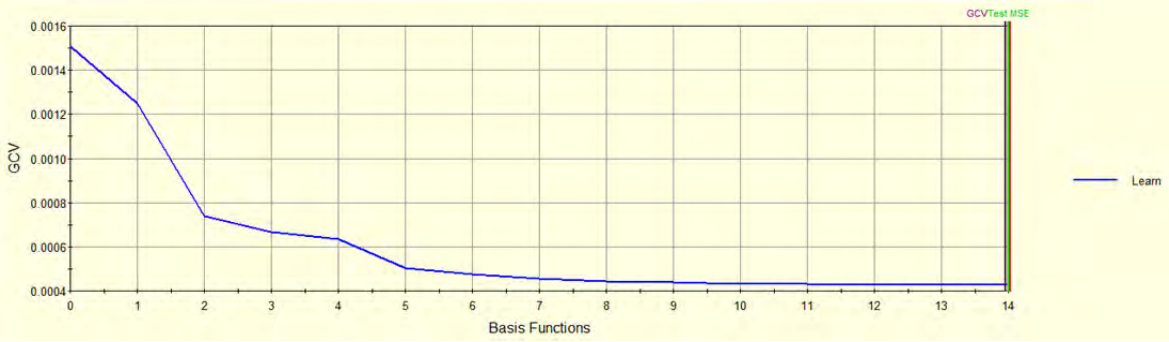
میزان R^2 مقدار خوبی را برای مدل نشان می‌دهد. مقادیر بین ۰/۸ تا ۰/۹ مقادیر خوب برای مدل محسوب می‌شود و مقدار کم‌تر از ۰/۷ دقت کم مدل را نشان می‌دهد. جدول ۲ شاخص‌های انتخاب مدل را نشان می‌دهد.

جدول۲. شاخص‌های انتخاب مدل CART

Table 2. Model selection criteria of CART

نام شاخص	مقدار شاخص
خطای جذر میانگین مربعات (RMSE)	۰/۰۱۴۸
خطای میانگین مربعات (MSE)	۰/۰۰۰۲
میانگین خطای مطلق (MAD)	۰/۰۱۱۳
مجذور میانگین خطای مطلق (MRAD)	۰/۲۶۲۶
مجموع مربعات متغیر هدف (SSY)	۱/۶۵۴
مجموع خطاهای مربع (SSE)	۰/۲۳۶۳
ضریب همبستگی (R^2)	۰/۸۵۷۱
ضریب همبستگی نرمال (R^2 Norm)	۰/۸۵۷۱
ضریب آکائیکه (AIC)	-۸۹۴۶/۳۶
ضریب آکائیکه تصحیح‌شده (AICc)	-۸۹۴۵/۴۸
انحراف (BIC)	-۸۸۴۱/۹۱
خطای نسبی (Relative Error)	۰/۱۴۲۸

نتایج داده‌کاوی به روش MARS به شرح ذیل است. این مدل، با تعداد حداکثر ۱۵ تابع پایه بر اساس شاخص GCV در مقدار بهینه ۱۴ به بیش‌ترین مقدار همبستگی بین متغیر وابسته و متغیرهای مستقل رسیده است که در شکل ۵ نشان داده‌شده است.



شکل۵. مقدار همبستگی بین متغیر وابسته و متغیرهای مستقل بر اساس شاخص GCV

Fig. 5. Value of correlation between dependent variable and independent variables based on GCV index

تعداد توابع پایه در مرحله نخست از روش MARS معادل ۱۵ تابع پایه است که در جدول ۳ مدل نهایی روش رگرسیون انطباقی چندمتغیره اسپیلاین آورده شده است.

جدول۳. مدل نهایی براساس روش رگرسیون انطباقی چندمتغیره اسپیلاین

Table 3. Final model based on multivariate adaptive regression spline method

گره	علامت	متغیرها	ضرایب	توابع پایه
			۰/۰٤	۰
۱/٤۸	+	SOIL	−۰/۰۸۲۵	۱
۱/٤۸	-	SOIL	۰/۰٦۹٦	۲
۲/۲۲	+	AET	−۰/۰۲۵۰	۳
۲/۲۲	-	AET	−۰/۰۳٤۳	٤
۹۲٦	+	DEM	۰/۰۰۰۱	۵
۹۲٦	-	DEM	−۰/۰۰۰۱	٦
−۳/۳	+	PDSI	۰/۰۳۹۸	۷
−۳/۳	-	PDSI	۰/۰۱۱۸	۸
۳/۳۹	+	VS	−۰/٦٦۵۹	۹
۳/۳۹	-	VS	−۰/۱۰۳۷	۱۰
۰/۱۷۷۸	+	SOIL	۰/۱۰۵۵	۱۱
۱۲۳۲	-	NDVI	۰	۱۳
۱۲۳۲	+	EVI	۰	۱٤
۳/۷۲	-	TMMN	−۰/۰۰٤۳	۱۵

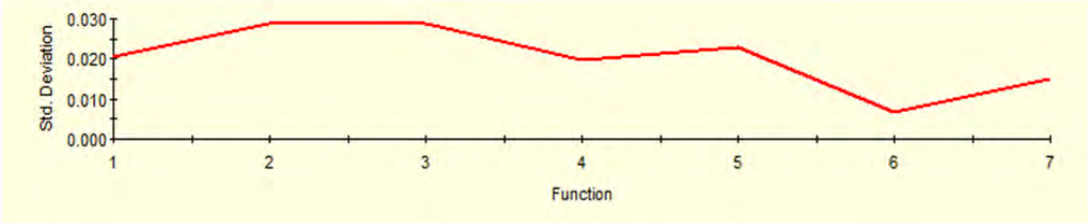
Basis Functions

BF1 = max(0, SOIL - 1.48889); BF2 = max(0, 1.48889 - SOIL); BF3 = max(0, AET - 2.22353); BF4 = max(0, 2.22353 - AET); BF5 = max(0, DEM - 926); BF6 = max(0, 926 - DEM); BF7 = max(0, PDSI + 3.30611); BF8 = max(0, -3.30611 - PDSI); BF9 = max(0, VS - 3.39); BF10 = max(0, 3.39 - VS); BF11 = max(0, SOIL - 0.177778); BF13 = max(0, NDVI - 1232); BF14 = max(0, 1232 - NDVI); BF15 = max(0, TMMN - 3.72778);

Y = 0.0454322 - 0.082459 * BF1 + 0.0695719 * BF2
- 0.0249663 * BF3 - 0.0343166 * BF4
+ 6.78743E-005 * BF5 - 6.88826E-005 * BF6
- 0.0397511 * BF7 + 0.0117701 * BF8
- 0.665928 * BF9 - 0.103687 * BF10
+ 0.105522 * BF11 + 1.21482E-005 * BF13
+ 3.17859E-005 * BF14 - 0.00432006 * BF15;

MODEL AOD = BF1 BF2 BF3 BF4 BF5 BF6 BF7 BF8 BF9 BF10 BF11 BF13 BF14 BF15;

خطای استاندارد مربوط به ترکیب هرکدام از روابط در شکل ٦ ارائه‌شده است. این نمودار نشان می‌دهد میزان خطای تابع به ازای حذف هر متغیر مرتبط چقدر است (جدول ٤). بر طبق این جدول، در توابع شماره ۲ و ۳ با حذف متغیرهای AET و DEM بیش‌ترین میزان خطا در مدل انطباقی چند متغیره اسپیلاین رخ می‌دهد. در مدل موردنظر، ۷ گام وجود دارد که در گام نخست سه تابع پایه موجود است و در کل ۱٤ تابع بهینه حاصل‌شده است که در بالا عنوان شد.



شکل۶. ارتباط بین خطای استاندارد و توابع (توابع بهینه و گام‌ها)

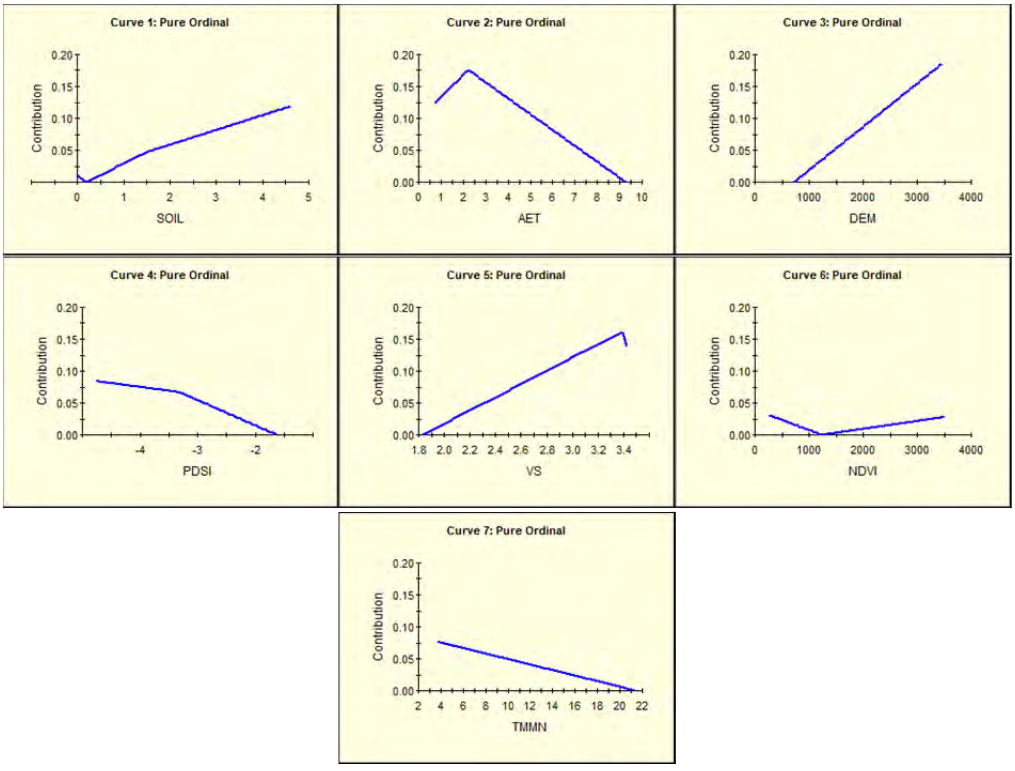
Fig. 6. Relationship between standard error and functions (optimal functions and steps)

جدول۴. توضیحات مبسوط گام‌ها، توابع پایه و میزان خطای حذف هرکدام

Table 4. Detailed description of the steps, basic functions and error amount of each deletion

متغیرها	مقدار پارامترهای مؤثر	تعداد توابع پایه	هزینه خطا	انحراف از معیار	تابع
SOIL	٦/٤٢٩	٣	٠/٠٠٠٥٧	٠/٠٢٠٦٠	١
AET	٤/٢٨٦	٢	٠/٠٠٠٦٦	٠/٠٢٨٦٣	٢
DEM	٤/٢٨٦	٢	٠/٠٠٠٤٨	٠/٠٢٨٧٣	٣
PDSI	٤/٢٨٦	٢	٠/٠٠٠٥٤	٠/٠١٩٦٠	٤
VS	٤/٢٨٦	٢	٠/٠٠٠٤٩	٠/٠٢٢٧٨	٥
NDVI	٤/٢٨٦	٢	٠/٠٠٠٤٥	٠/٠٠٦٧٤	٦
TMMN	٢/١٤٣	١	٠/٠٠٠٤٤	٠/٠١٥٠٥	٧

متغیرهایی که دارای گره یا به‌عبارت‌دیگر روابط غیرخطی بوده‌اند و هم‌چنین متغیرهایی که دارای روابط رگرسیونی خطی بوده‌اند نیز در شکل ۷ ارائه‌شده‌اند که حاکی از وجود روابط متقابل با درجه ۱ می‌باشند. هم‌چنین اهمیت نسبی متغیرها بر پیش‌بینی مقادیر عمق اپتیکی هواویزها در جدول ۵ و شاخص‌های انتخاب مدل MARS در جدول ٦ ارائه‌شده است.



شکل۷. متغیرهای مرتبط دارای گره و یا روابط رگرسیونی خطی روی متغیر عمق اپتیکی آئروسل با روش MARS

Fig. 7. Related variables with nodes or linear regression relationships on the aerosol optical depth variable with MARS method

جدول۵. اهمیت نسبی متغیرها روی پیش‌بینی مقادیر عمق اپتیکی آئروسل

Table 5. Relative importance of variables on predicting the aerosol optical depth

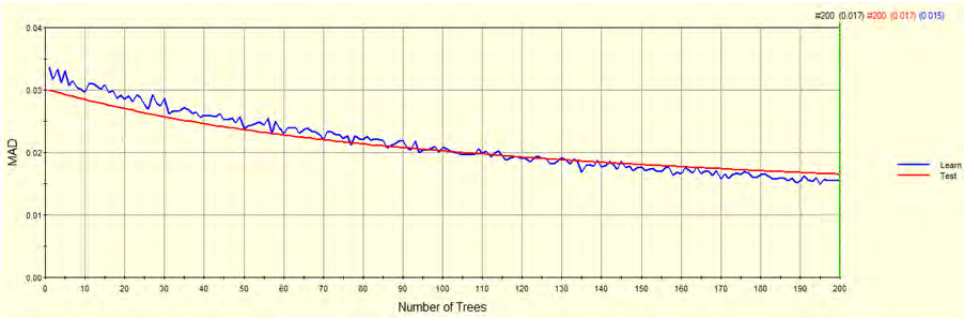
نام متغیر	مقدار اهمیت نسبی	شمای میزان اهمیت نسبی
AET	۱۰۰	
SOIL	۷۸/۷۷	
PDSI	۷۰/۱۶	
VS	۵۳/۶۴	
DEM	۴۵/۸۳	
NDVI	۳۲/۴۷	
TMMN	۲۰/۰۲	

جدول۶. شاخص‌های انتخاب مدل MARS

Table 6. Model selection criteria of MARS

نام شاخص	مقدار شاخص
خطای جذر میانگین مربعات (RMSE)	۰/۰۲۰۲۷
خطای میانگین مربعات (MSE)	۰/۰۰۰۴۱
اعتبارسنجی متقاطع تعمیم‌یافته(GCV)	۰/۰۰۰۴۳
میانگین خطای مطلق (MAD)	۰/۰۱۵۹۲
مجذور میانگین خطای مطلق (MRAD)	۰/۴۰۶۲۹
مجموع مربعات متغیر هدف (SSY)	۲/۲۵۸
مجموع خطاهای مربع (SSE)	۰/۶۱۶۶۰
ضریب همبستگی (R^2)	۰/۷۲۶۹۴
ضریب همبستگی نرمال (R^2 Norm)	۰/۷۲۶۹۴
ریشه مربعات اعتبارسنجی متقاطع تعمیم یافته (GCV R-Sq)	۰/۷۱۵۶۸

نتایج روش TreeNet به شرح زیر می‌باشد. مدل در درخت شماره ۲۰۰ به مقدار بهینه متغیر وابسته رسیده است که در شکل ۸ ارائه شده است. زیر شاخص‌های ارزیابی مدل و هم‌چنین اهمیت نسبی متغیرها در فرآیندمدل‌سازی در روش TreeNet در جدول‌های ۷ و ۸ ارائه شده است. میزان تأثیر هر کدام از متغیرها بر متغیر هدف یعنی عمق اپتیکی آئروسل (AOD) در شکل ۹ ارائه شده است.



شکل۸. نتایج روش TreeNet در خصوص مقدار بهینه متغیر وابسته و هم‌گرایی داده‌های یادگیری و آزمون

Fig.8. Results of the TreeNet method regarding the optimal value of the dependent variable and the convergence of learning and test data

جدول۷. زیر شاخص‌های ارزیابی مدل در روش TreeNet

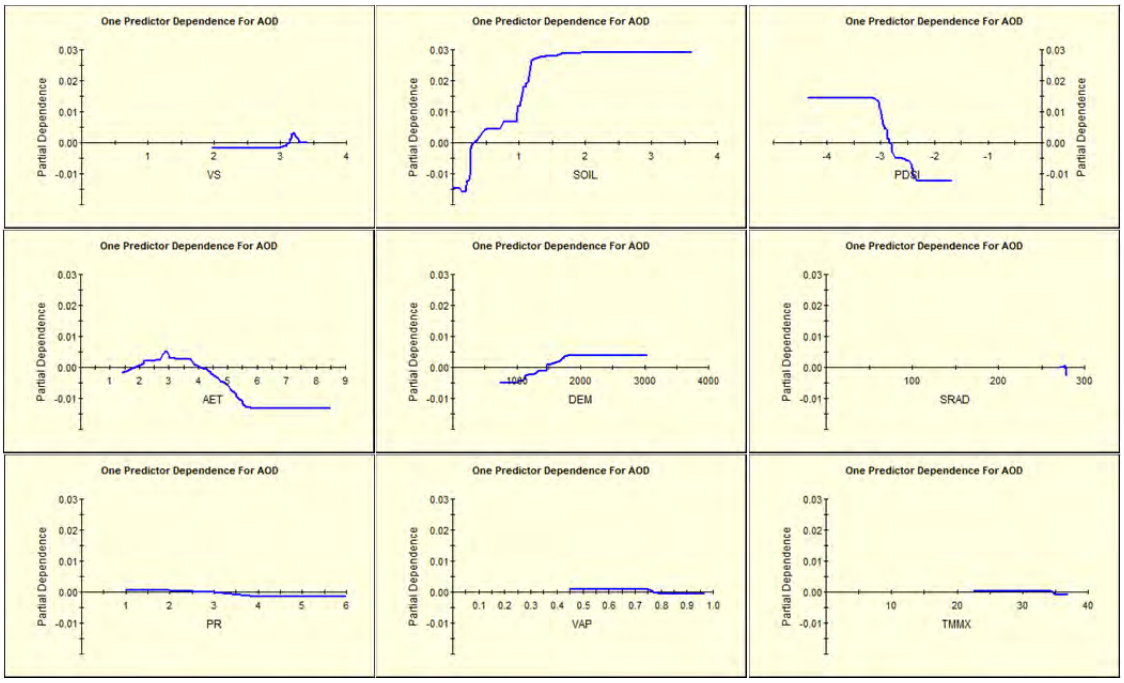
Table 7. Sub-indicators of model evaluation in TreeNet method

عنوان شاخص ارزیابی	مقدار شاخص
RMSE	۰/۰۱۹
MSE	۰/۰۰۰۳
MAD	۰/۰۱
MRAD	۰/۴۲
SSY	۱/۶۵
SSE	۰/۴۱
R2	۰/۷۵
R2 Norm	۰/۸
AIC	-۸۳۵۷/۲۸
AICc	-۸۳۵۶/۴
BIC	-۸۲۵۲/۸۴

جدول۸. اهمیت نسبی متغیرها در فرآیند مدل‌سازی در روش TreeNet

Table 8. The relative importance of variables in the modeling process in the TreeNet method

نام متغیر	مقدار نرمال‌شده اهمیت	شمای میزان اهمیت نسبی
SOIL	۱۰۰	|||||||||||||||||||||||||||||||||||||||||
PDSI	۸۳/۹۴	|||||||||||||||||||||||||||||||||||
AET	۶۲/۶۱	|||||||||||||||||||||||||
VS	۴۶/۷	|||||||||||||||||
DEM	۳۵/۹۹	|||||||||||||
SRAD	۲۳/۴۷	|||||||||
PR	۱۹/۳۲	|||||||
VAP	۱۶/۶۷	||||||
TMMX	۱۵/۸۹	||||||
DEF	۶/۴۱	||
TMMN	۶/۳۷	||
NDVI	۵/۵۱		
CARBON	۲/۹	
SLOPE	۲/۷	
BULK	۲/۳۲	



شکل۹. میزان تأثیر متغیرها بروی متغیر هدف (AOD) در روش TreeNet

Fig. 9. The effect of variables on the target variable (AOD) in the TreeNet method

بعد از پیش‌بینی مقادیر متغیر وابسته با استفاده از روش‌های داده‌کاوی یادشده، بر اساس ساختار رستری و ارزش‌های تخصیص داده‌شده به هر نقطه‌ی نمونه‌برداری شده (نقاط سمپل‌گیری شده از متغیرهای اقلیمی (متغیرهای مستقل) و عمق اپتیکی هواویزها (متغیر وابسته))، پهنه‌بندی ارزش‌ها در سامانه اطلاعات جغرافیایی و نرم‌افزار ArcGIS انجام گرفت. هم‌چنین مناطقی که بین سه روش مختلف داده‌کاوی به‌عنوان مناطق مشترک مطرح بودند با استفاده از توابع منطقی در قالب نقشه چهارم به همراه سه روش دیگر آورده شده است. پهنه‌بندی پتانسیل ایجاد گردوغبار در استان یزد به روش‌های MARS (ع)، CART (غ)، TreeNet (ف) و logical Intersect (ق) انجام شد که در شکل ۱۰ ارائه‌شده است. نتایج جمیع روش‌ها به حضور پهنه‌های با پتانسیل بالای گردوغبار در مناطق جنوبی و جنوب غربی استان یزد اذعان دارند و از پتانسیل عرصه‌های مختلف تولید گردوغبار به سمت شمال و شرق استان کاسته می‌شود. اگر روش CART را مبنای تصمیم‌گیری قرار دهیم؛ بر اساس بهترین مقادیر از شاخص‌های RMSE، MSE، MAD، MRAD و R^2 می‌توان عنوان نمود که مناطق غربی استان از سه کلاس مختلف با پتانسیل‌های متفاوت بر مبنای کلاس پتانسیل زیاد، متوسط و کم تشکیل‌شده است. درصد مساحت طبقات مختلف این کلاس‌ها در جدول ۹ ارائه‌شده است.

جدول۹. درصد مساحت پهنه‌های با پتانسیل مختلف گردوغبار در استان یزد

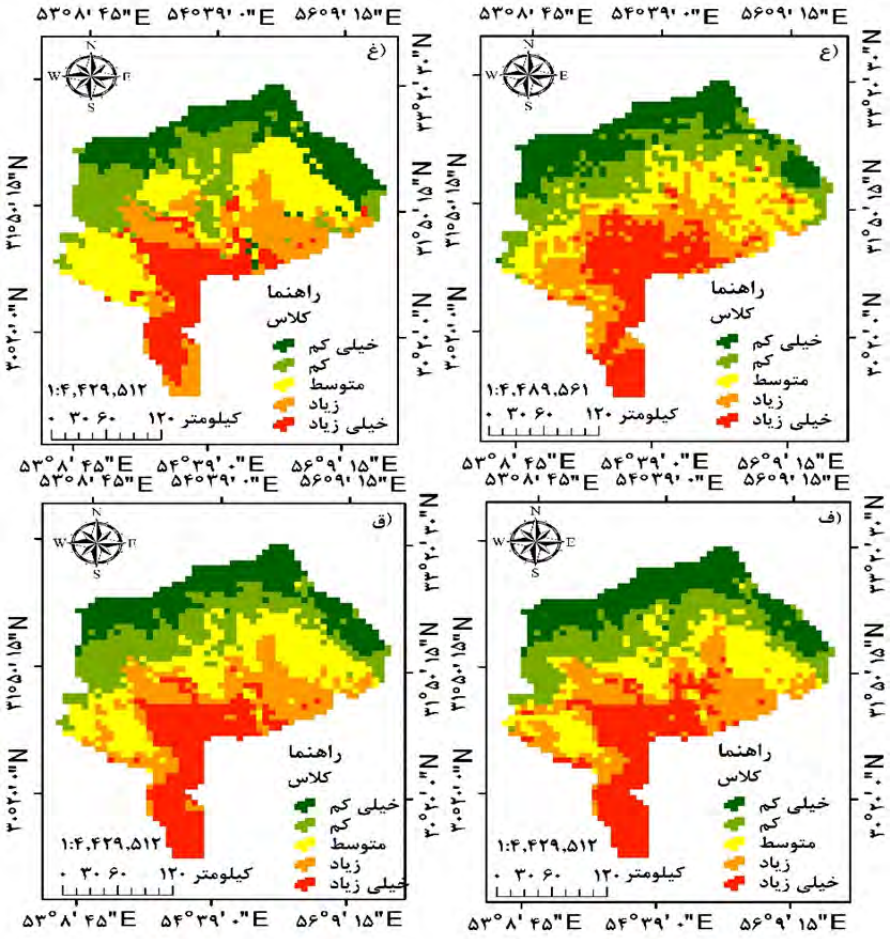
Table 9. Percentage of areas with different dust potentials in Yazd province

روش	پتانسیل				
	خیلی کم	کم	متوسط	زیاد	خیلی زیاد
CART	٪۱۹/۷۸	٪۱۹/۷۲	٪۲۵/۷۸	٪۱۸/۶۵	٪۱۶/۰۵
MARS	٪۱۹/۸۵	٪۲۰/۳۱	٪۲۰/۱۱	٪۱۹/۹۲	٪۱۹/۷۸
TreeNet	٪۱۹/۸۵	٪۲۰/۱۱	٪۱۹/۹۸	٪۱۹/۹۸	٪۲۰/۰۵
Intersect	٪۱۹/۱۲	٪۲۰/۵۸	٪۲۲/۹۸	٪۱۸/۸۵	٪۱۸/۴۵

بر طبق نتایج روش CART، بیش‌تر سطح استان متعلق به کلاس متوسط با سطح ۲۵/۷۸٪ است و پهنه‌های با پتانسیل خیلی زیاد با سطح ۱۶/۰۶٪ نسبت به بقیه کلاس‌ها کم‌ترین سطح را به خود اختصاص داده است. درنهایت، پهنه‌بندی پتانسیل عرصه‌های مختلف گردوغبار استان یزد بر اساس مدل‌های داده‌کاوی مختلف در شکل ۱۰ ارائه‌شده است.

تاکنون در ایران نسخه مشترکی برای به کار بستن بهترین مدل داده‌کاوی در خصوص داده‌های عمق اپتیکی هواویزها بکار گرفته نشده است. در مطالعه‌ای از بین سه مدل بیزی ساده (Naive Bayes)، نزدیک‌ترین همسایه (KNN) و درخت تصمیم (Decision tree)، اعلام شد که مدل درختان تصمیم با مقادیر همبستگی بیش‌تر، مجذور مربعات خطای کمتر و دقت (Accuracy) بالاتر نسبت به دو مدل دیگر عملکرد بهتری در پیش‌بینی متغیر هدف داشته است. هم‌چنین در آن مطالعه بیان داشتند مدل‌های داده‌کاوی درختان تصمیم به‌دلیل سادگی و تفسیر قابل‌فهم معمولاً رایج هستند (۲۵).

بر اساس نتایج به‌دست‌آمده،در پژوهش حاضر نیز دقت و عملکرد مدل CART نسبت به مدل‌های MART و MARS، بهتر بوده است. شایان یادآوری است که علاوه بر سه شاخصی که برای سنجش عملکرد در پژوهش‌های قبلی استفاده‌شده بود، در تحقیق حاضر از شاخص‌های دیگری همچون خطای میانگین مربعات، میانگین خطای مطلق، مجذور میانگین خطای مطلق، مجموع مربعات متغیر هدف، ضریب آکائیکه، ضریب آکائیکه تصحیح‌شده و انحراف مدل برای انتخاب بهترین مدل داده‌کاوی استفاده‌شده است. در مدل‌هایی که متغیر وابسته طبقه‌بندی‌شده است، از شاخص‌هایی نظیر منحنی‌های ROC (Receiver Operating Characteristic curve) و سطح زیر منحنی AUC (Area under the curve) استفاده می‌شود (۱۰ و ۲۲) و در متغیرهای وابسته‌ای که طبقه‌بندی نشده‌اند، از شاخص‌هایی نظیر ضریب همبستگی، مجذور میانگین مربعات خطا استفاده می‌شود (۲). در پژوهش حاضر، متغیر وابسته عمق اپتیکی هواویزها به شکل یک بازه پیوسته موردبررسی و داده‌کاوی واقع شد. در پژوهشی دیگر، از دو شاخص تورم واریانس (Variance inflation factor, VIF)) و ضریب رواداری (Tolerance coefficient, TC) برای بررسی میزان تأثیر متغیرهای مستقل که شامل متغیرهای اقلیمی و سطح زمین هستند بر متغیر وابسته (میزان گردوغبار) استفاده کردند (۱۰). در پژوهش یادشده متغیرهای سرعت باد و زمین‌شناسی به ترتیب بیش‌ترین و کم‌ترین اهمیت نسبی را برخیزش گردوغبار در استان خوزستان داشتند (۱۰). هم‌چنین در پژوهشی، متغیرهای شیب و کاربری اراضی بیش‌ترین و متغیر زمین‌شناسی کم‌ترین اهمیت نسبی را در خیزش گردوغبار در استان خراسان رضوی داشته‌اند (۵). از طرفی در پژوهش دیگری، متغیرهای شاخص پوشش گیاهی، ارتفاع و درجه شیب در شرق ایران را دارای بیش‌ترین تأثیر برخیزش گردوغبار معرفی کردند (۲۱). پژوهش حاضر در بین جمیع متغیرهای مستقل، به ترتیب رطوبت خاک، ارتفاع و تبخیر و تعرق مرجع و واقعی، بیش‌ترین اهمیت نسبی را در خصوص خیزش پدیده گردوغبار استان یزد به دست آورد. در پژوهش‌های داده‌کاوی در زمینه شناسایی مناطق مستعد خیزش گردوغبار، هرکدام مدل خاصی را به‌عنوان بهترین مدل، معرفی کرده‌اند. در این خصوص، مدل‌های داده‌کاوی ترکیبی و جنگل تصادفی و مدل درخت رگرسیون تقویت‌شده به‌عنوان مناسب‌ترین مدل‌های داده‌کاوی معرفی شدند (۵، ۱۰ و ۲۱).



شکل ۱۰. پهنه‌بندی پتانسیل تولید هواویز در عرصه‌های مختلف استان یزد بر اساس مدل‌های داده‌کاوی، ع) MARS، غ) CART، ف) TreeNet و ق) Intersect

Fig. 10. Zoning of the potential for aerosol production in different areas of Yazd province based on data mining models, MARS (A), CART (Gh), TreeNet (F) and Intersect (Q)

نتیجه‌گیری

پروداکت‌های اقلیمی و سطح زمین آیداهو و اوپن‌لند کمک شایانی به تحلیل محیط طبیعی می‌نمایند. روش‌های پیشرفته داده‌کاوی نیازمند گرفتن ارتباطات زمانی- مکانی، خودهمبستگی و خصوصیات پارامترها می‌باشند (۱٤). در پژوهش حاضر، بُعد مکانی محدوده گرد و غبارخیز استان یزد و مقادیر متغیرهای مختلف روی این پدیده در بازه زمانی ۱۷ ساله موردبررسی قرار گرفت. موارد خودهمبستگی بیش‌تر بر روی مباحث سری زمانی متغیرها مطرح می‌باشند و در این پژوهش بر بعد مکانی خصوصیات متغیرها بر پدیده گردوغبار تأکید شده است. با توجه به نتایج یادشده در مورد شناسایی تأثیرگذارترین متغیرها برگرد و غبار در مناطق مختلف، نمی‌توان یک یا چند متغیر را در پدیده خیزش گردوغبار برای همه مناطق، مشترک در نظر گرفت و این مهم از منطقه به منطقه‌ای دیگر تغییر می‌کند. کما اینکه متغیرهای زمین‌شناسی و کاربری اراضی در پژوهش حاضر جزء متغیرهایی بودند که هیچ‌گونه اثری بر متغیر وابسته یعنی حساسیت به گردوغبار نداشتند. در پژوهش حاضر، اشتراکات متغیرهای مستقل مهم و چرخه تصمیم‌گیری شامل تبخیر و تعرق واقعی، رطوبت خاک، شاخص خشک‌سالی پالمر، سرعت باد، ارتفاع، شاخص پوشش گیاهی و حداقل دمای روزانه بودند. هیچ‌کدام از پژوهش‌های مرتبط در مورد موضوع پژوهش، در انتخاب بهترین مدل

داده‌کاوی، همپوشانی نداشتند و مدل داده‌کاوی واحدی برای بررسی حساسیت مناطق مختلف به پدیده گردوغبار در ایران یافت نشد. شایان‌ذکر است، در این پژوهش مدل الگوریتم درختان رگرسیون و طبقه‌بندی انتخاب شد. پژوهش حاضر در نوع مدل‌های داده‌کاوی استفاده‌شده و متغیرهای مستقل با پژوهش‌های یادشده متفاوت بوده و با توجه به عدم همپوشانی نتایج انتخاب مدل برتر، نمی‌توان نسخه واحدی برای انتخاب بهترین مدل داده‌کاوی برای ایران در بحث گردوغبار ارائه نمود. لذا از بهترین مدل‌های منتخب در پژوهش‌های یادشده، برای داده‌کاوی پدیده گردوغبار در پژوهش‌های آتی استفاده و موردقیاس قرار گیرند. در مدل‌های داده‌کاوی نظیر درختان تصمیم، انتخاب تعداد بهینه گره‌ها و ابعاد درخت تصمیم بسیار مهم است و توصیه می‌گردد از مقدار بهینه این گره‌ها در تشکیل درخت تصمیم استفاده شود. در خصوص روش‌های ارزیابی نتایج مدل‌های داده‌کاوی، منحنی‌های ROC کمک شایانی به انتخاب بهترین مدل داده‌کاوی در مورد متغیر وابسته طبقه‌بندی‌شده می‌نماید و شاخص‌هایی نظیر خطای جذر میانگین مربعات، ضریب همبستگی، خطای میانگین مربعات و غیره کمک شایانی به انتخاب بهترین مدل داده‌کاوی در مورد متغیر وابسته طبقه‌بندی نشده می‌کند. هم‌چنین با پهنه‌بندی داده‌های خروجی از مدل‌های یادگیری ماشینی بر اساس پیکسل سایز ورودی داده‌های ماهواره‌ای در سامانه اطلاعات جغرافیایی، می‌توان به شناسایی مناطق حساس و خطرپذیر و نیز مهار آن اقدام نمود. قابل‌ذکر است در مورد پهنه‌های مختلف پتانسیل تولید گردوغبار در استان یزد، رده خطر خیلی زیاد با مساحت ۱۶/۰۵۶ درصد از سطح استان کم‌ترین مساحت را در قیاس با سایر طبقات خطر به خود اختصاص داد. پیشنهاد می‌گردد برای یک منطقه مطالعاتی از چندین مدل داده‌کاوی جهت انجام پیش‌بینی‌ها استفاده شود و با توجه به نتایج شاخص‌های ارزیابی عملکرد، بهترین مدل انتخاب گردد و صرفاً به نتایج یک مدل اکتفا نشود. قابل‌ذکر است اگر در مورد متغیر وابسته (عمق اپتیکی هواویزها) طبقه‌بندی استاندارد و مقبول جهانی وجود ندارد، جهت ارزیابی عملکرد مدل‌های پیش‌بینی داده‌کاوی، از شاخص‌هایی نظیر سطح زیر منحنی که مخصوص متغیر طبقه‌بندی‌شده است استفاده نشود و از شاخص‌هایی نظیر خطای جذر میانگین مربعات، ضریب همبستگی و غیره در مورد متغیر طبقه‌بندی نشده استفاده گردد. همچنین استفاده از داده‌های سنجش‌ازدور گوگل ارث انجین به لحاظ صرف زمان اندک تهیه و نیز دردسترس بودن داده‌ها، جهت انجام مطالعات محیطی نظیر شناسایی منابع گردوغبار پیشنهاد می‌گردد.

References

1. Ahmadlou M, Delavar M. 2015. Multiple land use change modeling using multivariate adaptive regression spline and geospatial information system. Journal of Geomatics Science and Technology, 5(2): 131-146. (In Persian).
2. Ali M, Asklany SA, El-wahab M, Hassan M. 2019. Data Mining Algorithms for Weather Forecast Phenomena Comparative Study. International Journal of Computer Science and Network Security, 19(9): 76-81.
3. Alibakhshi T, Azizi Z, Vafaeinezhad A, Aghamohammadi H. 2020. Survey of Area Changes in Water Basins of Shahid Abbaspour Dam Caused by 2019 Floods Using Google Earth Engine. Iranian Journal of Ecohydrology, 7(2): 345-357. (In Persian).
4. Bari Abarghuei H, Tabatabaei Aghda S, Tavakoli M, Najjar Hadashi N. 2006. The origin of Yazd storms and the damages caused by it. 1st National Conference on Wind erosion and dust storms. Paper presented at the 21 January, Yazd University, Yazd, Iran. (In Persion).
5. Boroughani M, Pourhashemi S. 2019. Susceptibility Zoning of Dust Source Areas by Data Mining Methods over Khorasan Razavi Province. Environmental Erosion Research Journal, 9(3): 1-22. (In Persian).
6. Danesh Shahraki M, Shahriari A, Gangali M, Bameri A. 2017. Seasonal and Spatial Variability of Airborne Dust Loading Rate over the Sistan plain cities and its Relationship with some Climatic Parameters. Journal of Water and Soil Conservation, 23(6): 199-215. (In Persian).
7. Ebrahimi-Khusfi Z, Ruhollah T-M, Maryam M. 2021. Evaluation of machine learning models for predicting the temporal variations of dust storm index in arid regions of Iran. Atmospheric Pollution Research, 12(1): 134-147. doi:https://doi.org/10.1016/j.apr.2020.08.029.
8. Friedman JH, Meulman JJ. 2003. Multiple additive regression trees with application in epidemiology.

Statistics in Medicine, 22(9): 1365-1381. doi:https://doi.org/10.1002/sim.1501.
9. Fridedman J. 1991. Multivariate adaptive regression splines (with discussion). Ann Stat, 19(1): 79-141.
10. Gholami H, Aliakbar M, Adrian LC. 2020. Spatial mapping of the provenance of storm dust: Application of data mining and ensemble modelling. Atmospheric Research, 233: 104716. doi:https://doi.org/10.1016/j.atmosres.2019.104716.
11. Gordon L. 2013. Using classification and regression trees (CART) in SAS® enterprise miner TM for applications in public health. SAS Global Forum 2013, San Francisco, California.
12. Halabian A, Javari M, Akbari Z, Akbari G. 2017. Evaluating the performance of decision tree model in estimating the suspended sediments of river (A case study on the basin of Meimeh river). Geography And Development Iranian Journal, 15(49): 81-96. (In Persian).
13. Hojati M. 2017. Artificial neural network based model to estimate dust storms PM10 content using MODIS satellite images. Journal of Environmental Studies, 42(4): 823-838. (In Persian).
14. Hunter H, Cervone G. 2017. Analysing the influence of African dust storms on the prevalence of coral disease in the Caribbean Sea using remote sensing and association rule data mining. International Journal of Remote Sensing, 38(6): 1494-1521. doi:https://doi.org/10.1080/01431161.2016.1277279.
15. Karimi K, Taheri Shahraiyni H, Habibi Nokhandan M, Hafezi Moghadas N. 2011. Identifying sources of origin for producing dust storms in Middle East using remote sensing. Journal of Climate Research, 2((7-8)): 57-72. (In Persian).
16. Khalighi Sigaroudi S, Shahbandari R, Dadfar R, Kamrani F. 2011. Investigation of the relationship between drought and dust storms (Case study: Yazd province). Paper presented at the 2nd National Conference on Wind Erosion and Dust Storms. Yazd University, Yazd, Iran. (In Persian).
17. Loh WY. 2011. Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery, 1(1): 14-23.
18. Mirakbari M, Ganji A, Fallah S. 2010. Regional bivariate frequency analysis of meteorological droughts. Journal of Hydrologic Engineering, 15(12): 985-1000. doi:https://doi.org/10.1061/(ASCE)HE.1943-5584.0000271.
19. Mohammad Khan S. 2017. The study of the status and trend of changes in dust storms in Iran during the period from 1985 to 2005. Irrigation and Watershed Management (Iranian Journal of Natural Resources) 2(3): 495-514. (In Persian).
20. Panahi M, Mirhashemi SH. 2015. Assessment among two data mining algorithms CART and CHAID in forecast air temperature of the synoptic station of Arak. Environmental Sciences, 13(4): 53-58. (In Persian).
21. Pourhashemi S, Amirahmadi A, Zangane Asadi MA, Salehi M. 2018. Identifying and determining the characteristics of dust centers in Khorasan Razavi province. Arid Regions Geography Studies, 9(34): 1-9. (In Persian).
22. Pourhashemi S, Boroghani M, Amirahmadi A, Zanganeh Asadi M, Salhi M. 2019. Dust source prioritization with using statistical models (Case study: Khorasan Razavi provience). Journal of Range and Watershed Managment, 72(2): 343-358. (In Persian).
23. Rashki A, Kaskaoutis D, Rautenbach CJW, Eriksson P, Qiang M, Gupta P. 2012. Dust storms and their horizontal dust loading in the Sistan region, Iran. Aeolian Research, 5(3): 51-62. doi:https://doi.org/10.1016/j.aeolia.2011.12.001.
24. Rezazadeh M, Irannejad P, Shao Y. 2013. Climatology of the Middle East dust events. Aeolian Research, 10: 103-109. doi:https://doi.org/10.1016/j.aeolia.2013.04.001.
25. Rokach L, Maimon OZ. 2014. Data mining with decision trees: theory and applications, vol 81. World scientific. 244 p.
26. Sharma H, Kumar S. 2016. A survey on decision tree algorithms of classification in data mining. International Journal of Science and Research (IJSR), 5(4): 2094-2097.
27. Sobhani B, Safarian Zengir V, Faizollahzadeh S. 2020. Modeling and prediction of dust in western Iran. Physical Geography Research Quarterly, 52(1): 17-35. (In Persian).
28. Soleimanpour S, Mesbah S, Hedayati B. 2018. Application of CART decision tree data mining to determine the most effective drinking water quality factors (case study: Kazeroon plain, Fars province). Iranian Journal of Health and Environment, 11(1): 1-14. (In Persian).
29. Tsolmon R, Ochirkhuyag L, Sternberg T. 2008. Monitoring the source of trans-national dust storms in north east Asia. International Journal of Digital Earth, 1(1): 119-129. doi:https://doi.org/10.1080/17538940701782593.
30. Zha W, Chan W-Y. 2005. Objective Speech Quality Measurement Using Statistical Data Mining. EURASIP Journal on Advances in Signal Processing, 2005(9): 721258. doi:10.1155/ASP.2005.1410.

RS & GIS for Natural Resources
(Vol. 12/ Issue 1) Spring 2021
Indexed by ISC, SID, Magiran, Noormags, Civilica, Google Scholar
journal homepage : www.girs.iaubushehr.ac.ir

Original paper

Identifying origins of atmospheric aerosols using remote sensing and data mining (Case study: Yazd province)

Mohamad Kazemi, Ali Reza Nafarzadegan, Fariborz Mohammadi, Ali Rezaei Latifi

Received: 13 January 2020 / Accepted: 19 September 2020
Available online 1 March 2021

Abstract

Background and Objective The Middle East is one of the most important regions in the world for dust production. Iran, located in the Middle East, is exposed to numerous local and trans-regional dust systems due to its location in the arid and semi-arid regions of the world. Dust storms, in addition to covering arable land and plants with wind-blown materials, destroy fertile lands and reduce biological production and biodiversity, and severely affect the survival of residents. Dust storms are involved in the transmission of dangerous pathogens to humans, air pollution, and damage to respiratory function. Dust storms in Yazd province are relatively common and the average number of days with dust storms in the province reaches 43 days a year. This phenomenon has caused many problems for the people of the province. The main indicators of air quality are the concentration of suspended particles and the aerosol optical depth (AOD) following the occurrence of dust events. Numerous studies have been conducted in the world to identify the centers of dust collection and their origin. However, to the best of the authors' knowledge, there is no study on the spatial zoning of dust conditions using three algorithms of CART, MARS and TreeNet algorithms as the predictive models. The purpose of this study is to forecast and zoning the potential of different areas for the production of dust aerosols using remote sensing data and data mining models as well as to specify the most important variables on this phenomenon in Yazd province.

Materials and Methods The Yazd province lies in a dry region of Central Iran. The province experienced average annual rainfall of about 57 mm and an average annual temperature of about 20 ºC. The maximum temperature experienced in the warmest month of the province is close to 46 ºC. The maximum wind speed in this province is up to 120 kilometres per hour. The Google Earth Engine (GEE) interface (Javascript editor) was applied to collect remote sensing data in order to form three data sets that contain features related to topography, climate, and land surface conditions. These features were employed as the independent variables of the models,

M. Kazemi [1], A. R. Nafarzadegan(✉) [2], F. Mohammadi [1,3], A. Rezaei Latifi [1,4]

1. Assistant Professor, Hormoz Studies and Research Center, University of Hormozgan, Bandar Abbas, Iran
2. Assistant Professor, Department of Natural Resources Engineering, Faculty of Agriculture and Natural Resources, University of Hormozgan, Bandar Abbas, Iran
3. Assistant Professor, Department of Water Sciences & Engineering, Minab Higher Education Complex, University of Hormozgan, Minab, Iran
4. Assistant Professor, Physics Department, Faculty of Sciences, University of Hormozgan, Bandar Abbas, Iran

e-mail: a.r.nafarzadegan@gmail.com

http://dorl.net/dor/20.1001.1.26767082.1400.12.1.4.5

which is built by taking advantage of three data mining algorithms, classification and regression tree (CART), multivariate adaptive regression splines (MARS), and TreeNet, to specify the potential of areas for dust production. The dependent variable (target variable) of the models was the aerosol optical depth (AOD), which was acquired from MOD04 AOD retrievals from the Moderate Resolution Imaging Spectroradiometer (MODIS) onboard NASA's Terra satellite. The outcomes of the three models for classifying areas with different dust potentials were evaluated under performance criteria, such as R-squared, mean absolute deviation (MAD), the mean square error (MSE), the mean relative absolute deviation (MRAD), and the root means square error (RMSE).

Results and Discussion The results showed the variables mostly affecting the dependent variable (AOD) in the MARS model were actual evapotranspiration, soil moisture, and the Palmer drought severity index. The values of R^2 and RMSE in the MARS model were equal to 0.72 and 0.02, respectively. Similarly, the features with the highest relative importance according to the TreeNet model were soil moisture, Palmer drought severity index, and actual evapotranspiration. The values of R^2 and RMSE in the TreeNet model were equal to 0.75 and 0.019, respectively. The results revealed that the CART model with R^2 = 0.85, MAD = 0.011, MSE = 0.002, MRAD = 0.262, and RMSE =0.014 had the best performance compared with the other two data mining models. The soil moisture, elevation, reference and actual evapotranspiration, minimum and maximum temperature, Palmer drought severity index, downward shortwave solar radiation, and wind speed were the most important variables in forecasting the potential of areas for dust production, respectively. Also, the areas with very high, high, moderate, low and very low susceptibility were occupied about 16%, 19%, 26%, 20% and 20% of the Yazd province, respectively.

Conclusion All three models, which were based on three data mining algorithms, CART, MARS, and TreeNet, had a good agreement in specifying the most important variables affecting the optical depth of the dust aerosols in the study area. However, these models indicated different priority order for the identified variables in terms of relative importance; Besides, there was a difference in their performance criteria. As mentioned above, the CART model was the best-performing model, of the current study, for specifying the potential of areas for the generation of dust aerosols. According to this model, 25.8% of the province was classified as the moderate-risk of aerosol production, 18.6% of the province as the high-risk of aerosol production, and 16.0% of the study region as the very high-risk of dust aerosols. The high-risk areas are mostly spread in the western and southwestern regions of the Yazd province.

Keywords: Aerosol optical depth, Spatial variables, Machine learning, Zoning

Please cite this article as: Kazemi M, Nafarzadegan AR, Mohammadi F, Latifi AR. 2021. Identifying Origins of Atmospheric Aerosols using Remote Sensing and Data Mining (Case study: Yazd province). Journal of RS and GIS for Natural Resources, 12(1): 63-85. http://dorl.net/dor/20.1001.1.26767082.1400.12.1.4.5