تحقیقات آب و خاک ایران، دوره ۵۰، شماره ۱۰، اسفند ۱۳۹۸ (ص ۲۵۳۸-۲۵۲۵) DOI: 10.22059/IJSWR.2019.280905.668198

Modelling and Prediction of Soil Classes Using Boosting Regression Tree and Random Forests Machine Learning Algorithms in Some Part of Qazvin Plain

SAYED ROHOLLA MOUSAVI[1], FERAIDON SARMADIAN[*1], ASGHAR RAHMANI[1]
1. Soil and Science Engineering Department, Faculty of Agricultural Engineering and Technology, University College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran.
(Received: May, 14, 2019- Revised: July. 3, 2019- Accepted: July. 15, 2019)

ABSTRACT

Appropriate selection of ancillary covariates have a specific important on digital soil mapping. Currently, use of machine learning algorithms for digital mapping and updating of conventional soil map has been developed in Iran. The current study has been done to compare the BRT and RF models for spatial prediction of subgroup and family classes with selection of axillary variables using VIF approach in some part of Qazvin Plain. 61 pedons were sampled based on stratified random, digged, described and classified with consideration of laboratory analysis up to family level. The most appropriate variables were selected among 15 Geomorphometry and Remote Sensing Indices using Variance Inflation Factor (VIF). Soil landscape modeling was conducted with RF and BRT learning algorithm in RStudio software based on Randomforest and C5.0 packages at subgroup and family levels. The results showed that six indices including CHA, DEM, STH, SI DVI and NDVI were selected as input variables. Assessment indices such as the Overall Accuracy (OA) and Kappa were obtained for BRT (35, 26%) and RF (70, 60%) at family level, respectively. Sensitivity analysis based on the mean decrease accuracy (MDA) revealed that the modified catchment area variable is the most relative important variable among the selected variables. Generally, by using feature selection innovative approach and effective learning algorithms, the spatial distribution of soil maps could be made even in low relief lands with acceptable accuracy.

Keywords: Digital Soil Mapping, Learning Algorithm, Random Forests Model, Boosting Regression Tree, Data mining

*Corresponding Author's Email: Fsarmad@ut.ac.ir

مدل‌سازی و پیش‌بینی مکانی کلاس خاک با استفاده از الگوریتم یادگیری رگرسیون درختی توسعه‌یافته و جنگل‌های تصادفی در بخشی از اراضی دشت قزوین

سیدروح اله موسوی[1]، فریدون سرمدیان[1*]، اصغر رحمانی[1]

۱. گروه مهندسی علوم خاک، دانشکده مهندسی و فناوری کشاورزی، پردیس کشاورزی و منابع طبیعی، دانشگاه تهران، کرج، ایران

(تاریخ دریافت: ۱۳۹۸/۲/۲۴- تاریخ بازنگری: ۱۳۹۸/۴/۱۲- تاریخ تصویب: ۱۳۹۸/۴/۲۴)

چکیده

انتخاب متغیرهای کمکی مناسب در روش‌های یادگیرنده ماشینی جهت نقشه‌برداری رقومی خاک از اهمیت ویژه‌ای برخوردار است. طی سال‌های اخیر در ایران استفاده از الگوریتم‌های یادگیرنده در نقشه‌برداری رقومی و بهنگام سازی نقشه‌های قدیمی توسعه یافته است. پژوهش حاضر در بخشی از اراضی دشت قزوین با هدف مقایسه جنگل‌های تصادفی (RF) و رگرسیون درختی توسعه‌یافته (BRT) در پیش‌بینی مکانی کلاس‌های زیرگروه و فامیل خاک بهمراه انتخاب متغیرهای کمکی با استفاده از شاخص تورم واریانس انجام شده است. ۶۱ خاکرخ به روش نمونه‌برداری تصادفی طبقه‌بندی‌شده حفر، تشریح و با تجزیه‌وتحلیل آزمایشگاهی تا سطح فامیل رده‌بندی گردید. مناسب‌ترین متغیرهای محیطی از میان ۱۵ متغیر ژئومورفومتری و شاخص‌های سنجش از دور با استفاده از فاکتور تورم واریانس انتخاب گردیدند. مدل‌سازی رابطه خاک – زمین‌نما در دو سطح زیرگروه و فامیل خاک با استفاده از دو الگوریتم یادگیرنده RF و BRT در نرم‌افزار RStudio بر اساس دو بسته "Randomforest" و "C5.0" اجرا گردید. نتایج انتخاب متغیرهای محیطی نشان داد که شش متغیر CHA، DEM ، STH، NDVI، SI و DVI به‌عنوان متغیر ورودی انتخاب گردیدند. شاخص‌های ارزیابی مدل‌ها شامل صحت کلی و شاخص کاپا به ترتیب برای الگوریتم BRT، ۳۵، ۲۶ درصد و برای الگوریتم RF، ۷۰، ۶۰ درصد در سطح فامیل خاک حاصل گردید. آنالیز حساسیت برمبنای شاخص میانگین حداقل صحت نشان داد که متغیر محیطی مساحت حوزه آبخیز اصلاح‌شده دارای بیشترین اهمیت نسبی در میان متغیرهای انتخاب شده است. به‌طورکلی با استفاده از رویکردهای نوین انتخاب متغیر و الگوریتم‌های یادگیرنده مؤثر می‌توان نقشه‌ی پراکنش مکانی خاک‌ها را حتی در نواحی با پستی‌وبلندی کم با صحت قابل‌قبول تهیه نمود.

واژه‌های کلیدی: نقشه‌برداری رقومی خاک، الگوریتم یادگیرنده، مدل جنگل تصادفی، درخت تصمیم توسعه‌یافته، داده-کاوی

مقدمه

دستیابی به اطلاعات خاک برای مدیریت پایدار اکوسیستم‌ها ضروری است. با توجه به اینکه در بسیاری از بخش‌های کشور ایران اطلاعات پایه در مورد خاک‌ها خارج از دسترس بوده و یا به‌سختی قابل دسترس است، نقشه‌برداری رقومی می‌تواند به‌عنوان ابزاری در زمانی که اطلاعات تفصیلی دقیق در مورد خاک‌ها وجود ندارد، برای مناطق فاقد اطلاعات، پیش‌بینی مکانی انجام دهد (Afshar *et al*., 2018). نقشه‌های مرسوم خاک به عنوان منبع اصلی اطلاعات خاک‌ها از لحاظ ارائه جزئیات مکانی و صحت دارای محدودیت می‌باشند، با این وجود این نقشه‌ها دارای دانشی با ارزش از روابط بین خاک‌ها و متغیرهای محیطی کمکی بوده و این دانش را می‌توان با استفاده از داده‌های محیطی باکیفیت و روش‌های نوین پردازش داده برای بهنگام سازی نقشه‌های مرسوم استخراج نمود (Yang *et al*., 2011). با توجه به اینکه نقشه‌برداری خاک‌ها به روش سنتی یک کار زمان‌بر و پرهزینه است، روش‌های آماری و ریاضی مختلفی برای پیش‌بینی مکانی خصوصیات و کلاس‌های خاک توسط محققین (Schloeder *et al*., 2001; Thomas *et al*., 2000 Yemefack *et al*., 2005) بکار گرفته شده است. رویکردی نوین توسط مک برتنی و همکاران، بر مبنای مدل کلارپت (Jenny, 1944) تحت عنوان معادله اسکورپن (شامل:

$$S_{c,s} = f(s,c,o,r,p,a,n) + e$$

ارائه شد که، $S_{c,s}$: ویژگی یا کلاس مربوط به خاک، s: مربوط به اطلاعات خاک یا پایگاه داده خاک یا از دانش کارشناس حاصل می‌گردد. c: اقلیم، o: جانداران از جمله جانوران و

* نویسنده مسئول: Fsarmad@ut.ac.ir

1 Scorpan

پوشش گیاهی، r: پستی و بلندی، p: مواد مادری، a: زمان و n: موقعیت مکانی و تابع f: یک طبقه‌بندی نظارت‌شده یا الگوریتم یادگیرنده نظارت‌شده را نشان می‌دهد. این معادله با هدف ایجاد ارتباط بین متغیرهای محیطی و متغیرهای وابسته (خصوصیت یا کلاس‌های خاک) بیان گردید (McBratney *et al*., 2003).

با افزایش دسترسی به داده‌های مدل رقومی ارتفاع و سنجش از دور، زمینه‌های لازم برای ارتقاء سطح کیفیت و نوآوری در پیش‌بینی مکانی خصوصیات و کلاس‌های خاک فراهم گردیده است که تحت عنوان "نقشه‌برداری رقومی خاک" نامیده می‌شود (Grunwald *et al*., 2011; McBratney *et al*., 2003). در این‌گونه مطالعات پیش‌بینی مکانی خصوصیات و کلاس‌های خاک با استفاده از شاخص‌های توپوگرافی و طیفی برگرفته‌شده از مدل‌های رقومی ارتفاع و تصاویر ماهواره صورت می‌پذیرد (Nauman and Thompson, 2014). در نقشه‌برداری رقومی، این ارتباط به‌وسیله مدل‌های پیش‌بینی کننده مختلف از طریق سامانه اطلاعات جغرافیایی، روش‌های آماری و پدولوژیستی برقرار می‌گردد. اخیراً الگوریتم‌های یادگیرنده نقش قابل‌توجهی را در پیش‌بینی مکانی روابط خاک- سرزمین به خود اختصاص داده‌اند که برخی از آنها شامل شبکه‌های عصبی مصنوعی، منطق فازی یا مدل‌های درختی می‌باشند که توسط روش‌های یادگیری ماشینی توسعه یافته‌اند (Grinand *et al*., 2008). الگوریتم طبقه‌بندی درخت رگرسیون[1] (CART) به منظور پیش‌بینی مکانی نقشه خاک با استفاده از داده‌ها و نقشه‌های یک منطقه مرجع به‌وسیله Lagacherie (1992) مورد استفاده قرار گرفت. اخیراً مدل‌های درخت تصمیم که همراه با توسعه روش‌های آماری است به‌عنوان مدل‌های توسعه‌یافته تصادفی ارتقاء یافته‌اند Freidman *et al*.,2000). رگرسیون درختی توسعه‌یافته[2] (BRT) یکی از روش‌های نوین است که بوسیله الگوریتم‌های جدید آماری با هدف ارتقاء کارایی یک مدل با برازش و ترکیب تعداد زیادی از درختان، به‌منظور پیش‌بینی مورد استفاده قرار می‌گیرد. مدل‌های توسعه‌یافته تصادفی، کارایی و دقت پیش‌بینی‌ها را از طریق کاهش بیش‌آموزش و بیش‌برازش که در مدل‌های درختی ساده رخ می‌داد را افزایش داده است. برازش تابع BRT ممکن است خطی[3]، منحنی[4] یا غیرخطی[5] باشد و از توزیع خطای نرمال، دو ارزشی و پواسن پیروی می‌کند (Death, 2007; Elith *et al*., 2008). اخیراً پژوهشگران از دو روش رگرسیون درختی و رگرسیون درختی توسعه‌یافته در مطالعه پراکنش مکانی سطوح گروه‌های بزرگ و زیرگروه خاک‌های شمال غرب استان کرمانشاه استفاده نموده و بیان داشتند که در هر دو سطح موردمطالعه، روش رگرسیون درختی توسعه‌یافته دارای مقادیر صحت عمومی و شاخص کاپای بالاتری است(Baghche Maryam and Shekaari, 2018). روش جنگل‌های تصادفی[6] (RF)، یک روش مبتنی بر رگرسیون درختی، دارای کارایی بالا در پیش‌بینی، مقاوم در مقابل نویز و عدم تأثیرپذیری از واریانس متغیرهای محیطی و خاک، موجب شده که به‌عنوان یک الگوریتم یادگیرنده مفید و دقیق برای پیش-بینی مکانی کلاس‌های خاک شناخته شود (Sreenivas *et al*. 2016). Khamoshi *et al*. (2019) برای پیش‌بینی مکانی کلاس-های فامیل خاک، کارایی بالای روش RF را در دشت قزوین گزارش نمودند. Mosleh *et al*., (2017) در مطالعه خود در دشت شهرکرد به مقایسه روش‌های مختلف نقشه‌برداری رقومی شامل RF، BRT،ANN و رگرسیون لجستیک چند متغیره[7] (MLR) پرداختند و گزارش نمودند که روش‌های مورداستفاده از توانایی یکسانی در مدل‌سازی سطوح تاکسنومیک برخوردار بودند. Mirakzehi *et al*. (2018) با استفاده از روش RF اقدام به نقشه‌برداری رقومی خاک در سطوح تاکسونومیک گروه بزرگ، زیرگروه و فامیل در اراضی دلتای سیستان نمودند و به ترتیب صحت کلی ۴۶، ۴۴ و ۴۶/۱ برای هر یک از این سطوح گزارش نمودند. با توجه به بررسی نتایج پژوهشگران ذکرشده تأکید بسیاری از آنان بر استفاده از متغیرهای محیطی مناسب برای پیش‌بینی مکانی کلاس‌های خاک است؛ بنابراین روش‌های متعددی در نقشه‌برداری رقومی خاک برای انتخاب بهترین متغیرها جهت مدل‌سازی از میان دسته گسترده‌ای از داده‌های کمکی وجود دارد. Tesfa *et al*. (2009) از روش همبستگی بین متغیرها برای تعیین پارامترهای بهینه جهت مدل‌سازی عمق خاک با روش RF استفاده نمودند. فاکتور دیگری که برای انتخاب بهترین متغیرها جهت مدل‌سازی استفاده می‌شود، فاکتور شاخص بهینه است که از واریانس و همبستگی بین نسبت‌های مختلف باندها استفاده می‌نماید (Chavez *et al*., 1982). در برخی مطالعات، انتخاب بر اساس نظر کارشناس و میزان دسترسی به اطلاعات در منطقه موردمطالعه بستگی دارد. Hengl *et al*. (2007) و Levi and Rasmussen (2014) از روش تجزیه مؤلفه‌های اصلی[8] (PCA) برای انتخاب مهمترین متغیرهای کمکی

1 Classification and Regression Tree Algorithm
2 Boosted Regression Tree
3 Linear
4 Curvilinear
5 Non-linear
6 Random Forests
7 Multi regression logistic
8 Principal Component Analysis

در مطالعه خود استفاده نمودند. در این پژوهش از رویکرد نوین فاکتور تورم واریانس[1] برای انتخاب مناسب‌ترین پارامترهای ورودی جهت مدل‌سازی (VIF) استفاده شده است. پژوهش حاضر با دو هدف (۱) مقایسه صحت دو الگوریتم یادگیرنده ماشینی RF و BRT در پیش‌بینی مکانی و پهنه‌بندی نقشه کلاس‌های خاک بر اساس سامانه رده‌بندی آمریکایی خاک و (۲) تعیین مهمترین متغیرهای کمکی مؤثر بر روی پراکنش مکانی خاک‌های منطقه انجام شده است.

مواد و روش‌ها

منطقه موردمطالعه

منطقه موردمطالعه با مساحت ۱۶۶۲۸ هکتار بخشی از اراضی پیرامون شهر آبیک از توابع استان قزوین است که در موقعیت‌های عرض‌های جغرافیایی ´ ۱ ° ۳۶ تا ´ ۹ ° ۳۶ و طول‌های جغرافیایی ´ ۱۴ ۵۰° تا ´ ۲۱ ° ۵۰ در جنوب شرق استان قزوین و شمال شرق شهر آبیک واقع گردیده است. واحدهای اصلی ژئومورفولوژی منطقه در سطح زمین‌نما[2] شامل تپه[3]، پنپلین[4]، پیدمونت[5] و دشت[6] است که به ترتیب واحدهای زمین‌نمای پیدمونت و دشت دامنه‌ای بیشترین مساحت منطقه را به خود اختصاص داده‌اند. متوسط ارتفاع منطقه ۱۲۸۷ متر نسبت به سطح دریای آزاد و دامنه تغییرات شیب صفر تا ۲۵ درصد است. میانگین بارندگی سالیانه ۲۸۴ میلی‌متر در دوره آماری ۲۰ ساله (۱۳۷۲-۱۳۹۲) و متوسط دمای هوا ۱۴ درجه سانتی‌گراد (Iran Meteorological Organization, 2013) به ترتیب دارای رژیم‌های رطوبتی و حرارتی زریک خشک[7]، اریدیک ضعیف[8] و ترمیک[9] می‌باشد (Van Wambeke, 2000). تشکیلات زمین شناسی شامل مواد مادری مربوط به دوره‌های ترشیاری و کواترنر عمدتاً شامل توف سبز، بازالت خاکستری، شیل و گدازه‌های آتشفشانی می‌باشد. کاربری غالب اراضی منطقه موردمطالعه شامل زراعت آبی کشت آبی، دیمزارها، مراتع و مراکز صنعتی و مسکونی است.

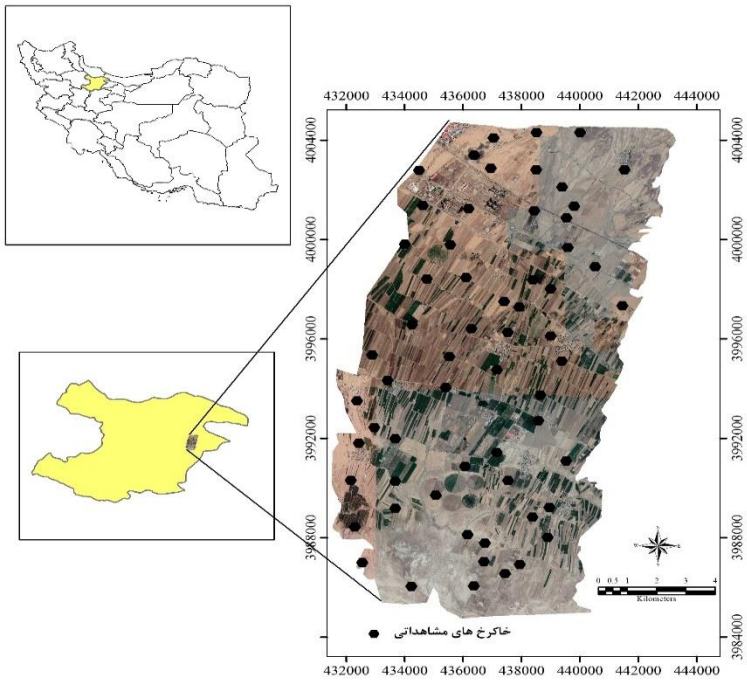
مطالعات میدانی و نمونه‌برداری

بر اساس تفسیر عکس‌های هوایی با مقیاس ۱:۴۰۰۰۰ (سازمان نقشه‌برداری کشور، ۱۳۷۶) و بازدیدهای میدانی و تصاویر ماهواره‌ی لندست ۸، مرزبندی‌های اولیه به منظور تفکیک واحدهای ژئوفرم منطقه بر اساس سیستم پیشنهادی Zinck (1988) صورت پذیرفت. این نقشه مبنای مطالعات خاکشناسی منطقه قرار گرفت (Mousavi *et al*., 2017). در مجموع موقعیت ۶۱ خاکرخ مشاهداتی بر اساس روش نمونه‌برداری تصادفی طبقه‌بندی‌شده با متوسط فاصله ۷۵۰ متر حفر گردید (شکل ۱). پس از تشریح خاکرخ‌های حفرشده از کلیه افق‌های ژنتیکی شناسایی‌شده بر اساس راهنمای تشریح و نمونه‌برداری خاک‌ها، نمونه‌برداری صورت گرفت (Schoeneberger *et al*., 2012). پس از نمونه‌برداری از خاکرخ‌های موردنظر، نمونه‌ها جهت هوا خشک شدن و انجام آنالیزهای لازم به آزمایشگاه منتقل گردیدند. pH گل اشباع با استفاده از pH متر، قابلیت هدایت الکتریکی عصاره اشباع با استفاده از هدایت‌سنج الکتریکی، بافت خاک به روش هیدرومتری (Gee and Bauder,1986)، کربن آلی به روش اکسیداسیون تر (Walkley and Black,1934)، کربنات کلسیم با روش تیتراسیون (Nelson, 1982)، گچ به‌وسیله روش نلسون و سومرز (Nelson, 1982) و ظرفیت تبادل کاتیونی با آمونیوم استات (Sumner and Miller, 1996) تعیین گردید. در نهایت بر اساس مشاهدات میدانی و نتایج تجزیه آزمایشگاهی خاکرخ‌های موردمطالعه تا سطح فامیل بر اساس سامانه رده‌بندی خاک آمریکایی (Soil Survey Staff, 2014) طبقه‌بندی گردیدند.

متغیرهای کمکی

متغیرهای کمکی[10] مورداستفاده در این پژوهش شامل مدل رقومی ارتفاع با قدرت تفکیک مکانی ۱۲/۵× ۱۲/۵ متر (ALOS-1 PALSAR L1.0 2007) و تصاویر سنجنده‌های (OLI/TIRS) لندست ۸ با قدرت تفکیک مکانی ۳۰×۳۰ متر است (U.S. Geology Survey 2014). کلیه متغیرهای ژئومورفومتری با استفاده از نرم‌افزارهای SAGA GIS[11] (Olaya, 2004) نسخه 7.2 و شاخص‌های سنجش از دور در محیط نرم‌افزار ERDAS IMAGINE نسخه 2014 تهیه گردیدند (جدول ۱). تغییر مقیاس مکانی[12] کلیه شاخص‌های سنجش از دور و متغیرهای ژئومورفومتری به قدرت تفکیک مکانی یکسان ۳۰ متر در نرم‌افزار ارداس انجام گردید.

1- Variance Inflation Factor
2- Landscape surface
3- Hilland
4- Peneplain
5- Piedmont
6- Plain
7- Dry Xeric
8- Weak Aridic
9- Thermic
10- Covariate
11- System for Automated Geoscientific Analyses
12- Resampling



شکل ۱: موقعیت منطقه موردمطالعه و خاکرخ‌های مشاهداتی

جدول ۱: متغیرهای محیطی مورداستفاده جهت پیش‌بینی مکانی کلاس‌های خاک

متغیر محیطی	ماهیت متغیر	نماد	نام متغیر
مدل رقومی ارتفاع	توپوگرافی	DEM	ارتفاع (Elevation)
		STH	ارتفاعات استانداردشده (Standardized Height)
		Slope	شدت شیب (Slope Gradient)
		RSP	موقعیت نسبی شیب (Relative Slope Position)
		Diffuinso	تابش پخشیده (Diffuse insolation)
		MCA	حوزه زهکشی اصلاح‌شده (Modified Catchment area)
		MSP	موقعیت میانی شیب (Midslope position)
		MrVBF	شاخص همواری دره با درجه تفکیک بالا (Multi-resolution valley bottom flatness index)
		CHA	مساحت حوزه آبخیز (Catchment area)
		NH	ارتفاعات نرمال شده (Normalized Height)
		SH	ارتفاعات شیبدار (Slope Height)
تصویر لندست ۸(OLI,TIRS)	انعکاس طیفی	NDVI	شاخص گیاهی تفاضلی نرمال شده (Normalized Difference vegetation Index)
		SI	شاخص شوری (Salinity Index)
		DVI	شاخص گیاهی تفاضلی (Difference Vegetation Index)
		RVI	شاخص گیاهی نسبی ((Relative Vegetation Index

انتخاب متغیرهای کمکی

در این مطالعه انتخاب متغیرهای کمکی بر اساس شاخص تورم واریانس[1](VIF) انجام گردید. فرآیند انتخاب متغیرها در روش اشاره‌شده، با انتقال متغیرها در قالب یک فایل عددی با فرمت (CSV) به محیط نرم‌افزار R Studio نسخه 1.1.456 و با استفاده از بسته "VIF " و روش داده‌کاوی (vifcor) صورت می‌پذیرد. این روش بر اساس یک رویکرد گام به گام اقدام به حذف آن دسته از متغیرهایی می‌نماید که در مجموعه متغیرها دارای بیشترین همبستگی با یکدیگر هستند (Dormann, C. F. *et al*., 2013).

مدل‌سازی خاک – زمین‌نما با استفاده از الگوریتم‌های درختی

یکی از الگوریتم‌های مورداستفاده در این پژوهش مدل جنگل‌های تصادفی (RF) است که برای مدل‌سازی پراکنش مکانی کلاس-های خاک در دو سطح فامیل و زیرگروه مورد استفاده قرار گرفت. مدل RF یک تکنیک یادگیرنده فعال است که توسط Breiman (2001) ارائه شده است. این مدل (RF) توسعه‌یافته از مدل طبقه-بندی و رگرسیون درختی (CART[2]) است. روش CART روشی است که داده‌ها را به‌طور تکراری برای به‌دست آوردن ارتباط بین متغیر پاسخ و متغیرهای مستقل و انجام تخمین جداسازی می-کند. در روش RF برخلاف سایر روش‌های درختی که تعداد محدودی درخت ترسیم می‌کنند، صدها یا هزاران درخت طبقه-بندی تولید می‌شود (Breiman and Cutler, 2004). این روش یک روش یادگیری گروهی است و برای طبقه‌بندی با ساختن تعداد درختان زیاد عمل می‌نماید (Breiman, 2001). اساس روش‌های یادگیرنده گروهی این است که گروهی از یادگیرنده‌های ضعیف، مجموعه‌ای از یادگیرنده‌های قوی را تشکیل می‌دهند. کلیه مراحل مدل‌سازی با استفاده از این روش یادگیری ماشینی با استفاده از بسته "Random Forest" به همراه کدنویسی در محیط نرم‌افزار R Studio صورت می‌پذیرد. در این تحقیق با استفاده از ۸۰٪ داده‌ها به ترتیب با ترسیم ۶۰۰ و ۷۰۰ درخت، حالت بهینه یا آموزش کامل برای دو سطح فامیل و زیر گروه خاک به دست آمد و با ۲۰٪ باقیمانده خاکرخ‌ها که برای ایجاد هر درخت استفاده نمی‌شوند (داده‌های بیرون از سبد (OOB)[3]اعتبار سنجی یا آزمودن مدل انجام پذیرفت؛ بنابراین نیازی به اعتبار سنجی جداگانه در این مدل نیست. تکنیک RF با استفاده از آنالیز حساسیت، اهمیت متغیرها در مدل‌سازی را نیز تعیین می‌کند. RF به روش میانگین کاهش حداقل[4](MDA) قادر به ارائه اهمیت متغیرهای مورداستفاده در فرآیند مدل‌سازی می‌باشد. در روش MDA، مقادیر درست متغیرها با مقادیری که به‌طور تصادفی برای هر درخت تولید شده است جایگزین می‌شود و اگر این جایگزینی اثری روی خطای اندازه‌گیری نداشته باشد اهمیت آن کم است و اگر مقدار خطای اندازه‌گیری افزایش یابد، آن متغیر مهم است (Breiman and Cutler, 2004).

رگرسیون درختی توسعه‌یافته

درختان تصمیم‌گیری از نسل جدید تکنیک‌های داده‌کاوی به‌شمار می‌آیند که در دو دهه اخیر توسعه زیادی یافته‌اند. یکی از انواع تکنیک‌های درختی، مدل رگرسیون درختی توسعه‌یافته (BRT) است که حاصل ترکیب دو تکنیک آماری رگرسیون درختی و بوستینگ است (Elith *et al*., 2008). به منظور آموزش مدل‌ها مجموعه خاکرخ‌ها (متغیرهای محیطی و کلاس‌های خاک) به صورت تصادفی به دو مجموعه با نسبت ۸۰:۲۰ تقسیم شدند. ۸۰ درصد داده‌ها برای آموزش مدل و ۲۰ درصد دیگر به‌عنوان داده-های اعتبارسنجی برای ارزیابی مورد استفاده قرار گرفت و برای جلوگیری از بیش‌برازش در روند مدل‌سازی از روش هرس پیش رشد درخت (که موجب کاهش پیچیدگی می‌شود) استفاده شد. در روش بوستینگ برای بهینه‌سازی الگوریتم C5.0 از ۸۰ درخت استفاده شد.

اعتبارسنجی مدل‌سازی

ارزیابی صحت مدل

در این مطالعه برای ارزیابی صحت، از روش اعتباربخشی یا نمونه آزمون استفاده شد. تقریباً ۸۰ درصد از نمونه‌ها (۴۸ نیمرخ برای طبقه‌بندی) برای آموزش و ۲۰ درصد نیمرخ‌ها برای ارزیابی صحت طبقه‌بندی (۱۳ نیمرخ) استفاده شد که به‌صورت تصادفی در مرحله مدل‌سازی انتخاب می‌شوند. صحت نقشه پیش‌بینی کلاس‌های خاک با استفاده از ماتریس خطا تعیین گردید (Congalton, 1991). از معیارهای صحت کلی طبقه‌بندی و ضریب کاپا برای ارزیابی صحت طبقه‌بندی استفاده شد که معادله آنها به شرح زیر است (Byrt *et al*. 1993):

$$OA = \sum_{i=1}^{n} Xij \,/\, N$$

(رابطه ۱)

(رابطه ۲)

$$Kappa = N\sum_{i=1}^{n} Xij - \sum_{i-1}^{n}(Xio - Xoi) \,/\, N^2 - \sum_{i=1}^{n}(Xio - Xoi)$$

1- Variance Inflation Factor
2 Classification and Regression Trees
3 Out of Bag
4 Mean Decrease in Accuracy

صحت کلی

در رابطه (۱) به ترتیب (OA) صحت کلی و N معرف کل پیکسهای طبقه‌بندی‌شده و $\sum_{i=1}^{n} Xij$ نمایه مجموع پیکسل‌های قطر اصلی ماتریس خطا (پیکسل‌های صحیح طبقه‌بندی‌شده) است. صحت کلی طبقه‌بندی ازجمله پارامترهای اندازه‌گیری است که فقط دقت کلی را گزارش می‌نماید و در مورد هر کدام از طبقات به‌طور مجزا اطلاعاتی ارائه نمی‌کند.

شاخص کاپا

ماتریس خطا معرف اختلاف بین توافق واقعی در داده‌های مرجع و یک طبقه‌بندی‌کننده خودکار و توافق احتمالی بین داده‌های مرجع و طبقه‌بندی‌کننده تصادفی است. در واقع این شاخص مقداری بین صفر و یک دارد که اگر کاپا برابر صفر باشد نشانه‌گر یک طبقه‌بندی کاملاً "تصادفی" و مقدار منفی نشان‌دهنده خطا در طبقه‌بندی و اگر این مقدار برابر یک باشد نشانگر یک طبقه-بندی کاملاً "صحیح" است (Rasouli, 2008).

در رابطه (۲) (n) تعداد ردیف‌ها در ماتریس، (Xij) تعداد مشاهدات در ردیف i و ستون j (درایه‌های قطر اصلی[1])، Xio و Xoi مجموع حاشیه[2]به ترتیب ردیف r و ستون i، N تعداد کل مشاهدات است. بر اساس طبقه‌بندی که Landis and Koch (1977) در مورد نتایج شاخص کاپا ارائه نمودند دقت آن به شرح جدول (۲) است.

جدول ۲. طبقه‌بندی شاخص کاپا

Kappa index	description
<۰/۰۱	Less than chance agreement
۰/۰۱-۰/۲	Slight agreement
۰/۲-۰/۴	Fair agreement
۰/۴-۰/۶	Moderate agreement
۰/۸-۰/۶	Substation agreement
۰/۸-۰/۹	Almost perfect agreement

صحت تولیدکننده[3]

(رابطه ۳)

$$PA = \frac{a_{tt}}{\sum_{i=1}^{N} a_{ik}}$$

در رابطه ۳: a_{tt} تعداد پیکسل‌های صحیح طبقه‌بندی‌شده بر روی قطر اصلی و $\sum_{i=1}^{N} a_{ki}$ جمع تعداد پیکسل‌هایی است که در آن ستون به‌عنوان نمونه‌های آموزشی آن طبقه‌بندی شده‌اند.

صحت کاربر[4]

(رابطه ۴)

$$UA = \frac{a_{ij}}{\sum_{i=1}^{N} a_{ik}}$$

در رابطه (۴): a_{ij} تعداد پیکسل‌های درست طبقه‌بندی‌شده بر روی قطر اصلی و $\sum_{i=1}^{N} a_{ki}$ نمایه جمع تعداد پیکسل‌هایی است که در آن ردیف به عنوان نمونه‌های آموزشی آن طبقه‌بندی شده‌اند.

دامنه تغییرات صحت تولیدکننده و صحت کاربر حد واسط ۰ و ۱ می‌باشد که تبع مقادیر بالاتر نشان‌دهنده عملکرد مناسب مدل می‌باشد (Behrens *et al*., 2010).

یافته‌ها و بحث

انتخاب متغیرهای کمکی (داده‌کاوی)

در جدول (۳) نتایج داده‌کاوی، جهت ورود مناسب‌ترین متغیرهای محیطی برای مدل‌سازی بر اساس آنالیز فاکتور تورم واریانس ارائه شده است. بر اساس این فاکتور از میان ۱۵ متغیر کمکی مورداستفاده، در نهایت شش متغیر کمکی به‌عنوان بهینه‌ترین پارامترها انتخاب گردیده‌اند که نسبتی برابر از میان متغیرهای ژئومورفومتری و شاخص‌های سنجش از دور را به خود اختصاص داده‌اند. از میان پارامترهای وابسته به توپوگرافی، مدل رقومی ارتفاع، ارتفاع‌های استاندارد و مساحت حوزه آبخیز دارای بیشترین اهمیت بودند. Mosleh *et al*. (2016) اظهار نمودند که پارامترهای مستخرج از مدل رقومی ارتفاع، حتی در مناطق دارای شدت پستی‌وبلندی کم، به‌عنوان متغیرهای محیطی مناسب در مدل‌سازی کلاس‌ها و خصوصیات خاک‌ها محسوب می‌شوند. Taghizadeh-Mehrjardi *et al*. (2015) پارامترهای ژئومورفومتری را به‌عنوان متغیرهایی مناسب جهت مدل‌سازی کلاس‌های خاک در سطح فامیل خاک گزارش نمودند. Afshar *et al*. (2018) بیان داشتند که ویژگی‌های مختلف مدل رقومی ارتفاع، هم به لحاظ منطقی و ریاضی و هم از دیدگاه تجربی، دارای همبستگی خوبی با کلاس‌های خاک می‌باشند و تا حد زیادی موجب افزایش صحت مطالعات نقشه‌برداری رقومی خاک‌ها می-شوند. همچنین شاخص‌های سنجش از دور مانند شاخص شوری (SI)، شاخص گیاهی تفاضلی نرمال شده (NDVI) و شاخص

1Diagonal Elements
2Marginal Totals
3Producers accuracy
4Users accuracy

گیاهی تفاضلی (DVI) دارای تأثیر قابل‌توجهی در مدل‌سازی و پیش‌بینی کلاس‌های خاک می‌باشند. Boettinger (2010) شاخص‌های مستخرج از تصاویر سنجش از دور را به‌عنوان مهمترین پارامترها در خصوص بارز نمودن نقش ارگانیسم (میزان و نوع پوشش گیاهی) در مدل‌سازی مکانی کلاس‌های خاک گزارش نمود.

جدول ۳: نتایج حاصل از فاکتور تورم واریانس جهت انتخاب متغیرهای بهینه

متغیر کمکی	نوع متغیر	توصیف	رفرنس
NDVI	انعکاس طیفی	$(NIR-R)/(NIR+R)$	Rouse *et al*., (1973)
SI	انعکاس طیفی	(G/NIR)	Abbas and Khan, (2007)
DVI	انعکاس طیفی	$(NIR-R)$	Tucker, (1979)
DEM	ژئومورفومتری	مدل رقومی ارتفاع	ALOS PLASAR, (2011)
CHA	ژئومورفومتری	$CHA=VolQdirect/Peff$	Wilson, (2018)
STH	ژئومورفومتری	$STH=NH*[(Zs-Z\min)-Z\min]$	Guo *et al*, (2019)

نکته: R,G, NIR به ترتیب باندهای ۵، ۴ و ۳ لندست ۸، Zs: ارتفاع مطلق، Zmin: ارتفاع حداقل، Vol Qdirect: حجم کل دبی مستقیم حوزه (m^3)، Peff: بارندگی مؤثر (m^3/m^2)

تشریح خاک‌های موردمطالعه

به‌طور کلی خاک‌های منطقه در سه رده انتی‌سولز، اینسپتی‌سولز و اریدی‌سولز قرار دارند. همچنین ۱۳ کلاس در سطح زیرگروه و ۲۲ کلاس در سطح فامیل شناسایی گردید (جدول ۴ و ۵). زیرگروه خاک ۱ (Fluventic Haploxerepts) با ۳۱/۵۲ درصد و زیرگروه کلاس ۴ (Lithic Xerorthents) با ۰/۲۰ درصد از کل اراضی منطقه به ترتیب بیشترین و کمترین مساحت را در بین خاک‌های شناسایی‌شده را شامل می‌شوند. فراوانی خاک‌های موجود در زیرگروه ۱ نشان‌دهنده این است که وقوع فرآیندهای فرسایش و رسوب منشأ اصلی تشکیل و تکامل خاک‌های منطقه موردمطالعه می‌باشند و از نظر پراکنش مکانی در سیمای اراضی دشت دامنه‌ای توزیع یافته‌اند. Farzamnia *et al*. (2015) در مطالعه خاکشناسی خاک‌های ارومیه گزارش نمودند که خاک‌های تشکیل‌شده به علت فرآیندهای متوالی فرسایش و رسوب‌گذاری دارای ضخامت زیاد و تحول پروفیلی کمی می‌باشند که منجر به رده‌بندی این خاک‌ها در رده غالب اینسپتی‌سول و زیرگروه Fluventic Haploxerepts و Typic Calcixerepts می‌گردند. خاک‌های منطقه موردمطالعه در سطح فامیل خاک نیز کلاس ۳ با رده‌بندی (Fine loamy, mixed, active, thermic Fluventic Haploxerepts) با ۲۵/۹۰ درصد بیشترین فراوانی را بین خاک-های شناسایی‌شده در این سطح به خود اختصاص داده است (جدول ۴).

جدول ۴- مساحت هر یک از زیرگروه‌های خاک با استفاده از مدل RF

کلاس خاک	رده‌بندی	مساحت (هکتار)	مساحت (درصد)
۱	Fluventic Haploxerepts	۵۲۴۱/۸۸	۳۱/۵۲
۲	Gypsic Aquisalids	۱۶۳/۱۸	۰/۹۸
۳	Gypsic Haplosalids	۱۶۳/۵۰	۰/۹۷
۴	Lithic Xerorthents	۳۲/۹۷	۰/۲۰
۵	Sodic Xeric Calcigypsids	۱۲۰۷/۵۸	۷/۲۶
۶	Sodic Xeric Haplocalcids	۱۷۶۰/۷۴	۱۰/۵۹
۷	Typic Calcixerepts	۳۲۲۷/۵۷	۱۹/۴۱
۸	Typic Haplocalcids	۷۳۴/۱۴	۴/۴۲
۹	Typic Haploxerepts	۸۸/۰۰	۰/۵۳
۱۰	Typic Xerorthents	۱۴۸۷/۴	۸/۹۵
۱۱	Xeric Calcigypsids	۵۴۵/۸۸	۳/۲۸
۱۲	Xeric Haplocalcids	۱۰۷۹/۶۴	۶/۴۹
۱۳	Xerofluvents Haplocambids	۸۹۸/۵۲	۵/۴۰
مجموع	---	۱۶۶۲۸	۱۰۰

جدول ۵- مساحت هر یک از کلاس‌های فامیل خاک با استفاده از مدل RF

لاس خاک	رده‌بندی	مساحت (هکتار)	مساحت(درصد)
۱	Fine, mixed, active, thermic, Xeric Calcigypsids	۴۰۰/۷۸	۲/۴۱
۲	Fine, mixed, active, thermic Sodic Xeric Calcigypsids	۳۳۱/۰۰	۱/۹۹
۳	Fine loamy, mixed, active, thermic Fluventic Haploxerepts	۴۳۰۶/۵۷	۲۵/۹۰
۴	Fine loamy, mixed, active,thermic Typic Calcixerepts	۵۵۲/۸۵	۳/۳۲
۵	Fine, mixed, active, thermic Xeric Haplocalcids	۹۰۰/۴۲	۵/۴۲
۶	Loamy skeletal, mixed, active, thermic Fluventic Haploxerepts	۳۴۶/۴۶	۲/۰۸
۷	Very Fine, mixed, active, thermic Xerofluventic Haplocambids	۱۰۴۶/۰۵	۶/۲۹
۸	Fine, mixed, active, thermic Xeric Haplocalcids	۱۰۹/۰۸	۰/۶۶
۹	Fine, mixed, subactive, thermic Typic Haplocalcids	۳۲۸/۷۰	۱/۹۸
۱۰	Fine, mixed, thermic,Gypsic Aquisalids	۱۰۷۱/۳۶	۶/۴۴
۱۱	Fine, mixed, active, thermic Sodic Xeric Haplocalcids	۲۱۹۰/۲۲	۱۳/۱۷
۱۲	Loamy over fragmental, mixed, calcareous, thermic Typic Xerorthents	۲۳۹۹/۱۸	۱۴/۴۳
۱۳	Loamy skeletal, carbonatic, thermic Typic Calcixerepts	۱۰۹۶/۲۷	۶/۵۹
۱۴	Fine loamy over fragmental, mixed, active, thermic Typic Calcixerepts	۵۷۶/۹۹	۳/۴۲
۱۵	Loamy skeletal, carbonatic, thermic Lithic Xerorthents	۵۲/۱۰	۰/۳۱
۱۶	Fine loamy over skeletal, carbonatic, thermic Typic Calcixerepts	۵۷/۵۳	۰/۳۵
۱۷	Fine, mixed, semiactive, thermic, Gypsic Haplosalids	۱۴۰/۱۴	۰/۸۴
۱۸	Coarse loamy over fragmental, mixed, subactive thermic, Typic Calcixerepts	۵۵/۳۹	۰/۳۳
۱۹	Loamy skeletal over clayey, mixed, active, themic Typic Calcixerepts	۳۲۵/۱۶	۱/۹۶
۲۰	Clayey over coarse loamy, mixed, active, thermic Xerofluventic Haplocambids	۹۸/۳۹	۰/۵۹
۲۱	Coarse loamy over fragmental, mixed, active, thermic Fluventic Haploxerepts	۱۶۶/۷۱	۱/۰۰
۲۲	Fine, mixed, active, thermic Typic Calcixerepts	۸۵/۶۷	۰/۵۲
مجموع	------	۱۶۶۲۸	۱۰۰

مقایسه دقت مدل‌های مورداستفاده

مقادیر صحت سنجی پیش‌بینی مکانی هر یک از کلاس‌های خاک بر اساس دو آماره صحت کلی و شاخص کاپا برای دو سطح زیرگروه و فامیل خاک (جدول ۶) نشان می‌دهد که مدل جنگل-های تصادفی نسبت به روش درخت تصمیم توسعه‌یافته به ترتیب با OA و Kappa، ۷۲ و ۶۵ در سطح زیرگروه و ۷۰ و ۶۰ در سطح فامیل دارای صحت بیشتری است. (Khamoshi *et al*. 2019) در نقشه‌برداری رقومی خاک‌های دشت قزوین با استفاده از مدل RF در سطح فامیل خاک، شاخص کاپا را ۸۳ درصد گزارش نمودند. شاخص صحت کاربر و تولیدکننده برای کلاس‌های خاک غالب در دو سطح زیرگروه و فامیل خاک در روش RF نشان داد که زیرگروه فلوونتیک هاپلوزرپتز دارای PA و UA ۱۰۰ درصد و کلاس فامیل سه، دارای PA و UA، ۸۶ و ۱۰۰ درصد می‌باشند که این نتیجه بیانگر این است که متغیرهای کمکی مناسبی جهت مدل‌سازی مورداستفاده قرار گرفته‌اند (Afshar *et al*., 2018). در روش درخت تصمیم مقادیر PA و UA برای زیرگروه فلونتیک هاپلوزرپتز ۶۰ و ۱۰۰ درصد و در سطح فامیل خاک با کلاس غالب منطقه به ترتیب PA و UA ۵۰ و ۸۰ درصد حاصل گردید

که نشان‌دهنده توانمندی بیشتر مدل RF در پیش‌بینی مکانی هر یک از کلاس‌های خاک به‌صورت منفرد نسبت به روش BRT بوده است (جدول ۷). دقت تولیدکننده و کاربر، به ما اجازه می‌دهد که سطح کم‌برآورد و بیش برآورد را در پیش‌بینی کلاس‌های خاک تخمین بزنیم (Lacoste *et al*., 2011). به‌طور کلی نتایج صحت مدل‌سازی نشان داد که از سطح زیرگروه به سمت فامیل خاک از میزان دقت پیش‌بینی مدل‌سازی در هر دو روش کاسته شده است. نتایج پژوهش Jafari *et al*. (2012) در جنوب ایران و Rad *et al*. (2016)در خاک‌های لسی شمال ایران نشان دادند که از سطح گروه بزرگ به سمت سری خاک، صحت کلی پیش‌بینی مکانی کلاس‌های خاک کاهش می‌یابد و در راستای نتایج پژوهش حاضر است. Zeraatpisheh *et al*. (2017) نیز در مطالعه خود بیان داشتند که روش جنگل تصادفی در سطح گروه بزرگ و زیرگروه خاک نسبت به سایر مدل‌های مورداستفاده در پیش‌بینی مکانی خاک‌های منطقه عملکرد بهتری داشته است. در مطالعات متعدد دیگری نیز به دلیل افزایش تنوع در کلاس‌های خاک در سطوح پایین‌تر تاکسونومیک، کاهش صحت پیش‌بینی مکانی گزارش شده است (Brungard *et al*., 2015; Taghizadeh-Mehrjardi *et al*., 2015). همچنین نتایج مدل RF نشان می‌دهد که مقدار خطای تخمین[1](OOB) با افزایش میزان صحت کلی و شاخص کاپا دارای یک رابطه معکوس می‌باشد به نحوی که مقدار OOB در سطح زیرگروه و فامیل به ترتیب ۶۲/۵ و ۷۵/۲۵ درصد مشاهده گردید (جدول ۶). Rad *et al*. (2015) خطای تخمین ۷۳ درصد را در سطح زیرگروه برای مطالعه خود با روش RF در استان گلستان گزارش نمودند. Brungard (2009) مشاهده کرد که کلاس‌های خاکی که دارای مشاهدات خاکرخ بیشتری هستند دارای خطای تخمین کمتری می‌باشند. Jafari *et al*. (2012) گزارش نمودند که کلاس‌های خاکی که نمونه کمتری دارند دارای خطای پیش‌بینی بیشتری هستند. بنابراین در این مطالعه نقشه پیش‌بینی مکانی کلاس‌های خاک بر اساس مدل RF که دارای بیشترین صحت نسبت به مدل BRT بود، تهیه و ارائه شده است (شکل ۲-الف وب).

جدول ۶- دقت پیش‌بینی سطوح تاکسونومیک زیرگروه و فامیل توسط الگوریتم‌های یادگیرنده

مدل یادگیری ماشین	شاخص صحت سنجی	زیرگروه	فامیل
RF	% Kappa	۶۵	۶۰
	OA	۷۲	۷۰
	%OOB	۶۲/۵	۷۵/۲۵
BRT	% Kappa	۳۷	۲۶
	OA	۴۸	۳۵

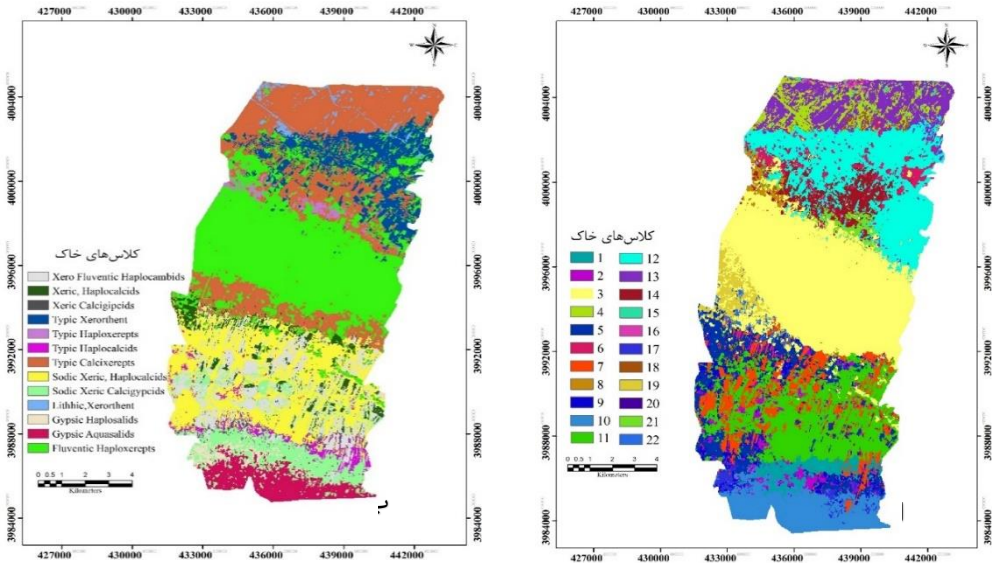
جدول ۷- فراوانی، صحت تولیدکننده و صحت کاربر برای فراوان‌ترین کلاس‌های خاک بر اساس مدل‌های برازش داده‌شده

کلاس خاک	درصد فراوانی	PA*[2]%		UA*[3]%	
		RF	BRT	RF	BRT
Fluventic Haploxerepts	۳۱/۲۵	۱۰۰	۶۰	۱۰۰	۱۰۰
Fine Loamy, mixed, active, thermic, Fluventic Haploxerepts	۲۵/۹	۸۶	۵۰	۱۰۰	۸۰

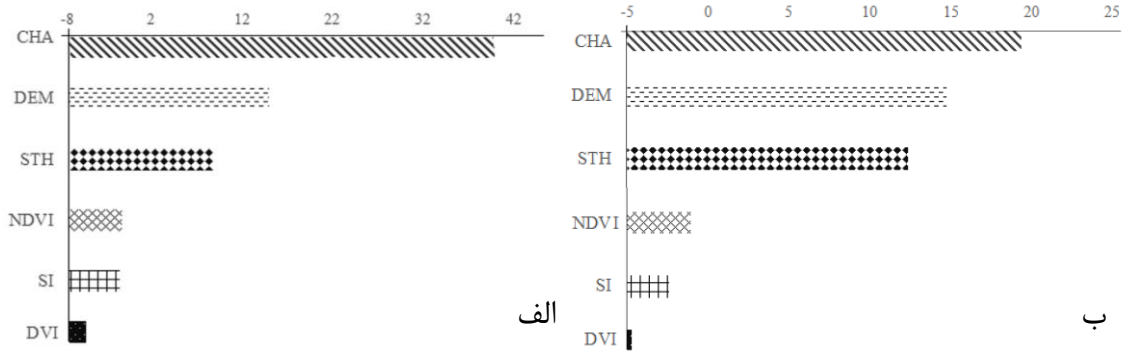
اهمیت نسبی متغیرهای کمکی

در این مطالعه جهت تعیین میزان ارجحیت هر یک از متغیرهای کمکی مورداستفاده از شاخص MDA استفاده گردید. RF به دو روش میانگین کاهش صحت[4](MDA) و میانگین کاهش جینی[5] (MDG) اهمیت متغیرها را نشان می‌دهد. در روش MDA، مقادیر درست متغیرها با مقادیری که به‌طور تصادفی برای هر درخت تولید شده است جایگزین می‌شود و اگر این جایگزینی اثری روی خطای اندازه‌گیری نداشته باشد اهمیت آن کم است و اگر مقدار خطای اندازه‌گیری افزایش یابد، آن متغیر مهم است. اهمیت نسبی متغیرهای کمکی مورداستفاده در مدل‌سازی مکانی زیرگروه و فامیل خاک در شکل (۳- الف و ب) ارائه شده است.

1- Out of Bag
2- Producer accuracy
3- Users accuracy
4- Mean Decrease in Accuracy
5- Mean Decrease in Gini



شکل ۲- پراکنش مکانی کلاس‌های فامیل خاک (الف) و زیرگروه‌های خاک (ب) با استفاده از مدل RF



شکل ۳- مقادیر عددی اهمیت متغیرهای کمکی در پیش‌بینی کلاس‌های فامیل خاک (الف) و زیرگروه خاک (ب) بر اساس روش آنالیز حساسیت MDA در مدل جنگل‌های تصادفی

بر اساس نتایج ارائه‌شده در شکل (۳-الف و ب) در دو سطح تاکسونومیک به ترتیب اهمیت نسبی متغیرهای کمکی CHA> DEM> STH> NDVI> SI> DVI از یک روند صعودی به نزولی برخوردار می‌باشند؛ بنابراین پارامترهای ژئومورفومتری اهمیت بیشتری نسبت به شاخص‌های سنجش دور در پیش‌بینی مکانی کلاس‌های خاک منطقه بر عهده داشته‌اند. از سایر نتایج قابل‌توجه در این تحقیق می‌توان بیان نمود علی‌رغم اینکه منطقه موردمطالعه از شدت پستی‌وبلندی چندانی برخوردار نیست (بیش از ۸۵ درصد سیمای اراضی دشت دامنه و دشت) با این حال پارامترهای ژئومورفومتریک بیشترین اهمیت در مدل‌سازی کلاس‌های خاک را در این منطقه دارند. در مناطق با شدت پستی‌وبلندی کم و متأثر از فرآیندهای فرسایش و رسوب ناشی از مناطق فوقانی شیب، متغیر کمکی CHA نشان داد که می‌تواند تغییرات خاک‌ها را حتی در مناطق مشاهده نشده با قدرت بالایی پیش‌بینی نماید (Rad *et al.*, 2014). در مطالعات دیگری (Mirakzehi *et al.*,2018 و Brungard *et al.*, 2015) شبکه آبراهه‌های زهکشی در حوزه‌های آبخیز موردمطالعه را به‌عنوان مهم‌ترین متغیر کمکی در پیش‌بینی مکانی کلاس‌های خاک در مناطق با شدت پستی‌وبلندی کم گزارش نمودند. Barthold *et al.* (2013) با استفاده از مدل RF به پیش‌بینی کلاس‌های خاک در چین پرداختند و گزارش نمودند که دو فاکتور اقلیم و کاربری اراضی از مهمترین عوامل تأثیرگذار بر روی پراکنش خاک‌های منطقه موردمطالعه می‌باشند. Mosleh *et al.*, (2017) در پژوهش خود به منظور مدل‌سازی کلاس‌های خاک از سطوح تاکسونومیک رده تا فامیل در دشت شهرکرد، پارامترهای ژئومورفومتری را مهمترین متغیرهای کمکی گزارش نمودند.

نتیجه‌گیری کلی

در این مطالعه از الگوریتم‌های یادگیرنده جنگل تصادفی و رگرسیون درختی توسعه‌یافته برای نقشه‌برداری رقومی کلاس-های خاک استفاده گردید. نتایج صحت سنجی دو آماره صحت عمومی و شاخص کاپا بدست آمده از ماتریس خطای ۲۰ درصد خاکرخ‌ها در دو سطح زیرگروه و فامیل خاک، بیانگر برتری مدل RF نسبت به BRT در پیش‌بینی مکانی می‌باشد. مدل جنگل تصادفی در مقایسه با رگرسیون درختی توسعه‌یافته به دلیل استفاده از تعداد درخت بیشتر و روش بهینه نمونه‌برداری بوت استراپ از متغیرهای کمکی و نقاط مشاهداتی صحت بالاتری را در پیش‌بینی و نقشه‌برداری مکانی کلاس‌های خاک ارائه داد. متغیرهای محیطی انتخاب‌شده توسط رویکرد نوین انتخاب متغیر VIF از دو مجموعه پارامترهای ژئومورفومتری و شاخص‌های انعکاس طیفی در پیش‌بینی مکانی پراکنش کلاس‌های زیرگروه و فامیل خاک خصوصاً در استفاده از الگوریتم یادگیرنده جنگل‌های تصادفی مؤثر واقع گردیدند و همچنین نتیجه آنالیز حساسیت مدل RF بیانگر اهمیت نسبی بیشتر پارامترهای ژئومورفومتری نسبت به شاخص‌های سنجش از دور بوده به طوری که در دو سطح پیش‌بینی کلاس خاک دارای روند مشابهی مشاهده گردید و متغیر کمکی مساحت حوزه آبخیز اصلاح‌شده[1]بیشترین تأثیر را در پیش‌بینی به خود اختصاص داد. لذا در مطالعات آتی نقشه-برداری در نواحی با پستی‌وبلندی کم خصوصاً در اراضی با کاربری کشاورزی با استفاده از پارامترهای ژئومورفومتری مناسب از قبیل متغیرVIF و الگوریتم یادگیرنده RF می‌توان نقشه‌ی پراکنش مکانی کلاس‌های فامیل خاک را با صحت قابل‌قبول جهت هرگونه بهره‌برداری تهیه نمود. همچنین پیشنهاد می‌گردد از رویکردهای نوین نمونه‌برداری مانند ابر مکعب لاتین مشروط برای انجام مطالعات نقشه‌برداری رقومی استفاده گردد.

REFERENCES

Afshar, F. A., Ayoubi, S., and Jafari, A. (2018). The extrapolation of soil great groups using multinomial logistic regression at regional scale in arid regions of Iran. *Geoderma*, 315, 36-48.

Baghche Maryam, M.M. and Shekaari, P.(2018). Soil Distribution Pattern Analysis in a Low Relief Area Using Decision Trees Algorithm. *Journal of Water and soil Research. Iran*. P 463-480.

Barthold, F. K., Wiesmeier, M., Breuer, L., Frede, H. G., Wu, J., and Blank, F. B. (2013). Land use and climate control the spatial distribution of soil types in the grasslands of Inner Mongolia. *Journal of Arid Environments*, *88*, 194-205.

Behrens, T., Zhu, A. X., Schmidt, K., and Scholten, T. (2010). Multi-scale digital terrain analysis and feature selection for digital soil mapping. *Geoderma*, 155(3-4), 175-185.

Boettinger, J. L. (2010). Environmental covariates for digital soil mapping in the western USA. In Digital Soil Mapping (pp. 17-27). *Springer*, Dordrecht.

Breiman, L. (2001) *Random forests*. Machine learning, 45(1), 5-32.

Breiman, L. and Cutler, A. (2004) *Random Forests*. Department of Statistics, University of Berkeley. http://www.stat.berkeley.edu/~breiman/Random Forests/cc_home.htm.

Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A. and Edwards Jr, T. C. (2015). Machine learning for predicting soil classes in three semi-arid landscapes. *Geoderma*, 239, 68-83.

Byrt, T., Bishop, J. and Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of clinical epidemiology*, 46(5), 423-429.

Chavez, P. S., Berlin, G. L. and Sowers, L. B. (1982) *Statistical method for selecting Landsat MSS. J. Appl*. Photogr. Eng, 8(1), 23-30.

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote sensing of environment*, 37(1), 35-46.

Death, G. (2007). *Boosted trees for ecological modeling and prediction*. Ecology 88 (1), 243–251.

Elith, J., Leathwick, J. R., and Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4), 802-813.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carre, G., and Munkemuller, T. (2013) Collinearity: *a review of methods to deal with it and a simulation study evaluating their performance*. Ecography, 36(1), 27-46.

Farzamnia, P., Manafi, S., and Momtaz, H. R. (2015). Evolution of soils formed on Quaternary sediments in some parts of Urmia Plain. *journal of soil management and sustainable production*. Vol (5.2).

Freidman, J., Hastie, T., and Tibshirani, R. (2000) *Additive logistic regression: a statistical view of boosting* (with discussion and a rejoinder by the authors). The annals of statistics, 28(2), 337-407.

Gee, G. W., and Bauder, J. W. (1986) *Particle-size analysis 1*. Methods of soil analysis: Part 1—Physical and mineralogical methods, (methodsofsoilan1), 383-411.

Grinand, C., Arrouays, D., Laroche, B. and Martin, M.P. (2008). Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. *Geoderma* 143, 180–190.

McBratney, A. B., Santos, M. M. and Minasny, B. (2003). On digital soil mapping. *Geoderma*, 117(1), 3-52. Abbas, A. and Khan, S. (2007) *Using remote sensing techniques for appraisal of irrigated soil salinity*. In: MODSIM 2007: International Congress on Modelling and Simulation: Land, Water and Environmental Management: Integrated Systems for Sustainability, pp. 2632–2638.

Grunwald, S., Thompson, J. A. and Boettinger, J. L. (2011). Digital soil mapping and modeling at continental scales: Finding solutions for global

[1] -Modified catchment area

issues. *Soil Science Society of America Journal*, *75*(4), 1201-1213.

Hengl, T., Toomanian, N., Reuter, H. I. and Malakouti, M. J. (2007). Methods to interpolate soil categorical variables from profile observations: lessons from Iran. *Geoderma*, 140(4), 417-427.

Iran Meteorological Organization. (2013). Climate Information, *Qazvin synoptic station*: Qazvin, Iran. Available at: http://www.irimo.ir/eng/index.php.

Jafari, A., Finke, P. A., Vande Wauw, J., Ayoubi, S. and Khademi, H. (2012). Spatial prediction of USDA-great soil groups in the arid Zarand region, Iran: comparing logistic regression approaches to predict diagnostic horizons and soil types. *European Journal of Soil Science*, 63(2), 284-298.

Jenny, H. (1994) *Factors of soil formation*: a system of quantitative pedology. Courier Corporation.

Khamoshi, S.E, Sarmadian, F and Keshavarzi A. (2019). Digital Soil Mapping Using Random Forests Model in Abyek, Qazvin Province. *Soil Research Journal*. No 3. P 394-403.

Lacoste, M., Lemercier, B. and Walter, C. (2011). Regional mapping of soil parent material by machine learning based on point data. *Geomorphology*, 133(1-2), 90-99.

Lagacherie, P. (1992) Formalisation des lois de distribution des sols pour automatiser la cartography pedologique a partir d'un secteur pris comme reference: cas de la petite region naturelle Moyenne Vallee de l'Héraut.

Landis, J. R. and Koch, G. G. (1977) *An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers*. Biometrics, 363-374.

Levi, M. R. and Rasmussen, C. (2014). Covariate selection with iterative principal component analysis for predicting physical soil properties. *Geoderma*, 219, 46-57.

Mirakzehi, K., Pahlavan-Rad, M. R., Shahriari, A. and Bameri, A. (2018). Digital soil mapping of deltaic soils: a case of study from Hirmand (Helmand) river delta. *Geoderma*, 313, 233-240.

Mosleh, Z., Salehi, M. H., Jafari, A., Borujeni, I. E. and Mehnatkesh, A. (2017). Identifying sources of soil classes variations with digital soil mapping approaches in the Shahrekord plain, Iran. *Environmental earth sciences*, 76(21), 748.

Mosleh, Z., Salehi, M. H., Jafari, A., Borujeni, I. E. and Mehnatkesh, A. (2016). The effectiveness of digital soil mapping to predict soil properties over low-relief areas. *Environmental monitoring and assessment*, 188(3), 195.

Mousavi, S. R., Sarmadian, F., Alijani, Z. and Taati, A. (2017). Land suitability evaluation for irrigating wheat by geopedological approach and geographic information system: A case study of Qazvin plain, Iran. *Eurasian Journal of Soil Science*, 6(3), 275.

Nelson RE (1982) Carbonate and gypsum. In: Page AL (ed) *Methods of soil analysis*. American Society of Agronomy, Madison, pp 181–197.

Olaya, V. I. C. T. O. R. (2004). *A gentle introduction to SAGA GIS*. The SAGA User Group eV, Gottingen, Germany, 208.

Rad, M. R. P., Khormali, F., Tomanian, N., Kiani, F. and Kamli, B. (2015). Digital soil mapping using Random Forest model in Golestan province. *Water and soil conservation Journal's*. 73-93.

Rad, M. R. P., Khormali, F., Toomanian, N., Brungard, C. W., Kiani, F., Komaki, C. B. and Bogaert, P. (2016). Legacy soil maps as a covariate in digital soil mapping: a case study from Northern Iran. *Geoderma*, 279, 141-148.

Rad, M. R. P., Toomanian, N., Khormali, F., Brungard, C. W., Komaki, C. B. and Bogaert, P. (2014). Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. *Geoderma*, 232, 97-106.

Rasouli, A. A. (2008) *Principles of Applied Remote Sensing with Emphasis on Satellite Image Processing*. Presses University of Tabriz.

Schloeder, C. A., Zimmerman, N. E. and Jacobs, M. J. (2001). Comparison of methods for interpolating soil properties using limited data. *Soil Science Society of America Journal*, *65*(2), 470-479.

Schoeneberger, P.J., Wysocki, D.A. and Benham, E.C. (2012) *Soil Survey Staff. Field book for describing and sampling soils*, 3nd version. Natural Resources Conservation Service. National Soil Survey Center, Lincoln.

Soil Survey Staff. (2014) *Keys to soil taxonomy*. 12th edn. USDANatural Resources Conservation Service, Washington, DC

Sreenivas, K., Dadhwal, V. K., Kumar, S., Harsha, G. S., Mitran, T., Sujatha, G., and Ravisankar, T. (2016). Digital mapping of soil organic and inorganic carbon status in India. *Geoderma*, 269, 160-173.

Sumner, M. E. and Miller, W. P. (1996). Cation exchange capacity and exchange coefficients. *Methods of soil analysis part 3*—chemical methods, (methodsofsoilan3), 1201-1229.

Taghizadeh-Mehrjardi, R., Nabiollahi, K., Minasny, B. and Triantafilis, J. (2015). Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. *Geoderma*, 253, 67-77.

Tesfa, T. K., Tarboton, D. G., Chandler, D. G. and McNamara, J. P. (2009). Modeling soil depth from topographic and land cover attributes. *Water Resources Research*, 45(10).

Thomas, P. J., Baker, J. C., Zelazny, L. W. and Hatch, D. R. (2000). Relationship of map unit variability to shrink–swell indicators. *Soil Science Society of America Journal*, *64*(1), 262-268.

U.S. Geology Survey. (2014). *Geology.com*/news/2010/free-lansatimages-from-USGS-2. http://glovis.usgs.gov.

Van Wambeke, A. R. (2000). *The Newhall Simulation Model for estimating soil moisture and temperature regimes*. Department of Crop and Soil Sciences. Cornell University, Ithaca, NY. USA.

Walkley, A. and Black, I. A. (1934) *An examination of the Degtjareff method for determining soil organic matter*, and a proposed modification of the chromic acid titration method. Soil science, 37(1), 29-38.

Yang, L., Jiao, Y., Fahmy, S., Zhu, A., Hann, S., Burt, J. E., and Qi, F. (2011). Updating conventional soil maps through digital soil mapping. Soil Science Society of America Journal, 75(3), 1044-1053.

Yemefack, M., Rossiter, D. G. and Njomgang, R. (2005). Multi-scale characterization of soil variability within an agricultural landscape mosaic system in southern Cameroon. *Geoderma*, *125*(1-2), 117-143.

Zeraatpisheh, M., Ayoubi, S., Jafari, A. and Finke, P. (2017). Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran. *Geomorphology*, 285, 186-204.

Zinck, J. A. (1988) *Physiography and soils*, ITC soil survey lecture notes. International Institute for Aerospace Survey and Earth Sciences, Enschede, 7.