پیش‌بینی کلاس‌های خاک با روش‌های درختان تصمیم گیری تصادفی و رگرسیون منطقی چند متغیره در استان گلستان

محمدرضا پهلوان راد [۱]، فرهاد خرمالی [۲]، نورایر تومانیان [۳]، فرشاد کیانی [۴]، بایرام کمکی [۵]
۱- استادیار بخش خاک و آب مرکز تحقیقات کشاورزی و منابع طبیعی سیستان، سازمان تحقیقات، آموزش و ترویج کشاورزی، ۲- استاد گروه علوم خاك دانشگاه علوم کشاورزی و منابع طبیعی گرگان، ۳- استادیار بخش خاک و آب مرکز تحقیقات کشاورزی و منابع طبیعی اصفهان، سازمان تحقیقات، آموزش و ترویج کشاورزی، ۴- استادیار گروه علوم خاك دانشگاه علوم کشاورزی و منابع طبیعی گرگان، ۵- استادیار گروه آبخیزداری و حفاظت خاک دانشگاه علوم کشاورزی و منابع طبیعی گرگان

چکیده
در این تحقیق دو روش درختان تصمیم‌گیری تصادفی و رگرسیون منطقی چند متغیره در پیش‌بینی کلاس‌های خاک در استان گلستان ارزیابی شدند. خصوصیات اراضی، نقشه واحدهای ژئومورفولوژی و شاخص پوشش گیاهی بهبودیافته و نقشه سری خاک موجود به عنوان متغیر های کمکی محیطی جهت تهیه نقشه کلاس‌های خاک در محدوده ۸۵۰۰۰ هکتار استفاده شدند. نتایج مقایسه دقت نقشه‌های تولید شده در مدل های مختلف نشان داد که در سطح زیرگروه، و سری روش درختان تصمیم‌گیری تصادفی برتری نسبی به روش رگرسیون منطقی چندتایی داشته و در سطح گروه بزرگ رگرسیون منطقی چند جمله‌ای دقت بیشتری داشته است. در سطح سری بالاترین دقت با درختان تصمیم‌گیری تصادفی به میزان ۴۰ درصد بدست آمد که ۱۰ درصد بیشتر از روش رگرسیون منطقی چندتایی بود. نتایج نشان داد که روش درختان تصمیم‌گیری تصادفی می‌تواند برای تهیه نقشه خاک های کارآمد در سطوح پایین‌تر طبقه‌بندی مورد استفاده قرار گیرد.

واژه‌های کلیدی: نقشه‌برداری رقومی خاک، درختان تصمیم‌گیری تصادفی، رگرسیون منطقی چند متغیره، استان گلستان

مقدمه
وجود نقشه‌های خاک از نیازهای اساسی در مطالعات علوم زمین می‌باشد. با وجودی که بیش از ۵۰ سال از شروع مطالعات خاکشناسی در ایران می‌گذرد، نقشه‌های خاک کشور به طور کامل تهیه نشده‌اند. در نقشه‌برداری مرسوم خاك دانش پدولوژی به طور دقیقی برای تولید نقشه‌های خاك استفاده می‌شود؛ اما هنگامی که با روش‌های در حال توسعه نقشه‌برداری رقومی خاک مقایسه می‌شوند زمان‌بر و هزینه‌بر نشان می‌دهند (Grunwald, ۲۰۱۰).
نقشه‌برداری رقومی خاك (DSM[۷۷]) شامل روش‌ها و مدل‌هایی است که بین توزیع خاک (کلاس‌ها یا خصوصیات خاك) و داده‌هایی که به آسانی و با قیمت ارزان از طریق روش‌های سنجش‌ازدور، تصاویر و عکس‌های ماهواره‌ای و داده‌های ژئومورفومتری بدست می‌آیند و تحت عنوان متغیرهای کمکی محیطی[۷۸] نامیده می‌شوند ارتباط برقرار می‌کند. اگر چه تکنیک‌های DSM بر روی محاسبات دیجیتالی استوار است ولی اساس آن بر روی معادلات تشکیل خاک (McBratney, et al., ۲۰۰۳ ؛ Jenny, ۱۹۴۱) قرار دارد.(۱۹۴۱) Jenny مدل معروف خود را که بر اساس آن خاك بر روی یك زمین‌نما تابعی از ۵ فاکتور محیطی است معرفی کرد. این مدل به صورت معادله $S= f (CL,O, R, P, T)$ تعریف شده است. (۲۰۰۳) McBratney, et al., مدل پدولوژیکی ClORPT را بازبینی کردند و پارامترهای خاک (کلاس ها و یا خصوصیات خاک) و موقعیت مکانی را به آن اضافه نموده و مدل جدیدی تحت عنوان مدل SCORPAN را برای کمی کردن ارتباط بین داده‌های دقیق مکانی و خاک ارائه کردند. اجزای این مدل تحت عنوان متغیرهای کمکی محیطی معرفی شده‌اند. مدل جدید به صورت $S= f (s, c, o, r, p, a, n)$ است که فاکتورهای s داده‌های صحرایی و یا آزمایشگاهی خاک در نقاط نمونه برداری یا در نقشه‌های خاک هستند و nموقعیت مکانی نقاط نمونه‌برداری می باشند که به پارامترهای مدل (۱۹۴۱) Jenny اضافه شدند.
مدل‌های مختلفی در نقشه‌برداری رقومی خاک مورد استفاده قرار می گیرند. مدل درختان تصمیم گیری تصادفی (RF) مدل توسعه یافته از مدل طبقه‌بندی و رگرسیون درختی (CART[۷۹]) می‌باشد. روش CART روشی است که داده‌ها را به طور تکراری برای بدست آوردن ارتباط بین متغیر پاسخ و متغیرهای مستقل و انجام تخمین جداسازی می‌کند.(۲۰۰۹) Brungard، تکنیک RF را برای پیش بینی مکانی کلاس‌ها و خصوصیات خاک استفاده کرد؛. وی دریافت که RF یک روش مفید برای پیش بینی کلاس‌های خاک، خصوصیات خاک و توزیع پوشش گیاهی است. (۲۰۱۳) Barthold et al., با استفاده از RF توزیع تیپ های مختلف خاک با روش رده‌بندی

[۷۷] Digital Soil Mapping
[۷۸] Environmental covariates
[۷۹] Classification and regression trees

WRB[80] را در سطحی حدود ۳۶۰۰ کیلومتر مربع با ۱۴۵ پروفیل در چین مورد ارزیابی قرار دادند و مشاهده کردند در مطالعه آنها مقدار خطای پیش بینی مدل برای سطح اول رده بندی روش WRB برابر ۶/۵۱ درصد بود.

مدل دیگر استفاده شده برای نقشه برداری کلاس خاک ها روش رگرسیون منطقی چند جمله‌ای[81] می‌باشد. روش رگرسیون منطقی چند جمله‌ای به طور موفقی برای پیش‌بینی خصوصیات خاک در مقایسه با پیش‌بینی کلاس‌های تاکسونومی خاک استفاده شده است. این روش می‌تواند برای پیش‌بینی کلاس‌های خاک مورد استفاده قرار گیرد (Marchetti ۲۰۱۱ et al.،). Jaffari (۲۰۱۲) et al.، روش رگرسیون منطقی دوتایی[82] و رگرسیون منطقی چند جمله‌ای را با ۱۲۶ پروفیل در سطحی ۹۰۰۰۰ هکتار برای پیش بینی توزیع گروه‌های بزرگ در منطقه زرند کرمان مورد استفاده قرار دادند و مشاهده کردند که در منطقه مورد مطالعه رگرسیون منطقی چند جمله‌ای دقت بیشتری نسبت به روش رگرسیون منطقی دوتایی داشت. (Marchetti ۲۰۱۱ et al.،) در مطالعه تخمین تیپ‌های خاک در ایتالیا مشاهده کردند که روش رگرسیون منطقی چند جمله‌ای یک روش تخمین سریع و قابل اعتماد برای متغیرهای اسمی[83] مثل کلاس‌های خاک می‌باشد.

این تحقیق با هدف بررسی دقت روش‌های RF و MLR در تخمین کلاس‌های خاک در بخشی از اراضی استان گلستان انجام گرفت.

مواد و روش‌ها

محدوده مورد مطالعه دارای مساحت حدود ۸۵۰۰۰ هکتار و در استان گلستان واقع شده است (شکل ۱). این محدوده با عرض ۱۹ کیلومتر از شمال شهر گرگان شروع و تا ۴۵ کیلومتر بطرف شمال ادامه دارد. ارتفاع محدوده بین ۱۵۸ متر در جنوب منطقه مورد مطالعه تا ۱۸ متر زیر سطح دریا می باشد. مواد مادری منطقه به‌طور عمده رسوبات لس و لس های منتقل شده با آب هستند.

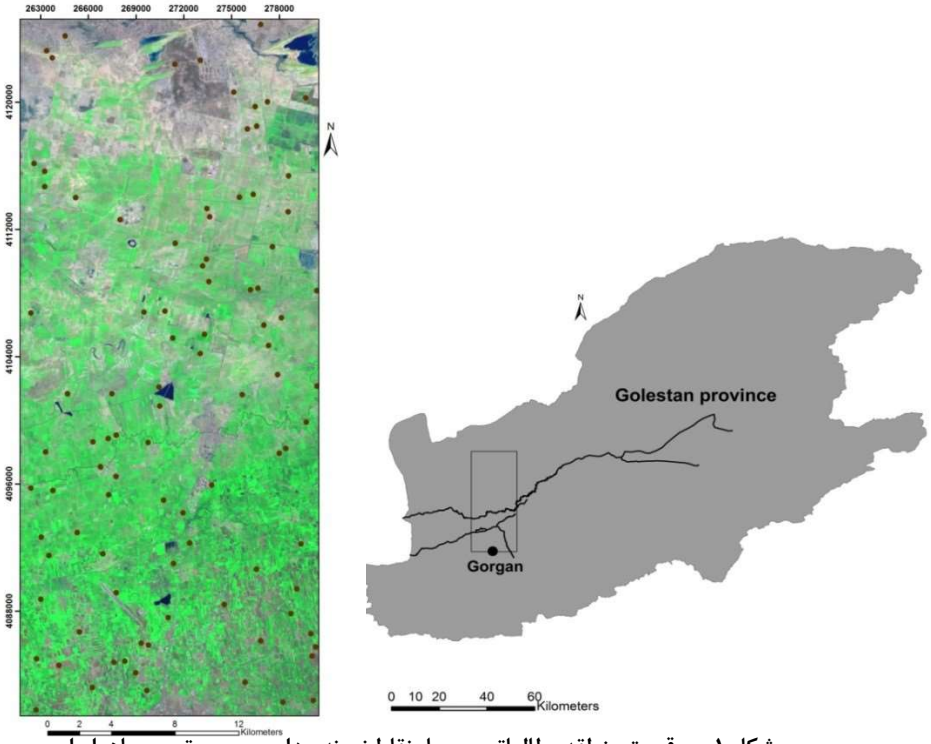
۱۳ خصوصیت متغیر محیطی که شامل خصوصیات زمین، شاخص پوشش گیاهی و واحدهای ژئومورفولوژی بودند در این مطالعه استفاده شد. خصوصیات زمین شامل شیب، ارتفاع، جهت، ، انحنای سطح[84]، انحنای نیمرخ[85]، شاخص حمل رسوب، شیب میانگین شیب بالایی[86]، طول جریان شیب پایینی[87]، بیشترین شیب به سمت پایین[88] و شاخص خیسی[89] از نقشه DEM با بزرگنمایی ۳۰ متر با استفاده از نرم افزارهای ArcGIS ۹.۳ و TAS[90] استخراج گردیدند. از SAVI به عنوان شاخص پوشش گیاهی استفاده شد چون به زمینه خاک حساسیت ندارد. همچنین از نقشه خاک موجود منطقه نیز به‌عنوان یک متغیر در مدل استفاده شد.

روش cLHS که توسط Minasny and McBratney (۲۰۰۶) معرفی شده است با استفاده از نرم افزار Matlab برای انتخاب محل‌های نمونه‌برداری پروفیل‌های خاک اجرا گردید. مقادیر همه متغیرهای کمکی در پیکسلهای ۳۰ در ۳۰ متر در نرم افزار ArcGIS بدست آمد و این مقادیر به نرم افزار Matlab منتقل و با تکنیک cLHS محل ۹۹ پروفیل تعیین و نمونه برداری انجام شد. پروفیل ها بر اساس دستور العمل نقشه‌برداری خاک و کلید طبقه‌بندی تا سطح فامیل طبقه بندی شدند. فامیل خاک به سری اصلی در نظر گرفته شد.

لایه‌های ۱۳ متغیر با پیکسل‌های ۳۰ متر مربعی و کلاس‌های خاک وارد نرم افزار (۲۰۱۳) R ۳.۰.۱ و R-Studio۰.۹۷.۵۵۱ گردید و مدل های RF و MLR اجرا گردید. در هر مدل نقشه نهاین بر اساس کمترین خطایی که هر مدل می دهد تولید شد. تعداد ۷۹ نمونه برای مدل سازی و ۲۰ نمونه برای اعتبارسنجی استفاده شد.

---

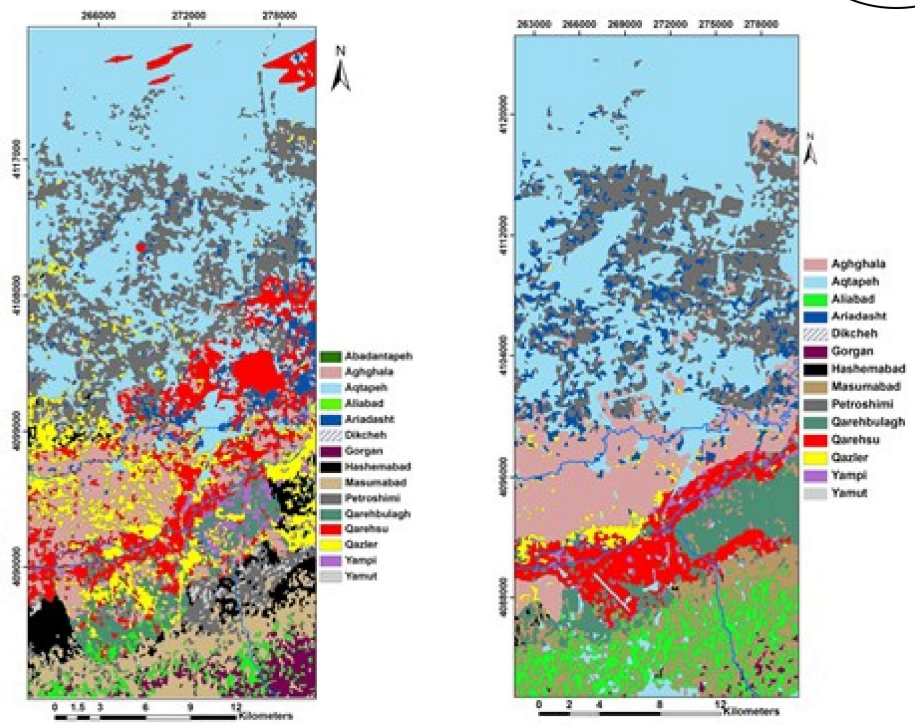
[80] World Reference Base for Soil Resources
[81] Multinomial logistic regression
[82] Binary logistic regression
[83] Categorical
[84] Plan Curvature
[85] Profile Curvature
[86] Mean upslope slope
[87] Downslope flowpath length
[88] Max downward slope
[89] Wetness index
[90] Terrain Analysis System



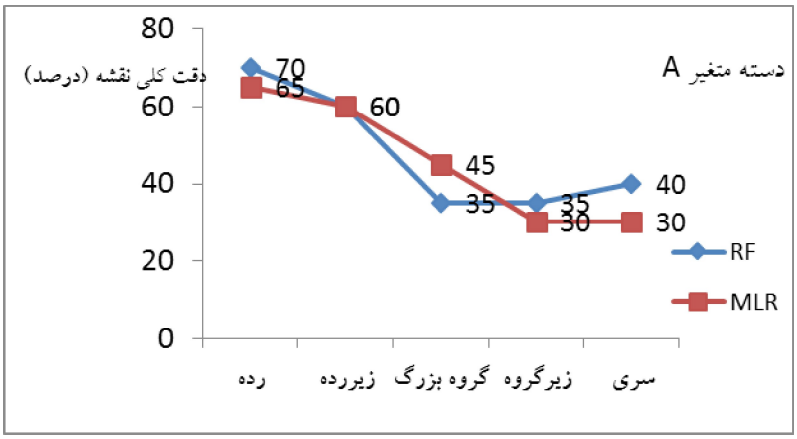
شکل ۱- موقعیت منطقه مطالعاتی و محل نقاط نمونه برداری بر روی تصویر ماهواره‌ای

نتایج و بحث

نقشه‌های تولید شده مربوط به سری خاک با کمترین خطای مربوط به هر مدل در شکل ۲ نشان داده شده است. مقایسه نتایج دقت کلی نقشه‌های تولید شده با هر دو روش درختان تصمیم‌گیری تصادفی و رگرسیون منطقی چند جمله‌ای در همه سطوح رده بندی در شکل ۳نشان داده شده است. نتایج نشان داد در سطح زیررده دقت هر دو روش یکسان بود. در سطح گروه‌بزرگ روش رگرسیون منطقی چند جمله‌ای نسبت به روش درختان تصمیم‌گیری تصادفی برتری داشت. در زیر گروه‌بزرگ دقت روش درختان تصمیم‌گیری تصادفی بیشتر بود . مقایسه نتایج دقت نقشه‌های تولید شده در سطح سری نشان داد که دقت نقشه در روش درختان تصمیم‌گیری تصادفی بیشتر بود و نشان می‌دهد که این تکنیک با افزایش تعداد کلاس‌ها تخمین بهتری نسبت به روش رگرسیون منطقی چند جمله‌ای می‌دهد و می‌تواند در سطوح پایین طبقه‌بندی مانند فامیل و سری به‌کار رود. مقدار دقت نقشه ۱۰درصد بیشتر از روش رگرسیون منطقی چند جمله‌ای بود که مقدار اختلاف قابل‌توجه‌ای در این تحقیق بود. بنابراین روش RF می تواند در تخمین نقشه‌های سطوح پایین‌تر رده‌بندی به‌کار رود. همچنین این نتایج نشان می‌دهد که در مطالعات تهیه نقشه خاک با روش رقومی خاک، نباید به یک تکنیک اتکا کرد و بهتر است از تکنیک‌های مختلف استفاده شود و نقشه نهایی بر اساس روشی که دارای دقت بالاتری است تهیه شود. Hueng et al., (۲۰۱۴) در نقشه‌برداری مواد مادری خاک با استفاده از درختان تصمیم‌گیری تصادفی مشاهده کردند که درختان تصمیم‌گیری تصادفی نقشه‌هایی با دقت بالا تولید کرد و یک روش موثر در ایجاد ارتباط بین توپوگرافی و مواد مادری خاک است.



شکل۲- نقشه تولید شده به روش **RF** (سمت راست) و **MLR** (سمت چپ)



شکل۳ -مقایسه دقت کلی روش‌های **RF** و **MLR** در سطوح مختلف رده‌بندی

منابع

Barthold, F.K., Wiesmeier, M., Breuer, L., Frede, H.G., Wu, J., Blank, F.B., ۲۰۱۳. Land use and climate control the spatial distribution of soil types in the grasslands of Inner Mongolia. Journal of Arid Environments ۸۸(۰), ۱۹۴-۲۰۵.

Brungard, C, W. ۲۰۰۹. Alternative Sampling and Analysis Methods for Digital Soil Mapping in Southwestern Utah. Thesis for Master of Science. Utah State University. USA.

Grunwald, S. ۲۰۱۰. Current State of Digital Soil Mapping and What Is Next. In: Boettinger, J. L., Howel, D. W., Moore, A. C., Hartemink, A. E., and Kienast-Brown, S., (eds.), Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer. Dordrecht Heidelberg London New York. Pp ۳-۱۲.

Hueng, R., Bulmer, C.B., Schmidt., M.G. ۲۰۱۴. Predictive soil parent material mapping at a regional-scale: A Random Forest approach. Geoderma. ۲۱۴-۲۱۵: ۱۴۱-۱۵۴.

Jafari, A., Finke, P.A., Van de Wauw, J., Ayoubi, S., Khademi, H., ۲۰۱۲. 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 ۶۳, ۲۸۴-۳۰۹.

Jenny, H. ۱۹۴۱. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, New York.

Marchetti, A., Piccini, C., Santucci, S., Chiuchiarelli, I., and Francaviglia, R., ۲۰۱۱. Simulation of soil types in Teramo province (Central Italy) with terrain parameters and remote sensing data. Catena. ۸۵: ۲۶۷-۲۷۳.

McBratney, A.B., Mendonça Santos, and M.L., Minasny, B. ۲۰۰۳. On digital soil mapping. Geoderma. ۱۱۷: ۳-۵۲.

Minasny, B., and McBratney, A.B. ۲۰۰۶. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences ۳۲(۹), ۱۳۷۸-۱۳۸۸.

Abstract

In this experiment Random forest and Multinomial logisic regression models were evaluated for predicting soil classes in Golestan province. Terrain attribute, SAVI, geomorphology map, land and conventional soil maps were used in ۸۵۰۰۰ ha. Comparison accuracy of the maps output in different models and category levels showed that in the level of subgroup and series random forest has more relative advantage and accuracy than multinomial logistic regression. In the level of great-group multinomial logistic regressions was more accurate. In the level of series, the most accuracy was obtained by random forest (۴۰%) and it is ۱۰% more than multinomial logistic regressions. Results showed that random forest could be used producing efficient maps with no large data in lower category levels.