استفاده از مدل میانگین (مدل تجمیعی) در نقشه برداری رقومی خاک

شاهرخ فاتحی [۱]

استادیار بخش تحقیقات خاک و آب، مرکز تحقیقات و آموزش کشاورزی و منابع طبیعی استان کرمانشاه، سازمان تحقیقات و آموزش ترویج کشاورزی، کرمانشاه

چکیده

یکی از راه‌های بهبود کارایی مدل‌سازی رقومی ویژگی‌های خاک که در نقشه‌برداری رقومی خاک مورد توجه قرار نگرفته‌است استفاده از مدل میانگین یا مدل تجمیعی می‌باشد. پژوهش حاضر، به منظور مقایسه مدل‌های پیش‌بینی منفرد مثل مدل رگرسیون خطی چندمتغیره، کوبیست و جنگل تصادفی با روش مدل میانگین از نوع ترکیب خطی تعمیم‌یافته انجام شده‌است. ابتدا در زیرحوضه آبخیز مرک واقع در استان کرمانشاه به وسعت ۲۴۰۰۰ هکتار، ۳۲۰ نمونه خاک سطحی به روش تصادفی جمع‌آوری گردید و پس از تعیین کربن آلی خاک، داده‌ها به دو دسته آموزشی (۲۴۰ نمونه) و اعتبارسنجی (۸۰ نمونه) تفکیک شدند. متغیرهای کمکی، مشتقات مدل رقومی ارتفاع و شاخص پوشش گیاهی بودند. ریشه میانگین مربعات خطا در مدل میانگین (۰/۵۵۸) پایین‌تر از مدل رگرسیون خطی چندمتغیره (۰/۵۸۶)، کوبیست (۰/۵۸۱) و جنگل تصادفی (۰/۵۷۵) به‌دست آمد. به طورکلی می‌توان بیان نمود استفاده از مدل میانگین در نقشه‌برداری رقومی خاک باعث افزایش صحت و دقت نقشه‌های پیش‌بینی می‌شود.

واژه های کلیدی: مدل‌سازی رقومی خاک، داده کاوی، مدل تجمیعی

مقدمه

نقشه‌برداری رقومی خاک را مک برتنی و همکاران (۲۰۰۳) با استفاده از معادله ینی (۱۹۴۱) که عوامل خاکسازی را توصیف می کند، صورت بندی نمودند. که اصطلاحاً مدل اسکورپن نامیده می‌شود و به صورت معادله زیر بیان می‌گردد:

$$S_c[x.y.\sim t].S_p[x.y.\sim t] = f(s[x.y.\sim t].c[x.y.\sim t].o[x.y.\sim t].r[x.y.\sim t].p[x.y.\sim t].a[x.y.\sim t].n) \quad (۱)$$

در این معادله، S_c= کلاس خاک، S_p= ویژگی‌های خاک،s= خاک‌ها (مشخصاتی از خاک در یک نقطه مشاهداتی)، c = اقلیم (خصوصیات اقلیمی محیط در یک نقطه مشاهداتی)، o= موجودات زنده (پوشش گیاهی، جانوران، یا فعالیت بشر)، r = توپوگرافی (مشخصات سیمای اراضی)، p= مواد مادری (سنگ شناسی)، a = سن (عامل زمان)، n = مکان (موقعیت مکانی)، t = زمان (اینجا منظور یک زمان تقریبی است)، x,y = مختصات مکانی و f = تابع یا تابع پیش‌بینی مکانی خاک می‌باشد. تابع پیش بینی مکانی در واقع روش‌های مختلف داده‌کاوی، یا زمین‌آماری و آماری هستند که برای استخراج روابط بین ویژگی‌های خاک و متغیرهای محیطی به کار می روند.

توابع پیش بینی مکانی یا همان روش‌های مدل‌سازی خاک بسته به ماهیت فرایند (خطی بودن و یا غیرخطی بودن روابط ویژگی‌های خاک و متغیرهای محیطی) و ویژگی داده‌ها (قابلیت دسترسی داده‌های کمّی و کیفی) به کار می‌روند. رگرسیون خطی چندمتغیره، مدل‌های خطی تعمیم یافته (McCullagh and Nelder, 1989)، درختان طبقه بندی و رگرسیون (Breiman et al., 1984)، شبکه عصبی مصنوعی (Anderson, 1995) ماشین بردار پشتیبان (Cortes and Vapnik, 1995) و الگوریتم ژنتیک (Koza, 1992) مدل‌های مرسوم و رایجی هستند که در نقشه‌برداری رقومی خاک مورد استفاده قرار می‌گیرند. برخی از این مدل‌ها علاوه بر پیچیدگی، خروجی‌هایی ارائه می‌دهند که تفسیر آنها کم و بیش مشکل است (مانند شبکه عصبی). امامدل‌های

دیگر، دارای پیچیدگی متفاوتی هستند مانند مدل کوبیست و جنگل تصادفی از خانواده مدل درختان طبقه‌بندی و رگرسیون، که به طور خودکار بر اساس عملکرد زیرمدل‌ها، متغیرهای پیش‌بینی کننده در آنها انتخاب می‌شود.

در پژوهش‌های نقشه‌برداری رقومی خاک معمولا از چند مدل آماری، داده‌کاوی یا زمین آماری برای تهیه نقشه پیش‌بینی ویژگی‌های خاک استفاده می‌شود و در نهایت یک مدل بر اساس نتایج اعتبار سنجی به عنوان مدل برتر و دقیق‌تر انتخاب می‌گردد. حال این سئوال مطرح می‌شود آیا استفاده ازترکیبی از مدل‌های مختلف نسبت به مدل‌های منفرد دقیق‌تر نخواهد بود؟

چنین وضعیتی در رشته‌های پژوهشی هیدرولوژی و اتمسفر مشاهده می‌شود دراین منظومه‌های علمی، پیش‌بینی‌های چندگانه از یک فرایند توسط تعدادی از مدل‌های پیش بینی کننده به دست می‌آید. هر مدل سهیم در پیش بینی، نقاط قوت و ضعف خاص خود را دارد. ترکیب این مدل ها مسلما بهتر از هرکدام از مدل پیش بینی به طور منفرد خواهد بود (Diks and Vrugt ,2010). در زمینه نقشه‌برداری رقومی خاک پادریان و همکاران (۲۰۱۴) و ملانو و همکاران (۲۰۱۴) از روش‌های مختلف مدل میانگین (مدل تجمیعی)[1] مانند مدل میانگین با اوزان معادل[2]، روش مدل میانگین واریانس وزن دار (روش باتز_گرانگر)[3]، میانگین روش مدل میانگین گرانگر-رامانتان[4] و روش مدل میانگین بیزی[5] برای پیش بینی ویژگی های خاک استفاده کرده‌اند. هدف از پژوهش حاضر استفاده از مدل میانگین یا مدل تجمیعی برای پیش‌بینی کربن آلی و مقایسه آن با با مدل‌های منفرد رگرسیون چند متغیره، کوبیست و جنگل تصادفی است.

مواد و روش‌ها

اراضی مورد مطالعه با وسعتی حدود ۲۴۰۰۰ هکتار در بیست کیلومتری جنوب شرقی شهرکرمانشاه و در بین مختصات جغرافیائی۴۷ درجه و ۴ دقیقه تا ۴۷ درجه و ۲۲ دقیقه طول شرقی و ۳۴ درجه تا ۳۴ درجه و ۹ دقیقه عرض شمالی قرارگرفته است. مطابق با آمار هواشناسی، متوسط درجه حرارت سالیانه‌ی هوا ۱۳/۲ درجه سانتیگراد و میانگین بارندگی سالیانه ۴۸۱/۴ میلیمتر است. بر اساس اقلیم نمای آمبرژه، اقلیم این ناحیه نیمه خشک و سرد می‌باشد. رژیم رطوبتی خاک زریک (xeric) و رژیم حرارتی خاک ترمیک (thermic) است. سنگ آهک، دولومیت، مارن، سنگ رس و ماسه سنگ، سنگ‌های اصلی سازندهای زمین شناسی محدوده مطالعاتی هستند. خاک‌های اینسپتی‌سولز، انتی‌سولز و ورتی‌سولز مهم‌ترین رده‌های خاک شناسایی شده در این ناحیه هستند (فاتحی،۱۳۸۷).

این پژوهش در طی مراحل زیر اجرا شده است:

جمع‌آوری داده‌های خاک

در این پژوهش محل ۳۲۰ نقطه مشاهداتی به روش تصادفی انتخاب گردید و نمونه‌های خاک سطحی (۳۰-۰ سانتیمتر) از این نقاط در سطح ۲۴ هزار هکتار منطقه برداشت شد و درصد کربن آلی آن‌ها به عنوان متغیر هدف به روش سوزانیدن تر اندازه‌گیری گردید.

متغیرهای محیطی کمکی

متغیرهای کمکی محیطی در این پژوهش به عنوان داده‌های ورودی پیش بینی کننده، با استفاده از مدل‌های رقومی ارتفاع با اندازه پیکسل ۱۰ متر، برخی مشتقات مراتب اول و دوم آن شامل ارتفاع، درصد شیب، تحدب، شاخص همواری درّه با درجه

[1] model averaging (ensemble model)
[2] Equal weights averaging (EW)
[3] Bates–Granger or variance weighted averaging (VW)
[4] Granger–Rama
There are no sources in the current document.nathan averaging (GRA)
[5] Bayesian model averaging (BMA)

تفکیک بالا[۶]، شاخص رسوب[۷]، شاخص خیسی[۸]، عمق دره[۹]، فاصله عمودی تا شبکه آبراهه[۱۰] و شاخص پوشش گیاهی نرمال شده با استفاده از نرم‌افزار ایلویس نسخه ۳/۸ و ساگا نسخه‌ی ۲/۲ تهیه گردید.

در این مطالعه از مدل‌های خطی چندمتغیره، کوبیست[۱۱] و جنگل تصادفی[۱۲] برای پیش‌بینی مقدار کربن آلی خاک استفاده شد و در مرحله بعد با استفاده از مدل خطی تعمیم یافته[۱۳] مدل‌های مذکور ترکیب و یک مدل میانگین یا مدل تجمیعی به دست آمد و کلیه مدل‌ها با روش اعتبارسنجی دوجانبه مورد ارزیابی قرار گرفت.

مدل های خطی چند متغیره

مدل رگرسیون خطی چند متغیره رابطه خطی بین متغیر هدف و حداقل دو متغیرپیش‌بینی کننده را نشان می‌دهد و اجرای آن نیازمند فرضیات مانند نرمال بودن توزیع و ثابت بودن واریانس متغیر هدف می‌باشد (McCullagh and Nelder, 1989).

مدل کوبیست

مدل کوبیست به دلیل توانایی آن در کشف روابط غیرخطی داده‌ها دارای ساختار عامه پسندی است و مسئله پیش‌بینی‌های محدودی که در دیگر مدل‌های درختان طبقه‌بندی و رگرسیون اتفاق می‌افتد را ندارد. در مدل کوبیست، ابتدا داده‌ها را با توجه متغیر هدف و متغیرهای کمکی به زیر مجموعه‌هایی باویژگی‌های مشابه، تفکیک می‌شوند. یک سری از قواعد زیرمجموعه‌های تفکیکی را معین می‌کنند. این قواعد به شکل سلسله مراتبی هستند. هر قاعده به شکل زیر است:

اگر{ شرایط واقعی است}

پس{ اجرای رگرسیون}

در غیر این صورت {اجرای قاعده بعدی}

شرایط ممکن است ساده و بر اساس یک متغیر کمکی باشد ولی اغلب تعدادی متغیر کمکی را شامل می‌شود. اگر شرایط واقعی باشد در مرحله بعد پیش‌بینی ویژگی خاک مورد نظر به وسیله اجرای مدل رگرسیون در درون هر زیر مجموعه تفکیکی انجام می‌گیرد. اگر شرایط درست (واقعی) نباشد بنابراین قاعده‌ای درگره بعدی درخت تعریف شده و توالی اگر، سپس، در غیر این صورت، تکرار می‌گردد. خروجی‌ها، حاصل از معادلات رگرسیونی هستند. مدل کوبیست براساس الگوریتم M5 کوینلان (۱۹۹۲) عمل می‌کند و در بسته نرم افزاری Cubist در محیط R اجرا گردید.

مدل جنگل تصادفی

بریمن(۲۰۰۱) الگوریتم جنگل تصادفی را پیشنهاد نمود. در درختان طبقه‌بندی و رگرسیون معمولی، هر گره با استفاده از بهترین تفکیک‌کننده از میان همه‌ی متغیرهای کمکی منشعب می‌شود امّا در جنگل تصادفی هر گره به طور تصادفی با استفاده از بهترین زیرمجموعه از بین زیرمجموعه‌های متغیرهای کمکی موجود درگره، منشعب می‌گردد. این یک راهبرد تا اندازه‌ای هوشمندانه است که در مقابل بیش برازش، قوی عمل می‌کند و به آسانی قابل استفاده است. مدل جنگل تصادفی دارای دو پارامتر اصلی است، یکی تعداد متغیرها کمکی موجود در زیرمجموعه تصادفی هر گره (mtry) و دیگری تعداد درختان در جنگل(ntree) است که معمولاً به مقدار آن‌ها حساس نیست. از شاخص‌های آنتروپی مانند شاخص جینی (gini index) برای کمّی سازی ناخالصی‌ها در هر گره استفاده می‌شود. در این روش ۳۷ درصد از مجموع کل داده‌ها درمرحله واسنجی مدل مورد استفاده قرار نمی‌گیرد که آنها را نمونه‌های بیرون از سبد (out of bag) می‌نامند. از این داده‌ها برای برآورد خطای پیش‌بینی مدل و به عبارت بهتر، اعتبارسنجی آن استفاده می‌شود که به آن خطای خارج از سبد (out of bag error) یا (OOB) می‌گویند.

---

[6] MRVBF
[7] Sediment index
[8] Topographic wetness index
[9] Valley Depth
[10] Vertical Distance to Channel Network
[11] cubist
[12] Random forest
[13] Generalized linear model

به همین دلیل معمولاً برای اعتبارسنجی مدل جنگل تصادفی از داده‌های مستقل استفاده نمی‌کنند. در این پژوهش برای اجرای مدل جنگل تصادفی از بسته‌ی نرم‌افزاری randoam forest در محیط R استفاده شد.

میانگین مدل یا مدل تجمیعی

معمولا در فرایند مدلسازی، پژوهشگران به منظور انتخاب بهترین مدل[14]، اقدام به استفاده از مدل‌های مختلف و ارزیابی و مقایسه آن‌ها می‌کنند. بهترین مدل، یک مفهوم ذهنی است که وابسته به هدف پژوهشگر از مطالعه می‌باشد. یک روش جایگزین برای مفهوم "بهترین مدل" استفاده از روش‌های مدل میانگین(مدل تجمیعی) است (Padarian et al., 2014). مدل میانگین (مدل تجمیعی) ترکیب نمودن پیش‌بینی مدل‌های مختلف می‌باشد. اصول بنیادین این روش ها می‌تواند توسط مدل ساده زیر توضیح داده شود:

$$Y_i=\sum_{k=1}^{Kk} W_k X_{ik} \quad (۲)$$

در این رابطه Y_i ترکیب خروجی مدل‌ها درنقطه‌i، k تعداد مدل‌های سهیم در پیش بینی، X_{ik} خروجی k امین مدل و W_k ضریب وزنی اختصاص یافته به k امین مدل می‌باشد. در بیشتر روش‌های مدل میانگین، جمع اوزان برابر با یک است. روش‌های مدل میانگین مختلفی از ساده‌ترین حالت که اوزان مدل‌های مختلف یکسان در نظر گرفته می شود تا مدل‌های پیچیده‌ای مانند مدل میانگین بیزی می‌توان استفاده نمود. در این پژوهش از مدل‌های خطی تعمیم یافته [15] برای ترکیب خطی مدل‌های مختلف و تولید مدل میانگین استفاده می‌شود که اصطلاحاً به این شیوه‌ی مدل میانگین، انبوهش مدل‌ها (stacking models) گفته می‌شود. چنین مدلی با استفاده از نرم افزار caret ensemble اجرا می‌گردد.

اعتبار سنجی

در این پژوهش، از روش اعتبارسنجی دوجانبه[16] استفاده شد. بدین منظور ۳۲۰ داده جمع‌آوری شده به دو دسته داده آموزشی (۲۴۰ نمونه) و اعتبار سنجی (۸۰ نمونه) تقسیم شد. خطای مدل‌ها بر اساس مرسوم‌ترین آن‌ها یعنی ریشه میانگین مربعات خطای تخمین[17] محاسبه گردید:

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}\left[z_{x_i}^{*}-z_{x_i}\right]^2}{n}} \quad (۳)$$

در این معادله، RMSE= ریشه میانگین مربعات خطای تخمین، $z_{x_i}^{*}$ =متغیر پیش‌بینی شده، z_{x_i}= متغیر اندازه‌گیری شده و n= تعداد نمونه می‌باشد. یک مدل خوب، دارای حداقل مقدار ریشه میانگین مربعات خطای تخمین می‌باشد (بروس و همکاران، ۲۰۱۱). معیار دیگر برای تخمین دقت مدل، ضریب تعیین(R^2) است که کارائی مدل در پیش‌بینی متغیر هدف با استفاده از متغیرهای کمکی را نشان می‌دهد.

$$R^2=\frac{\sum_{i=1}^{n}\left[z^{*}(x_i)-z^{*}(\bar{x})\right]^2}{\sum_{i=1}^{n}\left[z(x_i)-z(\bar{x})\right]^2} \quad (۴)$$

در معادله‌ی فوق $z(\bar{x})$ و $z^{*}(\bar{x})$ به ترتیب نشان دهنده‌ی میانگین داده‌های مشاهده‌ای و پیش‌بینی هستند.

[14] The best model
[15] Generalized linear models
[16] cross-validation
[17] Root Mean Squared Error

نتایج و بحث

در این تحقیق از ۹ متغیر کمکی شامل ارتفاع، درصد شیب، تحدب، شاخص همواری دره با درجه تفکیک بالا، شاخص رسوب، شاخص خیسی توپوگرافیک، عمق دره، فاصله عمودی تا شبکه آبراهه و شاخص پوشش گیاهی نرمال شده به عنوان متغیرهای محیطی کمکی و کربن آلی خاک به عنوان متغیر هدف استفاده شد. تمام مدل‌ها در ده تکرار اجرا شدند و ارزیابی درونی آن‌ها بر اساس اعتبار سنجی دو جانبه صورت گرفت و مدل بهینه بر اساس آن انتخاب گردید. همانطور که در جدول ۱ مشاهده می‌شود متغیرهای ارتفاع و شاخص پوشش گیاهی نرمال شده موثرترین متغیرهای ورودی در هر سه مدل هستند. ژائو و شی (۲۰۱۰) نشان دادند رابطه معناداری بین کربن آلی خاک و ارتفاع وشاخص پوشش گیاهی در مدل خطی چند متغیره وجود دارد که با نتایج پژوهش حاضر همخوانی دارد.

جدول ۱ - اهمیت نسبی متغیرهای کمکی در هر مدل

متغیرهای کمکی	رگرسیون خطی چند متغیره	مدل کوبیست	جنگل تصادفی
ارتفاع	۱۰۰	۱۰۰	۲۵/۶۱
شاخص پوشش گیاهی نرمال شده	۵۴/۷۱۸	۳۲/۸۸	۱۴/۱۴
شاخص رسوب	۴۳/۲۲	۲۱/۲۳	۶/۶
فاصله عمودی تا شبکه آبراهه	۳۸/۷۰۱	۲۵/۳۴	۱۱
شاخص همواری دره با درجه تفکیک بالا	۳۶/۸۵۸	۶/۸۵	۱۲/۵۹
شیب	۲۵/۰۶۸	۱۴/۳۸	۷/۷۴
شاخص خیسی توپوگرافیک	۲۲/۳۰۱	۴۰/۴۱	۹/۵
تحدّب	۲۲/۶۶۳	۰	۷/۷۱
عمق دره	۰	۰	۷/۹۴

از میان مدل‌های منفرد، مدل جنگل تصادفی با ضریب تبیین (۰/۱۴) و ریشه میانگین مربعات خطا (۰/۵۷۵) دارای بیشترین صحت و دقت برای پیش بینی کربن آلی خاک است (جدول ۲). مدل بهینه جنگل تصادفی برای پیش‌بینی مقدار کربن آلی خاک شامل ntree= ۵۰۰ و mtry= ۳ بود. اما نتایج اعتبارسنجی نشان می‌دهد مدل میانگین از نوع ترکیب خطی تعمیم‌یافته بیش از ۱۵ درصد تغییرات کربن آلی خاک در محدوده مورد مطالعه را توضیح می‌دهد که این مقدار در مدل های منفرد دیگر پایین‌تر است (ستون ضریب تبیین (R^2) در جدول ۲ مشاهده شود). ریشه میانگین مربعات خطا در مدل میانگین از نوع ترکیب خطی تعمیم‌یافته (۰/۵۵۸) پایین‌تر از مدل‌های منفرد رگرسیون خطی چندمتغیره (۰/۵۸۶)، کوبیست (۰/۵۸۱) و جنگل تصادفی (۰/۵۷۵) است (جدول ۲) که این نتایج خود بیانگر کارایی بیشتر مدل میانگین از نوع ترکیب خطی تعمیم‌یافته در پیش بینی مقدار کربن آلی خاک نسبت به سایر مدل های منفرد در این پژوهش می‌باشد. پادریان و همکاران (۲۰۱۴) ا ز سه مدل کوبیست، برنامه‌ریزی ژنتیکی و ماشین بردار پشتیبان و مدل‌های میانگین از نوع اوزان معادل و گرانگر-رامانتان برای پیش بینی ظرفیت رطوبت قابل استفاده خاک استفاده نمودند. نتایج آنها نشان دادکه مدل‌های میانگین عملکرد بهتری از مدل‌های منفرد دارند

جدول ۲ - نتایج اعتبار سنجی مدل‌های مختلف برای پیش بینی کربن آلی خاک

نوع مدل مورد استفاده	R^2	RMSE
خطی چند متغیره	۰/۰۹۶	۰/۵۸۶
کوبیست	۰/۱۱۹	۰/۵۸۱
جنگل تصادفی	۰/۱۴۱	۰/۵۷۵
مدل میانگین از نوع ترکیب خطی تعمیم‌یافته	۰/۱۵۴	۰/۵۵۸

به طور کلی می‌توان بیان نمود استفاده از مدل‌های میانگین در نقشه‌برداری رقومی خاک باعث افزایش صحت و دقت نقشه‌های پیش‌بینی نسبت به مدل های منفرد حتی بهترین آن‌ها می‌شود.

منابع

Anderson J.A. (1995). An Introduction to Neural Networks. Cambridge, USA: The MIT Press.

Breiman L., Friedman J.H, Olshen R.A., Stone & C.J.1984. Classification and regression trees.Wadsworth & Brooks. Monterey, CA.

Brus D. J., Kempen B., and Heuvlink G. B. M. 2011. Sampling for validation of digital soil maps. Eur. J. Soil Sci., 62: 394-407.

Cortes C. and Vaping V.1995. Support-vector networks. Machine learning, 20(3): 273–297.

Diks C.G.H., Vrugt J.A., 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stoch. Env. Res. Risk A., 24 (6): 809–820.

Fatehi Sh. 2010. Semi-Detailed Soil Survey of Merek Plain in Kharkeh River Basin. Kermanshah Agricultural and Natural Resources Research and Education Center, Agricultural Research, Education and Extension Organization (AREEO), Kermanshah, Iran, 68p. (In Persian).

Koza, J. (1992). Genetic Programming: On the Programming of Computers by Means of Natural Selection.The MIT Press.

Malone B., McBratney A., Odgers N.P., & Minasny B.2014. Using model averaging to combine soil property rasters from legacy soil maps and from point data. Geoderma, 232–234: 34–44.

McCullagh P. & Nelder J. 1989. Generalized Linear Model. Chapman & Hall, London.

Padarian J., Minasny B., and McBratney A.B.2014. The evolving methodology for global soil mapping. In: Arrouays, D., McKenzie, N., Hempel, J., Richer de Forges, A., McBratney, A.B. (Eds.), GlobalSoilMap: Basis of the Global Spatial Soil Information System. CRC Press, pp. 215–220.

Quinlan R.1992. Learning with continuous classes. Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, Hobart, Tasmania, pp. 343–348.

Zhao Y.C., and Shi X.Z.2010. Spatial Prediction and Uncertainty Assessment of Soil Organic Carbon in Hebei Province, China. In: Boettinger,J.L., Howell, D.W., More, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer, London, pp. 227–240.

Using model averaging (ensemble model) in soil digital mapping

Sh. Fatehi

Research Assistant Professor, Soil and Water Research Department, Kermanshah Agricultural and Natural Resources Research and Education Center, AREEO, Kermanshah

Abstract

Using model averaging (ensemble model) is a strong way to improve the performance of digital soil properties modeling. but these techniques in digital soil mapping aren't regarded as a good alternative. This study aimed to evaluate and compare the prediction accuracy of three models include multivariate linear regression, Cubist and Random forest with a model averaging approach for mapping the soil organic carbon contents. First, in Merck sub-watershed in Kermanshah province with an area of 24,000 hectares, 320 soil samples were collected randomly and then soil organic carbon content was measured. Soil data were split in two sets, training (240 samples) and validation (80 sample). NDVI and the derived surface parameters of DEM were used as covariates in the fitted models. The root mean square error (RMSE) in model averaging (0.558) was lower than multivariable linear regression (0.586), Cubist (0.581) and Random Forest (0.575). In general, it can be said application of model averaging to digital soil mapping increased the accuracy of the prediction maps.

.

Keywords: digital soil modeling, data manning, ensemble model