جنگل و فرآورده‌های چوب، مجلهٔ منابع طبیعی ایران
دورهٔ ۷۰، شمارهٔ ۳، پاییز ۱۳۹۶
ص ۴۴۱-۴۴۹
DOI: 10.22059/jfwp.2017.130158.650

بررسی عملکرد روش‌های گزینش متغیر در مدلسازی توان تولید رویشگاه راش شرقی

سمیه دهقانی‌نژاد[۱]، سید جلیل علوی[۲*]، سید محسن حسینی[۳]

۱. دانشجوی کارشناسی ارشد جنگلداری، دانشکدهٔ منابع طبیعی و علوم دریایی، دانشگاه تربیت مدرس، نور
۲. استادیار گروه جنگلداری، دانشکدهٔ منابع طبیعی و علوم دریایی، دانشگاه تربیت مدرس، نور
۳. استاد گروه جنگلداری، دانشکدهٔ منابع طبیعی و علوم دریایی، دانشگاه تربیت مدرس، نور

تاریخ دریافت: ۱۳۹۴/۰۶/۳۱، تاریخ پذیرش: ۱۳۹۵/۰۶/۰۲

چکیده

توان تولیدی رویشگاه، معیاری مهم برای مدیران جنگل در خصوص پیش‌بینی مقدار تولید، بهره‌برداری مجاز سالانه، رویش و انتخاب مناسب‌ترین گونه‌های درختی برای یک رویشگاه است. در مطالعهٔ حاضر قابلیت پیش‌بینی مدل‌های آماری با به‌کارگیری روش‌های مختلف گزینش متغیر برای ارتفاع غالب گونهٔ راش به‌عنوان شاخصی از توان تولید رویشگاه نسبت به متغیرهای خاک و فیزیوگرافی بررسی شده است. به این منظور، ۱۲۷ قطعه نمونه دایره‌ای به مساحت ۱۰۰۰ متر مربع در جنگل دانشگاه تربیت مدرس پیاده و در هر یک از آنها ارتفاع غالب گونهٔ راش اندازه‌گیری شد. عملکرد پنج روش گزینش متغیر در رگرسیون خطی چندگانه (انتخاب گام‌به‌گام متغیر با آزمون F و معیارهای اطلاعاتی آکائیک و بیزین، بهترین زیرمجموعه‌ها با استفاده از معیارهای اطلاعاتی آکائیک و بیزین) و روش رگرسیون درختی بررسی شد. به‌منظور مقایسهٔ کارایی روش‌های یادشده، اعتبارسنجی متقابل با ۲۵۰۰ تکرار استفاده شد. نتایج نشان داد که در بین پنج روش مبتنی بر رگرسیون خطی چندگانه، اختلاف اندکی در قابلیت پیش‌بینی وجود دارد. عملکرد روش‌های گام‌به‌گام مشابه الگوریتم‌های بهترین زیرمجموعه است و انتخاب معیارهای مقایسهٔ مدل‌ها (شامل معیار اطلاعاتی آکائیک، معیار اطلاعاتی بیزین یا آمارهٔ F) تأثیر اندکی بر قابلیت پیش‌بینی دارد. در تحقیق حاضر، روش مبتنی بر درخت رگرسیون، قابلیت پیش‌بینی کمتری داشت. استدلال این است که بهترین روش انتخاب متغیر وجود ندارد و هر یک از روش‌های مبتنی بر رگرسیون که در این تحقیق استفاده شده، می‌تواند مدل‌های پیش‌بینی مفید را ایجاد کند.

واژه‌های کلیدی: بهترین زیرمجموعه، توان تولید رویشگاه، روش گام‌به‌گام، گزینش متغیر، معیار اطلاعاتی آکائیک، معیار اطلاعاتی بیزین.

مقدمه

برآورد دقیق توان تولید جنگل، که توانمندی رویشگاه را در زمینهٔ کارکردهای مختلف جنگل همانند تولید چوب، ترسیب کربن، حفظ خاک، کنترل سیلاب، حفظ تنوع زیستی و تأمین آب نشان می‌دهد، از مؤلفه‌های بسیار مهم در مدیریت صحیح منابع جنگلی است. از این‌رو یکی از وظایف خطیر مدیران جنگل کمی‌سازی و سنجش اختلافات و تفاوت‌ها در رویشگاه‌ها [۱] و برآورد توان تولید رویشگاه به‌طور غیرمستقیم و با مدلسازی رابطهٔ بین توان تولید رویشگاه و متغیرهای محیطی، ویژگی‌های منطقه و ویژگی‌های پوشش‌های گیاهی زیراشکوب است [۲].

در مدلسازی متغیر پاسخ با استفاده از متغیرهای تبیینی

نویسندهٔ مسئول، تلفن:۰۹۱۱۱۵۸۰۰۹۷
Email: j.alavi@modares.ac.ir

متعدد، انتخاب متغیر و برآورد ضرایب در مدل رگرسیونی اساسی‌ترین بخش در مدلسازی است. از دیدگاه عملی، انتخاب متغیر به تعیین این موضوع معطوف می‌شود که کدام متغیر بیشترین تأثیر را بر متغیر وابسته دارد؛ در صورتی که از دیدگاه آماری روشی را برای دستیابی به توازن بین نکویی برازش و اصل صرفه‌جویی ارائه می‌کند. به‌عبارتی، با شناسایی مؤثر زیرمجموعه‌ای از متغیرهای مهم، انتخاب متغیر می‌تواند هم تفسیرپذیری مدل را تسهیل کند و هم صحت پیش‌بینی را بهبود بخشد [۳].

معیارهای بسیاری برای مقایسهٔ آماری مدل‌های چندمتغیری وجود دارد. اکولوژیست‌ها به‌تازگی به معیارهای تئوری اطلاعات نظیر معیار اطلاعات آکائیک[1] (AIC) و معیار اطلاعات بیزی شوارتز[2] (BIC) توجه نشان داده‌اند [۴]. برای مثال مازرول (۲۰۰۶) اشاره می‌کند که آمارهٔ AIC در انتخاب مدل (برای مثال انتخاب متغیر) نسبت به رویکردهای آزمون فرضیه برتری چشمگیری دارد [۵]. قدیمی‌ترین الگوریتم برای انتخاب متغیرهای تبیینی مراحل گام‌به‌گام هستند. الگوریتم‌های جدیدتر شامل جست‌وجوی کامل زیرمجموعه‌های بیشتری از متغیرهای پیشگو نسبت به رویکرد گام‌به‌گام است که همین موضوع موجب تردید اکولوژیست‌ها دربارهٔ سودمندی روش انتخاب گام‌به‌گام متغیر شده است. برای مثال، ویتینگام و همکاران (۲۰۰۶)، به استفادهٔ گسترده از رویکردهای گام‌به‌گام در مجله‌های مرتبط با بوم‌شناختی به‌دلیل اریب‌ها و کاستی‌های مسلم رگرسیون گام‌به‌گام چندگانه انتقاد کرده‌اند [۶].

محققان بسیاری، روش‌های گزینش متغیر را بررسی و ارزیابی کرده‌اند [۷]. رافالوویچ و همکاران (۲۰۰۸) مرور جامعی در این زمینه انجام دادند. آنها قابلیت روش‌های متعدد در لحاظ متغیرهای مهم و حذف متغیرهای نامربوط را ارزیابی کردند و دریافتند که روش‌های مبتنی بر رگرسیون گام‌به‌گام و BIC بهترین عملکرد را دارند، درصورتی که روش‌های مبتنی بر AIC کارایی نامناسبی داشتند و از آنها باید اجتناب ورزید [۸]. شبیه‌سازی تنها روشی است که در آن مدل واقعی شناخته می‌شود، اما نتایج مطالعات شبیه‌سازی، به روش ایجاد داده‌های شبیه‌سازی و معیارهای انتخاب‌شده برای مقایسهٔ تکنیک‌های مختلف مدلسازی وابسته است. برای مثال مورتاف (۱۹۹۸) دریافت که رویکردهای متکی بر آزمون F، آمارهٔ مالو و BIC فراوانی یکسانی از تصمیمات درست در خصوص ورود یا حذف متغیرهای تبیینی ارائه می‌دهد. وی همچنین مشاهده کرد که تفاوت اندکی در قابلیت روش‌های یادشده وجود دارد [۹]. رویکردهای مبتنی بر درخت طبقه‌بندی و رگرسیون یا CART[3]، نسبت به روش‌های رگرسیونی، عملکرد ضعیف‌تری دارند؛ اما مورتاف (۱۹۹۸) با توجه به اینکه داده‌ها براساس یک مدل رگرسیون خطی شبیه‌سازی شده بودند، دربارهٔ اعتبار مقایسه تردید نشان داد [۹].

ارزیابی توان تولید رویشگاه‌های جنگلی مقولهٔ بسیار مهمی در مطالعات جنگلداری در جهان و یکی از مؤلفه‌های بسیار مهم در مدیریت صحیح منابع جنگلی است که در ایران کمتر به آن توجه شده است. یکی از معیارهای مهم متأثر از کیفیت رویشگاه، ارتفاع غالب است که به‌صورت ارتفاع ۱۰۰ اصله از قطورترین درختان در هکتار بیان می‌شود [۱۰]. این معیار برخلاف قطر برابرسینه، کمتر تحت تأثیر عملیات پرورشی در جنگل قرار می‌گیرد و رابطهٔ نزدیکی با حجم درختان دارد و به همین دلیل معیار مناسبی برای بررسی کیفیت رویشگاه است [۱۱].

در تحقیق حاضر، عملکرد شش روش گزینش متغیر در مدلسازی ارتفاع غالب گونهٔ راش، به‌عنوان یک معیار توان تولید رویشگاه، نسبت به متغیرهای محیطی بررسی شد. ازآنجا که مدل مناسب و شناخته‌شده‌ای از متغیر پاسخ

1. Akaike's information criterion
2. Schwarz's Bayesian information criterion
3. Classification and Regression Tree

برحسب متغیرهای تبیینی وجود ندارد، از اعتبارسنجی متقابل برای ارزیابی توانایی پیش‌بینی مدل استفاده می‌شود که می‌تواند مقایسهٔ منصفانه‌ای از روش‌های انتخاب متغیر باشد.

مواد و روش‌ها

منطقهٔ تحقیق

پژوهش حاضر در جنگل آموزشی و پژوهشی دانشکدهٔ منابع طبیعی دانشگاه تربیت مدرس صورت گرفت. مساحت منطقه ۱۸۱۷ هکتار و ارتفاع از سطح دریا ۱۵۰ تا ۱۷۵۰ متر است. این منطقه بین عرض جغرافیایی "۳۶°۲۹'۲۳ تا "۳۶°۳۲'۵۶ و طول جغرافیایی "۵۱°۴۳'۲۰ تا "۵۱°۴۷'۳۹ قرار گرفته است. مقدار بارندگی در این ناحیه براساس داده‌های ایستگاه هواشناسی نوشهر ۱۳۰۸ میلی‌متر در سال است. گرم‌ترین ماه‌های سال تیر و مرداد با میانگین دمای ۲۹/۲ و سردترین ماه سال، بهمن با میانگین دمای۲/۶ درجهٔ سانتی‌گراد است. همچنین میانگین دمای سالانه ۱۵/۹ درجهٔ سانتی‌گراد ثبت شده است. تیپ جنگل در منطقه، آمیخته‌ای از راش به‌همراه گونه‌های ممرز، انجیلی، افرا، شیردار، نمدار، ون و بلوط است. به‌طور میانگین گونهٔ راش حدود ۵۰ درصد از کل حجم سرپا و حدود ۴۰ درصد از کل تعداد درختان منطقه را در بر می‌گیرد [۱۲].

روش بررسی

به‌منظور جمع‌آوری اطلاعات لازم از ۱۲۷ قطعه نمونهٔ دایره‌ای با مساحت ۱۰۰۰ مترمربع به روش تصادفی سیستماتیک در توده‌هایی که در آنها گونهٔ راش غالب بود استفاده شد. پس از پیاده نمودن قطعات نمونه، در داخل هر یک از آنها، قطر برابرسینه و ارتفاع تمام درختان راش با قطر بیشتر از ۷/۵ سانتی‌متر اندازه‌گیری و ویژگی‌های عمومی رویشگاه مثل جهت جغرافیایی، ارتفاع از سطح دریا و درصد شیب نیز ثبت شد. ارتفاع غالب که در اینجا به‌صورت میانگین ارتفاع سه اصله از قطورترین درختان در هر قطعه نمونه تعریف می‌شود، به‌طور معمول شاخصی از کیفیت رویشگاه به‌شمار می‌رود. جهت جغرافیایی با استفاده از رابطهٔ زیر به شاخص تابش خورشیدی[1] تبدیل شد که در آن θ مقدار آزیموت جهت جغرافیایی برحسب درجه است [۱۳].

$$TRASP = \left[۱ - \cos\left((\pi/۱۸۰)(\theta - ۳۰)\right)\right]/۲ \qquad (۱)$$

در مرحلهٔ بعد در هر قطعه نمونه، نمونه‌برداری خاک از عمق صفر تا ۱۰ سانتی‌متر صورت گرفت. پس از خشک کردن نمونه‌ها در هوای آزاد و جدا کردن ریشه‌ها، سنگ‌ها و سایر ناخالصی‌ها از آن، کلوخ‌ها خرد و از الک ۲ میلی‌متری عبور داده شد. در بررسی فاکتورهای خاک، مجموعه‌ای از خصوصیات فیزیکی و شیمیایی خاک از جمله وزن مخصوص ظاهری به روش کلوخ و پارافین، بافت خاک به روش هیدرومتری، pH خاک با دستگاه pH متر الکتریکی، درصد مادهٔ آلی به روش والکی و بلک، فسفر قابل جذب به روش السون، پتاسیم به‌وسیلهٔ دستگاه طیف‌سنج اتمی و نیتروژن به روش کجلدال اندازه‌گیری شد.

رویکردهای گزینش متغیر

روش‌های گزینش متغیر برای گنجاندن در مدل رگرسیون خطی چندگانه عبارت‌اند از: ۱. روش‌های گام‌به‌گام، که در آنها برخی از معیارهای کمّی برای مقایسهٔ مدل‌های رگرسیون با و بدون متغیر پیشگوی خاص استفاده می‌شوند و افزودن یا حذف متغیرهای تبیینی به‌صورت پی‌درپی ادامه دارد تا زمانی که به مقدار معیار برسیم؛ ۲. بهترین زیرمجموعه‌های متغیرهای انتخابی، که در آن مجموعه‌ای از تمام گروه‌های ممکن متغیرهای تبیینی جست‌وجو شده و زیرمجموعه‌ای از متغیرهای پیشگو که مطلوب‌ترین مقادیر معیارهای کمّی را دارند شناسایی می‌شود.

معیارهای مورد استفاده برای مقایسهٔ مدل‌های رگرسیون چندگانه عبارت‌اند از:

۱. سطح معنی‌داری یا p-value. چنانچه سطح معنی‌داری برای متغیر تبیینی کمتر از مقدار ۰/۰۵ باشد، آن متغیر، مؤثر قلمداد شده و وارد مدل رگرسیونی می‌شود.

1. Radiation Index

۲. معیار اطلاعاتی آکائیک. این معیار در مدل رگرسیون خطی چندگانه به‌صورت زیر نوشته می‌شود [۱۳]:

$$AIC = n\ln(RMSE)+2p \quad (۲)$$

۳. معیار اطلاعاتی بیزین شوارتز. این معیار در مدل رگرسیون خطی چندگانه به‌صورت زیر نوشته می‌شود [۱۳]:

$$AIC = n\ln(RMSE)+p\ln(n) \quad (۳)$$

n تعداد مشاهدات، p تعداد ضرایب رگرسیون و RMSE مجذور میانگین مربعات خطاست.

درخت طبقه‌بندی و رگرسیون یا CART یک روش مدلسازی متفاوت با روش رگرسیون بحث‌شده در بالاست. CART یکی از روش‌های رگرسیون ناپارامتری است که درخت تصمیم‌گیری را براساس الگوریتم تقسیم‌بندی دودویی رشد می‌دهد. برخلاف روش‌های رگرسیون که در آنها بین پیچیدگی مدل و تغییرات توجیه‌شده توازنی برقرار است، در روش درخت طبقه‌بندی و رگرسیون، انتخاب مدل توسط هرس کردن درختان پیچیده به درختان ساده‌تر با استفاده از اعتبارسنجی متقابل انجام می‌گیرد [۱۴].

انتخاب متغیر

قبل از برازش مدل، متغیرهای تبیینی مورد بررسی به‌صورت جداگانه از نظر وجود دادهٔ پرت بررسی شدند. به‌منظور بررسی هم‌خطی بین متغیرهای مستقل از عامل تورم واریانس[۲] (VIF) و مقدار آستانهٔ ۱۰ استفاده شد. مقدار VIF برای متغیرهای درصد شن و درصد کربن و مادهٔ آلی بیشتر از ۱۰ بود، در نتیجه این دو متغیر از تجزیه‌وتحلیل حذف شدند و مدلسازی با استفاده از باقی‌ماندهٔ متغیرها صورت گرفت.

در مطالعهٔ حاضر شش روش انتخاب متغیر و ساخت درخت تصمیم در نرم‌افزار آماری R اعمال شد. روش‌های برازش مدل در مطالعهٔ حاضر به‌شرح زیر است:

۱. انتخاب گام‌به‌گام متغیر با آزمون F : p-value توسط مدل‌های با و بدون متغیر تبیینی خاص با استفاده از آزمون F مجموع مربعات محاسبه می‌شود. در تحقیق حاضر از مقدار P=۰/۰۵ به‌عنوان حد آستانه برای گنجاندن یا حذف متغیرهای پیشگو استفاده شد.

۲. انتخاب گام‌به‌گام متغیر با استفاده از AIC: ابتدا با مدلی که فاقد متغیر پیشگو است شروع شده و مراحل زیر تکرار می‌شود: ۱. متغیر پیشگویی که کمترین AIC را دارد به مدل اضافه می‌شود؛ ۲. اگر هیچ یک از متغیرهای تبیینی سبب کاهش AIC نشوند، حذف می‌شوند. روش زمانی به پایان می‌رسد که هیچ متغیر اضافه‌شده یا حذف‌شده‌ای که AIC را کاهش دهد، وجود نداشته باشد.

۳. انتخاب گام‌به‌گام متغیر با استفاده از معیار بیزین شوارتز: همانند روش ۲ است، اما از معیار BIC استفاده می‌شود.

۴. بهترین زیرمجموعه‌ها با استفاده از AIC: تابع regsubsets در بستهٔ leaps در نرم‌افزار آماری R با استفاده از این معیار بهترین زیرمجموعه از متغیرهای پیشگو شناسایی می‌شود.

۵. بهترین زیرمجموعه‌ها با استفاده از BIC: همانند روش ۴ است، با این تفاوت که رتبه‌بندی برحسب BIC انجام می‌گیرد.

۶. هرس کردن درخت رگرسیون و طبقه‌بندی با استفاده از پارامتر پیچیدگی: برای برازش درخت رگرسیون از بسته rpart در نرم‌افزار آماری R استفاده شد. سپس به‌منظور جلوگیری از بیش‌برازش[۱]، هرس کردن با استفاده از پارامتر پیچیدگی و اعتبارسنجی متقابل 10-fold صورت پذیرفت. این پارامتر توازنی بین پارامترهای دقت و پیچیدگی درخت برقرار می‌کند.

اعتبارسنجی متقابل روش‌های انتخاب متغیر

قابلیت پیش‌بینی هر یک از روش‌های بالا، با استفاده از اعتبارسنجی متقابل ارزیابی می‌شود. برای هر یک از روش‌های یادشده، مراحل زیر ۲۵۰۰ بار تکرار شد:

۱. دادهٔ به‌کاررفته در این تحقیق، به‌صورت تصادفی به

1. Overfitting

یک زیرمجموعهٔ آموزشی شامل ۷۵ درصد مشاهدات و یک زیرمجموعهٔ اعتبارسنجی شامل ۲۵ درصد مشاهدات تقسیم می‌شود؛ ۲. برای هر یک از روش‌های انتخاب متغیر، برای ساخت مدل پیش‌بینی از داده‌های آموزشی استفاده می‌شود؛ ۳. مدل حاصل از مرحلهٔ ۲ (ضرایب رگرسیون یا درخت رگرسیون) به متغیرهای تبیینی موجود در مشاهدات زیرمجموعهٔ اعتبارسنجی اعمال شده و مقدار متغیر پاسخ برای زیرمجموعهٔ اعتبارسنجی پیش‌بینی می‌شود.

مقدار مجذور میانگین مربعات خطا برای اعتبارسنجی متقابل از طریق زیر محاسبه و معیار ارزیابی روش‌های انتخاب متغیر در نظر گرفته می‌شود:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Q_i - \hat{Q}_i\right)^2}{n}} \qquad (۴)$$

که در آن RMSE مجذور میانگین مربعات خطا، n تعداد مشاهده‌ها برای زیرمجموعهٔ اعتبارسنجی متقابل، Q_i مقدار مشاهده‌شده برای مورد i، و $\hat{Q}_i$ مقدار پیش‌بینی‌شده براساس مدل حاصل از داده‌های آموزشی است. در مواردی که هیچ متغیری برای مجموعهٔ خاصی از داده‌های آموزشی انتخاب نشود، میانگین مقدار متغیر پاسخ برای دادهٔ آموزشی به‌عنوان پاسخ پیش‌بینی‌شده برای تمام آیتم‌های زیرمجموعه اعتبارسنجی استفاده می‌شود. برای هر یک از روش‌های انتخاب متغیر، ۲۵۰۰ مقدار اعتبارسنجی متقابل از RMSE به‌منظور خلاصه‌ای از قابلیت پیش‌بینی ارائه می‌شود.

نتایج و بحث

آماره‌های توصیفی مربوط به متغیرهای خاکی و فیزیوگرافی مربوط به ۱۲۷ قطعه نمونهٔ برداشت‌شده، در جدول ۱ ارائه شده است.

در جدول ۲ برای هر یک از روش‌های انتخاب متغیر، متوسط تعداد متغیرهای تبیینی موجود در مدل‌های برازش‌یافتهٔ حاصل از مجموعه داده‌های آموزشی ارائه شده است. میانگین تعداد متغیرها در روش درخت رگرسیون، از دیگر روش‌های گزینش متغیر، کمتر است. با این‌حال، از آنجا که یک متغیر پیشگو ممکن است در بیش از یک گره از درخت رگرسیون رخ دهد، مقایسهٔ تعداد متغیرهای پیشگو در رگرسیون درختی در مقایسه با مدل‌های رگرسیون معمولی، مقایسهٔ ناقصی از پیچیدگی مدل را فراهم می‌کند.

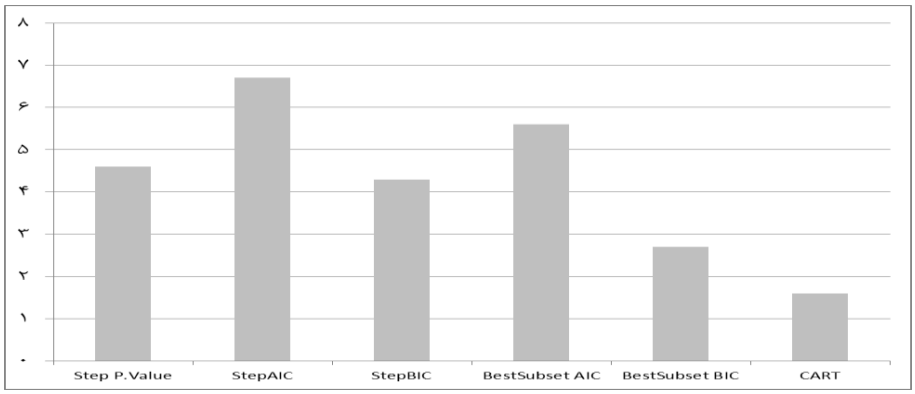
شکل ۱ نمودار خلاصهٔ گرافیکی از نتایج ارائه‌شده در جدول ۲ را نمایش می‌دهد. به‌طور کلی، روش‌های مبتنی بر AIC (۲ و ۴) مدل‌های بزرگ‌تری را ایجاد کردند، زیرا AIC جریمهٔ به‌نسبت کوچکی را برای پیچیدگی مدل در نظر می‌گیرد. مدل‌های حدواسط با استفاده از روش مبتنی بر آزمون F (۱) و BIC (۳ و ۵) ایجاد شدند. از طرف دیگر، RMSE برای مدل‌های مبتنی بر BIC بیشتر از AIC است.

جدول ۱. مقادیر حداقل، حداکثر، میانگین و انحراف معیار برای متغیرهای تبیینی

متغیرهای تبیینی	حداقل	حداکثر	میانگین	انحراف معیار
ارتفاع از سطح دریا	۱۰۶۷	۱۴۴۵	۱۲۲۹	۸۰/۶۳
درصد شیب	۳/۳۰	۷۲/۶۰	۲۶.۲۰	۱۲/۳۷
تابش خورشیدی	۰	۱	۰/۳۰۳۵	۰/۲۸
درصد رس	۰	۶۴	۳۵/۸۵	۱۲/۰۴
درصد سیلت	۱۶	۵۶	۳۶/۸۱	۷/۶۸
وزن مخصوص ظاهری	۱/۰۱	۲/۰۵	۱/۵۲	۰/۲۵
نیتروژن	۰/۱۳	۰/۶۹	۰/۳۲۷۹	۰/۱۲
نسبت کربن به نیتروژن	۴/۴۶	۲۰/۱۸	۱۰/۳۰	۳/۲۶
پتاسیم	۶	۲۲۴	۹۹/۲	۶۳/۴۷
فسفر	۴	۳۷/۸۴	۱۴/۱۳	۸/۷۹
اسیدیته	۵	۷/۵۳	۶	۰/۵۴

جدول ۲. میانگین تعداد متغیرهای انتخاب‌شده، ضریب تبیین و RMSE با استفاده از شش روش ۱. Step P.Value: انتخاب گام‌به‌گام متغیر با آزمون F؛ ۲. StepAIC: انتخاب گام‌به‌گام متغیر با استفاده از AIC؛ ۳. StepBIC: انتخاب گام‌به‌گام متغیر با استفاده از BIC؛ ۴. BestSubset AIC: بهترین زیرمجموعه‌ها با استفاده از AIC؛ ۵. BestSubset BIC: بهترین زیرمجموعه‌ها با استفاده از BIC؛ ۶. CART: درخت رگرسیون و طبقه‌بندی.

روش انتخاب متغیر	۱	۲	۳	۴	۵	۶
میانگین تعداد متغیر انتخاب‌شده	۴/۶	۶/۷	۴/۳	۵/۶	۲/۷	۱/۶
ضریب تبیین برای دادهٔ مدلسازی	۳۲/۸۴	۳۶/۷۷	۳۲/۰۳	۳۶/۲۳	۳۰/۳۳	۲۹/۱۱
RMSE برای داده مدلسازی	۳/۴۸	۳/۳۸	۳/۵۱	۳/۴۰	۳/۵۶	۳/۶۶
ضریب تبیین برای دادهٔ اعتبارسنجی	۲۰/۶۸	۲۰/۸۵	۲۰/۹۸	۲۰/۵۶	۲۰/۴۹	۱۱/۴۶
RMSE برای دادهٔ اعتبارسنجی	۳/۹۳	۳/۹۵	۳/۹۲	۳/۹۳	۳/۹۱	۴/۲۲



شکل ۱. میانگین تعداد متغیرهای انتخاب‌شده با استفاده از شش روش

درخت رگرسیون از دیگر روش‌های گزینش متغیر، ضریب تبیین کمتری داشته است. روش‌های مبتنی بر AIC (۲ و ۴) همانند مرحلهٔ تعداد متغیر، ضریب تبیین بیشتری داشته‌اند. مدل‌های بینابینی با استفاده از تکنیک‌های مبتنی بر آزمون F (۱) و BIC (۳ و ۵) ایجاد شدند.

در روش مبتنی بر درخت (۶) به‌طور کلی مجذور میانگین مربعات خطا بیشتر از روش‌های مبتنی بر رگرسیون (۵-۱) است. در روش مبتنی بر درخت، برای تعدادی از تکرارها، در اعتبارسنجی متقابل هیچ متغیر تبیینی انتخاب نشده بود؛ با وجود این نمی‌توان گفت که درختان رگرسیون بدون شاخه، ابزار پیش‌بینی مفیدی هستند؛ اما الگوریتم‌های هرس کردن که به این درختان تک‌ریشه منجر شده‌اند، آنها را از بیش‌برازش و تورم واریانس محافظت می‌کنند. با مراجعه به جدول ۲ مشاهده می‌شود که ضریب تبیین برای روش مبتنی بر درخت به‌طور محسوسی کمتر از روش‌های دیگر است. مقایسهٔ مجذور میانگین مربعات خطا و ضریب تبیین برای روش‌های مبتنی بر رگرسیون حکایت از شباهت مقادیر مرتبط با رویکردهای مبتنی بر رگرسیون (۵-۱) دارد که نشان می‌دهد متوسط قابلیت پیش‌بینی هر پنج روش به‌نسبت یکسان بوده است. این نتایج با اظهارات برخی از بوم‌شناسان در مورد ضعف روش گام‌به‌گام مغایرت دارد [۹، ۱۵].

همان‌طور که اشاره شد روش‌های مبتنی بر معیار اطلاعات آکائیک، مدل‌هایی با بیشترین تعداد متغیر پیشگو را ایجاد کردند. ازآنجا که توان پیش‌بینی این مدل‌ها، بیشتر از مدل‌های حاصل از روش آزمون F و BIC نبوده است، انتخاب تکنیک‌های گزینش متغیر باید براساس ملاحظات دیگر صورت گیرد. همان‌طور که مورتاف (۱۹۹۸) اشاره می‌کند محقق ممکن است یک روش گزینش متغیر بیشتر

محافظه‌کارانه‌تر (F-test یا BIC) یا روش‌های آزادانه‌تر (AIC) را ترجیح دهد [۹]. بسیاری از متخصصان آمار اشاره دارند که تفسیر p-value متغیرهای تبیینی در مدل رگرسیون مشکل یا ناممکن است، زیرا p-value عدم قطعیت در ساخت مدل را توضیح نمی‌دهد [۱۶]. برخی از محققان استفاده از آزمون F در انتخاب متغیر را بد جلوه داده‌اند [۱۵، ۱۷]، با این حال، آمارهٔ F می‌تواند رویکرد ارزشمندی در بیان توازن بین تغییرات توجیه‌شده و پیچیدگی مدل باشد. در واقع F-to-enter، ضریب تبیین تعدیل یافته، و AIC را می‌توان مواردی خاص از شکل آمارهٔ مالو Cp تلقی کرد [۱۶]. با اعمال روش انتخاب گام‌به‌گام (روش ۱) با توجه به مقدار p-value ۰/۰۵ میانگین تعداد متغیر ۴/۶ به‌دست می‌آید. در صورت استفاده از آستانهٔ ۰/۰۱، مدل به‌طور متوسط ۴/۲ متغیری ایجاد می‌شود. در نهایت، با استفاده از آستانهٔ ۰/۱۰ مدل ۵/۶ متغیری ایجاد می‌شود. از این‌رو هنگامی که استفاده از AIC یا BIC به‌تنهایی امکان‌پذیر نیست، سطح معنی‌داری آزمون F را می‌توان پارامتر تنظیم جریمهٔ پیچیدگی مدل در نظر گرفت [۱۵]. در این زمینه رد کردن روش انتخاب گام‌به‌گام متغیر با استفاده از آزمونF به‌عنوان ابزار مناسب برای گزینش، تنها به‌دلیل استفاده از آزمون فرضیه ناعادلانه به‌نظر می‌رسد [۶، ۱۸]. مورتاف در سال ۲۰۰۹ نیز دریافت که عملکرد روش‌های گام‌به‌گام و بهترین زیرمجموعه مشابه است و انتخاب معیار برای مقایسهٔ مدل‌ها (AIC، BIC و آمارهٔ F) تأثیر اندکی بر قابلیت پیش‌بینی دارد [۱۹]. نتایج تحقیق حاضر که در آن بهترین روش برای انتخاب مدل آماری وجود ندارد، با مطالعات مورتاف (۲۰۰۹) مطابقت دارد.

نتیجه‌گیری

مقایسهٔ روش‌های گزینش در تحقیق حاضر براساس قابلیت پیش‌بینی بوده است. چنانچه هدف از مدلسازی، شناسایی روابط علت و معلولی باشد، معیار مقایسهٔ روش‌ها ممکن است با معیارهایی که در اینجا در نظر گرفته شده متفاوت باشد. باید گفت در زمان ساخت مدل، بین دو هدف اصلی باید تمایز قائل شد. هدف اول پیش‌بینی است که در آن ساختار مدل از ملاحظهٔ کمی برخوردار است و هدف دوم، تبیین است که در آن تلاش می‌شود متغیرهای پیشگوی مهم شناسایی شده و بینشی نسبت به رابطهٔ بین متغیر پیشگو و متغیر پاسخ از طریق ساختار مدل حاصل شود. در خصوص پیش‌بینی، برازش مدل و میانگین مربعات خطای پیش‌بینی معیارهای اصلی برای کفایت مدل هستند. نکتهٔ مهم دیگر این است که استفاده از داده‌های واقعی در تحقیق حاضر، به‌جای دادهٔ شبیه‌سازی، استنباط و نتیجه‌گیری را دشوار می‌کند. از طرف دیگر، نمونه‌برداری از مجموعهٔ داده‌ها که نتایج را به جمعیت بزرگ‌تر تعمیم دهد، نیز کار دشواری است. در تحقیق حاضر از ۱۲۷ قطعه نمونه استفاده شد. افزایش شدت نمونه‌برداری ممکن است تا حدی بر مسائل یادشده فایق آید. شبیه‌سازی تنها روش مشخص برای مقایسهٔ تکنیک‌های مدلسازی است که در آن مدل درست شناخته می‌شود. شایان ذکر است که گسترهٔ وسیعی از روش‌های گزینش متغیر وجود دارد که هر کدام از آنها می‌توانند مدل‌های ارزشمندی ایجاد کند که برای کاربرد خاصی در نظر گرفته شود. مفیدترین مدل با استفاده از روش به‌دست آمدن مدل‌ها تعیین نمی‌شود، بلکه بیشتر از طریق متناسب بودن آن مدل برای آن هدف تعیین می‌شود.

References

[1]. Vanclay, J.K. (1992). Assessing site productivity in tropical moist forests: a review. Forest Ecology and Management, 54(1): 257-287.

[2]. Bergès, L., Chevalier, R., Dumas, Y., Franc, A., and Gilbert, J.M. (2005). Sessile oak (*Quercus petraea* Liebl.) site index variations in relation to climate, topography and soil in even-aged high-forest stands in northern France. Annals of Forest Science, 62(5), 391-402.

[3]. Hoeting, J.A., Davis, R.A., Merton, A.A., and Thompson, S.E. (2006). Model selection for geostatistical models. Ecological Applications, 16(1): 87-98.

[4]. Ward, E.J. (2008). A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. Ecological Modelling, 211(1): 1-10.

[5]. Mazerolle, M.J. (2006). Improving data analysis in herpetology: using Akaike's information criterion (AIC) to assess the strength of biological hypotheses. Amphibia-Reptilia, 27(2): 169-180.

[6]. Whittingham, M.J., Stephens, P.A., Bradbury, R.B., and Freckleton, R.P. (2006). Why do we still use stepwise modelling in ecology and behaviour?. Journal of Animal Ecology, 75(5): 1182-1189.

[7]. Sauerbrei, W., Royston, P., and Binder, H. (2007). Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Statistics in Medicine. 26(30): 5512-5528.

[8]. Raffalovich, L.E., Deane, G.D., Armstrong, D., and Tsao, H.S. (2008). Model selection procedures in social research: Monte-Carlo simulation results. Journal of Applied Statistics, 35(10): 1093-1114.

[9]. Murtaugh, P.A. (1998). Methods of variable selection in regression modeling. Ccommunications in Statistics- Simulation and Computation, 27(3): 711-734.

[10]. Zobeiry, M. (2009). Forest Inventory (Measurement of Tree and Forest). University of Tehran Press. Tehran.

[11]. Hogg, B.W., and Nester, M.R. (1991). Effect of stocking rate on predominant height of young Caribbean pine plantations in coastal Queensland. Australian Forestry, 54(3): 134-138.

[12]. Fakoor, E. (2014). Evaluating the site productivity of Oriental Beech using data mining techniques (decision tree). Master's thesis, Tarbiat Modares University, 79 p.

[13]. Aertsen, W., Kint, V., Van Orshoven, J., Ozkan, K., and Muys, B. (2010). Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. Ecological Modeling, 221(8): 1119-1130.

[14]. De'ath, G., and Fabricius, K.E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology, 81(11): 3178-3192.

[15]. Mundry, R., and Nunn, C.L. (2008). Stepwise model fitting and statistical inference: turning noise into signal pollution. The American Naturalist, 173(1): 119-123.

[16]. Miller, A. (2002). Subset Selection in Regression, 2nd edition, Chapman & Hall/CRC, Boca Raton, Florida.

[17]. Burnham, K.P., and Anderson, D.R. (2002). Model selection and multimodel inference: A practical information-theoretic approach, 2nd edition. Springer, New York.

[18]. Lukacs, P.M., Thompson, W.L., Kendall, W.L., Gould, W.R., Doherty Jr, P.F., Burnham, K.P., and Anderson, D.R. (2007). Concerns regarding a call for pluralism of information theory and hypothesis testing. Journal of Applied Ecology, 44(2): 456-460.

[19]. Murtaugh, P.A. (2009). Performance of several variable-selection methods applied to real ecological data. Ecology Letters, 12(10): 1061-1068.

Evaluating the performance of variable selection methods in modeling the site productivity of Oriental Beech (*Fagus Orientalis* Lipsky)

S. Dehghaninejad; M.Sc. Student in Forestry, Faculty of Natural Resources, Tarbiat Modares University, Nour, I.R. Iran

S. J. Alavi*; Assist. Prof., Faculty of Natural Resources, Tarbiat Modares University, Nour, I.R. Iran.

S. M. Hosseini; Prof., Faculty of Natural Resources, Tarbiat Modares University, Nour, I.R. Iran

(Received: 22 September 2015, Accepted: 23 August 2016)

ABSTRACT

Forest site productivity is an important criterion for forest managers. In this study, the predictability of statistical models was studied by applying various methods of variable selection for beech dominant height as an indicator of forest site quality in relation to edaphic and physiographic factors. For this purpose, 127 0.1 ha circular sample plots were established in the forests of Tarbiat Modares University and within each plot, the dominant height of beech trees was calculated. The performance of five variable selection methods was evaluated in multiple linear regression, and regression trees. In order to compare the performance of the above methods, Cross-validation, involving repeated splits of the dataset into training and validation subsets (2500 times) was used to obtain honest estimates of predictive ability. The results showed that there is little differences in the predictive ability of five methods based on multiple linear regression. Stepwise methods performed similarly to exhaustive algorithms for subset selection, and the choice of criterion for comparing models (Akaike's information criterion, Schwarz's Bayesian information criterion or F statistics) had little effect on predictive ability. In this study method based on regression trees yielded with substantially lower predictive ability. It is concluded that there is no best method of variable selection and that any of the regression-based approaches discussed here is capable of yielding useful predictive models.

Keywords: Subset selection, Site quality, Stepwise method, Variable selection, Akaike's information criterion (AIC), Schwarz's Bayesian information criterion (BIC).

* Corresponding author, Email: j.alavi@modares.ac.ir, Tel: 09111580097