Predicting site index from biophysical variables: An evaluation of two modeling approaches

Charles O. Sabatia
Harold E. Burkhart
Predicting site index of plantation loblolly pine from biophysical variables

Charles O. Sabatia *, Harold E. Burkhart

Department of Forest Resources and Environmental Conservation, Virginia Polytechnic Institute and State University, 319 Cheatham Hall, 310 West Campus Drive, Blacksburg, VA 24061-0324, United States

ARTICLE INFO
Article history:
Received 20 January 2014
Received in revised form 17 April 2014
Accepted 18 April 2014
Available online 14 May 2014

Keywords:
Pinus taeda
Random forest
Factor analysis
Climate change

ABSTRACT
Concerns of the effect of climate change on forest productivity have impelled the need to accurately predict forest productivity from climate, physiographic and edaphic variables (biophysical variables). We fitted and evaluated random forest models and nonlinear least squares regression models for predicting plantation loblolly pine (Pinus taeda L) site index from biophysical variables. Tree and stand location data were provided by the Virginia Tech Forest Modeling Research Cooperative. Climate data for each stand location were computed using the Oakridge National Laboratories' daily surface weather prediction models, while soils data were extracted from the USDA Natural Resource Conservation Service SSURGO GIS database using GIS data extraction techniques. Separate models were fitted for non-intensively managed (Non-IMP) and intensively managed (IMP) loblolly pine plantations. Variable selection methods in both modeling approaches showed that the number of biophysical variables that were important in predicting site index of IMP loblolly pine was smaller than the number for Non-IMP stands. The non-parametric random forest models had better fit and prediction statistics than the least squares parametric models but exhibited the potential to give illogical predictions under extrapolation. Site index predictions from both modeling approaches exhibited a regression towards the mean.
Background

- Need to use empirical models to evaluate effect of climate change on forest productivity
- Need for a biophysical variables measure of site productivity
- Parametric approach:
  - Productivity = $f$(precipitation, temp, soils, …)
- Non-parametric approach
  - Random forests (based on Breiman’s ensemble approach on regression/classification trees)
Objective

- Compare predictive performance of the random forests approach to a parametric approach, on plantation loblolly pine data
Approach

- Fit the models on site index and biophysical variables data from 2 region-wide plantation loblolly studies

- Compare:
  - Model fit
  - Model prediction errors

- Evaluate logic of predictions
Data

IMP Plantations ★ Non-IMP Plantations

75% : 25% fitting:validation split
Data

- Site Index – GADA site index equation
- Climate – 32 yr. temp. & precipitation from DAYMET
- Elevation
- Soils - USDA NRCS SSURGO soil map unit weighted averages
  - Available water storage in depth 0 – 150cm
  - Total depth of biotic soil layers (200 cm maximum)
  - Sand %
  - Clay %
  - Silt %
  - Organic Matter %
Data

- 32-year average climate data:
  1. Length of the Growing Season
  2. Growing Season Precipitation
  3. Number of Days in Growing Season with Precipitation ≥ 13mm
  4. Mean Growing Season Temperature
  5. Annual Precipitation
  6. Mean Annual Temperature
  7. Summer Mean Maximum Temperature
  8. Summer Precipitation
  9. January Mean Maximum Temperature
  10. July Mean Maximum Temperature
  11. Annual Growing Degree Days
  12. Growing Season Growing Degree Days
  13. Summer Growing Degree Days
  14. Summer Dryness Index
  15. Growing Season Dryness Index
  16. Late Summer Precipitation
  17. Summer-Winter Temperature Differential

Total of 24 predictors (soils, elevation, climate)
Random forests variable selection

- Random forest model by *cforest* algorithm
  - $SI = f(24 \text{ biophysical variables})$

- Rank variables based on variable importance (VI) values from *cforest*
  
- Backward elimination using *randomForest* algorithm
  - $SI = f(24 \text{ biophysical variables})$
  - $SI = f(23 \text{ biophysical variables})$
  - $SI = f(22 \text{ biophysical variables})$
  - ......
  - $SI = f(2 \text{ biophysical variables})$ (Highest VI predictors)

- Graph of *randomForest* pseudo $R^2$ to determine effect of variable removal on model fit
Random forests variable selection

1. Annual pptn.
2. Soil depth
3. Soil available water capacity
5. Elevation

1. Summer pptn.
2. Late summer pptn.
3. Mean max. summ. temp
4. Elevation
Random forests models

- **Non-IMP**
  \[ SI = \text{rtf}(\text{Annual pptn.}, \]  
  \[ \text{Soil depth,} \]  
  \[ \text{Soil available water capacity,} \]  
  \[ \text{Growing seas. dry. ind.,} \]  
  \[ \text{Elevation}) \]

- **IMP**
  \[ SI = \text{rtf} (\text{Summer pptn.}, \]  
  \[ \text{Late summer pptn.} \]  
  \[ \text{Mean max. summer temp.,} \]  
  \[ \text{Elevation}) \]

**NB:** rtf – regression tree function
Parametric variable selection

24 Biophysical variables

Factor analysis

Underlying factors

Examine & interpret factor loadings

Important factors – Non-IMP
1. Heat index
2. Precipitation
3. Drought index
4. Soil water availability
5. Soil texture

Important factors – IMP
1. Heat index
2. Precipitation
3. Drought index
4. Soil water availability
Parametric models

**Base model**

\[
SI = \theta_0 \times \left( \frac{T^{\theta_1} \cdot \exp(\theta_2 T) + PR^{\theta_3} \cdot \exp(\theta_4 PR) + AWS^{\theta_5} \cdot \exp(\theta_6 AWS)}{DRY^{\theta_7}} \right) + \epsilon
\]

**Non-IMP fitted model**

\[
SI = \exp(\alpha_0) \times \left( \frac{LGS^{\alpha_1} + AP^{\alpha_2} \times \exp(\alpha_3 AP) + AWS^{\alpha_4} \times \exp(\alpha_5 AWS)}{SDI^{\alpha_6} \times \%sand^{\alpha_7}} \right)
\]

**IMP fitted model**

\[
SI = \exp(\beta_0) \times \left( \frac{LGS^{\beta_1} + AP^{\beta_2} \times \exp(\beta_3 AP) + AWS^{\beta_4} \times \exp(\beta_5 AWS)}{SDI^{\beta_6}} \right)
\]

---

**LGS** – Length of grow. seas.  
**SDI** = summ. dry. ind. \(\times 10\)  
**AP** = ann. pptn. \(\div 10\)  
**AWS** = Available water stor. cap.
### Fit and Prediction Statistics

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>$\text{FI}_{adj}$ (%)</th>
<th>MB (m)</th>
<th>PRMSR (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random forest</td>
<td>Non-IMP</td>
<td>80.39</td>
<td>0.0107</td>
<td>1.980</td>
</tr>
<tr>
<td></td>
<td>IMP</td>
<td>84.71</td>
<td>-0.0028</td>
<td>2.641</td>
</tr>
<tr>
<td>Parametric</td>
<td>Non-IMP</td>
<td>33.58</td>
<td>0.0103</td>
<td>2.409</td>
</tr>
<tr>
<td></td>
<td>IMP</td>
<td>41.61</td>
<td>0.0004</td>
<td>3.106</td>
</tr>
</tbody>
</table>

- $\text{FI}_{adj}$ and MB are from model fitting data
- PRMSR is from validation data
Model prediction behavior

Predicted Vs Observed Site Index

Random forest

Parametric
Model prediction behavior

Under a 20% decrease in annual precipitation (black dots)

Random forest

Parametric
Model prediction behavior

Under a 20% increase in summer dryness index (black dots)

Random forest

Parametric
Summary/Discussion points

- Site index predictions from biophysical variables regressed towards the mean
- Random forest model performed poorly under extrapolation
- Parametric model exhibited logical behavior under extrapolation
Acknowledgements

Forest Modeling Research Cooperative at Virginia Tech

PINEMAP