A comparison of a non-parametric and a parametric approach for predicting loblolly pine site index from biophysical variables

Charles O. Sabatia and Harold E. Burkhart

Postdoctoral Research Associate and University Distinguished Professor
Virginia Polytechnic Institute and State University, Department of Forest Resources and Environmental Conservation
Blacksburg, VA 24061

INTRODUCTION

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). Previous attempts to predict site index from biophysical variables have applied parametric methods such as linear regression (e.g. Monserud et al. 2006) and also non-parametric methods (e.g. Crookston et al. 2010). Either approach has strengths and weaknesses especially when it comes to predicting site index that can be input in a stand growth and yield projection system for plantation loblolly pine. In the current study, we compared the performance of a non-parametric approach called regression trees to that of a parametric approach that involves the use of a nonlinear function fitted by least squares regression.

DATA & METHODS

Study Area and Stand Data

The Virginia Tech Forest Modeling Research Cooperative loblolly pine regionwide thinning studies (Figure 1) formed the basis of the current study. The non-intensively managed plantations (non-IMP) studies were established in 1980 to 1982 in genetically unimproved stands that were 8 to 25 years old. These studies had 8 measurements completed at 3-year intervals. The intensively managed plantations (IMP) studies were established in 1997 to 2000 in genetically improved stands that were 3 to 8 years old. These studies were measured at 2 year intervals and at 3 year intervals at later ages. Plot measurement data for these studies were available for the period since plot establishment to 2011. Each study location in the non-IMP and IMP studies had 3 0.04 to 0.08 ha plots under an unthinned, a light thin, or heavy thin treatment. For the current study, only data from the unthinned treatment is considered. The analysis dataset was made up of dominant height measurements from 172 non-IMP locations and 136 IMP locations and also elevation data. Site index (base age 25) (SI25) for each non-IMP location was computed using the equation

$$Site\ Index = \frac{\alpha + X_0}{1 + \frac{\beta}{X_0} \times 25^{\gamma}} \qquad [1]$$

where $X_0 = 0.5\left(H_d - \alpha + \sqrt{(H_d - \alpha)^2 + 4 \times \beta H_d Age^{\gamma}}\right)$; H_d is the dominant height (m) at the age closest to 25; and α, β, and γ are parameters that were computed separately for each data set .

Soils Data

Soils data for each study location were extracted from the USDA Natural Resource Conservation Service SSURGO GIS database using GIS data extraction techniques. The soils data processed for each location included soil depth (to a 2-meter maximum); percent clay, sand, silt, and organic matter; and soil available water storage in the depth 0 to 150 cm.

Climate Data

Daily climate records for each location for the period 1980 to 2011 were obtained using the Oakridge National Laboratories' daily surface weather prediction models (Thornton et al. 2012). For each location and year, the daily climate data were processed to provide 17 seasonal and annual climate characteristics such as summer precipitation, annual mean temperature, summer dryness index, length of the growing season, etc. The location level value of each climate characteristic was computed as an arithmetic mean of the values of the characteristic over the 32-year climate record period.

The regression trees approach of predicting site index from biophysical variables

This non-parametric approach was implemented using the randomForest algorithm in R software. The algorithm was programmed to fit regression trees (random forest) models that predict site index from the biophysical variables and identify the most influential biophysical variables through backward elimination of predictors and an observation of when random forest model R^2 no longer changed. Nine biophysical variables were identified as the most influential for non-IMP data and 7 as the most influential for IMP data (Table 1). A random forest model based only on the most influential predictors was then fitted for each of the datasets.

Table 1: Biophysical factors identified as most influential in predicting site index

Model	Data	Influential biophysical factors
Non-parametric random forest	IMP	Growing season precipitation frequency; Average growing season temperature; Maximum January temperature; Total summer precipitation; Summer dryness index; Late summer(August) precipitation; January-July temperature differential
	NonIMP	Growing season precipitation; Growing season precipitation frequency; Annual precipitation; summer precipitation; Growing season dryness index; Late summer precipitation; Summer dryness index; Soil depth; Elevation
Parametric nonlinear regression	IMP and NonIMP	Mean annual temperature; Total annual precipitation; Growing Season dryness index; Available soil water storage (0-150cm depth)

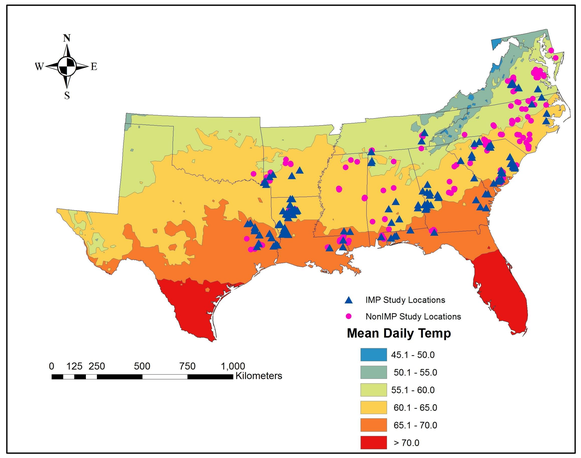

Figure 1: Locations of the Virginia Tech Forest Modeling Research Cooperative regionwide thinning studies superimposed on the map of the annual mean daily temperature across the study region

The factor analysis and nonlinear regression approach of predicting site index from biophysical variables

In this parametric approach, a factor analysis, using SAS software, was performed on the 24 biophysical variables. For both the non-IMP and IMP data, the results of the factor analysis indicated that four factors (accounting for 87 and 86% of the variability respectively) were most responsible for the location to location differences in the biophysical factors. Examination of the factor loadings suggested that the location to location variability in biophysical factors could be accounted for by measures of: **1) temperature, 2) precipitation, 3) dryness index,** and **4) soil water**. The biophysical variables found to best represent these measures, in the nonlinear model used, are given in Table 1. The four biophysical predictors were then incorporated in the following Type II combined power exponential function:

$$SI = \exp(\alpha_1) \times \left(\frac{T^{\alpha_1}.\exp(-\alpha_2 T) + PR^{\beta_1}.\exp(-\beta_2 PR) + AWS^{\gamma_1}.\exp(-\gamma_2 AWS)}{DRY^{\delta}}\right) + \varepsilon \qquad [2]$$

This function assumes that the effects of temperature (*T*), precipitation (*PR*), and available water storage (*AWS*) are additive and that site index would increase, with increase in these factors, to a maximum then decline. The model also assumes that increased dryness would decrease site index to an asymptotic minimum. This model exhibits biologically logical trends with change in the predictor biophysical factors. It was fitted to the non-IMP and IMP data using the Model procedure in SAS software.

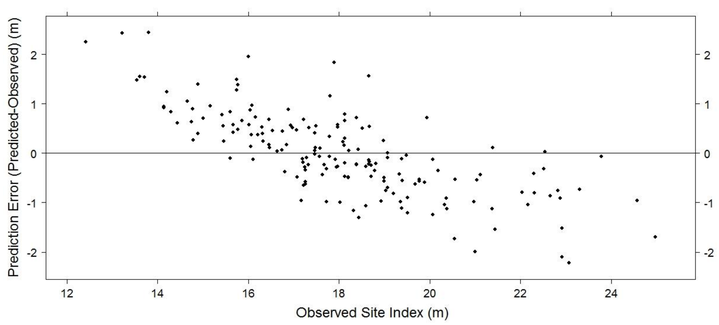

Figure 2: Plot of residuals versus "observed" site index from the biophysical variables parametric model

RESULTS

The parameter estimates when model 2 was fitted to the non-IMP and IMP data are given in Table 2. The parameter estimates were logical and produced expected trends when each factor was varied while holding others constant. The model fit statistics are given in Table 3. When the model residuals were plotted against the observed site index, the trend for non-IMP data appeared as shown in Figure 2. A similar trend was obtained with IMP data.

The model fit statistics for the non-parametric random forest model are also given in Table 3. The trends when the residuals were plotted against the observed site indices were also similar to that in Figure 2.

Table 2: Parameter estimates when model 2 was fitted to the Non-IMP and IMP data

Data	Parameter estimates and their standard errors							
	λ	α_1	α_2	β_1	β_2	γ_1	γ_2	δ
Non-IMP	-19.886 (5.807)	7.327 (1.945)	-	6.006 (1.490)	-0.047 (0.011)	9.4707 (2.912)	-0.379 (0.156)	0.333 (0.121)
IMP	-10.057 (6.003)	5.251 (2.161)	-0.102 (0.048)	4.193 (1.590)	-0.043 (0.016)	7.073 (3.166)	-0.382 (0.184)	0.704 (0.139)

Note: All the parameters, except the λ parameter for IMP data, are significant ($p \leq 0.04$). The standard error of each parameter estimate is given in parenthesis below the estimate. The α_2 parameter was not significant in the non-IMP data hence was excluded from the non-IMP stands model

Table 3: Fit statistics for the non-parametric and parametric models fitted to predict loblolly pine site index from biophysical variables

Model	Data	Fit Index	RMSE (m)
Non-parametric	Non-IMP	0.8759	0.87
	IMP	0.8649	0.98
Parametric	Non-IMP	0.2331	2.11
	IMP	0.1805	2.29

Key Observations

The non-parametric approach resulted in a closer fit to the data used for predicting forest productivity from biophysical variables. It is, however, difficult to provide a biological interpretation to the model. In addition, the model does not provide an explicit mathematical equation that can be easily applied.

The parametric approach resulted in a model with a poorer fit to the data but one that is biologically logical and contains an explicit mathematical function that can be applied.

Both approaches for predicting site index of planted loblolly pine from biophysical factors showed similar trends in the residuals, with over predictions for low quality sites and under predications for high quality sites.

CONCLUSION

Predicting loblolly pine site index from biophysical variables can be done by parametric or non-parametric approaches. The non-parametric approach might result in more accurate predictions but the biological logic of the predictions might need to be evaluated given the black box nature of how the randomForest algorithm handles the predictors. Predictions by either approach may be okay under average site conditions. However, one should be aware of the tendency to over predict for poor sites and under predict for high quality sites.

Literature Cited

Crookston, N.L., G.E. Rehfeldt, G.E. Dixon, and A.R. Weiskittel. 2010. Addressing climate change in the forest vegetation simulator to assess impacts on landscape forest dynamics. For. Ecol. Manag. 260:1198-211

Monserud, R.A. 2006. Predicting lodgepole pine site index from climatic parameters in Alberta. For. Chron. 82:562-571.

Thornton, PE, MM Thornton, BW Mayer, N Wilhelmi, Y Wei, RB Cook 2012. Daymet: Daily surface weather on a 1 km grid for North America,1980 - 2008. Acquired online (http://daymet.ornl.gov/) on 08/Dec/2012 from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A.

Acknowledgements: Stand data for this study were provided by the Forest Modeling Research Cooperative at Virginia Tech. Financial support for data analysis was provided by the USDA NIFA PINEMAP project.