Predicting forest attributes from climate data using a recursive partitioning and regression tree algorithm

Greg C. Liknes[1], Christopher W. Woodall[2], and Charles H. Perry[2]

ABSTRACT: *Climate information frequently is included in geospatial modeling efforts to improve the predictive capability of other data sources. The selection of an appropriate climate data source requires consideration given the number of choices available. With regard to climate data, there are a variety of parameters (e.g., temperature, humidity, precipitation), time intervals (e.g., 30-year normal, seasonal average), and summary statistics (e.g., mean, minimum) which can be selected. In this study, we propose a technique for evaluating the combination of climate parameters that are most closely related to ground observations of forest attributes. Using data from the Forest Inventory and Analysis (FIA) program of the U.S. Forest Service as response variables, recursive partitioning and regression tree analysis was applied using a suite of climate variables from the Daymet database as predictor data. Although model improvement scores for climate variables were modest, the technique provides opportunities for deciding among a wide array of possible climate predictors.*

KEYWORDS: Daymet, climate, forest inventory, data mining

Introduction

Forest composition and structure are, in part, a function of local climatic conditions (Bailey 1995, Whittaker 1975). Geospatial modeling predictions of forest attributes may use climate information to augment topographic and remote sensing data. However, the full range of climate data possibilities are rarely considered because of the daunting number of possible combinations of time interval, descriptive statistics, and spatial resolution. For example, what is the appropriate length of time over which to assess typical rainfall in a given area? Should we consider mean, maximum, or minimum values of climate inputs (e.g., precipitation, solar radiation) and state variables (e.g., temperature, humidity)?

[1]Research Physical Scientist, USDA Forest Service, Northern Research Station, 1992 Folwell Avenue, St. Paul, MN 55108; Corresponding author email: gliknes@fs.fed.us
[2]USDA Forest Service, Northern Research Station, 1992 Folwell Avenue, St. Paul, MN 55108

In: McWilliams, Will; Moisen, Gretchen; Czaplewski, Ray, comps. 2009. 2008 Forest Inventory and Analysis (FIA) Symposium; October 21-23, 2008: Park City, UT. Proc. RMRS-P-56CD. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 1 CD.

Ohmann and Gregory (2002) predicted tree species composition and structure in coastal Oregon using a variety of data. They found Landsat Thematic Mapper satellite imagery and climate information to have the first and second most explanatory power, respectively, followed by location, topography, ownership, and geology. The authors transformed temperature and precipitation data from the Parameter-elevation Regressions on Independent Slopes Model (PRISM) climate dataset to create eight predictor variables. Their climate predictors captured seasonality, variability, growing season conditions, and continentality, but due to the limitations of the PRISM dataset, other factors such as radiation were not included.

Previous work by Liknes and Woodall (2007) began to assess the climate factors that have the most predictive power relative to forest attributes by examining correlations between forest inventory data and a variety of parameters in the Daymet[3] climate database. In this study, we aimed to build on our previous work by re-examining the Daymet dataset with an improved technique.

Data

Data were analyzed for the states of Michigan, Minnesota, and Wisconsin. These states cover 49 million hectares, of which 21 million hectares are sub-boreal and temperate forests.

Forest Inventory Data

Field data were collected between 2001 and 2006 on nearly 20,000 forested or partially-forested plots by the Forest Inventory and Analysis (FIA) program of the U.S. Forest Service. The 0.4-ha plots in the study area are re-visited every five years, allowing calculation of growth and mortality for a subset of plots visited in both 2001 and 2006.

Climate Data

Climate data used in this study were taken from the Daymet climate database (Thornton et al. 1997). The Daymet raster datasets provide full coverage of the conterminous United States at 1-km resolution for a suite of climate parameters including temperature, precipitation, humidity, and radiation. Additionally, measures of climate variability (interannual standard deviation and day-to-day variability) are available for each parameter, as well as selections of a single-year

[3] Daily Surface Weather and Climatological Summaries (http://www.daymet.org)

average, 18-year annual average, or an 18-year average for a specific month of the year. The 18-year average datasets are for the period between 1980 and 1997.

We considered a subset of the Daymet database, with all selected variables averaged over the 18-year period. See the Appendix for a list of the climate variables used in this study.

Methods

Forest inventory data for individual trees were aggregated to calculate biomass, basal area, and growth for each plot. Climate values at each plot were extracted using a Geographic Information System operation, which assigned each plot to a climate pixel based on proximity to the nearest pixel center.

Recursive Partitioning and Regression Tree Analysis

Although various data mining techniques are available, we chose to examine recursive partitioning and regression tree (*rpart*) analysis (Therneau 1983). The *rpart* algorithm shares many similarities with other data mining techniques, particularly classification and regression tree analysis. For more information on classification and regression tree analysis, refer to Breiman et al. (1983). The following advantages for analyzing multiple predictor variables are provided by *rpart*: it is compatible with the prediction of continuous variables, the resulting models can be presented as intuitive binary trees, and it requires relatively few input choices by the user.

We implemented the *rpart* algorithm using R statistical software[4] and the *rpart* package[5]. More detail on the parameters used with *rpart* appears in the Appendix. Six separate model runs were conducted using the suite of climate parameters as input predictors; biomass, basal area, and growth were response variables. Two predictor sets were used. The first set contained a suite of climate variables representing minimum, maximum, or mean daily values or mean annual total values. The second set included all of the first predictor set and additional climate data related to variability (see Appendix). Model improvement for node splits was used as a basis for determining which climate variables had the strongest relationship to forest attributes where improvement is defined as

$$1 - (SS_{right} + SS_{left}) / SS_{parent} \quad (1)$$

and *SS* is the sum-of-squares, *right* and *left* refer to the sides of the split, and *parent* refers to the node that is split.

[4] The version of R used was 2.8.0. The software is available at http://www.r-project.org.
[5] The version of the *rpart* package used was 3.1-42.

Results

For both basal area and biomass, the top node splits were related to minimum daily air temperature, while for growth the top node splits were related to precipitation frequency (Table 1). Figure 1 shows the top three node splits that resulted when an *rpart* model was created using the second predictor dataset (climate and variability, see Appendix) to predict forest biomass. The primary split for biomass occurred on minimum daily December temperature at a value of -15.375 degrees Celsius. Including interannual and day-to-day variability (second predictor set) did not change the results for biomass, but it did change the results for basal area and growth; interannual variability appeared among the top node splits in both cases.

Table 1: Results from six model runs predicting forest attributes using selected climate variables and recursive partioning and regression tree analysis in Minnesota, Wisconsin, and Michigan.

Response variable	Predictor set	Top three node splits	Model improvement
Basal area	1	January minimum daily air temperature	0.02145327
		Mean daily total shortwave radiation	0.01690332
		Mean precipitation event size	0.006725006
Basal area	2	Interannual variability in December mean daily minimum temperature	0.02569515
		Mean daily total shortwave radiation	0.01612992
		Mean precipitation event size	0.003692035
Biomass	1	December minimum daily air temperature	0.03854474
		Mean daily total shortwave radiation	0.03828900
		Minimum daily air temperature	0.01672963
Biomass	2	December minimum daily air temperature	0.03854474
		Mean daily total shortwave radiation	0.03828900
		Minimum daily air temperature	0.01672963
Growth	1	Precipitation frequency (<0.255)	0.009561471
		Precipitation frequency (<0.275)	0.0035711810
		Cooling degree days	0.005642840
Growth	2	Precipitation frequency	0.009561471
		Interannual variability in December mean daily minimum temperature	0.005102520
		Interannual variability in total precipitation	0.006068496

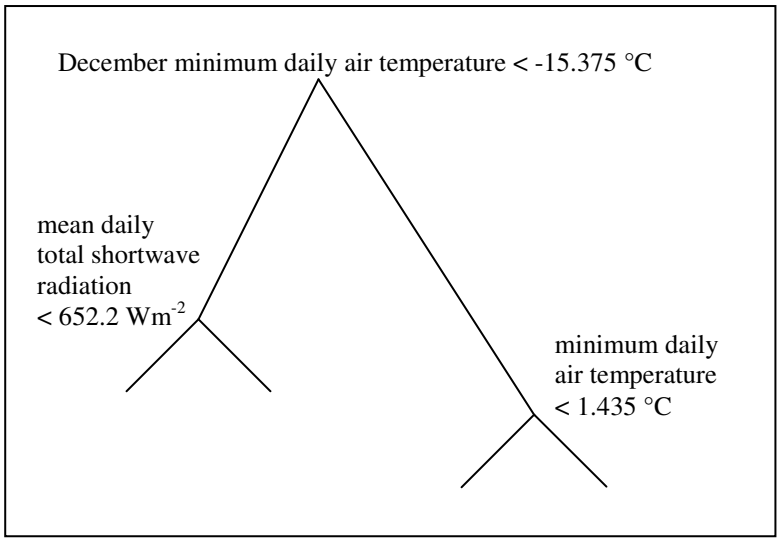

Figure 1: A regression tree resulting from a recursive partitioning and regression tree model of forest biomass in Minnesota, Wisconsin, and Michigan using Daymet climate variables as predictors.

Model improvement values are generally useful as a relative indication of how predictors perform. Improvement of the biomass model due to the predictors at top node splits was approximately three times that of the growth model (0.03 vs. 0.009).

Discussion

Analysis of climate variables with the *rpart* algorithm resulted in modest model improvements. More importantly, the technique allowed for selection of the most influential climate variables from a large set of potential factors. The technique is straightforward to implement and highlights the possible importance of less conventional climate variables (such as mean daily minimum temperatures). The inclusion of some measure of interannual variability of temperature and precipitation may also be warranted for some forest attributes. Furthermore, the predictors selected for model inclusion should depend on the attribute to be modeled. The results presented should not be considered as definitive guidance for the selection of climate variables used in the prediction of forest attributes. For example, future work could include an analysis using *rpart* and the entire Daymet database over a larger geographic area, as well as comparisons of results using other data mining or exploratory data analysis techniques.

Many issues require further exploration to improve the utility of this technique with FIA plot data. For example, there is a mismatch in the areal extent of a

Daymet raster cell (100 ha) and the 0.4-ha FIA plot. The mismatch may have a sizeable impact on the relationship between our response and predictor variables. Additionally, if climate variables are used in conjunction with remotely-sensed data in a model, there may be interactions that are not accounted for in this approach.

A temporal mismatch also may have affected our results. Forest attribute data were obtained between 2001 and 2006, but climatic data were averaged over the period between 1980 and 1997. Contemporaneous climate data may improve model performance, and these data would be available by aggregating monthly PRISM data[6]. However, PRISM has a smaller set of parameters relative to Daymet (e.g., total shortwave radiation is excluded). Although questions surrounding spatial and temporal resolution need to be addressed, *rpart* analysis using climate variables holds promise for improving variable selection and model prediction.

References

Bailey, R.G. 1995. Description of the ecoregions of the United States, 2nd ed. Misc. Publ. 1391. Washington, DC: U.S. Department of Agriculture, Forest Service.

Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 1983. Classification and Regression Trees. Belmont, CA: Wadsworth.

Liknes, G.C; Woodall, C.W. 2007. An evaluation of climate parameters in relation to forest inventory data: towards guidance for satellite-based forest attribute modelling. Proceedings of ForestSat 2007 - Forests and Remote Sensing: Methods and Operational Tools; 2007 Nov 5-7; Montpellier, France. [CD-ROM].

Ohmann, J.L.; Gregory, M.J. 2002. Predictive mapping of forest composition and structure with direct gradient analysis and nearest neighbor imputation in coastal Oregon, U.S.A. Canadian Journal of Forest Research. 32: 725-741.

Therneau, T.M. 1983. A short introduction to recursive partitioning. Orion Technical Report. Palo Alto, CA: Stanford University, Department of Statistics.

Thornton, P.E.; Running, S.W.; White, M.A. 1997. Generating surfaces of daily meteorological variables over large regions of complex terrain. Journal of Hydrology, 190: 214-251.

Whittaker, R.H. 1975. Communities and ecosystems, 2nd ed. New York and London: Macmillan.

[6] PRISM Group (http://www.prism.oregonstate.edu/)

Appendix

rpart parameters used:
minisplit = 100
minibucket = 100
cp = 0.0001
method = anova

Sample *rpart* implementation in R:

```
fit <- rpart(Climate$BasalArea ~ ta_a_pa + tx_a_pa + tn_a_pa + td_a_pa + tc_a_pa +
        tf_a_pa + tn_a_pm12 + tn_a_pm01 + tn_a_pm02 + pf_a_pa + pe_a_pa +
        hv_a_pa + rt_a_pa, control=list(minisplit=100, minibucket=100, cp=0.0001),
        method="anova")
```

Appendix Table: Climate variables used as predictors of basal area, biomass, and growth using the *rpart* algorithm.

Climate variable description	Daymet abbreviation	Predictor Set
Mean daily air temperature	ta_a_pa	1,2
Minimum daily air temperature	tn_a_pa	1,2
Maximum daily air temperature	tx_a_pa	1,2
Growing degree days	td_a_pa	1,2
Cooling degree days	tc_a_pa	1,2
Freezing degree days	tf_a_pa	1,2
January minimum daily air temperature	tn_a_pm01	1,2
February minimum daily air temperature	tn_a_pm02	1,2
December minimum daily air temperature	tn_a_pm12	1,2
Mean precipitation event size	pe_a_pa	1,2
Precipitaion frequency	pf_a_pa	1,2
Mean daily total shortwave radiation	rt_a_pa	1,2
Mean daily water vapor pressure	hv_a_pa	1,2
Day-to-day variability in mean daily temperature	tva_a_pa	2
Day-to-day variability in mean daily minimum temperature	tvn_a_pa	2
Day-to-day variability in mean daily maximum temperature	tvx_a_pa	2
Day-to-day variability in mean daily water vapor pressure	hvv_a_pa	2
Day-to-day variability in total shortwave radiation	rvt_a_pa	2
Interrannual variability in mean daily minimum temperature	tn_s_pa	2
Interannual variability in January mean daily minimum temperature	tn_s_pm01	2
Interannual variability in February mean daily minimum temperature	tn_s_pm02	2
Interannual variability in December mean daily minimum temperature	tn_s_pm12	2
Interannual variability in mean daily maximum temperature	tx_s_pa	2
Interannual variability in total precipitation	pt_s_pa	2
Interannual variability in mean precipitation event size	pe_s_pa	2
Interannual variability in precipitation frequency	pf_s_pa	2