An Emperical Diameter Growth Model for Trees in the US Northern Forest

Neal Maker

October, 2019

Introduction

Numerous tree diameter growth models have been developed for the forests of the Northeast, which vary in their approaches, accuracies, and applicability. The goal of this analysis is to make full use of the US Forest Service's Forest Inventory and Analysis (FIA) data to improve on the accuracy of existing models without appreciably increasing inventory costs (by using easily observed predictors). A particular focus is put on accounting for the interactions between predictors, so that subtle differences can be modeled between individual trees in more complex arrangements. The model will be useful for informing management decisions on multi-species, multi-age, and irregularly structured forests in particular.

An individual tree model (which predicts growth for individual trees rather than on a per area basis) was chosen to better account for complex forests. Many forests in the Northeast have been affected by multiple disturbances of varying intensity (opportunistic logging chief among them) and are now irregularly structured and compositionally diverse (Teck and Hilt 1991). Management schemes focused on growing high quality logs must be responsive to these variations; and individual tree models do a better job accounting for such heterogeneity than stand-level models, which are better suited to even-aged, monospecific stands (Peng 2000).

Distance-dependent modeling can help address heterogeneity as well, by accounting for the spatial relationships between individual trees to more accurately estimate competition between them. The cost of obtaining geographic information for individual trees is still high, though, and a distance-independent model is easier to use with existing forest inventory data. Also, competition indices can be obtained from conventional (non-spatial) inventory techniques, and in many cases they can be used to derive growth estimates with accuracies comparable to those derived from spatially-explicit competition indices (Kuehne, Weiskittel, and Waskiewicz 2019).

A number of distance-independent, individual tree diameter growth models have been developed for use in forest management planning in the Northeast. Teck and Hilt (1991) used data from 14 Northeastern states to predict the potential diameter growth for separate species, based on tree diameter at breast height (dbh) and a measure of site class (aka productivity). Overtopping basal area (a measure of competition for individual trees) was then used to modify potential growth downward and obtain actual growth predictions. These species-specific models were incorporated into the NE-Twigs and FVS forest growth simulators.

Westfall (2006) used a similarly broad geographic extent, but employed a mixed-effects model, which allowed different species to be modeled together, overcoming sample size limitations common to species-specific models. A greater number of predictors were used, which included crown ratio (the percent of a tree's height with a live crown), basal area (a measure of forest stocking), latitude, longitude, and elevation.

A. Weiskittel et al. (2016; see also A. Weiskittel et al. 2019) recognized that the existing models were biased in the New York Adirondacks, and developed a more targeted model based on data from five experimental forests in the region. Theirs is also a mixed-effects model, with model coefficients varying by species. Like Teck and Hilt, they used only four predictors: species, dbh, overtopping basal area, and site class.

The model described here is built on a considerably larger sample than previous models, which was drawn from the US Northern Forest region (Maker 2019). The Northern Forest covers a fairly broad geographic area while still representing a coherent socio-ecological unit–in which trees can be expected to follow a similar set of patterns.

A random forest algorithm was chosen to train the model, for a number of reasons. first, random forests can handle a large number of predictors, unlike generative models (which would require many parameters and be subject to overfitting) or nearest neighbor-type algorithms (which suffer the "curse of dimensionality"). Second, they are computationally efficient, especially when being used for prediction. Because the model will be used by practitioners "in the field", this will be a major asset. Finally, random forest algorithms can account for the interactions between numeric and categorical predictors, like those between species and latitude, or between stand basal area and forest type. Interactions like these are not accounted for in most existing growth models, and they can help to describe some of the variation in diameter growth rates, increasing the accuracy.

Data & Analysis

FIA data are collected by the US Forest Service across the country and across ownerships. Data are collected from a stratified random sample of permanent plots, which are periodically reinventoried so that changes to the country's forests can be observed. They are stored in a publicly available relational database[1] that includes information about site characteristics, individual trees, and growth rates.

This analysis was carried out within the statistical computing environment R,[2] and FIA remeasurement data were obtained from a Northern Forest dataset that was built to support forest growth modeling (Maker 2019).[3] Its data were collected during periodic inventories between 1999 and 2019, from plots that were remeasured every 5.1 years, on average.

The growth rates and predictive variables used in this analysis came from the remeasurement of 363,352 individual trees, located on a total of 10,307 plots spread relatively evenly across the Adirondacks, northern Vermont, northern New Hampshire, and northern Maine (see Maker 2019 for an explanation of the study area). For comparison, Westfall (2006) used 32,547 observations from 2,370 FIA plots across 13 northeastern states; and A. Weiskittel et al. (2016) used 25,438 observations from 577 plots on five different properties in the Adirondacks.

Sixteen predictors were used in the model, including site-specific variables (site class, slope, aspect, latitude, longitude, elevation, landscape position, forest type, stocking, and basal area) and tree-specific variables (species, dbh, crown ratio, crown class, tree class, and overtopping basal area). They are described by Maker (2019). A random subset of 20% of the observations was reserved for testing the final growth model and The remaining 80% was used for exploratory analysis and model training.

Exploratory Analysis

Diameter growth rates in the region range from 0 to 0.52 inches per year, measured at breast height (4.5 feet above the ground), with a mean growth rate of 0.076 inches per year. Growth rates above 0.4 inches per year are very uncommon, and are all from larger than average white and red pines with crown ratios greater than 60%.

As has been demonstrated in previous studies (Teck and Hilt 1991; Pacala et al. 1996; Lessard, McRoberts, and Holdaway 2000; Bragg 2005; A. Weiskittel et al. 2016), diameter growth is highly correlated to dbh, with distinct growth curves for individual species (figure 1). Red pine, white pine, and red oak have the highest growth rates overall, while striped maple and Norway spruce have the lowest (table 1).

Factors that account for competition between trees appear to be correlated with diameter growth as well. Crown ratio exhibits a positive, linear relationship to diameter growth (figure 2), basal area exhibits a negative, linear relationship (figure 3), overtopping basal area exhibits a negative relationship that levels out above about 300 ft^2/ac (figure 4), and diameter growth slows as forest stocking increases (figure 5) and as trees' crowns become more impacted by their neighbors (figure 6).

[1] https://www.fia.fs.fed.us/
[2] The R Foundation: https://www.r-project.org/
[3] https://github.com/nealmaker/fia-data-nf

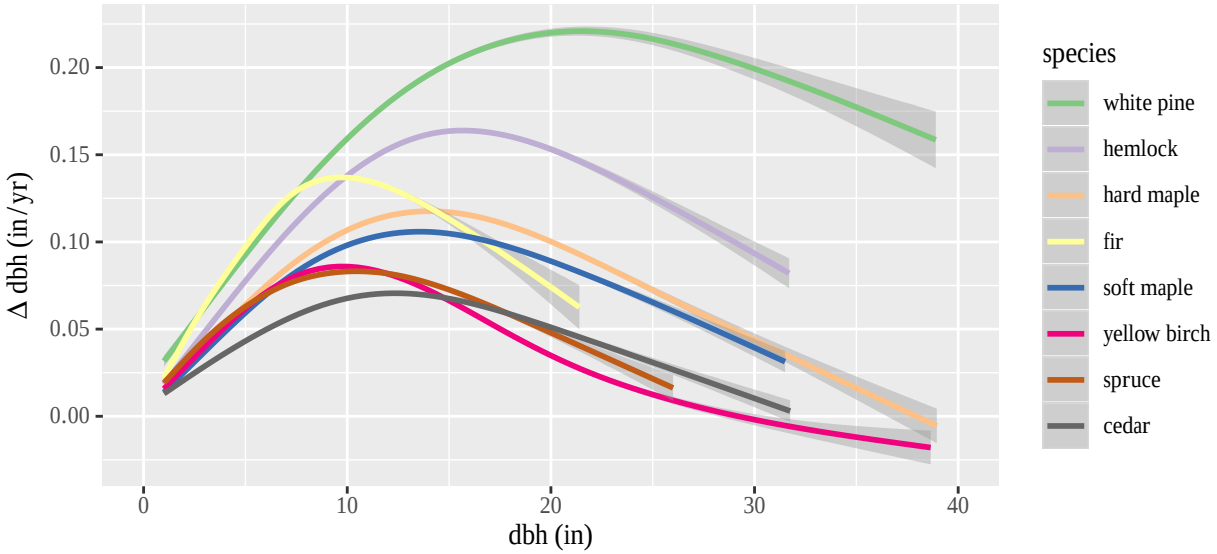

Figure 1: Diameter growth trends for a selection of species groups, smoothed using generalized addative models. Shaded regions show 95% confidance intervals.

Table 1: Sample size (n), mean diameter growth in inches per year (Δdbh), and standard deviation of growth (sd) for species groups in the training data.

species group	n	Δdbh	sd	species group	n	Δdbh	sd
red pine	1012	0.160	0.060	soft maple	36256	0.070	0.032
white pine	9513	0.146	0.067	black cherry	2163	0.065	0.029
red oak	3199	0.144	0.057	yellow birch	16247	0.064	0.028
cottonwood	28	0.142	0.098	spruce	42080	0.064	0.032
hickory	291	0.139	0.055	elm	723	0.060	0.030
hemlock	13816	0.115	0.049	paper birch	15794	0.057	0.024
butternut	35	0.089	0.027	cedar	19536	0.055	0.020
aspen	7808	0.089	0.044	other softwood	87	0.050	0.030
ash	7080	0.087	0.046	hophornbeam	1837	0.045	0.025
hard maple	19955	0.085	0.036	tamarack	1418	0.044	0.021
white oak	185	0.082	0.030	scots pine	156	0.043	0.023
fir	62323	0.080	0.049	other hardwood	4940	0.042	0.029
beech	18770	0.080	0.037	norway spruce	598	0.035	0.018
basswood	532	0.071	0.028	striped maple	4298	0.022	0.015

Several unexpected relationships stand out. The first is the finding that "deep sands" and "beaver ponds" are the landscape types with the highest average diameter growth rates, even though they are generally associated with poorer sites. (Their growth rates average 0.102 and 0.095 inches per year, respectively.) In the case of the deep sands, the reason is that they support an abundance of white and red pine. Some 30 percent of the sampled trees in deep sands are pines, compared to just 3 percent region wide; and white and red pine are the fastest growing species.

Beaver ponds tell a different story. While they support species with closer to average growth rates (fir and soft maple mostly), their forests have the lowest average basal area of any landscape type (100 ft^2/ac compared to a regional average of 132 ft^2/ac), and their trees have the highest average crown ratio as a result (50 percent compared to a regional average of 41 percent). Apparently when trees do get established in beaver ponds, they have relatively few neighbors, get a lot of sun, and grow well.

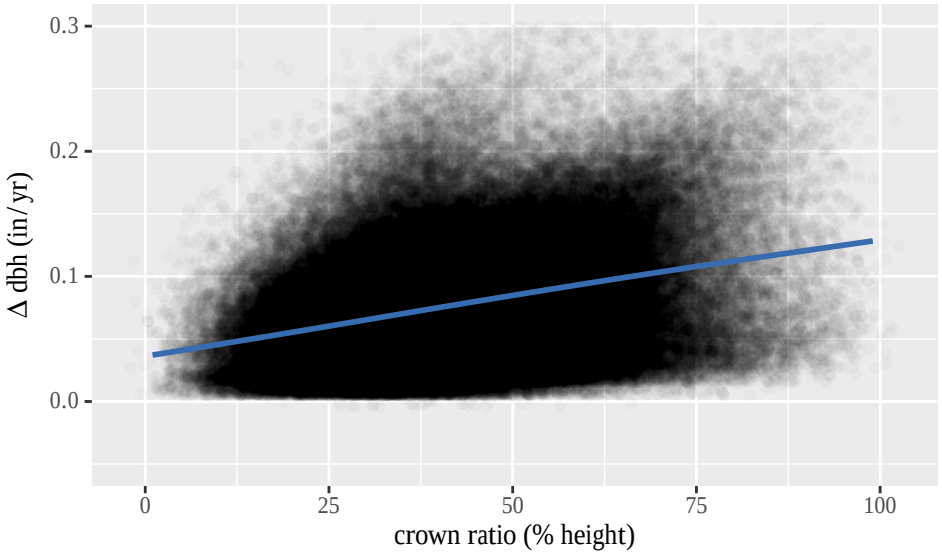

Figure 2: Crown ratio and diameter growth of individual trees. Observations are displayed with random vertical and horizontal offset and partial transparency so that their relative concentration can be visualized. Darker areas show a greater concentration of observations. Trend line in blue calculated using generalized additive model.

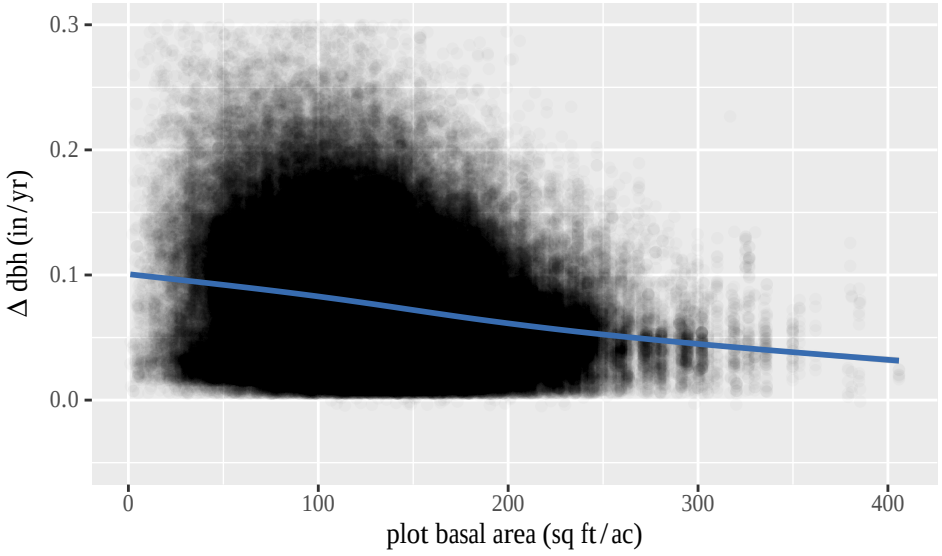

Figure 3: Plot basal area and diameter growth of individual trees. Observations are displayed with random vertical and horizontal offset and partial transparency so that their relative concentration can be visualized. Darker areas show a greater concentration of observations. Trend line in blue calculated using generalized additive model.

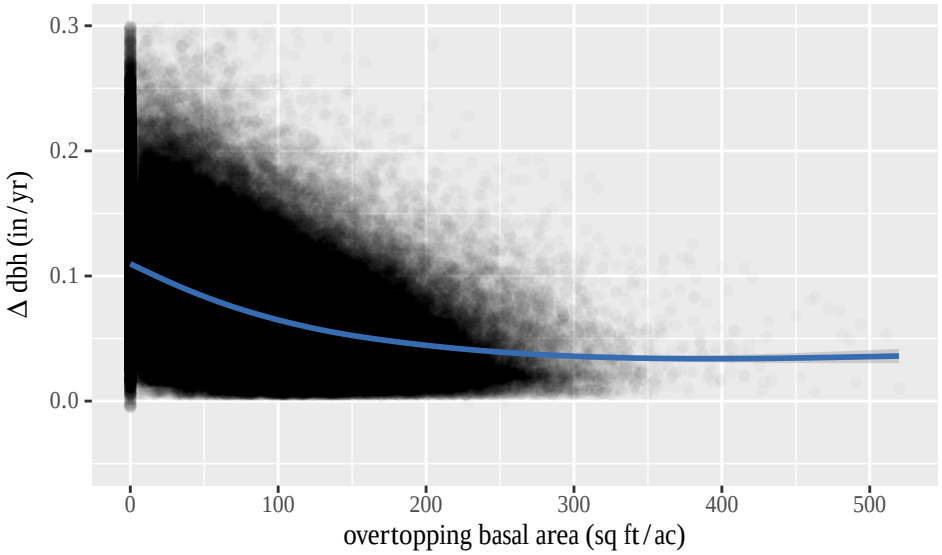

Figure 4: Overtopping basal area and diameter growth of individual trees. Observations are displayed with random vertical offset and partial transparency so that their relative concentration can be visualized. Darker areas show a greater concentration of observations. Trend line in blue calculated using generalized additive model.

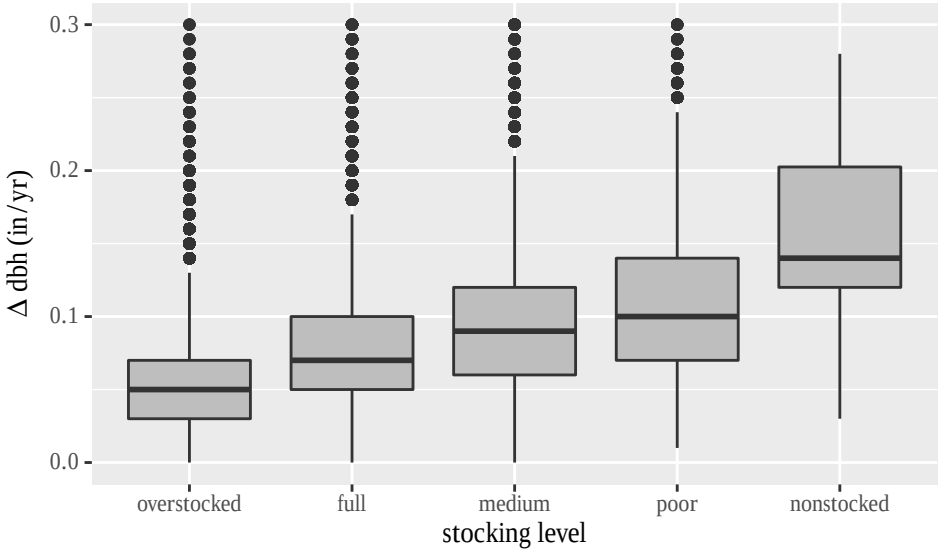

Figure 5: Diameter growth distributions at different stand stocking levels.

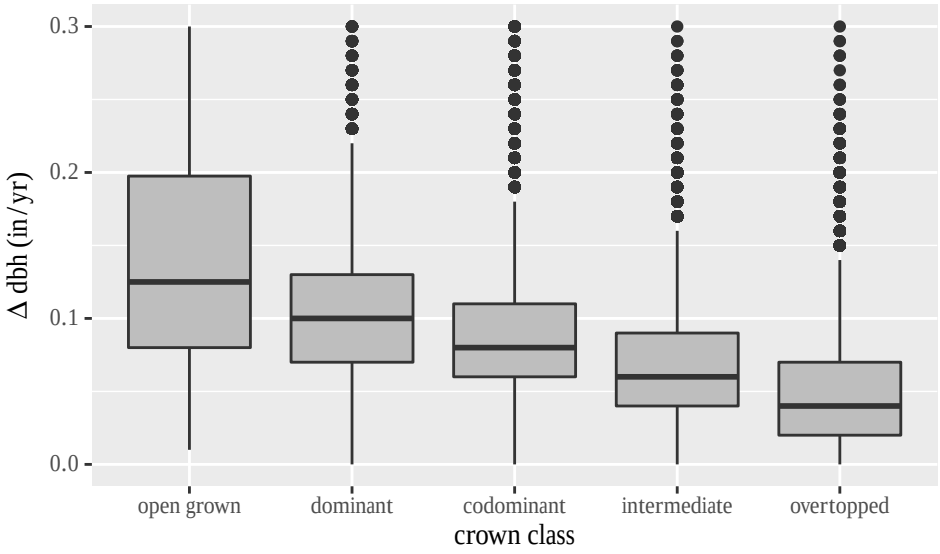

Figure 6: Diameter growth distributions by crown class.

Another notable relationship, which highlights the importance of accounting for interactions, is seen in the geographic variation of growth rates. On their own, latitude and longitude explain only a small amount of the variation in diameter growth. When latitude and longitude are assessed in tandem, however, a much stronger relationship becomes apparent, with the relationship between latitude and diameter growth varying based on longitude. The Black River valley west of the Adirondacks, the Champlain Valley on the border of New York and Vermont, and the southern Maine coast have above average growth rates; while the central Adirondack plateau, New Hampshire's White Mountains, and (most notably) north central Maine have below average rates (figure 7). These geographic differences probably relate in part to climatic factors, which are not otherwise accounted for in the model; and in part to variable species compositions, site classes, and management regimes, which are already accounted for (as management regimes chiefly affect stocking, species composition, and tree diameters).

Pre-Processing & Model Formulation

Prior to training the model, all numeric predictors were centered around zero, scaled based on standard deviations, and transformed using Yeo-Johnson transformations (Yeo and Johnson 2000). Predictors' distributions were standardized in this way to increase the efficiency of model training and to increase the model's accuracy.

Initially, all of the predictors in the training data were retained and used to train a random forest model for which their variable importance could be calculated. Examination of numeric predictors' correlation coefficients did not uncover any very highly correlated features that would warrant eliminating predictors prior to training.

The "ranger" implementation of the random forest was used for this and subsequent model training (**???**), with 200 regression trees in each forest, splits within each tree determined based on estimated response variances, and minimum node sizes of five observations. The optimal number of random predictors to use for each tree was determined using cross validation. Variable importance was calculated for each predictor as impurity, based on the variance of the responses.

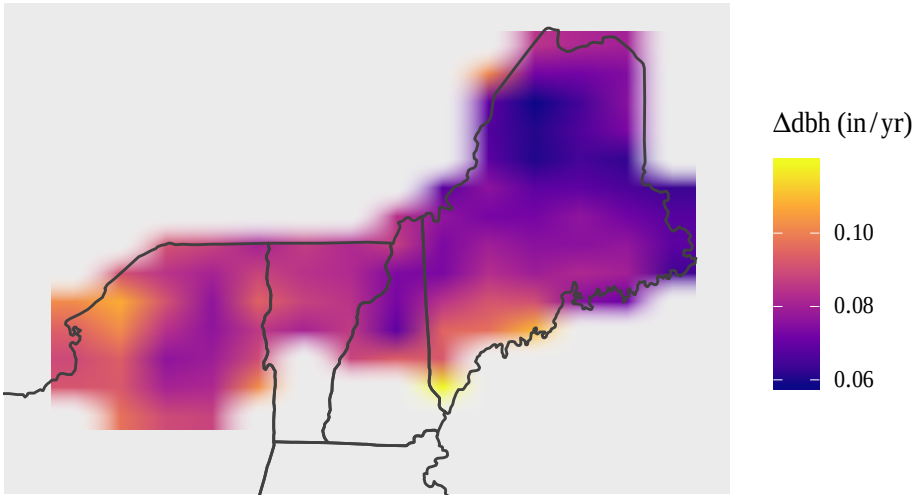

Figure 7: Geographic variation in diameter growth, depicted using two dimensional bin smoothing and interpolation.

Table 2: Scaled variable importance of predictors in full model. Larger values indicate more important predictors.

predictor	importance
dbh	100.00
species	51.33
crown ratio	26.36
basal area	20.71
overtopping basal area	5.09
forest type	2.84
latitude	1.44
longitude	1.33
elevation	1.04
stocking	0.57
aspect	0.26
slope	0.26
tree class	0.13
crown class	0.11
site class	0.04
landscape position	0.00

As expected based on the exploratory analysis, dbh and species were the most important predictors of diameter growth in the initial (full) model, followed by crown ratio, plot basal area, and overtopping basal area (table 2). Landscape position, site class, crown class, tree class, aspect, slope, stocking, and elevation all had low variable importance. They were removed and a second (operational) random forest model was fit to the remaining data, with the hope of simplifying inventory procedures without much loss of predictive accuracy.

Results

The overall accuracies of the two models were similar. The root mean square error (RMSE) of the full model was estimated at 0.00836 inches per year, while that of the operational model was estimated at 0.00848 inches per year, based on testing against the independent test data. Coefficients of determination were similar as well, with the full model explaining an estimated 96.49 percent of the variation in diameter growth and the operational model explaining an estimated 96.39 percent.

Both models have errors narrowly distributed around zero (figure 8), showing that the model is unbiased for the Northern Forest region as a whole. Predictions are also unbiased for many species groups, although several do show limited bias (figure 9). The most notable is cottonwood, whose growth rates are under-predicted by approximately 0.023 inches per year on average. Predictions are much less biased for more common species. Mean absolute errors for the eight most common species groups in the region (fir, spruce, soft maple, hard maple, cedar, beech, yellow birch, and paper birch) are all less than 0.001 inches per year.

Conclusions

The loss of accuracy in moving from the full model to the operational model looks to be very minimal, and the operational model should keep inventory costs down because technicians will not have to record attributes like crown class, slope, and aspect to allow for accurate diameter growth predictions.

The operational model described here does appear to have increased the accuracy of diameter growth predictions in the Northern Forest region over previous models. For example, its species specific normalized RMSEs ($RMSE/\mu$) range from 0.06 (for butternut) to 0.64 (for Scots pine) and average 0.18; and compare

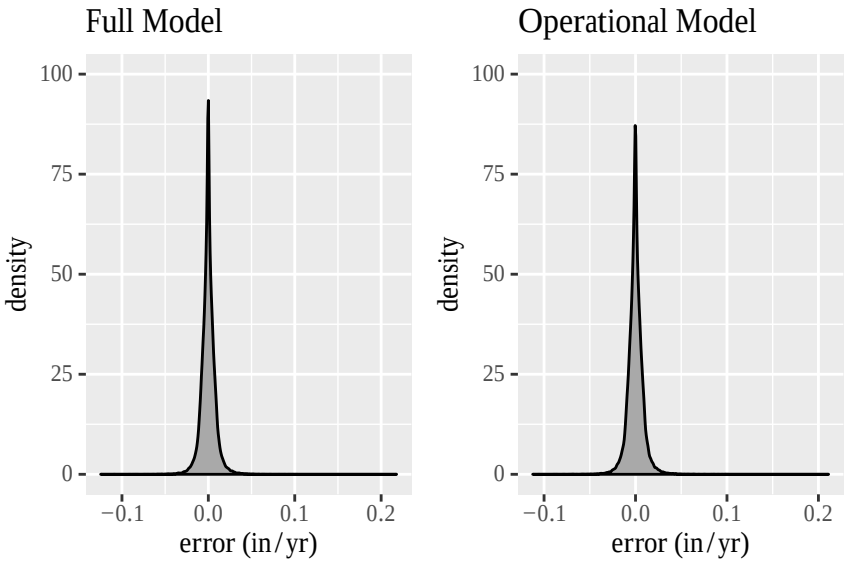

Figure 8: Kernel density estimate of error distributions of full and operational models, fit to test data. Negative errors are underpredictions.

favorably with those of A. Weiskittel et al. (2016), which range from 0.48 to 0.73 and average 0.61; and those of Teck and Hilt (1991), which range from 0.57 to 1.03 and average 0.74.

The model also improves somewhat on species-specific bias, with mean absolute errors for the ten of the most common species ranging from 0.003 to 0.011 inches per year and averaging 0.006 inches per year. A. Weiskittel et al. (2016) report absolute mean biases for the ten common species that range from 0.002 to 0.019 inches per year and average 0.010 inches per year.

It is clear from this study and from other studies in the Northeast that dbh and species are two of the most important predictors of diameter growth. Inter-tree competition can also explain some of the variability in growth rates, though different studies in the region disagree on the best metrics to use. Teck and Hilt (1991) and A. Weiskittel et al. (2016) favored overtopping basal area; Kiernan, Bevilacqua, and Nyland (2008) found that tree-specific measures of competition were not useful and used plot-level basal area alone to describe the effects of competition; and Westfall (2006) found Crown ratio to be important, even though Kiernan, Bevilacqua, and Nyland (2008) and A. Weiskittel et al. (2016) rejected it as superfluous. In the models developed here, crown ratio was the most important competition index, followed by plot-level basal area and overtopping basal area. Crown class and stocking explained very little diameter growth variability and were excluded from the operational model.

It makes intuitive sense that crown ratio and plot basal area would contribute differently to diameter growth, as they describe fundamentally different attributes. Plot-level metrics like basal area describe a tree's access to external resources (light and water, for example), while tree-level metrics like crown ratio describe a tree's internal resources (the size of its growth engine). Part of the success of this model does seem to stem from its accounting for multiple, functionally different competition indecies.

This model also benefits from relatively fast prediction, which is especially important for use in regular forest planning. A k-nearest neighbor algorithm would probably be more accurate and much faster to train, for example, but it would be slow to make predictions with, making it a poor choice.

Perhaps the greatest detriment of this model is its lack of transparency and portability. Previous, parametric models can be expressed using equations with species-specific coefficients, and are easily reproduced by foresters with a working knowledge of any spreadsheet program. The random forest algorithm, on the other hand, cannot be expressed as a simple formula and will be inaccessible to many working foresters. Foresters

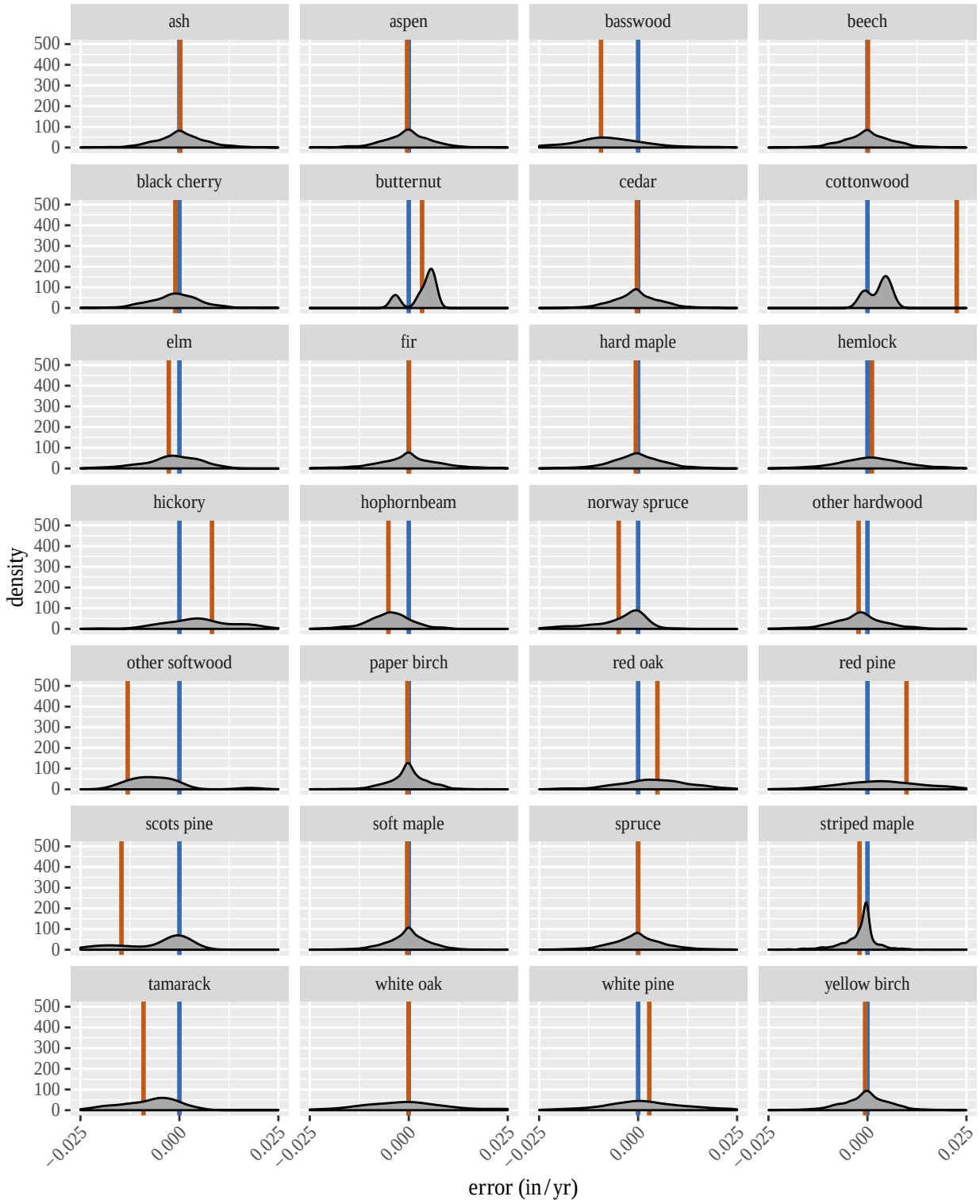

Figure 9: Kernal density estimates of error distributions for individual species groups in the operational model. Vertical blue lines show zero (no error) and vertical brown lines show species groups' average errors. Negative errors are underpredictions.

who do already work in R can obtain the code from GitHub[4] and incorporate the model into their workflows.

While opaque, the random forest algorithm offers many benefits for growth modeling. It is non-parametric, and does not depend on any assumptions about the distributions of the various factors. This is a major benefit when working with forestland attributes, which often have skewed distributions and some of which have poorly understood distributions. Also, the random forest does not presuppose the forms which growth relationships will take. It builds the optimal forms itself based on the data, limiting the bias introduced by humans.

This latter point is both a blessing and a curse. By being purely empirical, non-parametric models like the random forest can limit bias and increase accuracy, but they also limit the model's ability to extrapolate. Models that fail to capture the underlying processes in forest growth are poorly suited for predicting growth in novel scenarios. These models are well suited to predicting the growth of Adirondack trees over relatively short time scales, but they are inappropriate for modeling the effects of climate change on growth, or for modeling the outcomes of new management systems.

Bibliography

Bragg, Don C. 2005. "Optimal Tree Increment Models for the Northeastern United States." *In: Proceedings of the Fifth Annual Forest Inventory and Analysis Symposium; 2003 November 18-20; New Orleans, LA. Gen. Tech. Rep. WO-69. Washington, DC: U.S. Department of Agriculture Forest Service. 222p.* 069. https://www.fs.usda.gov/treesearch/pubs/14282.

Kiernan, Diane H., Eddie Bevilacqua, and Ralph D. Nyland. 2008. "Individual-Tree Diameter Growth Model for Sugar Maple Trees in Uneven-Aged Northern Hardwood Stands Under Selection System." *Forest Ecology and Management* 256 (9): 1579–86. doi:10.1016/j.foreco.2008.06.015.

Kuehne, Christian, Aaron R. Weiskittel, and Justin Waskiewicz. 2019. "Comparing Performance of Contrasting Distance-Independent and Distance-Dependent Competition Metrics in Predicting Individual Tree Diameter Increment and Survival Within Structurally-Heterogeneous, Mixed-Species Forests of Northeastern United States." *Forest Ecology and Management* 433 (February): 205–16. doi:10.1016/j.foreco.2018.11.002.

Lessard, Veronica C., Ronald E. McRoberts, and Margaret R. Holdaway. 2000. "Diameter Growth Models Using FIA Data from the Northeastern, Southern, and North Central Research Stations." *In: McRoberts, Ronald E.; Reams, Gregory A.; van Deusen, Paul C., Eds. Proceedings of the First Annual Forest Inventory and Analysis Symposium; Gen. Tech. Rep. NC-213. St. Paul, MN: U.S. Department of Agriculture, Forest Service, North Central Research Station: 37-42* 213. https://www.fs.usda.gov/treesearch/pubs/14374.

Maker, Neal. 2019. "Development of an FIA Dataset to Model Tree-Level Changes in the Northern Forest." https://github.com/nealmaker/fia-data-nf.

Pacala, Stephen W., Charles D. Canham, John Saponara, John Silander, Richard Kobe, and Eric Ribbens. 1996. "Forest Models Defined by Field Measurements: Estimation, Error Analysis and Dynamics." *Ecological Monographs* 66 (1): 1–43. https://msu.edu/~kobe/docs/pacala%20et%20al%2096%20ecol%20monographs.pdf.

Peng, Changhui. 2000. "Growth and Yield Models for Uneven-Aged Stands: Past, Present and Future." *Forest Ecology and Management* 132: 259–79.

Teck, Richard M., and Donald E. Hilt. 1991. "Individual Tree-Diameter Growth Model for the Northeastern United States." *Res. Pap. NE-649. Radnor, PA: US. Department of Agriculture, Forest Service, Northeastern Forest Experiment Station. 11 P* 649. doi:10.2737/NE-RP-649.

Weiskittel, Aaron, Christian Kuehne, John Paul McTague, and Mike Oppenheimer. 2016. "Development and Evaluation of an Individual Tree Growth and Yield Model for the Mixed Species Forest of the Adirondacks

[4] https://github.com/nealmaker/dbh-growth-nf

Region of New York, USA." *Forest Ecosystems* 3 (1). doi:10.1186/s40663-016-0086-3.

Weiskittel, Aaron, Christian Kuehne, John McTague, and Mike Oppenheimer. 2019. "Correction to: Development and Evaluation of an Individual Tree Growth and Yield Model for the Mixed Species Forest of the Adirondacks Region of New York, USA." *Forest Ecosystems* 6 (December). doi:10.1186/s40663-019-0182-2.

Westfall, James A. 2006. "Predicting Past and Future Diameter Growth for Trees in the Northeastern United States." *Canadian Journal of Forest Research 36:1551-1562* 36. https://www.fs.usda.gov/treesearch/pubs/15883.

Yeo, In-Kwon, and Richard A. Johnson. 2000. "A New Family of Power Transformations to Improve Normality or Symmetry." *Biometrika* 87 (4): 954–59. doi:10.1093/biomet/87.4.954.