An Emperical Diameter Growth Model for Trees in the Adirondacks

Neal Maker

May 28, 2019

Introduction

The principal goal of this analysis was develop a diameter growth model for trees in New York state's Adirondack region, which can be incorporated into Pekin Branch Forestry's[1] forest inventory, analysis and planning toolbox. The model will be used to predict the growth of existing forest stands to aid in management planning.

An individual tree model (which predicts growth for individual trees rather than on a per area basis) will be most useful. 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 too, 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 will be easier to use with existing forest inventory data. Also, competition indicies 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 indicies (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. They are applicable to a wide geographic area, but lack the accuracy of models with narrower focus.

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 these broad-area models are biased in the Adirondacks, and developed a more targeted model based on data from four 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.

This analysis was conducted to further increase the accuracy of diameter growth predictions for the Adirondacks, by making full use of the potential predictor variables available, by better accounting for the interactions that occur between predictors, and by taking advantage of the large Forest Inventory and Analysis (FIA) dataset that is available for the region.

[1] www.pekinbranch.com

Data & Analysis

FIA data are collected by the US Forest Service across the country and across ownerships. Data are collected from 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[2] that includes information about site characteristics, individual trees, and growth rates. This analysis was carried out within the statistical computing environment R[3] and FIA data were obtained using the laselva tool that was developed for R by Chamberlain (2018).

In the Adirondacks, FIA data include diameter growth rates for 83,660 individual trees, measured on a total of 2,355 plots. A number of plots lack condition data, however, including key predictive variables like site class, stand stocking, and forest type. Data from these plots were removed from the analysis, and a total of 65,748 observations were retained, representing 52 unique species, collected from 1,860 different plots. The remaining plots are uniformly distributed across every Adirondack county. For comparison, A. Weiskittel et al. (2016) used 25,438 diameter growth observations from 577 total plots on five different properties.

Several changes were made to the data to make them compatible with Pekin Branch Forestry's inventory protocols. Species were grouped into functional categories (combining some species within genera and grouping uncommon species and species without commercial value) and forest types were recombined to match Pekin Branch's categorization scheme. In the reformatted dataset, trees are grouped into 26 species groups and 17 forest types.

Factors used as predictors in the final model include site-specific variables (site class, slope, aspect, latitude, longitude, landscape position, forest type, forest stocking, and basal area) and tree-specific variables (species, dbh, crown class, and crown ratio). Site class is a discrete ordinal measure of the site's inherent productivity. Landscape position is a categorical factor that classifies sites according to various site and soil conditions and moisture regimes; with levels that include "deep sands", "dry tops", "moist slopes & coves", and "flatwoods", among others. Forest type is defined by the species composition of the current forest on a site. And crown class is a discrete ordinal measure of a tree's access to light, ranging from "open grown", to "overtopped".

Once the data was cleaned and pre-processed, and the variables of interest identified, a random subset of 20% of the observations was reserved for testing the final growth model. The remaining 80% was used for exploratory analysis and model training.

Exploratory Analysis

Diameter growth rates in the region range from 0 to 0.58 inches per year, measured at breast height (4.5 feet above the ground), with a mean growth rate of 0.083 inches per year. Growth rates above 0.3 inches per year are very uncommon, and are almost all from white pine trees with crown ratios greater than 50%.

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). Cottonwood, white pine, and red oak have the highest growth rates overall, and norway spruce has the lowest (table 1).

Factors that account for competition between trees are correlated with diameter growth as well. Crown ratio looks to have the clearest relationship (figure 2), with growth increasing at a decreasing rate as crown ratios rise. Plot level stocking (which depends on the forest type) also shows a slight correlation to growth, with slower growing trees in more densely stocked areas (figure 3). Growth seems to slow again slightly in nonstocked plot areas; probably because they include poorer sites that do not support the growth of closed canopy forests.

[2] https://www.fia.fs.fed.us/

[3] The R Foundation: https://www.r-project.org/

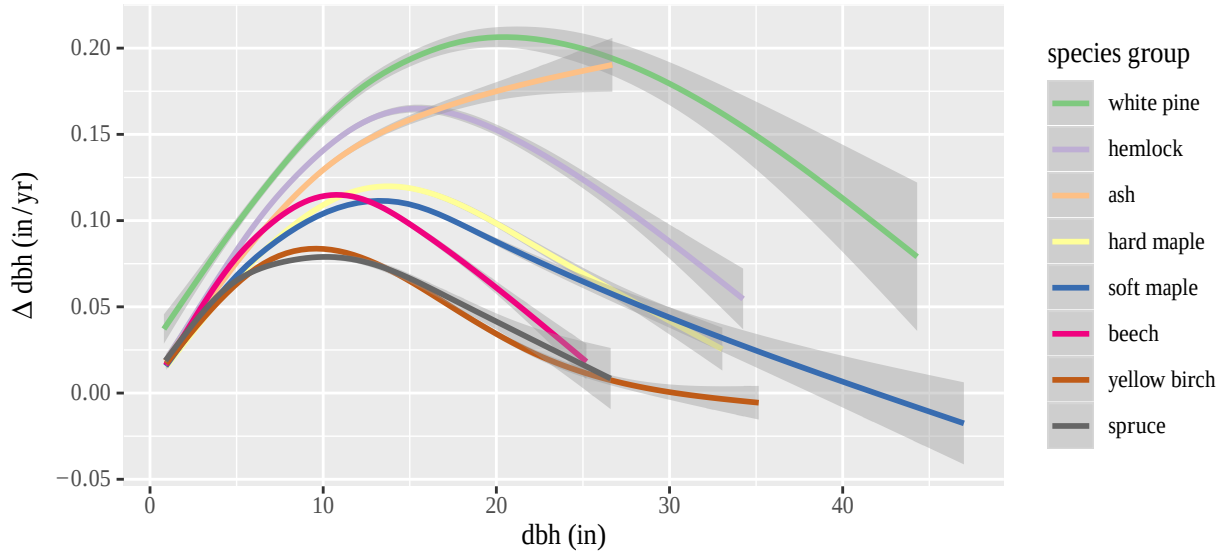

Figure 1: Diameter growth trends for common species groups, smoothed using generalized addative models. Shaded regions show 95% confidance intervals.

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
cottonwood	7	0.216	0.047	beech	7926	0.072	0.037
white pine	2062	0.145	0.071	black cherry	1444	0.070	0.029
red oak	539	0.144	0.063	basswood	422	0.069	0.029
red pine	378	0.138	0.054	yellow birch	3228	0.062	0.027
hickory	338	0.125	0.053	paper birch	1162	0.062	0.019
hemlock	3117	0.119	0.046	elm	587	0.061	0.027
aspen	759	0.108	0.051	spruce	5160	0.059	0.032
fir	5541	0.094	0.052	other hardwood	433	0.058	0.030
ash	2197	0.094	0.047	cedar	1067	0.056	0.025
butternut	14	0.091	0.029	tamarack	258	0.044	0.024
hard maple	6537	0.088	0.036	scots pine	201	0.042	0.019
white oak	128	0.086	0.029	other softwood	51	0.041	0.011
soft maple	8711	0.082	0.031	norway spruce	329	0.032	0.010

Other measures of inter-tree competition have more complex relationships to diameter growth because of interactions with other factors. Plot basal area, like stocking, tends to be negatively correlated to growth, but the relationship is obscured at the ends of the basal area spectrum (figure 4). Plots with basal areas greater than 400 square feet per acre show a rise in per tree diameter growth because they are all white pine forests (one of the fastest growing species) with relatively large trees, located on productive sites. Very low basal area plots show a marked decrease in per tree diameter growth because they tend to be less productive sites with small trees.

Likewise, crown class shows a clear relationship to diameter growth, with more dominant trees growing faster, except that open grown trees are among the slowest growing (figure 5). Like with basal area, this results from the fact that open grown trees in the data are disproportionately small. While the mean tree diameter in the data overall is 7.39 inches, among open grown trees it is only 3.23 inches.

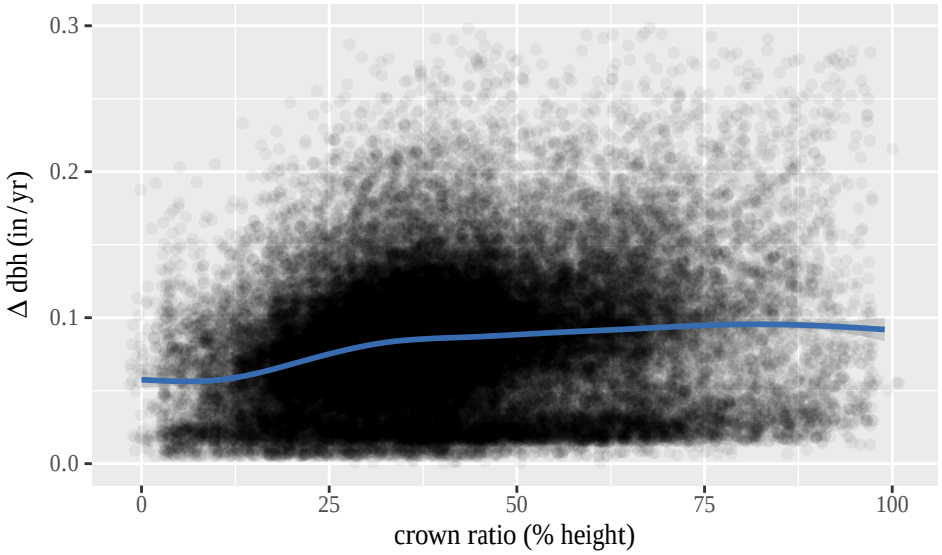

Figure 2: Crown ratio and diameter growth of individual trees. Observations are displayed with random vertical and horizontal offset and partial transparency so that their relative concentration can be visualized. Darker areas show a greater concentration of observations. Trend line in blue calculated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

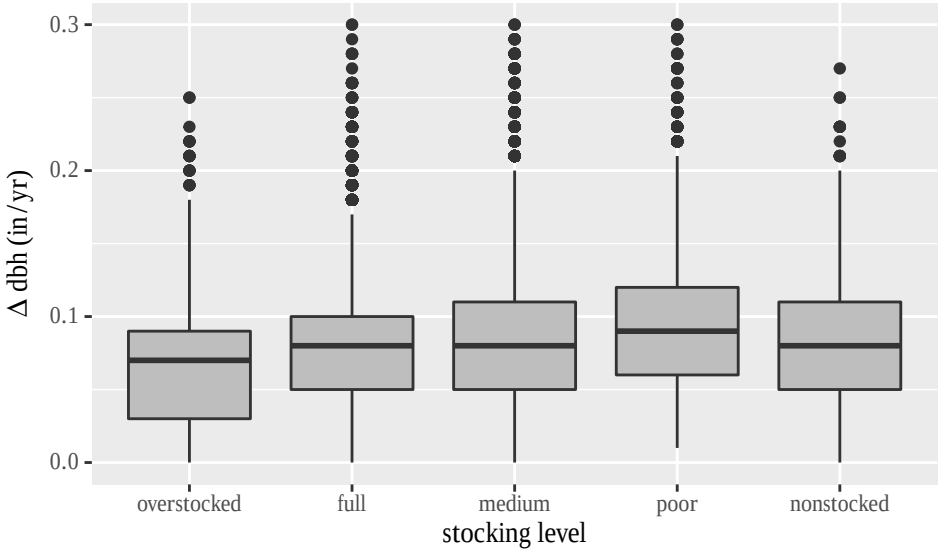

Figure 3: Plot level forest stocking and diameter growth of individual trees.

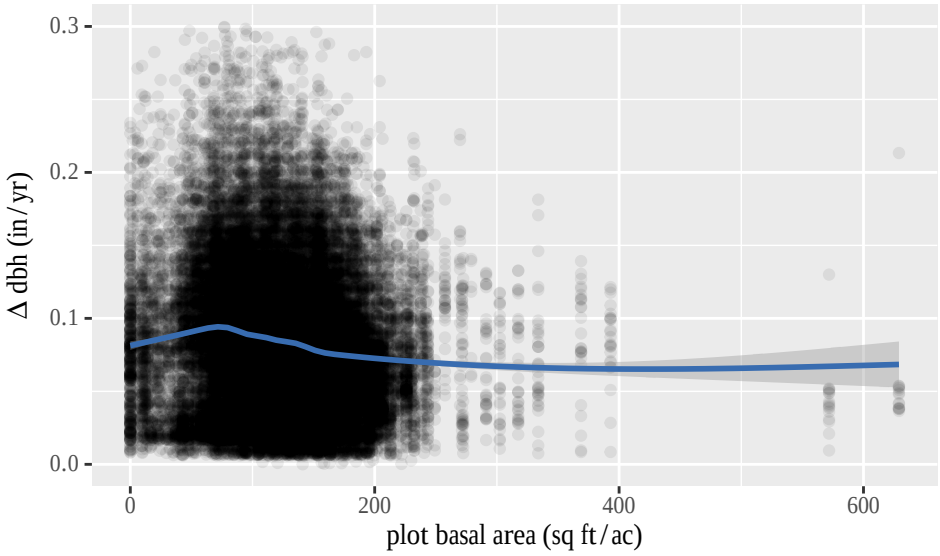

Figure 4: Plot basal area and diameter growth of individual trees. Observations are displayed with random vertical offset and partial transparency so that their relative concentration can be visualized. Darker areas show a greater concentration of observations. Trend line in blue calculated using generalized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

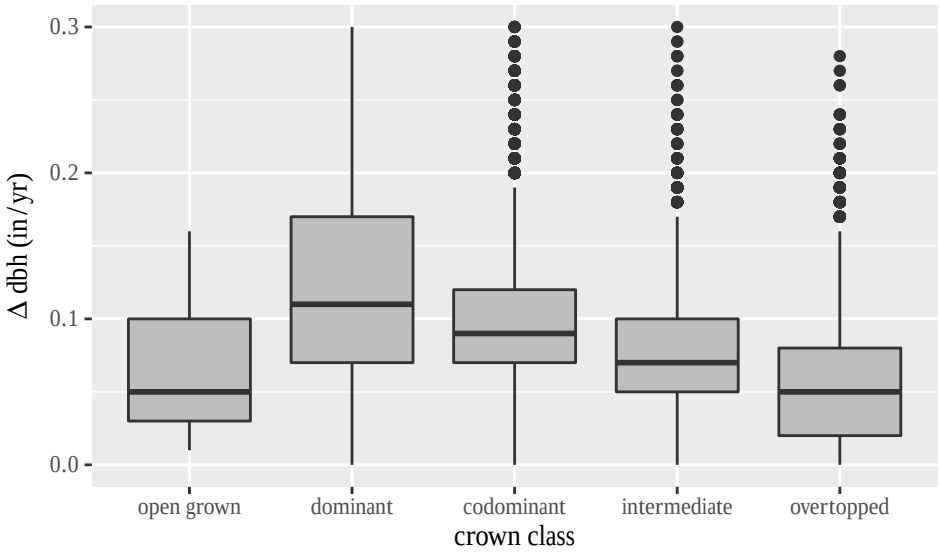

Figure 5: Crown class and diameter growth of individual trees.

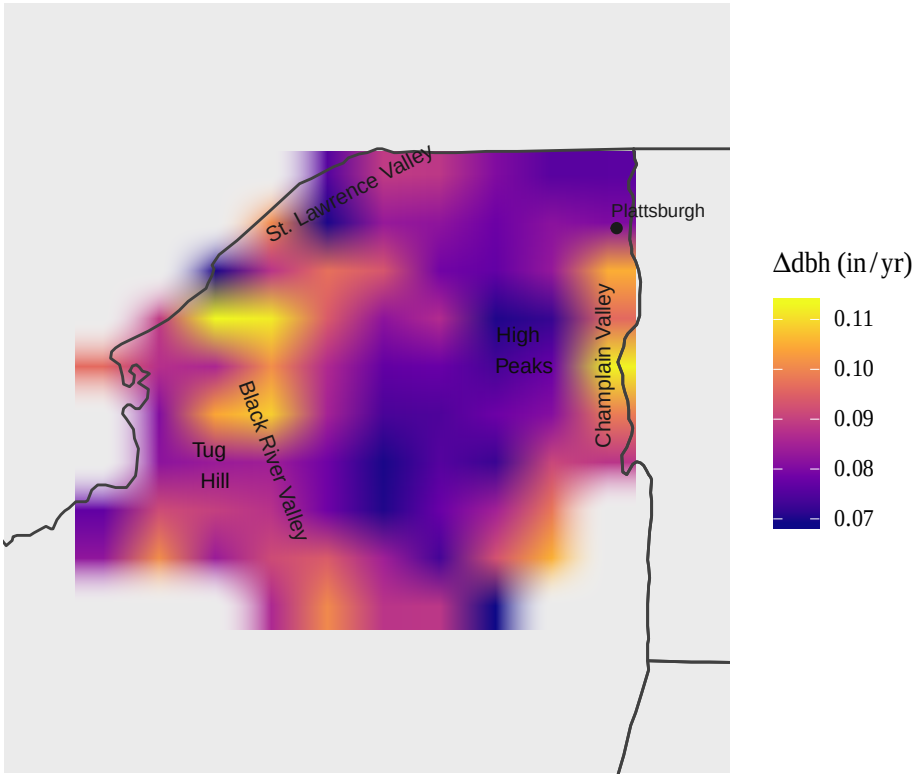

Figure 6: Geographic variation in diameter growth, depicted using two dimensional bin smoothing and interpolation.

Another notable interactive effect in the data is seen in the geographic variation of growth rates. On their own, latitude and longitude explain only a very small amount of the variation in diameter growth, with trees in the south growing slightly faster than those in the north, and trees in the east and west growing somewhat faster than mid-longitudinal trees. When latitude and longitude are assessed in tandem, however, a much stronger relationship becomes apparent. The lower Black River valley in the west and the Champlain valley near Vermont clearly have the highest growth rates, while the central Adirondack plateau has the lowest (figure 6). The Saint Lawrence Valley in the north also shows relatively fast growth rates, but growth is much slower in the northeast part of the region (in Clinton County north of Plattsburgh).

Model Formulation

Two different approaches were combined to construct the diameter growth model. A naive Bayes approach was used to modify the mean diameter growth (μ) by the average growth residuals for individual categories in the categorical predictors. This sub-model included terms for the simple effects of species grouping (b_s), forest type (b_f), and landscape position (b_l). Regularization was used on the three simple effects to downgrade the strength of any effects trained on small sample sizes. Optimal regularization parameters λ_s, λ_f, and λ_l were determined for each effect, respectively, using k-fold cross validation to prevent overfitting. Simple effect terms took the form

$$b_j = \frac{\sum_{i=0}^{n_j}(\Delta dbh_{ji} - \mu)}{n_j + \lambda}$$

where b_j is the effect for category j, i is the individual growth observation, n is the sample size, Δdbh is the diameter growth rate, μ is the mean growth rate across all the data, and λ is the regularization parameter.

The optimal regularization parameters were determined to be 2, 6, and 2, respectively.

To account for interactions between quantitative variables, a random forest algorithm was fitted to the residual growth rates after applying all the simple terms. Discrete ordinal factors (site class, stocking, and crown class) were treated as numeric predictors. Other predictors included slope, aspect, basal area, latitude, longitude, crown ratio, and dbh. Accuracy was greatly reduced when categorical variables (species, forest type, and landscape position) were included in the random forest, so they were excluded. The forest was made up of 200 regression trees, and the minimum node size was determined using k-fold cross validation on the training data.

The final model can be expressed as

$$\Delta dbh = \mu + b_s + b_f + b_l + rf + \epsilon$$

where rf is the random forest effect, ϵ is an error term, and other variables have been defined previously.

The model was parameterized using the training data alone, and the overall accuracy was estimated using the test data that had been set aside previously. Finally, the model was refit to all the available data (training and test sets combined), using the pre-determined parameters.

Results

The overall mean diameter growth rate μ is 0.0834 inches per year, and simply predicting the mean for every instance yields a root mean square error (RMSE) of 0.0454 inches per year. Accounting for the simple effects of categorical variables yields moderate decreases in RMSE. Accounting for species group reduces the RMSE to 0.0394 inches per year; accounting for species and forest type reduces it to 0.0393 inches per year; and accounting for species, forest type, and landscape position reduces it to 0.0392 inches per year.

The random forest brings significant gains to the model's accuracy, lowering the RMSE (evaluated with the training data alone) to 0.0137 inches per year. Results from the training show that, of the numeric predictors, dbh is by far the most important, followed by plot basal area and crown ratio. Latitude and longitude together are similar in importance to either one of the competition indicies above, while slope, aspect, stocking, site class, and crown class are less important. Variable importance values for these predictors are reported in table 2.

Table 2: Variable importance of numeric predictors in random forest model.

predictor	importance
dbh	100.00
basal area	17.32
crown ratio	17.21
longitude	7.45
latitude	6.20
slope	4.08
aspect	2.87
stocking	1.51
site class	0.54
crown class	0.00

When tested against the independent testing data, the overall model returns a RMSE of 0.0194 inches per year.

Errors are narrowly distributed around 0 (figure 7), showing that the model is unbiased for the Adirondack region as a whole. Predictions are also unbiased for many species groups, although several do show limited

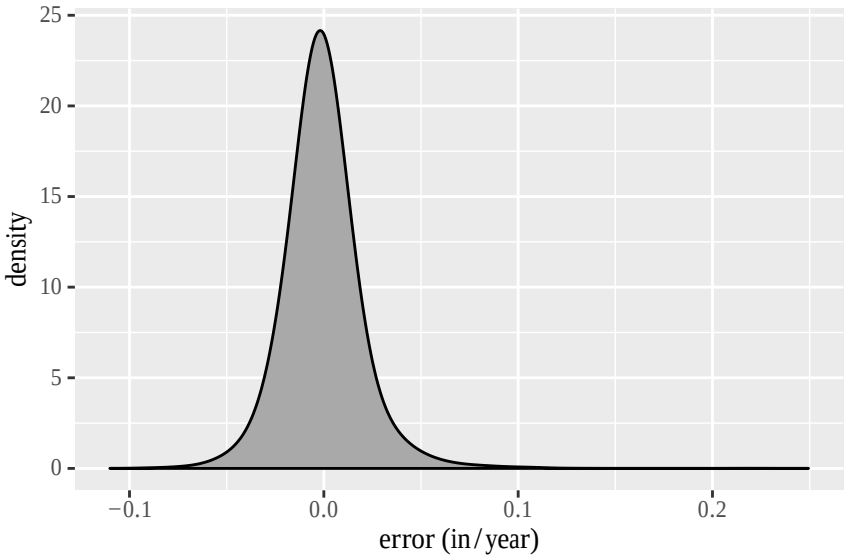

Figure 7: Kernel density estimate of error distribution of final model, fit to test data. Negative errors are overpredictions.

bias (figure 8). The most notable are white and red oak, whose growth rates are overpredicted by approximately 0.016 and 0.014 inches per year, respectively. Predictions are much less biased for more common species. Absolute mean error rates for the ten most common species groups in the region (soft maple, beech, hard maple, fir, spruce, yellow birch, hemlock, ash, white pine, and black cherry) are all less than 0.005 inches per year, with the exception of hard maple, whose growth rates were overpredicted by an average of 0.008 inches per year.

Conclusions

This model does appear to have increased the accuracy of diameter growth predictions in the Adirondacks over previous models. Its normalized RMSE ($RMSE/\mu$) of 0.233 compares favorably with the species-specific normalized RMSEs from A. Weiskittel et al. (2016), which range from 0.48 (hemlock) to 0.73 (other hardwoods) and average 0.61. The reduction is probably partly to do with the larger, more representative dataset used here, but there do appear to be significant gains made by modeling the interactions between numeric predictors, which were neglected in previous studies. These interactions are clearly important in the Adirondack's structurally diverse, mixed species forests, and help to account for geographic variation in growth trends.

The model also improves on species-specific bias, with absolute mean biases below 0.005 inches per year for all but one of the ten of the most common species. 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, which was not used in this study, but could have been incorporated. 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. Crown ratio was an important competition index in this model and for Westfall (2006), yet Kiernan, Bevilacqua, and Nyland (2008) and A. Weiskittel et al. (2016) rejected the measure as superfluous. Perhaps as models grow more sophisticated at assessing the interactions between variables we will gain a more nuanced understanding of how competition indices overlap and where they stand on their own.

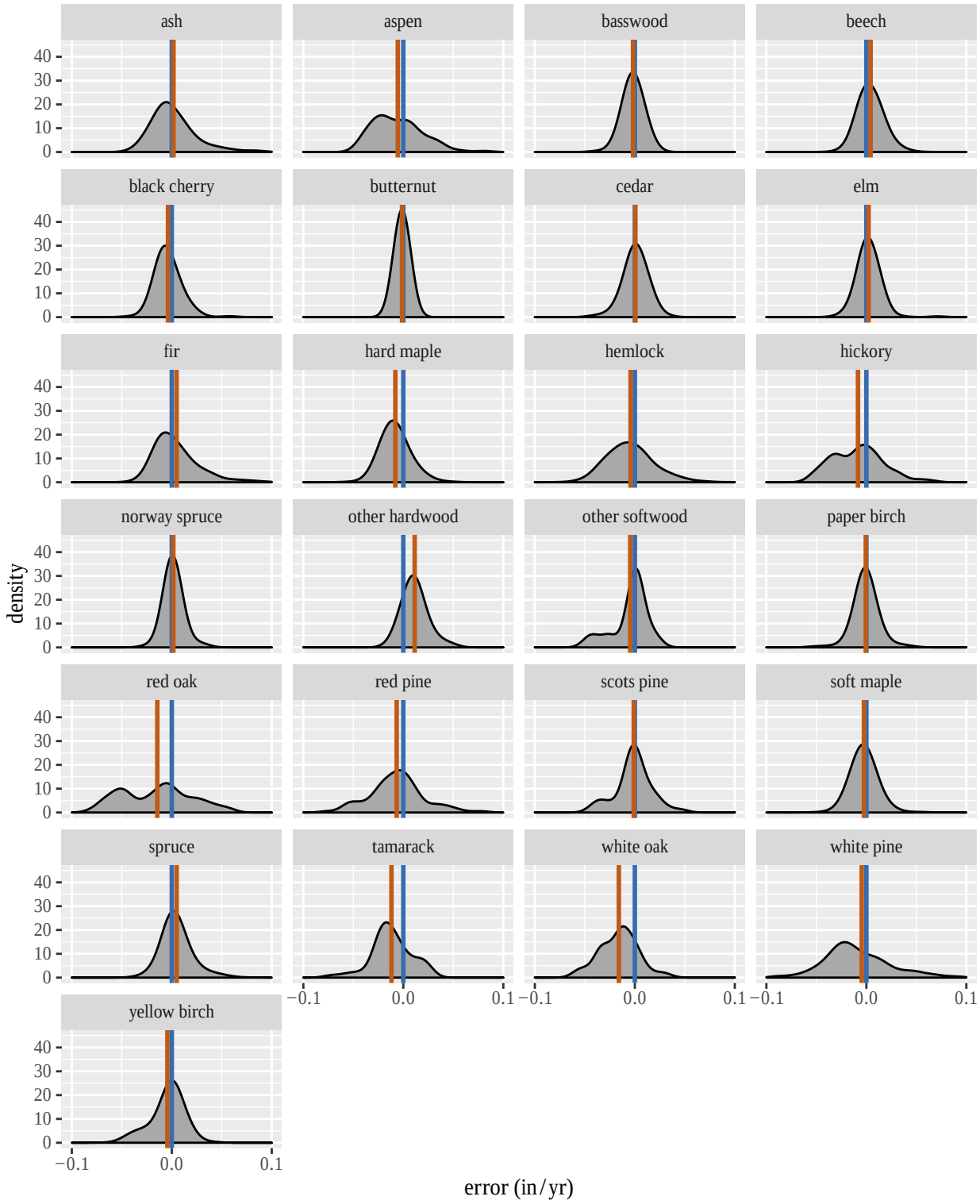

Figure 8: Kernal density estimates of error distributions for individual species groups. Vertical blue lines show 0 (no error) and vertical brown lines show species groups' average errors. Negative errors are overpredictions. No cottonwoods were present in the testing data, so they have been omitted.

In this model, it seemed that one tree-specific index (crown ratio) and one plot-level index (basal area) were enough to account for the bulk of the competition-related variability, and that the remaining indices (crown class and stocking) were mostly redundant. This makes intuitive sense, as the tree-level and plot-level metrics describe fundamentally different attributes. Plot-level metrics like basal area and stocking 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).

Basal area, latitude, and longitude were the only plot-level factors that contributed meaningfully to growth predictions. The effects of landscape position, forest type, aspect, slope, and site class were minimal. In aggregate, those factors bring a marginal improvement to the model's accuracy, but they could be removed from the analysis to save on inventory costs without a significant impact.

Potential improvements to the model include the incorporation of other predictors like overtopping basal area, which could be derived from the available data, and an incorporation of interactions between categorical and numeric predictors (especially between species and dbh). Different algorithms or different formulations of the random forest algorithm may do a better job incorporating categorical and numeric variables, which could improve the model's accuracy and streamline the prediction process. The model used here does benefit from fast prediction, though, 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 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 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. This model is well suited to predicting the growth of Adirondack trees over relatively short time scales, but it is inappropriate for modeling the effects of climate change on growth, or for modeling the outcomes of new management systems.

References

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.

Chamberlain, Scott. 2018. "Laselva: FIA Data." https://github.com/ropenscilabs/laselva.

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

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

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.

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 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.