

An FIA dataset for modeling tree-level changes in the Northern Forest

Neal Maker

October 13, 2021

Introduction

The intent of this project is to put together a dataset of remeasured trees from the US Forest Service's Forest Inventory and Analysis (FIA) records for the Northern Forest region, which will allow tree-level changes to be modeled accurately. Specifically, the dataset should support the development of unbiased models of dbh increment, height, height increment, crown ratio increment, and survival; and the determination as to which variables are the most important predictors of those outcomes. Potential predictors to be kept in the dataset are those that (1) are widely available in the FIA data in the region (to maintain large sample sizes), and (2) can be recorded in forest inventories or remotely without large increases in inventory costs.

The Northern Forest region was chosen because it covers a fairly large geographic extent while still representing a coherent ecological region, in which trees can be expected to follow a similar set of behaviors. Models developed with the dataset should be relatively unbiased for individual forests within the region, but will still allow for streamlined analyses across disparate ownerships. The US Northern Forest is defined here as including Oswego, Oneida, Lewis, Jefferson, Saint Lawrence, Herkimer, Fulton, Hamilton, Franklin, Essex, Clinton, and Warren Counties in New York; Franklin, Orleans, Essex, Chittenden, Lamoille, Caledonia, Washington, Addison, Orange, and Grand Isle Counties in Vermont; Coos, Grafton, and Carroll Counties in New Hampshire; and Oxford, Franklin, Somerset, Androscoggin, Kennebec, Waldo, Hancock, Washington, Penobscot, Piscataquis, and Aroostook Counties in Maine.

Methods

FIA data were downloaded from the FIA DataMart¹ in the form of state-specific csv files, which were generated by the Forest Service from the FIA Oracle database tables. These data are current as of October 13, 2021. In the future, the dataset can be re-built using updated csv files to incorporate new remeasurement data.

Records for individual trees were joined to plot and condition data to add site and stand attributes, and data from remeasured plots were joined to data from previous inventories to add starting and ending measurements. Records for trees without remeasurement data were discarded, along with records for trees that were already dead at their starting measurement, trees that were incorrectly inventoried during starting or ending inventories, and seedlings with diameters measured at the root collar instead of at breast height. FIA plot designs varied in the past in different years and locations, and we only used inventories that employed the current, standard plot design. This design started being used in the mid 1990s, and allows easy comparison between inventories from different times and places.

Some of the variables retained from the FIA tables were recorded in the field, while others were determined remotely by the FIA Program. A number of variables were also calculated after the fact in this analysis.

¹<https://apps.fs.usda.gov/fia/datamart/datamart.html>

These include plot basal area and tree overtopping basal area, which were calculated by grouping trees into their respective plots and subplots; species-specific plot basal area and overtopping basal area; species richness within subplots; coefficients of variation of dbh and height as measures of subplot structural diversity; and diameter, height, and crown ratio growth rates, which were calculated using remeasurement data. Diameter and height growth rates are also reported in the FIA database, but the FIA Program estimates diameter rates using a model instead of calculating them from the remeasurement data, making them unsuitable for training new models.

In addition to the previously well defined variables mentioned above, a new competition index was created in hopes that it would better account for competitive effects at the neighborhood scale. It is based very loosely on the sail light index used by Charles Canham and others in the Sortie-ND simulation software (see www.sortie-nd.org/help/manuals/help/data/light_behaviors/sail_light.html) and is called the porous sail index, or PSI. Unlike the sail light index, which maps the locations of individual tree crowns to determine their effect on incoming solar radiation, the PSI quantifies the overall light-blocking ability of individual trees and uses that as a proxy for competitive effect, regardless of trees' specific locations within the plot.

To calculate PSI, a porous sail value is first assigned to every live tree in the plot. The porous sail value is proportional to the crown's total light blocking area (B), calculated as the sectional area of a tree's crown (A), based on its height (h) and width (w), multiplied by the crown density (d), which is the proportion of light blocked through the crown's section.

$$B = Ad = \pi d \frac{h}{2} \frac{w}{2}$$

Since crown height can be roughly derived from tree height (H) and crown ratio (r , expressed as a proportion) as $h = Hr$, the relationship can be rewritten as

$$B = \frac{\pi d H r w}{4}$$

Crown density and crown width are unknown, so two crude assumptions are made. The first is that the crown density is directly proportional to crown ratio, such that $d = c_1 r$, where c_1 is a constant. This relationship came from an examination of FIA data, and is roughly accurate. The second assumption is that the ratio of crown width to crown height is fixed across all trees, such that $w = c_2 h = c_2 H r$, where c_2 is a constant. This is a more problematic assumption, as different species have inherently different crown shapes, which will cause the PSI to systematically overestimate the influence of some species and underestimate the influence of others. Still, it is the best we can do without knowing crown widths, and it is no more problematic than overtopping basal area, which suffers from the same faulty assumption and is widely used. Incorporating these assumptions,

$$B = \frac{\pi c_1 H r^2 w}{4} = \frac{\pi c_1 c_2 H^2 r^3}{4}$$

The porous sail value (P) only needs to be proportional to the light blocking area, so

$$P = H^2 r^3 \propto B$$

The porous sail index for a given tree is the sum of neighboring trees' porous sail values, weighted by their heights relative to the target tree's height.

$$PSI_t = \sum_{i=1}^{n-1} \left(P_i \frac{H_i}{H_t} \right)$$

where PSI is the porous sail index, t denotes the target tree, n is the number of trees in the plot, and i denotes one of the $n - 1$ neighboring trees.

Starting and ending values were retained for variables that naturally change from one measurement to another, and midpoint values were calculated for some variables by averaging the starting and ending values. Midpoint values were recorded to better reflect average conditions during the remeasurement period.

Some of the ostensibly fixed variables like slope, aspect, and site class were found to change from one measurement to another in a minority of instances. For example, aspect was recorded differently in nine percent of remeasurements, slope was recorded differently in 17 percent of remeasurements, and site class was recorded differently in four percent of remeasurements, despite the fact that they were measured on the same plots and should have remained constant. The differences between starting and ending values are generally small, however, and are probably measurement errors. In the case of slope, the mean absolute difference of deviating measurements is only four percent. Among erroneous site class measures, the average is only one site class. Aspect errors tend to be higher, averaging 89 degrees, but they can be attributed to the difficulty of determining aspects in relatively flat terrain. If only plots with slopes over 20 percent are considered, the mean absolute aspect error falls to 33 degrees. All these discontinuities were assumed to be random measurement errors, and starting values were arbitrarily retained in the dataset while ending values were discarded.

Variables in the dataset were renamed and FIA codes for the levels of categorical variables were replaced with descriptive strings, to make them more intuitive and user-friendly. Tree species were also grouped into species groups and FIA forest types into more general forest types; so they match common inventory protocols and to facilitate the incorporation of uncommon species and forest types into growth models. For example, most species in the genus *Populus* are combined into a single “aspen” group, although cottonwoods (*Populus deltoides*) are kept in their own group because they exhibit very different growth characteristics. Similarly, the FIA forest types “balsam fir”, “white spruce”, “red spruce”, “red spruce/balsam fir”, and “black spruce” were combined into a single “Spruce-fir” group, but “northern white-cedar” was kept in its own “Cedar” group.

Missing Values

Examination of the data showed that some 23.6% of the values were missing. The missing values were associated with only a fraction of the variables, depicted in the following table.

Variable	Missing	% Missing
psi1_mid	448823	96.77
psi1_s	423494	91.31
psi_ash	423494	91.31
psi_aspen	423494	91.31
psi_basswood	423494	91.31
psi_beech	423494	91.31
psi_black.cherry	423494	91.31
psi_butternut	423494	91.31
psi_cedar	423494	91.31
psi_cottonwood	423494	91.31
psi_elm	423494	91.31
psi_fir	423494	91.31
psi_hard.maple	423494	91.31
psi_hemlock	423494	91.31
psi_hickory	423494	91.31
psi_hophornbeam	423494	91.31
psi_norway.spruce	423494	91.31
psi_other.hardwood	423494	91.31
psi_other.softwood	423494	91.31
psi_paper.birch	423494	91.31
psi_red.oak	423494	91.31
psi_red.pine	423494	91.31
psi_scots.pine	423494	91.31
psi_soft.maple	423494	91.31

Variable	Missing	% Missing
psi_spruce	423494	91.31
psi_striped.maple	423494	91.31
psi_tamarack	423494	91.31
psi_white.oak	423494	91.31
psi_white.pine	423494	91.31
psi_yellow.birch	423494	91.31
psi1_e	386639	83.36
ht_var_mid	381971	82.35
ht_var_s	364382	78.56
ht_var_e	243741	52.55
ht_mid	180768	38.97
ht_rate	180768	38.97
HT_MIDPT	144833	31.23
ht_s	140923	30.38
HT_BEGIN	140923	30.38
ht_e	135201	29.15
HT_END	135201	29.15
ball_mid	64442	13.89
ball_e	64438	13.89
cr_mid	64430	13.89
cr_e	64430	13.89
cr_rate	64430	13.89
crown_class_e	64430	13.89
ba1_mid	64232	13.85
ba1_e	64232	13.85
tree_class_e	48689	10.50
dbh_e	45604	9.83
dbh_rate	45604	9.83
DIA_END	45604	9.83
dbh_var_mid	7987	1.72
dbh_var_e	7527	1.62
forest_type_e	869	0.19
stocking_e	869	0.19
dbh_var_s	855	0.18
dbh_s	23	0.00
ball_s	11	0.00
ba_cedar	4	0.00
bal_cedar	4	0.00
ba_beech	3	0.00
ba_fir	3	0.00
ba_soft.maple	3	0.00
bal_beech	3	0.00
bal_fir	3	0.00
bal_soft.maple	3	0.00
ba_hard.maple	1	0.00
bal_hard.maple	1	0.00

Most of these missing values are associated with porous sail indexes and heights. The FIA program only records heights on a random subset of observations, and the porous sail index can only be calculated on plots in which all live trees have height measurements. So while these missing values limit the amount of data available for training growth models, they are random omissions and should not introduce bias.

Many of the remaining missing values are mid-interval or end-of-interval values that are missing because the trees died. Across all the data, 64232 trees died during the remeasurement interval, and these correspond perfectly with the mid-interval and end-of-interval *bal* missing values. Mid-interval and end-of-interval crown ratio values are also missing for all trees that died, though they are also missing for a small number of trees that lived. End-of-interval *dbh* values are missing for many of the trees that died, but are present for some of them because they were estimated from stump diameters by the FIA program.

A small number of observations are missing values for other variables too, like ending forest types and stocking, and these appear to be random recording errors that can be safely removed without introducing bias.

Organization

The final dataset contains 462,919 unique tree records, which were tallied across 11,339 plots evenly distributed throughout the region. Tallied trees belong to 28 different species groups and were located in 17 different forest types in 14 different physiographic (landscape) positions. Remeasurement periods ranged from 2.91 to 7.82 years and averaged 5.14 years. Eighty-six percent of tallied trees lived through the remeasurement period and the remaining fourteen percent died.

A description of each variable in the final dataset and its source is provided below. Fields from the FIA database are referenced by their Oracle table and field names, in the form *TABLE\$FIELD*. Some of the variables account for more than one column in the dataset. Variable names amended with *_s* are measurements taken at the start of the remeasurement period; those amended with *_e* are measurements taken at the end of the remeasurement period; those amended with *_mid* are estimates of mid-period values, calculated by averaging the starting and ending measurements; and those amended with *_rate* are annual rates of change, averaged over the remeasurement period. Positive rates are increasing values, and negative rates are decreasing values.

spp

Species or species group. Adapted from *TREE\$SPCD*.

dbh

Diameter at breast height (4.5' above ground), measured in inches. From *TREE\$DIA*. Note that *dbh_rate* is calculated as $(dbh_e - dbh_s) / interval$ and is the preferred variable for model formulation. *dbh_rate_fia* is from *TREE_GRM_COMPONENT\$ANN_DIA_GROWTH* and is estimated using an existing diameter growth model. It is included for reference only and should not be used to train new models.

cr

Compacted crown ratio (percent of tree height supporting live crown). From *TREE\$CR*.

crown_class

Tree canopy position. From *TREE\$CCLCD*:

- 1 Open grown (crown has received full light for most or all of its life)
- 2 Dominant (crown extends above main canopy and receives full light from above and partly from sides)
- 3 Codominant (crown in main canopy and receives full light from above, but little from sides)
- 4 Intermediate (crown extends into main canopy, but receives little direct light)
- 5 Overtopped (crown entirely below main canopy level, receiving no direct light)

tree_class

General quality of a live tree. From *TREE\$TREECLCD*:

- 2 Growing-stock (of commercial species and meeting minimum merchantability standards)
- 3 Rough-cull (sound wood, but does not meet minimum merchantability standards)
- 4 Rotten-cull (does not meet minimum merchantability standards and more than half of cull is rotten)

spp_rich

Subplot species richness, calculated as the total number of unique species occurring on each subplot.

dbh_var

Coefficient of variation of trees' diameters at breast height within each subplot, used as a measure of structural diversity. Calculated by dividing the standard deviation of diameters by their mean.

ht_var

Coefficient of variation of trees' heights within each subplot, used as a measure of structural diversity. Calculated by dividing the standard deviation of heights by their mean.

ba

Plot basal area, measured in square feet per acre of all live trees, 1" dbh or greater. Calculated by computing individual trees' per acre basal areas ($ba * tpa$), then summing those basal areas within subplots. Variables whose names end with a species name (such as *ba_ash*) are plot basal areas of only the species indicated.

bal

Overtopping basal area, measured in square feet per acre. Calculated by computing individual trees' per acre basal areas ($ba * tpa$), then for each tree summing the per acre basal areas of other trees in in the same subplot with larger diameters. Variables whose names end with a species name (such as *bal_ash*) are overtopping basal areas that only account for overtopping trees of the species indicated.

psi

Porous sail index, whose computation is described in the 'Methods' section above. Variables whose names end with a species name (such as *psi_ash*) are porous sail indexes that only account for trees of the species indicated.

ht

Total tree height, measured in feet. From *TREE\$HT*. For trees with broken tops, heights are estimated by FIA program.

forest_type

Forest type defined by the species dominating stocking. Adapted from *COND\$FORTYPCD*. Note that FIA does not recognize a "mixedwood" forest type, so plots with greater than half of their basal area in softwood species are generally considered softwood types, and those with greater than half of their stocking in hardwoods are considered hardwood types. The exceptions are the "Pine-hardwood" and "Cedar-hardwood" types. The forest types used here do not always coincide well with available stocking charts. Types in the Northern Forest region include:

Northern hardwood
Transition hardwood
Oak-hickory
Cottonwood
Pine-hardwood
Cedar-hardwood
Spruce-fir
Cedar
Hemlock
Larch (includes tamarack)
Norway spruce
White pine
Red pine
Scots pine
Mixed softwood
Other
Nonstocked

stocking

Plot-level stocking of all live trees 1" dbh and larger. From *COND\$ALSTKCD*:

1 Overstocked

- 2 Fully stocked
- 3 Medium stocked
- 4 Poorly stocked
- 5 Nonstocked

landscape

Physiography. From *COND\$PHYSCLCD*. Depends on land form, topographic position, and soil type.

Classes include:

dry tops
dry slopes
deep sands
other xeric
flatwoods
rolling uplands
moist slopes & coves
narrow floodplains/bottomlands
broad floodplains/bottomlands
other mesic
swamps/bogs
small drains
beaver ponds
other hydric

site_class

Site productivity class. From *COND\$SITECLCD*. Defined by potential wood growth in cubic feet per acre per year:

- 1 225+ ft³ac⁻¹yr⁻¹
- 2 165-224 ft³ac⁻¹yr⁻¹
- 3 120-164 ft³ac⁻¹yr⁻¹
- 4 85-119 ft³ac⁻¹yr⁻¹ (equivalent to class I in VT)
- 5 50-84 ft³ac⁻¹yr⁻¹ (equivalent to class II in VT)
- 6 20-49 ft³ac⁻¹yr⁻¹ (equivalent to class III in VT)
- 7 0-19 ft³ac⁻¹yr⁻¹ (equivalent to class IV in VT)

slope

Slope in percent. From *COND\$SLOPE*.

aspect

Aspect in degrees. From *COND\$ASPECT*.

lat

Plot latitude in decimal degrees (random offset is applied to protect landowners' privacy). From *PLOT\$LAT*.

lon

Plot longitude in decimal degrees (random offset is applied to protect landowners' privacy). From *PLOT\$LON*.

elev

Plot elevation in feet above mean sea level. From *PLOT\$ELEV*.

date

Inventory date. Calculated from *PLOT\$MEASYEAR*, *PLOT\$MEASMON*, and *PLOT\$MEASDAY*.

interval

Length of remeasurement period in years. Calculated as *date_e* – *date_s*.

status_change

Change in tree status during remeasurement period. Based on *TREE\$STATUSCD*. One of:

lived

died (natural mortality)

cut

plot

A unique identifier for the plot the tree was recorded on. Corresponds to *PLOT\$CN* attribute for the ending inventory.