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

Neal Maker

April 13, 2020

## 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 DataMart[1] 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 April 13, 2020. 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. We also calculated a number of variables ourselves after the fact. These

[1] https://apps.fs.usda.gov/fia/datamart/datamart.html

include plot basal area and tree overtopping basal area, which we calculated by grouping trees into their respective plots and subplots; and diameter, height, and crown ratio growth rates, which we 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.

We retained starting and ending values for variables that naturally change from one measurement to another, and calculated midpoint values 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 87 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 34 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.

We renamed the variables in the dataset and replaced FIA codes for the levels of categorical variables with descriptive strings, to make them more intuitive and user-friendly. We also grouped tree species into species groups and FIA forest types into more general forest types, so they match common inventory protocols. For example, most species in the genus *Populus* are combined into a single "aspen" group to make data collection easier, 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.

# Organization

The final dataset contains 421,011 unique tree records, which were tallied across 10,364 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 7.67 years and averaged 5.1 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)

**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.

**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.

**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+ $ft^3 ac^{-1} yr^{-1}$
*2* 165-224 $ft^3 ac^{-1} yr^{-1}$
*3* 120-164 $ft^3 ac^{-1} yr^{-1}$
*4* 85-119 $ft^3 ac^{-1} yr^{-1}$ (equivalent to class I in VT)
*5* 50-84 $ft^3 ac^{-1} yr^{-1}$ (equivalent to class II in VT)
*6* 20-49 $ft^3 ac^{-1} yr^{-1}$ (equivalent to class III in VT)
*7* 0-19 $ft^3 ac^{-1} yr^{-1}$ (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.