Journal of Forestry, 2022, 607–632
https://doi.org/10.1093/jofore/fvac022
Research Article - forest ecology
Received October 12, 2021; Accepted July 22, 2022
Advance Access publication September 15, 2022

OXFORD

Research Article - forest ecology

TreeMap 2016 Dataset Generates CONUS-Wide Maps of Forest Characteristics Including Live Basal Area, Aboveground Carbon, and Number of Trees per Acre

Karin L. Riley,[1,*] Isaac C. Grenfell,[2] John D. Shaw,[3] and Mark A. Finney[4]

[1]Fire Sciences Laboratory, Rocky Mountain Research Station, USDA Forest Service, 5775 W. Hwy. 10, Missoula, MT 59808, USA. (karin.l.riley@usda.gov). [2]Fire Sciences Laboratory, Rocky Mountain Research Station, USDA Forest Service, 5775 W. Hwy. 10, Missoula, MT 59808, USA. (isaac.c.grenfell@usda.gov). [3]Forest Inventory and Analysis, USDA Forest Service, 507 25th St., Ogden, UT 84401, USA. (john.d.shaw@usda.gov). [4]Fire Sciences Laboratory, Rocky Mountain Research Station, USDA Forest Service, 5775 W. Hwy. 10, Missoula, MT 59808, USA. (mark.a.finney@usda.gov).

*Corresponding author email: karin.l.riley@usda.gov

Abstract

The TreeMap 2016 dataset provides detailed spatial information on forest characteristics including number of live and dead trees, biomass, and carbon across the entire forested extent of the continental United States at 30 × 30m resolution, enabling analyses at finer scales where forest inventory is inadequate. We used a random forests machine learning algorithm to assign the most similar Forest Inventory Analysis (FIA) plot to each pixel of gridded LANDFIRE input data. The TreeMap 2016 methodology includes disturbance as a response variable, resulting in increased accuracy in mapping disturbed areas. Within-class accuracy was over 90% for forest cover, height, vegetation group, and disturbance code when compared to LANDFIRE maps. At least one pixel within the radius of validation plots matched the class of predicted values in 57.5% of cases for forest cover, 80.0% for height, 80.0% for tree species with highest basal area, and 87.4% for disturbance. A new feature of the dataset is that it includes linkages to select FIA data in an attribute table included with the TreeMap raster, allowing users to map summaries of 21 variables in a GIS. TreeMap estimates compared favorably with those from FIA at the state level for number of live and dead trees and carbon stored in live and dead trees.

Study Implications: TreeMap 2016 provides a 30 × 30 m resolution gridded map of the forests of the continental United States. Attributes of each grid cell include a suite of forest characteristics including biomass, carbon, forest type, and number of live and dead trees. Users can readily produce maps and summaries of these characteristics in a GIS. The TreeMap also includes a database containing, for each pixel, a list of trees with the species, diameter, and height of each tree. TreeMap is being used in the private sector for carbon estimation and by land managers in the National Forest system to investigate questions pertaining to fuel treatments and forest productivity as well as Forest Plan revisions.

Keywords: forest mapping, carbon, biomass, basal area, trees per acre, machine learning, random forests, Forest Inventory Analysis, LANDFIRE

© The Author(s) 2022. Published by Oxford University Press on behalf of the Society of American Foresters. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Detailed measurements from a network of forest plots are often used for estimating forest characteristics, such as the number of live and dead trees by size and species, biomass, and carbon at national, regional, and state levels. However, when plot networks are sparse relative to the area of interest, they are inadequate to provide accurate estimates, especially in spatially variable landscapes, or to accurately represent heterogeneity at finer spatial scales. For example, the Uinta Mountains and vicinity in northern Utah and southern Wyoming includes nearly 8.8 million acres and all or part of three National Forests and is sampled by 1,464 forested plots measured by the USDA Forest Service's Forest Inventory and Analysis (FIA) program. The lodgepole pine forest type, which is rare in this area yet important for a variety of management objectives, is represented by only sixty-seven plots, and only twenty in the area of most active management. This limits capability for planning and management based on FIA plot data alone. Conversely, existing gridded maps of vegetation characteristics (e.g., LANDFIRE) provide modeled information for all locations on the landscape but lack the level of detail needed for estimating many forest characteristics. To model detailed measurements at all forested locations on the landscape, we employ machine learning: imputation is a statistical technique which assigns detailed measurements at plot networks to gridded landscape data, offering continuous coverage of fine-scale forest characteristics.

In this publication, we present a new edition of a tree-level model of the forests of the conterminous United States, the TreeMap 2016. The TreeMap project uses machine learning to assign or impute forest plot data measured by the FIA program to landscape maps of vegetation, disturbance, and biophysical parameters from the LANDFIRE project (Riley, Grenfell, and Finney 2016; Riley Grenfell, Finney, and Wiener 2021). As a result, detailed data based on attributes measured on FIA plots are now available at the 30 × 30m pixel resolution.

The FIA program collects a comprehensive inventory of the forests of the United States, with plots measured at a density of approximately one per 6,000 acres (Bechtold and Patterson 2005). FIA analyzes the plot data to report on the status of US forests and also trends in forested area, its location, ownership, tree growth, mortality, and harvest (Forest Inventory Analysis 2022b). Although the error bars around FIA estimates are narrow at scales such as state or national, they can become quite wide at smaller scales due to the sparseness of plots noted above, leading to a set of statistical procedures known as small area estimation in an effort to makes inferences (Moisen, Blackard, and Finco 2004). Conversely, gridded landscape data like that from the LANDFIRE project provides complete spatial coverage for landscape attributes such as percent forest cover but does not provide detailed observations needed for many applications, such as lists of trees. LANDFIRE is a multiagency project that produces nationally consistent gridded maps of wildland vegetation and fuels (Rollins 2009). LANDFIRE's maps are the basis for many fire behavior modeling efforts and forest planning efforts undertaken by the USDA Forest Service (Forest Service). The TreeMap project integrates data from LANDFIRE and FIA, two consistent and well-maintained sources of landscape and forest data, using FIA data to add further detail on forest characteristics to LANDFIRE data. The TreeMap process leverages the strengths of these two existing datasets to produce a third dataset without some of the limitations of the other two. By combining attributes from these two datasets with a machine learning approach, we produced a tree-level gridded model of the forests of the conterminous United States at 30 × 30 m resolution (Riley et al. 2019). The TreeMap can thus provide detailed information for the areas between plots as well as adding detail to existing vegetation maps, with the result that estimation of forest characteristics (e.g., number of trees, live biomass, live and dead carbon) can be provided at a variety of scales, ranging from a 10,000-acre private timberland, for example, to the county, state, or regional level.

The original TreeMap dataset covered the western United States circa 2008 and is available online (Riley et al. 2018) along with an updated version for the conterminous United States (CONUS) circa 2014 (Riley et al. 2019). The primary differences between these two editions were the year they represent (2008 versus 2014) and the geographic scope (western United States versus CONUS). The new version described here offers several improvements in addition to updates for forest changes and disturbances circa 2016. This CONUS-wide version is also currently available online (Riley, Grenfell, Finney, and Shaw 2021). The TreeMap 2016 update was primarily completed to provide newer maps as new LANDFIRE editions become available. In addition, we responded to the needs of users who apply the TreeMap to compute the height and density of dead trees to prevent hazard to fire responders, improving the accuracy with which we map disturbed areas (Riley et al. 2022, Risk Management Assistance 2021).

TreeMap is distinct from other imputed forest vegetation products in that it provides an FIA plot identifier to each pixel whereas other datasets provide forest characteristics such as live basal area (e.g., Ohmann and Gregory 2002; Pierce Jr et al. 2009; Wilson, Lister, and Riemann 2012). The FIA plot identifier can be linked to the hundreds of variables and attributes recorded for each tree and plot in the FIA DataMart, FIA's public repository of plot information (Forest Inventory Analysis 2022a). As in previous versions, for convenience, we provide a Tree Table with information drawn from the FIA DataMart on the size, species, and status (live or dead) of each tree. The Tree Table is provided in SQL and .csv formats, and is a standalone table not linked to the raster because of the one-to-many relationship with plot identifier (i.e., each plot contains multiple trees) (Riley, Grenfell, Finney, and Shaw 2021). Combined with the distance and azimuth recorded for each tree in FIA's databases, the TreeMap contains the information needed to map tree-level spatial realizations of forest structure used in applications including computational fluid dynamics modeling of fire behavior (Russ Parsons, personal communication, October, 2021). (The most current version of the FIA database has redacted the distance and azimuth measurements, but users may rely on earlier copies of the databases they may have downloaded to obtain these parameters.)

The TreeMap 2016 included two substantial improvements over the previous version: (1) the inclusion of a raster attribute table that summarizes forest characteristics and enables mapping of these, and (2) improvement in the accuracy with which FIA plots are assigned to recently disturbed areas. Whereas in previous versions of the dataset, we included only the plot identifier in the gridded map, leaving it to the user to link to the FIA databases to retrieve attributes such as forest type or carbon, as a new feature of the 2016 dataset release, we have provided a raster attribute table that includes several commonly used FIA variables and derived attributes. Users may simply access the attribute table of the TreeMap raster in a GIS to generate maps of number of dead trees, number of live trees, forest type, stand size, live basal area, live canopy cover, stand height, live aboveground carbon, and several others. This new feature increases the ease of use of the TreeMap dataset and the accessibility of FIA data to users. Users may still query the FIA DataMart directly via SQL statements to retrieve additional attributes. We have also improved the accuracy with which we assign disturbed plots to recently disturbed areas, thus enhancing the estimates of snag hazard now used during wildland fire incidents (Riley et al. 2022, Risk Management Assistance 2021). In addition to improved snag hazard mapping, recent new uses of the TreeMap datasets include fuel reduction planning (Gunnar Carnwath and John Shaw, personal communication, virtual meeting 2021) and carbon inventory (EarthShot and Spatial Informatics Group, personal communication, virtual meeting). TreeMap is currently being integrated into FIA state-level reports in some western states.

In this article, we detail the changes in methodology and demonstrate the capability to map and summarize forest characteristics using the new attribute table feature of TreeMap 2016. We present validation of the dataset using three methods.

Methods

Random Forests Imputation

As in earlier TreeMap datasets, we used machine learning to assign data collected on FIA plots to a gridded map of CONUS at 30 × 30m resolution. We used a random forests methodology to assign the most similar FIA forest plot to each pixel mapped as forested by LANDFIRE. Forested pixels were defined as those with ≥10% live tree cover as per LANDFIRE's Existing Vegetation Cover raster (LANDFIRE 2020a). TreeMap does not include areas mapped by LANDFIRE's Existing Vegetation Type as "Recently disturbed forest", as these pixels do not have information on the type of forest, a requirement for the imputation process; many but not all of these areas fall below the criterion of 10% live forest cover.

Random forests requires two sets of data known as reference data and target data (Crookston and Finley 2008): (1) detailed observations (such as the tree-level measurements at FIA forest plots) referred to as the "reference data", and (2) landscape data with more general information for all ground locations (often derived from satellite imagery, as in the case of LANDFIRE), referred to as the "target data" (Riley, Grenfell, and Finney 2016). During the modeling process, the reference data (FIA plots) are assigned to the target data (LANDFIRE) to produce detailed measurements for all locations.

The reference data for this project consisted of a set of 100% accessible, single-condition, 100% forested plots (n = 65,652). We used only single-condition plots because they are more homogenous and thus more suited to convey conditions at a single pixel for which a single set of observations is obtained from LANDFIRE; multicondition plots can include multiple land uses (forest and nonforest), reserved status, forest type, stand-size class, regeneration status, stand density, or ownership on a single plot (Burrill et al. 2018). Furthermore, we used only plots that fell within the boundaries of CONUS and had over 10% live tree cover (as determined by the StrClass keyword in the Forest Vegetation Simulator [FVS]) to be consistent with the definition of "forested" across both the reference and target data (Crookston and Stage 1999; Riley, Grenfell, and Finney 2016; Riley, Grenfell, Finney, and Wiener 2021).

The target data consisted of a set of gridded maps from the LANDFIRE project circa 2016. These depict location, topography, vegetation, disturbance, and biophysical parameters (Table 1) (LANDFIRE 2020a).

We used the random forests algorithm in the *yaimpute* package of the statistical software R with default settings (Crookston and Finley 2008, R Foundation 2019). Random forests constructs a "forest" of decision trees. For each decision tree, a number of plots from the reference data equal to the number available is randomly chosen with replacement, with the result that not all plots are included in each decision tree, and plots that are left out can be used to assess model accuracy, computed via the out-of-bag error rate. Predictor variables must be available in both the target and reference data. We chose a suite of predictor variables related to location, topography, vegetation, disturbance, and biophysical parameters. Random forests can use more than one response variable, and here we chose four vegetation variables to optimize predictions for carbon (our ultimate research goal). Each decision tree is assigned one response variable. At each branch (or "node") of a decision tree, a number of variables is randomly chosen that equals the square root of the number of predictor variables available (we had seventeen predictors, so four would be chosen at each node for evaluation) (Table 1) (Riley, Grenfell, and Finney 2016). Each of the chosen variables at a node is evaluated to see which reduces the variance in the response variable the most when the observations are split into two groups (or "buckets"); the variable that reduces variance the most is chosen for that partition. (Thus, the function of the response variable here is to partition the tree.) Partitioning continues until further partitioning would not reduce the variance significantly or would reduce the number of observations in a bucket to less than five. That decision tree is then complete, and the program completes another until it reaches the number specified by the user. We used four hundred decision trees per LANDFIRE zone, or one hundred for each of our four response variables (Table 1). Random forests selects the plot identifier that appears most often in the terminal bucket with the observations at a given pixel for assignment to that pixel (Riley, Grenfell, and Finney 2016; Riley, Grenfell, Finney, and Wiener 2021). For more detail on methods and diagrams of project workflow, we refer the interested reader to the text of Riley, Grenfell, and Finney (2016) and Riley, Grenfell, Finney, and Wiener (2021), particularly Figure 2 in the former and Figure 1 in the latter.

We chose a suite of predictor variables that appeared in both the reference (FIA forest plot) and target (gridded LANDFIRE) data, as is required by random forests (Table 1). Predictor variables included location (latitude and longitude), topography (slope, aspect, and elevation), vegetation (Existing Vegetation Cover [EVC] in percent, Existing Vegetation Height [EVH], and Existing Vegetation Group [EVG]), biophysical variables (maximum temperature, minimum temperature, relative humidity, photosynthetically active radiation, vapor pressure deficit, and precipitation), and disturbance (disturbance year and code, where "0" corresponded to no disturbance since 1999 and "1" corresponded to disturbance by either fire or insect/disease). Following the custom of LANDFIRE data, EVC is divided into nine classes (10–20%, 20–30%, 30–40%, 40–50%, 50–60%, 60–70%, 70–80%, 80–90%, and 90–100%), whereas EVH is partitioned into five classes (0–5 m, 5–10 m, 10–25 m, 25–50 m, and > 50 m). Response variables included EVC, EVH, EVG, and disturbance code. Variables can serve as both predictor and response variables in random forests, which boosts accuracy of these variables. Because our original research goal was to predict carbon, for the first two editions of the dataset, we chose response variables related to tree volume: EVC, EVH, and EVG. Here we added a fourth response variable, disturbance code, to improve the model's performance in disturbed areas. Good performance in disturbed areas was considered critical to predicting snag hazard to increase firefighter safety during response to wildland fires (Riley et al. 2022). Predictor and response variables were computed for the target and reference data as described in Table 1.

Derived Characteristics and Data Linkages

After the modeling process described above has been performed, each pixel of TreeMap has been assigned

Table 1. Variables used in imputation must be present in both reference and target data. All variables listed are predictor variables in the random forests model; response variables are listed in bold. Section numbers refer to variable descriptions in the FIA Database user's manual (Burrill et al. 2018).

Variable name	Variable source for reference (FIA) Data	Variable source for target (LANDFIRE) data
Latitude	FIA program through a memorandum of cooperation	Pixel centroid
Longitude	—	—
Slope	FIA DataMart, Condition table, section 2.4.33	LANDFIRE 2016 Remap (LF 2.0.0) (LANDFIRE 2020a)
Aspect	FIA DataMart, Condition table, section 2.4.34	—
Elevation	FIA DataMart, Plot table, section 2.3.22	—
Existing Vegetation Cover	Estimated by FVS StrClass keyword	—
Existing Vegetation Height	Estimated by FVS StrClass keyword, "Stratum_1_Nom_Ht" field	—
Existing Vegetation Group	Keyed via LANDFIRE AutoKey routine (see Riley, Grenfell, Finney, and Wiener (2021) for description)	Existing Vegetation Type raster from LANDFIRE 2016 Remap (LF 2.0.0) (LANDFIRE 2020a)
Maximum temperature	Obtained by spatial overlay of plot coordinates with LANDFIRE Biophysical Gradient Raster(LANDFIRE 2020b)	LANDFIRE Biophysical Gradient Raster (LANDFIRE 2020b)
Minimum temperature	—	—
Relative humidity	—	—
Photosynthetically active radiation	—	—
Vapor pressure deficit	—	—
Precipitation	—	—
Disturbance type	Derived from disturbance codes in FIA DataMart, sections 2.4.38, 2.4.40, and 2.4.42 (see Riley, Grenfell, Finney, and Wiener (2021) for specifics)	Derived from LANDFIRE disturbance rasters (LANDFIRE 2020a) (see Riley, Grenfell, Finney, and Wiener (2021) for specifics)
Disturbance year	Derived from disturbance years in FIA DataMart, sections 2.4.39, 2.4.41, and 2.4.43 (see Riley, Grenfell, Finney, and Wiener (2021) for specifics)	—

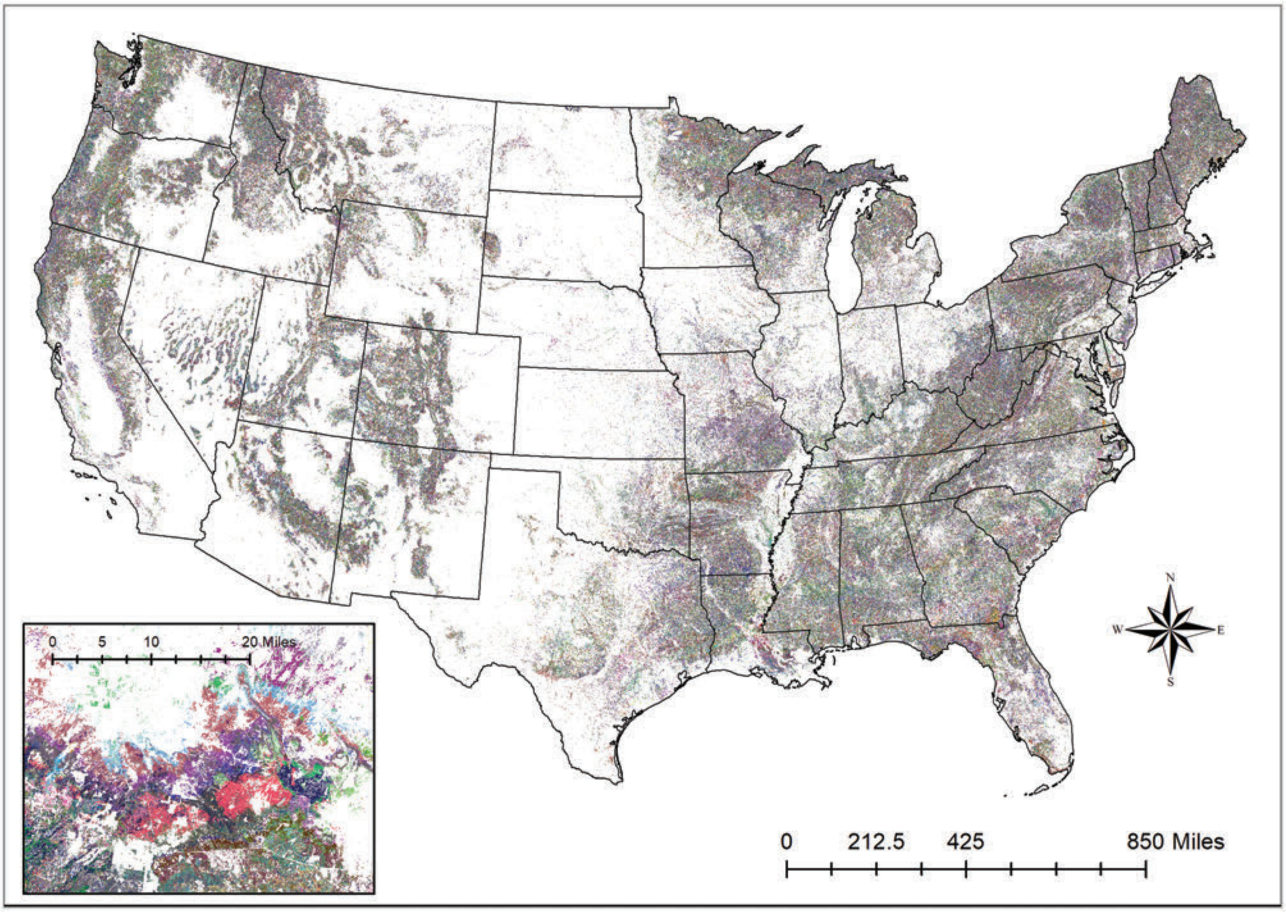

Figure 1 Continental-scale map of TreeMap 2016. Each unique color corresponds to a unique FIA forest plot. The inset window shows a portion of the TreeMap for the Mogollon Rim in Arizona; color bands highlight biophysical gradients that drive plot assignments.

an imputed FIA plot identifier ("CN"). The TreeMap methodology was designed to enable linkages between these gridded plot identifiers and the FIA databases (FIADB) (Forest Inventory Analysis 2022a) to access this rich dataset of attributes available for the plot and its individual trees. In addition, the FIADB linkage provides direct access to FIA data in FVS-ready format (Shaw and Gagnon 2020), allowing users to not only obtain estimates of forest attributes for their area of interest but also project future forest conditions. As FVS is used frequently by users in the National Forest System and in the private sector, this linkage can enable estimates of tree growth, amount of wood in harvested products, etc.

We provide two tables with the TreeMap 2016: (1) the Tree Table, an aspatial table of individual tree characteristics for each plot, including the species, status (live or dead), and diameter of each tree, and (2) an attribute table attached to the raster that summarizes a set of commonly used attributes (e.g., number of live trees) drawn from the FIA data (Riley, Grenfell, Finney, and Shaw 2021). As for previous editions of TreeMap, we used the CNs to extract basic tree-level information for each plot from the FIA DataMart and store it in the TreeMap Tree Table (Table 2). The Tree Table (produced via the TreeMap-FIA database linkage) allows users to create customized summaries of the tree-level data (such as the percentage of stand composition of certain species based on basal area or number of trees above a chosen diameter) that can then be joined back to the TreeMap raster via the CN value found in the raster attribute table to produce maps of the custom-summarized attribute.

As a new feature of TreeMap, we also include a set of commonly used plot-level descriptors in the raster attribute table (Table 3). Most of these are derived from tree-level data, either as part of the FIA data compilation process or using modifications of standard FIA queries on the published tables to generate plot-level instead of population-level data. In some cases, the variables are recorded directly by crews in the field. For example, forest type (FORTYPCD) is based on

Table 2. Attribute definitions and sources for the tree table ("TreeMap2016_tree_table.csv" in the Research Data Archive (Riley, Grenfell, Finney, and Shaw 2021)). All fields derived from FIA tables unless otherwise noted. Variable section references correspond to FIA Database documentation (Burrill et al. 2018).

Column name	Description	Source (table/column/query)
tm_id	TreeMap record number; corresponds to "Value" field in the TreeMap raster ("TreeMap2016.tif")	Native to TreeMap dataset
CN	Control number for unique TreeMap pixel assignments	Corresponds to PLOT.CN and TREE.PLT_CN in FIADB, and is used to link to the tables and query results listed below.
STATUSCD	A code that indicates whether the tree is live (1) or dead (2) when the plot was measured	From TREE.STATUSCD (Sec.3.1.15)
TPA_UNADJ	Number of trees per acre, based on the scale of the plot design	From TREE.TPA_UNADJ (Sec. 3.1.111)
SPCD	Tree species code	From TREE.SPCD (Sec. 3.1.16)
COMMON NAME	Common name of tree species	From REF_SPECIES.COMMON_NAME (Sec. 9.5.2)
SPECIES_ SYMBOL	USDA PLANTS Database code	From REF_SPECIES.SPECIES_SYMBOL (Sec.9.5.7)
DIA	Tree diameter (inches), measured at 4.5 feet above ground for timber species and at the ground line for woodland species	From TREE.DIA (Sec. 3.1.18)
HT	Tree's total height (ft) from the tip of the apical meristem to the ground. The actual height may be different when the main stem is broken, in which case the total height is estimated	From TREE.HT (Sec. 3.1.20)
ACTUALHT	The height of the tree (ft) from the ground level to the highest point of the tree still present	From TREE.ACTUALHT (Sec. 3.1.22)
CR	Compacted crown ratio, or the percent of the bole with healthy foliage	From TREE.CR (Sec. 3.1.24)
SUBP	Subplot number (plots are composed of four subplots)	From TREE.SUBP (Sec. 3.1.9)
TREE	Tree number, which corresponds to a unique identifier on a subplot	From TREE.TREE (Sec. 3.1.10)
AGENTCD	Cause of death, including insect, fire, etc., where applicable	From TREE.AGENTCD (Sec. 3.1.27)

Table 3. Attribute definitions and data sources for the output TreeMap ("TreeMap2016.tif" in the Research Data Archive, (Riley, Grenfell, Finney, and Shaw 2021)). All fields derived from FIA tables unless otherwise noted. Variable section references correspond to FIA Database documentation (Burrill et al. 2018).

Column name	Description	Source (table/column/query); Sec = FIADB manual section
CN	Control number for unique TreeMap pixel assignments	Corresponds to PLOT.CN and TREE.PLT_CN in FIADB and is used to link to the tables and query results listed below
Count	Count of pixels having CN in TreeMap	Native to TreeMap
Value	TreeMap record number; corresponds to "tm_id" field in the tree table ("TreeMap2016_tree_table.csv") and raster attribute table	Native to TreeMap
FORTYPCD	Forest type code, assigned by FIA algorithm	From COND.FORTYPCD (Sec. 2.4.16)
ForTypeName	Forest type description, assigned by FIA algorithm	From REF_FOREST_TYPE.MEANING (Sec. 9.3.2) using FORTYPCD = VALUE (Sec.9.3.1)
FLDTYPCD	Forest type code, assigned by FIA field crew	From COND.FLDTYPCD (sec. 2.4.17)
FldTypeName	Forest type description, assigned by FIA field crew	From REF_FOREST_TYPE.MEANING (Sec. 9.3.2) using FLDTYPCD = VALUE (Sec.9.3.1)
STDSZCD	Stand size code, assigned by FIA algorithm	From COND.STDSZCD (Sec. 2.4.20)
FLDSZCD	Stand size code, assigned by FIA field crew	From COND.FLDSZCD (Sec. 2.4.21)
BALIVE	Live tree basal area (sq ft)	From COND.BALIVE (Sec. 2.4.51)
CANOPYPCT	Live canopy cover (percent)	From Forest Vegetation Simulator StrClass calculation
STANDHT	Height of dominant trees (ft)	From Forest Vegetation Simulator StrClass calculation
ALSTK	All live tree stocking (percent)	From COND.ALSTK (Sec. 2.4.53)
GSSTK	Growing-stock stocking (percent)	From COND.GSSTK (Sec. 2.4.54)
QMD_RMRS	Stand quadratic mean diameter	From COND.QMD_RMRS (Sec. 2.4.136)
SDIPCT_RMRS	Stand density index (percent of maximum)	From COND.SDIPCT_RMRS (Sec. 2.4.139)
TPA_LIVE	Number of live trees per acre (DIA > 1")	Query; Sum TREE.TPA_UNADJ WHERE (((COND.COND_STATUS_CD) = 1)AND ((TREE.STATUSCD) = 1) AND ((TREE.DIA)>=1))
TPA_DEAD	Number of standing dead trees per acre (DIA > = 5")	Query; Sum TREE.TPA_UNADJ WHERE (((COND.COND_STATUS_CD) = 1) AND ((TREE.STATUSCD) = 2) AND ((TREE.DIA)>=5) AND ((TREE.STANDING_DEAD_CD) = 1))
VOLCFNET_L	Volume, live, cubic feet per acre	Query Sum VOLCFNET*TPA_UNADJ WHERE (((COND.COND_STATUS_CD) = 1) AND ((TREE.STATUSCD) = 1))
VOLCFNET_D	Volume, standing dead, cubic feet per acre	Query Sum VOLCFNET*TPA_UNADJ WHERE (((COND.COND_STATUS_CD) = 1) AND ((TREE.STATUSCD) = 2) AND ((TREE.DIA)>=5) AND ((TREE.STANDING_DEAD_CD) = 1))
VOLBFNET_L	Volume, live, sawlog board feet per acre (log rule:Int'l ¼ inch)	Query Sum VOLBFNET * TPA_UNADJ WHERE (((TREE.TREECLCD) = 2) AND ((COND.COND_STATUS_CD) = 1) AND ((TREE.STATUSCD) = 1))

Table 3. Cotinued

Column name	Description	Source (table/column/query); Sec = FIADB manual section
DRYBIO_L	Dry live tree biomass, above ground (tons per acre)	Query; Sum (DRYBIO_BOLE, DRYBIO_TOP, DRYBIO_STUMP, DRYBIO_SAPLING, DRYBIO_WDLD_SPP)/2000*TPA_UNADJ WHERE (((COND.COND_STATUS_CD) = 1) AND ((TREE.STATUSCD) = 1))
DRYBIO_D	Dry standing dead tree biomass, above ground (tons per acre)	Query; Sum (DRYBIO_BOLE, DRYBIO_TOP, DRYBIO_STUMP, DRYBIO_SAPLING, DRYBIO_WDLD_SPP)/2000*TPA_UNADJ WHERE (((COND.COND_STATUS_CD) = 1) AND ((TREE.STATUSCD) = 2) AND ((TREE.DIA)>=5) AND ((TREE.STANDING_DEAD_CD) = 1))
CARBON_L	Carbon, live above ground (tons per acre)	Query; Sum (DRYBIO_BOLE, DRYBIO_TOP, DRYBIO_STUMP, DRYBIO_SAPLING, DRYBIO_WDLD_SPP)/ 2/2000*TPA_UNADJ WHERE (((COND.COND_STATUS_CD) = 1) AND ((TREE.STATUSCD) = 1))
CARBON_D	Carbon, standing dead (tons per acre)	Query; Sum (DRYBIO_BOLE, DRYBIO_TOP, DRYBIO_STUMP, DRYBIO_SAPLING, DRYBIO_WDLD_SPP)/ 2/2000*TPA_UNADJ WHERE (((COND.COND_STATUS_CD) = 1) AND ((TREE.STATUSCD) = 2) AND ((TREE.DIA)>=5) AND ((TREE.STANDING_DEAD_CD) = 1))
CARBON_DWN	Carbon, down dead (tons per acre)	From COND. CARBON_DOWN_DEAD (Sec. 2.4.67)
tm_id	TreeMap record number; duplicates “Value” field	native to TreeMap

FIA's forest type classification algorithm (Arner et al. 2001), and the field forest type (FLDTYPCD) is based on crew observation. For each tree- and plot-level variable, the "Source" columns in Tables 2 and 3 provide the detailed information needed to understand the derivation of each variable, whether it is extracted directly from an FIADB table, computed using a query, or obtained by other means.

We refer to "standard" FIA queries above because FIA queries, online tools, and tables in state reports typically adhere to national standards; that is, they use common diameter minima and other filtering criteria that allow for consistent comparisons across geographic units. For example, standard FIA queries use a 1 inch minimum for number of live trees and a 5 inch minimum for dead trees because the measurement of dead saplings (1.0–4.9 inches diameter) is a relatively new addition to FIA protocols, and therefore these data are not available for all plot visits. Because of this change, FIA queries cannot produce dead sapling estimates to include trees < 5 inches diameter at breast height/diameter at root collar for all inventory periods or geographic units. Because TreeMap is based on the same plot data that are used in FIA queries, it necessarily uses the same query standards to produce correct and comparable estimates. This is also useful for comparisons between TreeMap- and FIA-based population estimates, such as those we present in this article.

Methods for plot-level summaries of attributes presented in the raster attribute table are given below and in Table 3. From the tree-level data, we calculated the number of live and dead trees, live basal area, and aboveground carbon (carbon in standing live and dead trees) (Table 3). Number of live trees per acre was calculated by subsetting the tree records to those with live status (STATUSCD = 1) and multiplying each individual tree record by the number of trees it represents per acre, as per the FIA plot design (TPA_UNADJ field), then summing for each plot. Number of dead trees per acre was calculated in the same fashion, except by subsetting to dead trees (STATUSCD = 2) greater than or equal to 5 inches in diameter. Mean tree density for a unit was calculated by summing the number of trees and then dividing by the number of forested pixels.

Live and dead aboveground biomass and carbon calculations were based on querying biomass components in the FIA databases. Aboveground biomass of trees > 5.0 inches in diameter is divided into three components: the main stem (DRYBIO_BOLE), the top and limbs (DRYBIO_TOP), and the stump (DRYBIO_STUMP), whereas for saplings and woodland species, the biomass is calculated for the tree as a whole (DRYBIO_SAPLING and DRYBIO_WDLD_SPP). For these components, biomass is stored in FIADB in pounds per acre, whereas biomass and carbon are usually reported in tons or tons per acre. Thus we used the following calculations to calculate tons per acre of biomass at the plot level:

$$\text{Above-ground biomass}_i = \left(\frac{\sum_i (\text{DRYBIO\_BOLE}, \ \text{DRYBIO\_TOP}, \ \text{DRYBIO\_STUMP}, \ \text{DRYBIO\_SAPLING}, \ \text{DRYBIO\_WDLD\_SPP}) \cdot TPA\_UNADJ_i}{2000} \right)$$

where i = the i[th] tree on a plot, STATUSCD = 1 and DIA ≥1.0 for live trees, and STATUSCD = 2 and DIA ≥5.0 for dead trees. Biomass for all trees on a plot is then summarized to obtain the total for each plot. Per-acre figures can be converted to per-pixel values (900 m^2) by multiplying by a factor of 0.222395 acres/pixel.

To obtain carbon, standard FIA queries use the simple conversion of carbon = biomass/2 and we followed this convention. In addition, we calculated the mean live tree carbon loading by dividing the total live tree carbon loading by the number of forested pixels within a unit. This calculation was repeated to calculate the mean dead tree carbon loading for each unit.

Validation

To assess the 2016 TreeMap product we used a multipronged validation process. We have described the out-of-bag error rate above; it is a calculation internal to the random forests model. Additionally, as part of the modeling process we (1) compared forest characteristics at FIA plot locations to the imputed TreeMap raster, and (2) compared the imputed TreeMap raster data to the LANDFIRE target data. In addition, we conducted what might be considered a post hoc validation, in which we compare TreeMap-based metrics against FIA-based estimates at various geographic scales. These three types of validation are intended to show, at a coarse scale, how users might initially assess the performance of TreeMap in their own areas of interest. It is not an exhaustive assessment, partly because space does not allow for such detail, but mostly because users' specific results will vary by their specific analysis requirements, range of forest conditions, and geographic scope.

Comparison of Characteristics of Plot Location with Imputed Data

The first validation method involved a spatial comparison of characteristics of the TreeMap imputed data (EVC, EVH, EVG, disturbance class, and the two tree species with highest basal area) with the characteristics of FIA plots at their locations. Four of these five characteristics are the response variables chosen for the random forests model (EVC, EVH, EVG, and disturbance class) and the fifth is a characteristic derived independently from the FIA tree data (two tree species with highest basal area). We have chosen these variables for assessment because they are expected to convey suitability of the dataset for many applications regarding tree volume and carbon. To identify the two tree species with highest basal area, we computed the basal area of each tree on a plot and summed by species (Riley, Grenfell, Finney, and Wiener 2021). This validation method indicates how well the imputation performed relative to ground truth (although the FIA plot measurements may contain some error as well, and pixel locations in a national map projection will also have error).

For this endeavor, we used a set of 2,937 multicondition plots measured by FIA in 2016, the same date as the target LANDFIRE data. These plots represent all multicondition plots measured by FIA in 2016 that were 100% forested and 100% accessible for measurement. These validation plots are located across CONUS and were not used to train the random forests model or available for imputation. We elected to use multicondition plots not used in the imputation rather than holding out single-condition plots so as not to reduce the number of plots available for imputation (as this number was already smaller than ideal) and to avoid losing plots from rare vegetation types, which can be represented by less than five available plots. FIA's plot design includes a set of four circular subplots of radius 24 ft, with three of the subplots arrayed around the center, one at a distance of 120 ft between subplot centers (Burrill et al. 2018), resulting in a plot footprint that approximately occupies 4–9 pixels at the 30 × 30 m pixel scale. Because of this relationship between TreeMap pixel size and FIA plot footprint, many validation plots intersect with TreeMap pixels with several different CNs assigned, and some intersect with nonforest pixels with null values.

For each validation plot, we identified all forested pixels of the TreeMap whose centroids fell within the plot radius. As measures of accuracy, we checked how often at least one imputed pixel within the plot radius matched the plot value for each of the response variables (EVC, EVH, EVG, disturbance). We also repeated this analysis using basal area of tree species in lieu of vegetation group, as tree species are a measurable characteristic of forest plots and are not affected by any uncertainty in the classification system used to assign EVG. Specifically, we checked whether at least one pixel within plot radius was assigned a plot with at least one of the same top two species as the validation plot. In addition, we checked how often the average cover value of all pixels within the plot radius was within 10% of the plot value and the average height value of all pixels within the plot radius was within 16.4 ft. These values were chosen because tree cover classes are 10% wide and the narrowest of the height classes is 16.4 ft wide.

Comparison of Imputed and Target Data

The second validation method compares characteristics of the imputed TreeMap data with the gridded LANDFIRE target data. For each of the 2,699,430,013 pixels, we compare agreement between the LANDFIRE target grids and the imputed TreeMap grids for each of the four response variables, EVC, EVH, EVG, and disturbance code. This method indicates how well the model performed with respect to the input maps, or in other words, how closely the random forests model was able to find a plot that matched each LANDFIRE pixel. This is partially a reflection of how well the FIA plots cover the span of combinations of EVC, EVH, EVG, and disturbance code found in the LANDFIRE data, and the number of times only a less-than-perfect match was the best available. If the input LANDFIRE maps were perfect and indicated ground truth, this would be the only method needed; however, some error is present. Agreement between target and imputed data was reported using the producer's accuracy (the proportion of pixels in which a category in the target data was correctly assigned in the imputed map) and user's accuracy (the proportion of pixels in which a class in the imputed data was correctly assigned to the target data).

Comparison of Area-Level Summaries Derived from TreeMap and FIA

We summarized a number of live and dead trees as well as live and dead carbon derived from the TreeMap, calculating values for each pixel as described above and dividing geographic regions using state, National Forest, and National Park Service boundaries from the Protected Areas Database of the United States 2.1 (U.S. Geological Survey Gap Analysis Project 2020). FIA plot-based estimates for the same areas were performed using the EVALIDator online tool (USDA Forest Service 2022). EVALIDator permits users to query data belonging to various "evaluations", which are groups of plots collected over enough years to represent a full sample and therefore can be used to create population estimates. Evaluations typically include 5–7 years of continuous inventory for eastern states and 10 years for western states, in contrast with LANDFIRE and TreeMap, which are based on remote sensing data from a single year. Because of this mismatch in timespan, TreeMap and FIA estimates cannot be compared precisely, especially in cases where rapid change is occurring. For example, if a large fire causes high tree mortality within its perimeter in a western state in 2015, the LANDFIRE 2016 maps and the TreeMap 2016 will include the effects of this fire in its entirety. However, the FIA estimates are likely to reflect only one-tenth of this change in 2016, as only one-tenth of the plots will have been revisited and remeasured in that year. This means FIA estimates will lag the pace of rapid landscape changes. We therefore elected to use the most current evaluations available in Evalidator for each state, which may span, for example, 2010–2020 in the western United States and 2013–2020 or 2015–2020 in the eastern United States and capture more of these types of rapid change for the period centered on 2016 than an evaluation ending in 2016.

For each metric, we obtained both the totals and means (i.e., values per unit area). We illustrate these comparisons with graphs and using the R^2 statistic for agreement. One reason for including mean estimates in this step is that the definitions of "forested" differ somewhat between the LANDFIRE, TreeMap, and FIA. As noted above, pixels were included in TreeMap if they met the definition of at least 10% live tree cover in 2016 and had an EVG assigned (meaning that some recently severely disturbed areas were excluded). FIA, on the other hand, defines forest as land having had a minimum of 10% live tree cover at some point in the past. As a result, a recently cut or severely burned FIA plot location will count as forest land in FIA estimates but will not likely be assigned a plot CN in TreeMap. Totals will be more greatly affected by these differences in the definition of "forested" than means.

Results

Of 65,652 plots available for imputation, 55,609 plots were selected at least once by the random forests model for imputation to 2,699,430,013 forested pixels in CONUS (Figure 1). The random forests model exhibited extremely high model accuracies, with out-of-bag error rates ranging from 0.0011 to 0.0092 for EVG, 0.0001 to 0.0009 for EVH, 0.0030 to 0.0137 for EVC, and 0 for disturbance code. As described in methods, we used the FIA plot CN assignments, the summation of pixel counts, and the pixel area expansion factors to produce maps and summaries of several of the derived products, including number of live and dead trees, forest type, live basal area, and live and dead carbon; we present results below. Results presented here are not exhaustive and represent only a small fraction of the types of analyses users may conduct.

Validation

Comparison of Characteristics of Plot Location with Imputed Data

We compared the canopy cover (EVC), height (EVH), vegetation group (EVG), the dominant two tree species (by basal area), and the disturbance code at the location of 2,749 multicondition FIA plots to the characteristics of the plot imputed at these locations (Table 4). These plots excluded those that did not coincide with LANDFIRE forested areas. Canopy cover of at least one pixel within the plot's radius matched that of the imputed value 57.5% of the time, while the mean cover of the pixels within the plot's radius was within 10% of the plot value in 60.9% of cases. This represents an improvement of about 12%–13% over the TreeMap 2014 edition for these two metrics. The height class of at least one pixel within the plot radius matched that of the plot in 80.0% of cases, a decline in accuracy of about 6% from the TreeMap 2014. However, when measured by a second metric, accuracy in height increased: the percentage of plots where the mean height of pixels within the plot radius was within 16.4 ft of the plot value increased by about 3% to 73.0%. The vegetation group of at least one pixel within the plot radius matched the plot value in 51.8% of cases; this metric was not reported for TreeMap 2014. The species of at least one of the two tree species with highest basal area matched that of at least one pixel within the plot radius in 80.0% of cases, an increase of about 3% from the previous version. The disturbance code of at least one pixel within the plot's radius matched that of the plot in 87.4% of cases, a slight decrease of about 3% from the TreeMap 2014.

Comparison of Imputed and Target Data

When the imputed values of each pixel were compared to the corresponding pixel in the LANDFIRE target data, TreeMap 2016 had accuracies above 90%

Table 4. Summary of spatial accuracy assessment for TreeMap 2016. Accuracy is reported in percent.

	At least one pixel within plot radius matched plot value in…		Weighted cover/height of pixels within plot radius was within 10%/ 16.4 ft of plot value in …	
Response variable	TreeMap 2014	TreeMap 2016	TreeMap 2014	TreeMap 2016
Cover (EVC)	44.0%	57.5%	48.7%	60.9%
Height (EVH)	85.7%	80.0%	70.3%	73.0%
Vegetation group (EVG)	not tabulated	51.8%	NA	NA
Tree species with two highest basal areas	76.7%	80.0%	NA	NA
Disturbance code	90.6%	87.4%	NA	NA

Table 5. Within-class accuracy for Existing Vegetation Cover (EVC), Existing Vegetation Height (EVH), Existing Vegetation Group (EVG), and disturbance code for TreeMap 2014 and TreeMap 2016.

	EVC	EVH	EVG	Disturbance code
TreeMap 2016	97.7%	99.6%	94.8%	99.98%
TreeMap 2014	97.2%	99.2%	93%	90.3%
Difference	0.5%	0.4%	1.8%	9.68%

for all four response variables (EVC, EVH, EVG, and disturbance code) at the national scale. This represents an improvement for all response variables over TreeMap 2014 (Table 5). The greatest improvement in accuracy was achieved in the disturbance code variable, which was a direct reflection of adding it as one of the response variables, which resulted in accuracy jumping from 90.3% to 99.98% for this variable (Table 5).

Overall accuracy for EVC was 97.7% (Tables 5 and S1). Confusion was most common across adjacent classes, meaning relative differences were within 10%. Overall, the proportion of the landscape falling into the various cover classes in the FIA plots (reference data), LANDFIRE pixels (target data), and imputed dataset was similar across the three datasets (Figure S1). Because the FIA plots are a systematic sample of the landscape, they are expected to capture the proportion of these classes fairly accurately, so this similarity indicates that LANDFIRE and the TreeMap do as well. As in previous datasets, the producer's accuracy increased with the number of plots available for imputation in a cover class (Figure S2).

Overall within-class accuracy in EVH was 99.6% at the national scale (Tables 5 and S2) and confusion was most frequent across adjacent classes. In general, the proportion of the landscape falling into each height class was similar across the FIA plot (reference) data, LANDFIRE (target) data, and the imputed dataset (Figure S3). Accuracy generally increased with the number of plots available for imputation in a class; however, because all four height classes were well-populated with plots, accuracies were close to 100% for all classes, and the increase in accuracy with plot number was small (Figure S4).

Overall accuracy in EVG was 94.8% (Tables 5 and S3). Most EVGs with low accuracy account for only a small proportion of the landscape. In general, the proportion of the landscape falling into each vegetation group was similar in the FIA plot (reference) data, the LANDFIRE (target) data, and the imputed dataset (Figure S5). Accuracy increased with the number of plots in an EVG (Figure S6). Errors in classification often involved classification to a similar vegetation type or one with close associations (e.g., limber pine and juniper).

Disturbance code had the highest accuracy of any response variable at 99.98% nationally (Tables 5 and S4). The proportion of the landscape in the disturbed class was likely underestimated by LANDFIRE, whereas the undisturbed class was conversely overestimated (Figure S7). Mapping insect- and disease-affected areas is challenging at the national scale, and when we compared the locations of FIA plots affected by insect and disease to those locations in

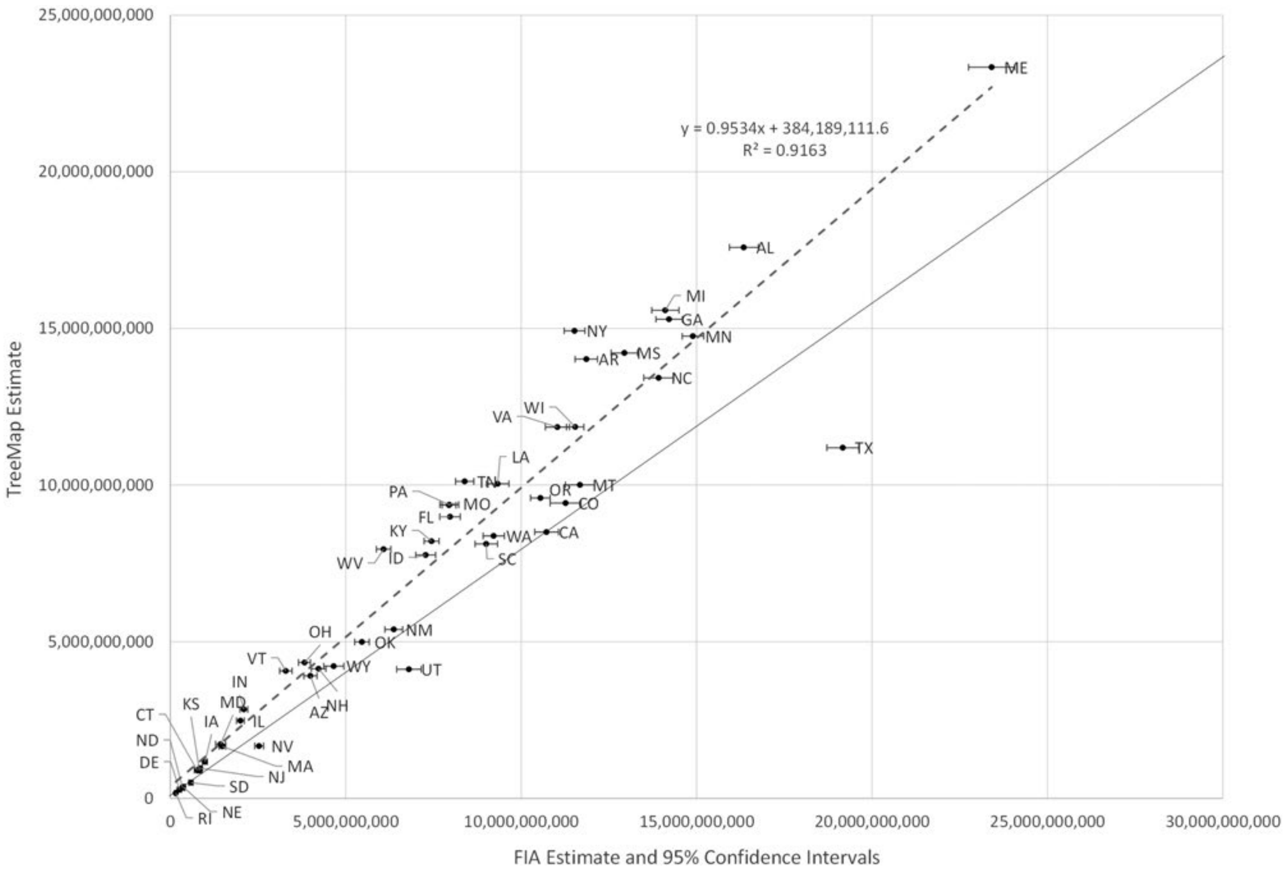

Figure 2 Comparison of number of live trees, estimated from TreeMap and FIA by state. The 95% confidence intervals are displayed for FIA estimates.

the LANDFIRE disturbance rasters, these locations were not always mapped as disturbed. This underestimation of insect- and disease-affected areas propagates to the imputed raster. Producer's accuracy increased with the number of plots in a disturbance class (Figure S8).

Comparison of Area-Level Summaries Derived from TreeMap and FIA

Numbers of trees by state. With the notable exception of Texas, TreeMap and FIA estimates of total numbers of live trees are tightly clustered around the 1:1 diagonal and the fitted relationship ($R^2 = 0.916$) (Figure 2). Inspection of TreeMap in Texas revealed that errors of omission (live trees on the ground, but no plot CN assigned) tend to occur in sparsely wooded land (e.g., mesquite and juniper) that is on the margin of the FIA 10% cover threshold. This assessment is consistent with the fact that many of the other states that lie appreciably below the diagonal (e.g., Colorado and Utah) also have considerable areas of sparse woodlands.

State-level estimates from TreeMap are mapped in Figure 3 and given in Table S5. As noted above, in previous versions of TreeMap, calculations for number of live and dead trees had to be accomplished by summarizing either the Tree Table or linking to the FIA DataMart, and then using the plot identifier (CN) field to join these values to the TreeMap raster. Now, users can use the "TPA_LIVE" (number of live trees per acre) and "TPA_DEAD" (number of dead trees per acre) fields of the raster (Table 3) and the acres-to-pixel expansion factor to conduct spatial summaries by state, region, or other unit in a GIS, which is how we created these summaries.

In contrast, although the relationship between TreeMap and FIA estimates of dead trees is also close along a fitted line ($R^2 = 0.917$), TreeMap generally underestimates the total number of dead trees (Figure 4). This is an expected result, for two reasons. The first is the aforementioned difference in the area classified as forest by TreeMap and FIA. An FIA plot sampling 100% mortality, or nearly so, due to fire or insects will still count as forest land in FIA's estimates, and the dead trees tallied on the plot will contribute to the dead tree estimate, whereas areas with < 10% live tree cover will not be included in TreeMap estimates. The

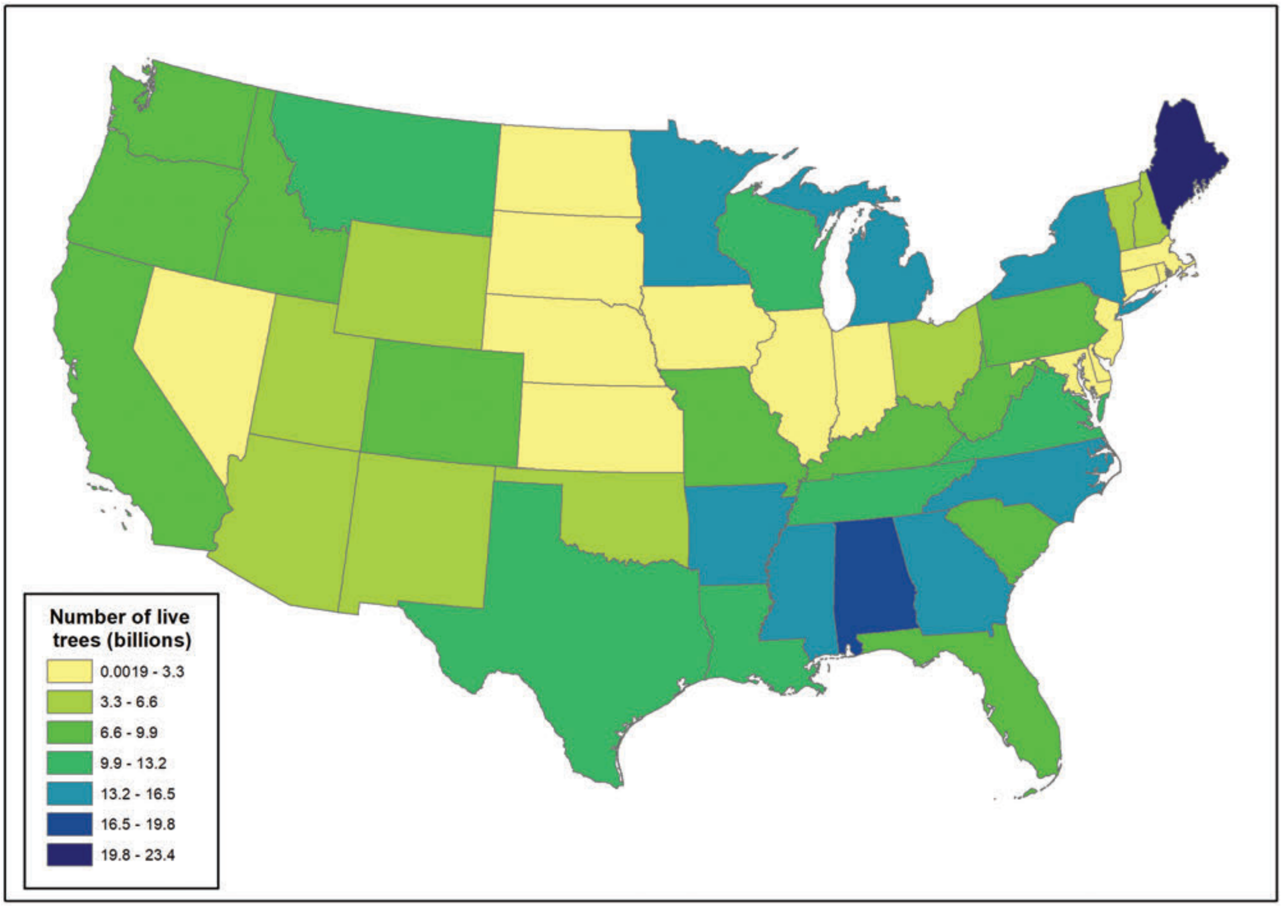

Figure 3 Total number of live trees by state, estimated from TreeMap.

second reason is the temporal asynchrony between the single-year TreeMap snapshot and the multiyear sampling scheme that has also been described above as an inherent characteristic of the continuous, annualized FIA inventory. In Figure 4, note that the states with the highest numbers of dead trees and largest deviations from the diagonal are western states, where the inventory evaluations span longer cycles and mortality rates have generally increased during the past decade.

TreeMap estimates of total number of dead trees are displayed by state in Figure 5 and Table S5. Although the highest number of trees occur in eastern states, dead trees are concentrated in the northwest and Colorado. We display the total carbon in live and dead standing trees by state estimated from TreeMap in Figure 6 and Table S6; figures are highest along the Pacific coast. Examining state-level maps of number of live trees, number of dead trees, and carbon highlights different patterns in these forest attributes.

Although the source of difference in the number of dead trees by state can be explored by detailed, location-specific analysis, given the large variety of situations present among the CONUS states it is beyond the scope of this article. However, it is possible to partly mitigate the differences attributable to the forest definition issue by normalizing estimates using the respective forest areas defined by TreeMap and FIA. When compared on a per-acre basis (Figure 7) the estimates of live trees do not change considerably from the pattern of totals seen in Figure 2. Although the fitted relationship departs from the 1:1 diagonal slightly more, it has slightly lower variation ($R^2 = 0.925$ versus 0.916). However, for dead trees, the comparison improves substantially (Figure 8). The variation in the relationship between TreeMap and FIA again improves slightly ($R^2 = 0.930$ vs 0.917) but more importantly, there is a much greater alignment with the diagonal. For example, in the case of Montana, which had the highest number of dead trees and where the TreeMap estimate was less than 60% of the FIA estimate, the TreeMap estimate is just 5% higher than the FIA estimate on a per-acre basis. Comparable changes also occur for the next four states that are highest-ranked for total number

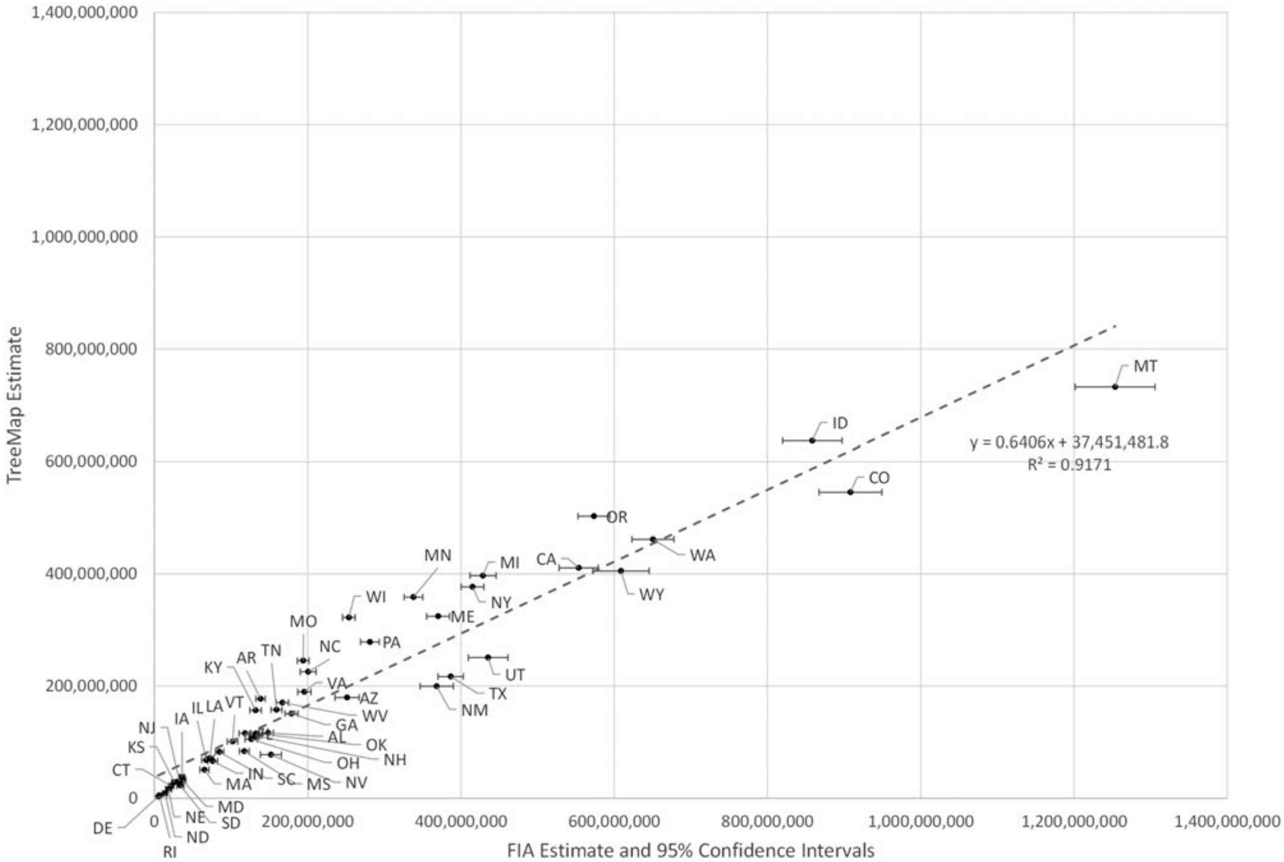

Figure 4 Total number of dead trees by state, comparing TreeMap 2016 estimates to FIA plot-based estimates and 95% confidence intervals.

of dead trees: Colorado, Idaho, Washington, and Wyoming.

To illustrate a different set of metrics and range of scales, we compare aboveground live tree carbon and aboveground dead tree carbon for National Forests across CONUS. The Protected Areas Database (PAD) includes 124 Forest Service units with at least some TreeMap coverage, including smaller units and proclaimed National Forests that are recorded as administratively combined units in the FIA database. Here we report on the eighty-six units that are aligned in PAD and FIA. Because we have already illustrated the differences when comparing population totals versus per acre estimates, for this comparison we only show the per-acre comparisons. Volume-based metrics, such as wood volume, biomass, and carbon, add a new dimension to TreeMap assessment because accuracy is not only a function of the number of trees but also their size.

Results for carbon generally reflect the pattern seen for numbers of trees; that is, relatively close matches between TreeMap and FIA for live-tree carbon (Figure 9) and lesser agreement for dead-tree carbon (Figure 10). As the area estimated grows smaller (e.g., state to National Forest scale), the possibility increases that the estimates from TreeMap could be more locally accurate than the FIA estimates for those areas, especially where conditions are highly variable across the landscape and the FIA plots may not represent the area sufficiently.

Derived Characteristics and Data Linkages: Linking Imputed Dataset to Forest Attributes

So that users can readily generate maps of attributes such as forest type and carbon via the raster attribute table, we performed the linkage of the forest plot identifiers (CN) to a subset of attributes in the FIA DataMart and applied some simple calculations to compute plot-level summaries to associate forest characteristics with the plot CNs found in the TreeMap. A complete list of attributes in the raster attribute table

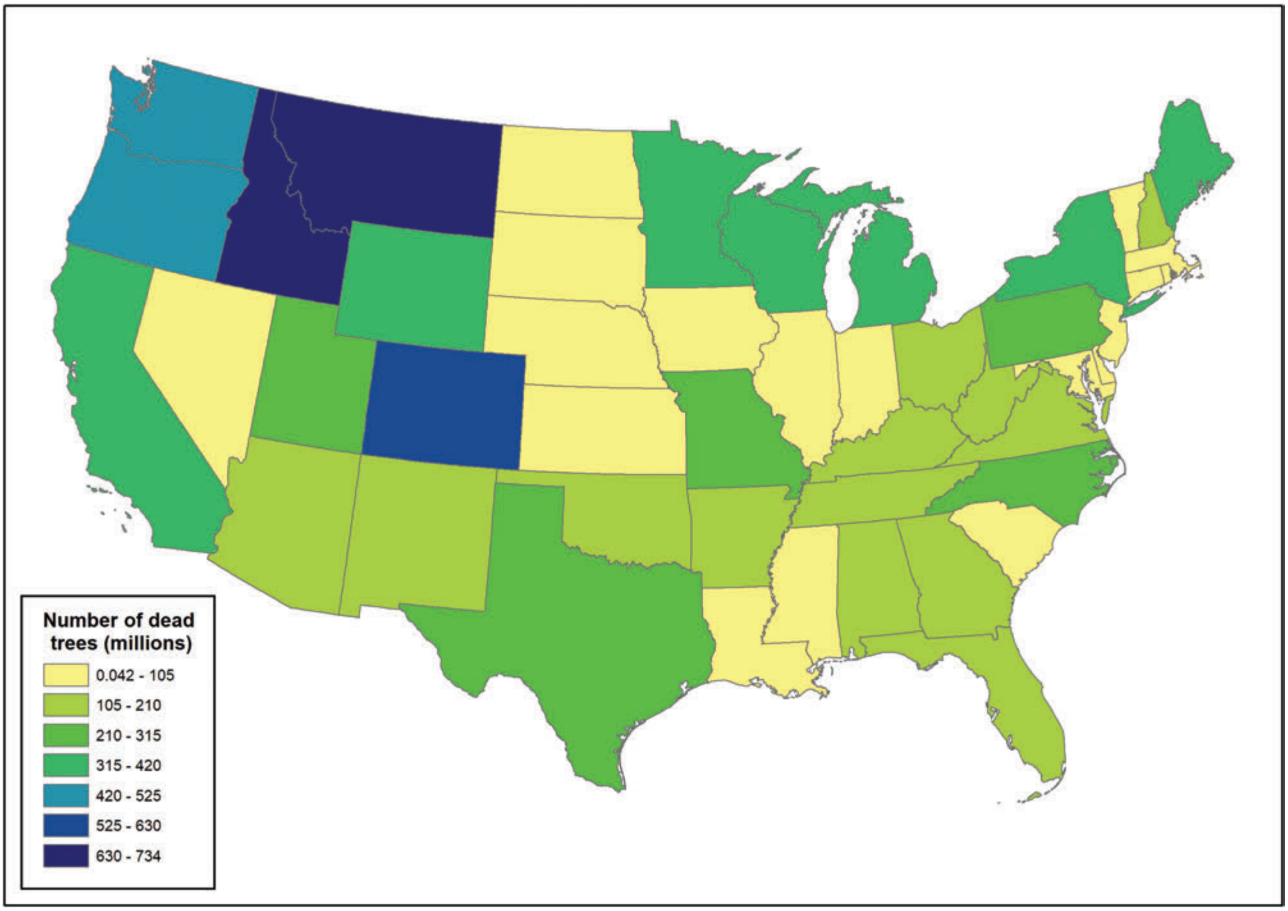

Figure 5 Total number of dead trees by state, estimated from TreeMap.

is included in Table 3. A set of tree-level attributes can also be accessed via the Tree Table included with the dataset, which includes a row for each tree on each of the 55,609 plots included in the dataset. These attributes include diameter, species, height, status (live or dead), and compacted crown ratio for each tree; a complete list of tree-level attributes is found in Table 2. These tree-level records can be used to generate any number of summary metrics including number of live and dead trees per acre or two tree species with highest basal areas, for example, as we have demonstrated in this article.

Because a subset of the variables in the FIA DataMart are provided in the attribute table of the TreeMap 2016 raster, maps of these attributes can now be readily generated by users without linking to the FIA DataMart state tables. We display maps of several of these variables here. Below, we show the unique plot identifiers or CNs for a subset of the Mogollon Rim of Arizona (Figure 11a), as well as the live basal area (Figure 11b), the forest type (Figure 11c), the number of live trees per acre (Figure 11d), the number of dead trees per acre (Figure 11e), and the live-tree carbon (Figure 11f). We note that fifty different forest types appeared in this area according to the TreeMap dataset. On investigation, we found that some of these forest types were not expected to appear in the Mogollon Rim, including balsam fir (1 out of 10,972,740 pixels), knobcone pine (216 pixels), and blue oak (210,302 pixels or 1.9%). All told, pixels from forest types not expected to appear in the area totaled about 2.6% of pixels. Some EVGs appear in broad distributions across several regions, so these types of assignments can occur where the EVG of the pixel included tree species or forest types from out of the area; those assignment were technically correct based on the EVG but can include tree species that are not endemic. These assignments were not so much a misclassification as a consequence of the fact that some EVGs are broad in spatial distribution. Because latitude and longitude are predictor variables, there is a penalty for assigning plots that are from a long distance away; plots are infrequently assigned from more than 300 miles away (Riley, Grenfell, and Finney 2016). Where the random forests model

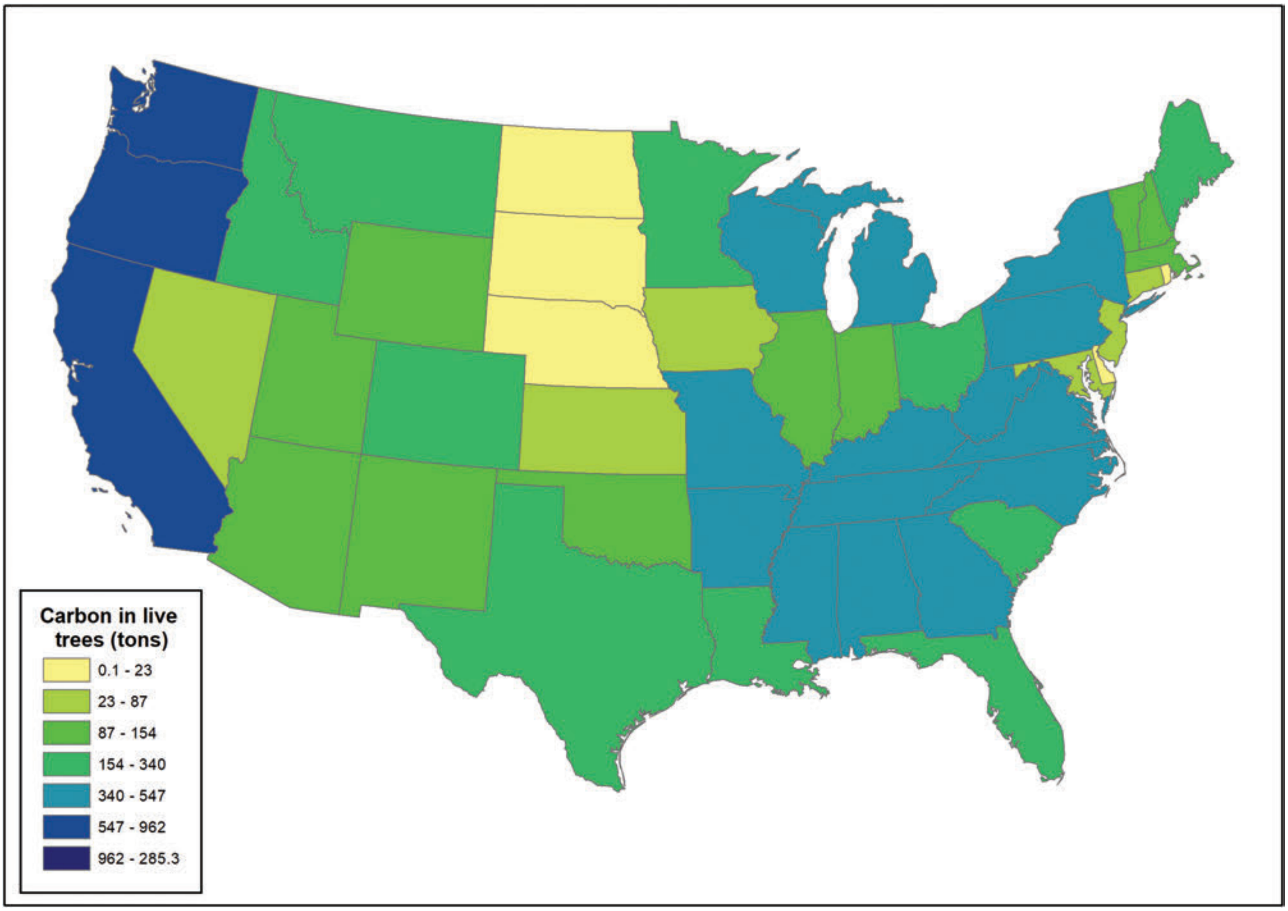

Figure 6 Total carbon in live standing trees by state.

has assigned a plot from out of the area, no good matches were available locally. Because approximately 85%–99% of pixels have forest types expected to appear in local areas, these assignments are not expected to affect many analyses; these assignments should have little effect on analyses concerning basal area and biomass, for example, whereas analyses investigating species types will be affected.

Discussion

When comparing the characteristics of the TreeMap to the LANDFIRE target data on a pixel-by-pixel basis, accuracies for all four response variables were higher than 90% at the national scale, meaning that the random forests model is able to assign plots that closely match these characteristics at each LANDFIRE pixel. When comparing the TreeMap to the locations of a set of FIA plots not included in the model, accuracies for each response variable and two tree species with highest basal areas range from 52% to 87%; when comparing LANDFIRE to FIA plots, results are similar (Riley, Grenfell, Finney, and Wiener 2021). These statistics vary somewhat from zone to zone. Assuming that FIA plots (and the way we have classified their cover, height, and vegetation group) represent the landscape with a high level of accuracy, taken together, this suggests that (1) LANDFIRE has moderate to high accuracy in representing the response variables, (2) TreeMap has high accuracy in representing the response variables as per LANDFIRE's mapping, and therefore, (3) LANDFIRE's moderate to high accuracy thus propagates to TreeMap when compared to conditions on the ground. However, the scale of the forest plots means that a single plot may correspond to anywhere between four and nine 30 × 30m pixels, and we may not be able to accurately identify these pixels due to the national projection of the datasets (which means a pixel may not accurately assign to its exact location on the ground and compass directions are distorted). Some amount of error is thus present in this plot-to-pixel accuracy assessment.

We are in the process of conducting additional accuracy assessments for variables not reported in this

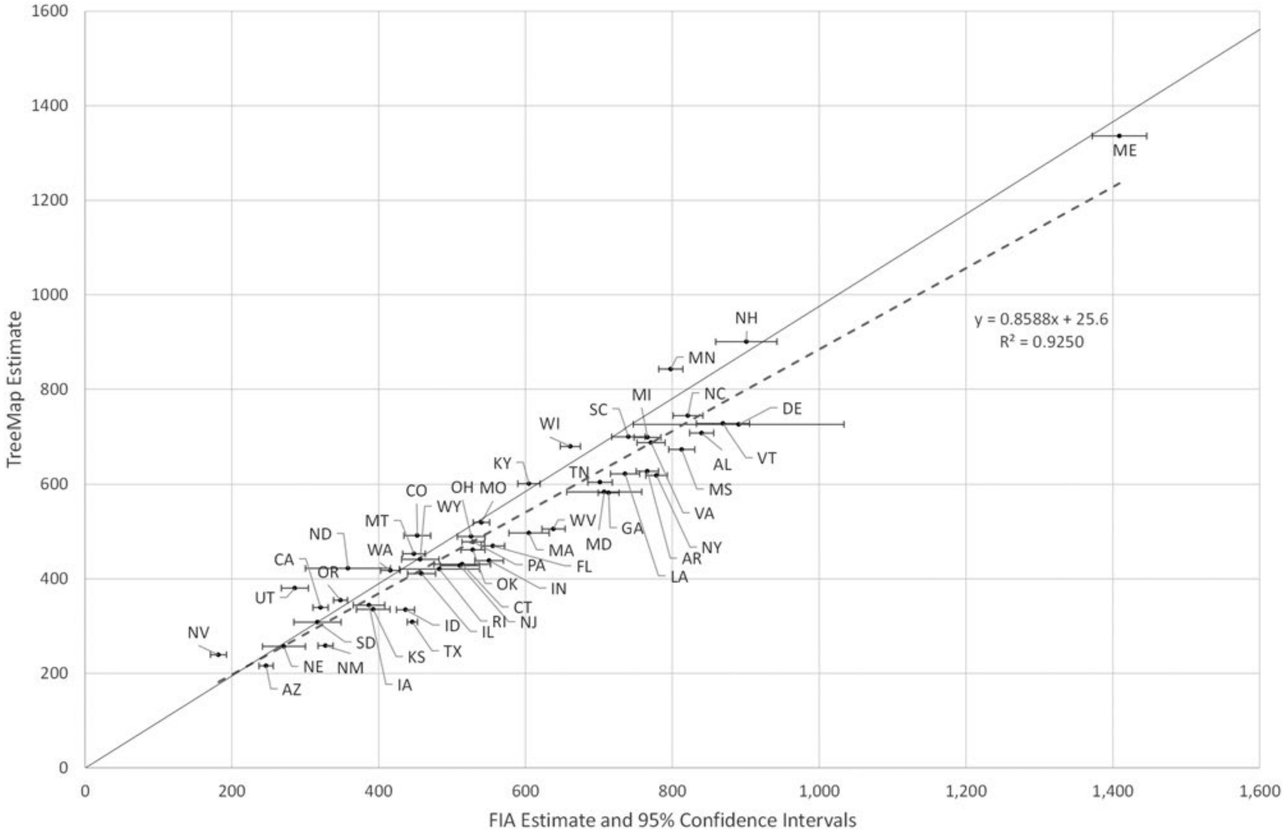

Figure 7 Number of live trees per acre by state, comparing TreeMap 2016 estimates to FIA plot-based estimates and 95% confidence intervals.

publication and at varying scales and multiple locations to aid users with proper application of the dataset in land management problems. Using multicondition plots as validation plots means that the validation plots may be somewhat different in character from the single-condition plots used in the imputation, but as all validation plots are 100% forested and 100% accessible to ground measurements, they serve as a useful benchmark.

The accuracy of matching disturbed plots to disturbed pixels greatly increased from 90.3% in TreeMap 2014 to 99.98% in TreeMap 2016 because disturbance was now included as a response variable. However, the accuracy of mapping disturbed areas in the target data declined slightly by about 3%, likely due to additional areas being affected by insect and disease infestation and these not being mapped in the LANDFIRE disturbance rasters. In future versions of the TreeMap, we hope to include new mapping of insect- and disease-affected areas to increase our accuracy in capturing these. We are investigating the Landscape Change Monitoring System, which uses an ensemble of forest disturbance maps to identify areas of change (Healey et al. 2016).

We found that assignments from forest types not expected to occur in an area (as with the example given above for blue oak in the Mogollon Rim) appear in an estimated 1%–15% of pixels, depending on the region. These assignments will have minimal impact on most analyses because they appear in a small percentage of pixels. However, these assignments can be disconcerting for users, so in future versions of TreeMap we plan to make some changes to methods that will eliminate these assignments. In the interim, users might consider substituting similar species, perhaps from the same genus, for affected trees that are native to the area. We note that because the TreeMap is designed for predicting carbon, the current methodology is likely superior for that end. In making changes to the methodology in future versions to ensure only species endemic to an area are imputed, the random forests model will need to choose a plot that is not as good a match for the cover and height, degrading carbon predictions somewhat. However, because the change is expected to affect only 1%–15% of pixels, carbon calculations shouldn't suffer very much, whereas vegetation group and species mix predictions will improve, an improvement important to many applications.

Whereas other imputed datasets assign summary attributes, TreeMap is distinct in that it provides a

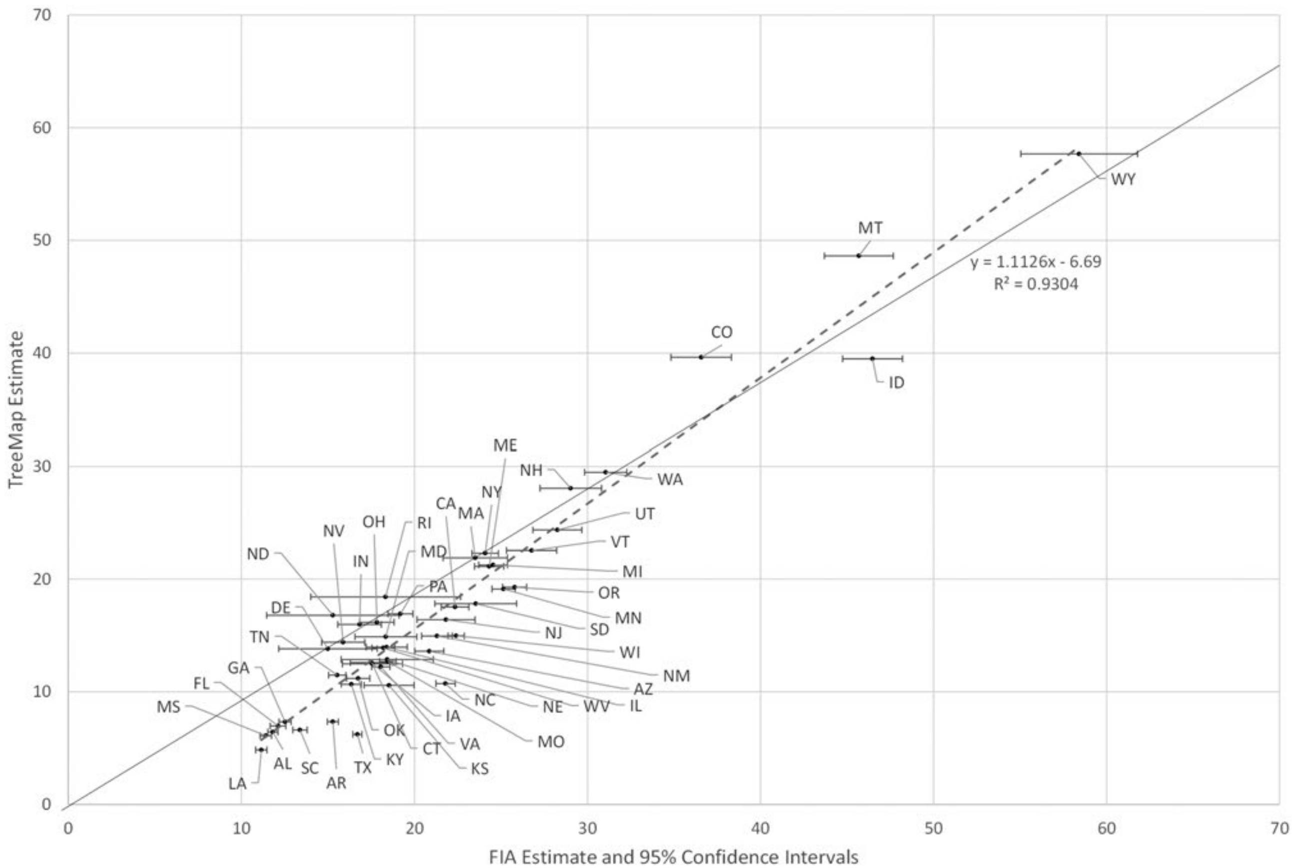

Figure 8 Number of dead trees per acre by state, comparing 2016 TreeMap estimates to FIA plot-based estimates and 95% confidence intervals.

linkage via the plot identifier to hundreds of attributes measured on FIA forest plots, stored in multiple tables containing data from plot locations, stands, and individual trees. Although the plot identifier was an intermediate product of Wilson et al. (2012), it is not included in the published tree species dataset and therefore users cannot explore the data beyond the published metrics. Other products, including Wilson et al. (2012), impute a single characteristic (e.g., live basal area) often via analyses of time series of Landsat or MODIS satellite data or produce separate imputations of single characteristics via the mean or mode of the *k* nearest plots in the feature space (e.g., live basal area in one product and tree species in another) (Ohmann and Gregory 2002; Pierce Jr et al. 2009; Wilson, Lister, and Riemann 2012; Wilson, Woodall, and Griffith 2013; Wilson, Knight, and McRoberts 2018). One possible limitation of this method by which individual attributes are imputed is that coherence between different attributes is not guaranteed and could lead to combinations at any given pixel that are unlikely or could not occur in nature. Investigation of the other strengths and weaknesses of these two methodologies in comparison to each other is warranted and is a next step in this research.

The Tree Table and raster attribute table included in the TreeMap 2016 product comprise only a small portion of FIADB. The attributes selected for inclusion represent a relatively conservative set of commonly used metrics. These attributes include all of those shown in this article and other attributes that are closely related (e.g., carbon is derived from biomass and various other volumetric values, forest type classifications). Although we have done nominal evaluation of these attributes, similar to what has been shown for the attributes selected for illustration in this article, accuracy is ultimately a location- and application-specific characteristic. This sort of in-depth analysis is beyond the scope of this article, but guidelines for accuracy and uncertainty assessment for project-scale applications are in development.

Users who desire to take advantage of FIA variables beyond what has been provided with the

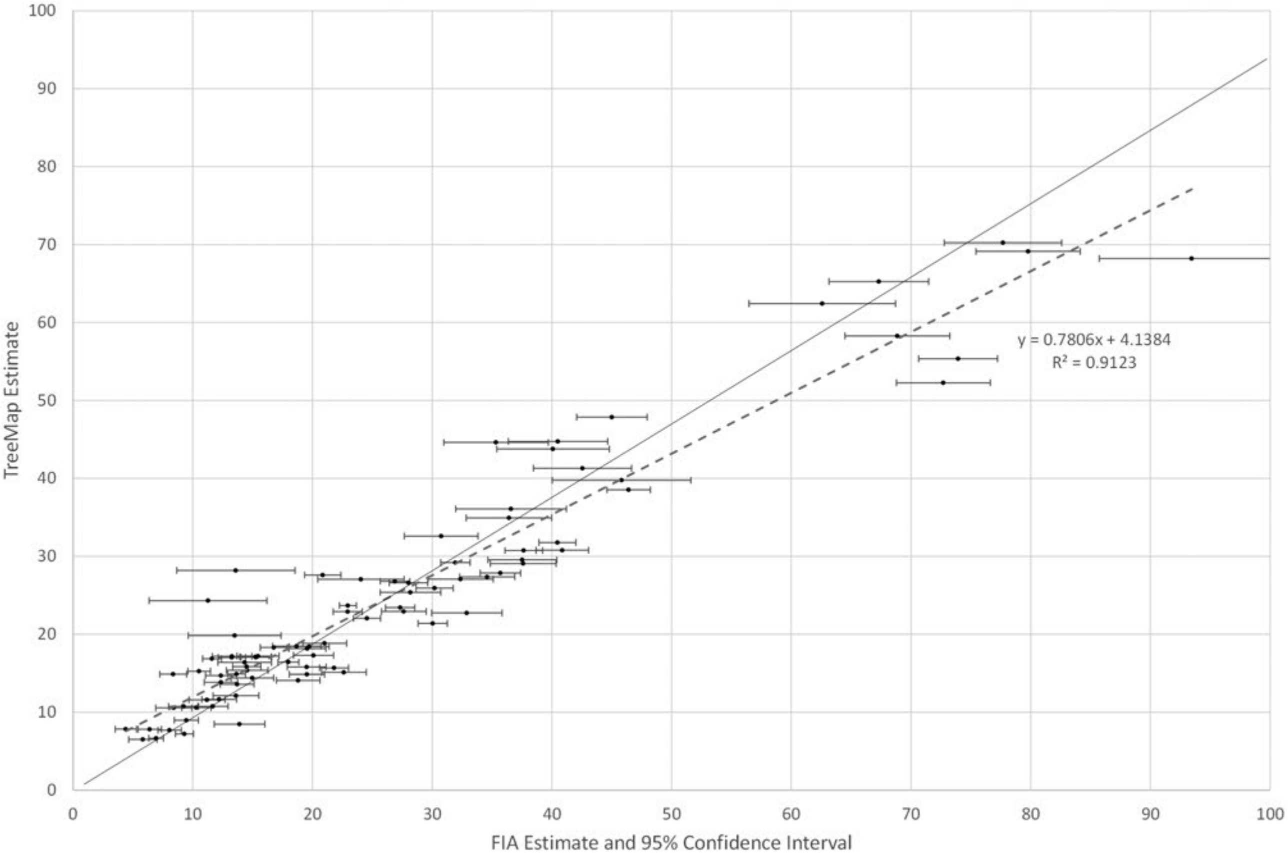

Figure 9 Mean aboveground live tree carbon in tons per acre for 86 National Forest units.

TreeMap 2016 should follow a common sense approach and consider additional variables cautiously. Although the linkage between the TreeMap and FIADB allows users to access the entirety of FIADB, the nature of the imputation process is such that not all available variables are suitable for use in their imputed locations. For example, though the modeling process considers elevation as a predictor variable, it is unlikely that TreeMap will accurately reproduce topography using elevations from imputed plots. Similarly, soil is not included in the modeling process, so in areas where site quality is strongly influenced by soil characteristics, FIADB variables such as site index or habitat type may not be accurate for all imputed locations. These considerations are especially important when using imputed plot locations in FVS, because FVS growth equations are sensitive to variables such as elevation and site index. Because of the large number of variables present in FIADB and the wide variety of potential user applications, it is impractical to evaluate the suitability for every variable in every location. Therefore, users should perform their own validation process when considering the use of variables included with TreeMap in novel applications or when drawing additional variables from FIADB.

The current version does not include an assessment of uncertainty, which may be helpful in assisting users with appropriate application of the dataset for questions including those regarding carbon estimation. Uncertainty estimation for random forests differs from that of traditional frequentist models and sample-based estimates such as those conducted by FIA, a topic we will explore during imputation of the next TreeMap version. To cover the issue of error in the model, we have offered three different assessments of the dataset and calculated model error. Model error, as expressed by the out-of-bag error rate, ranged from 0 to 0.0137 depending on the response variable and zone; uncertainty from the random forests imputation itself is likely to be low, therefore, with larger sources being (1) unavailability of forest plot data suitable for some pixels and (2) inaccuracies in mapping of landscape data by LANDFIRE. Accuracy is expected to be higher for the variables that served as both predictor and response (this includes the EVC, EVH, EVG, and

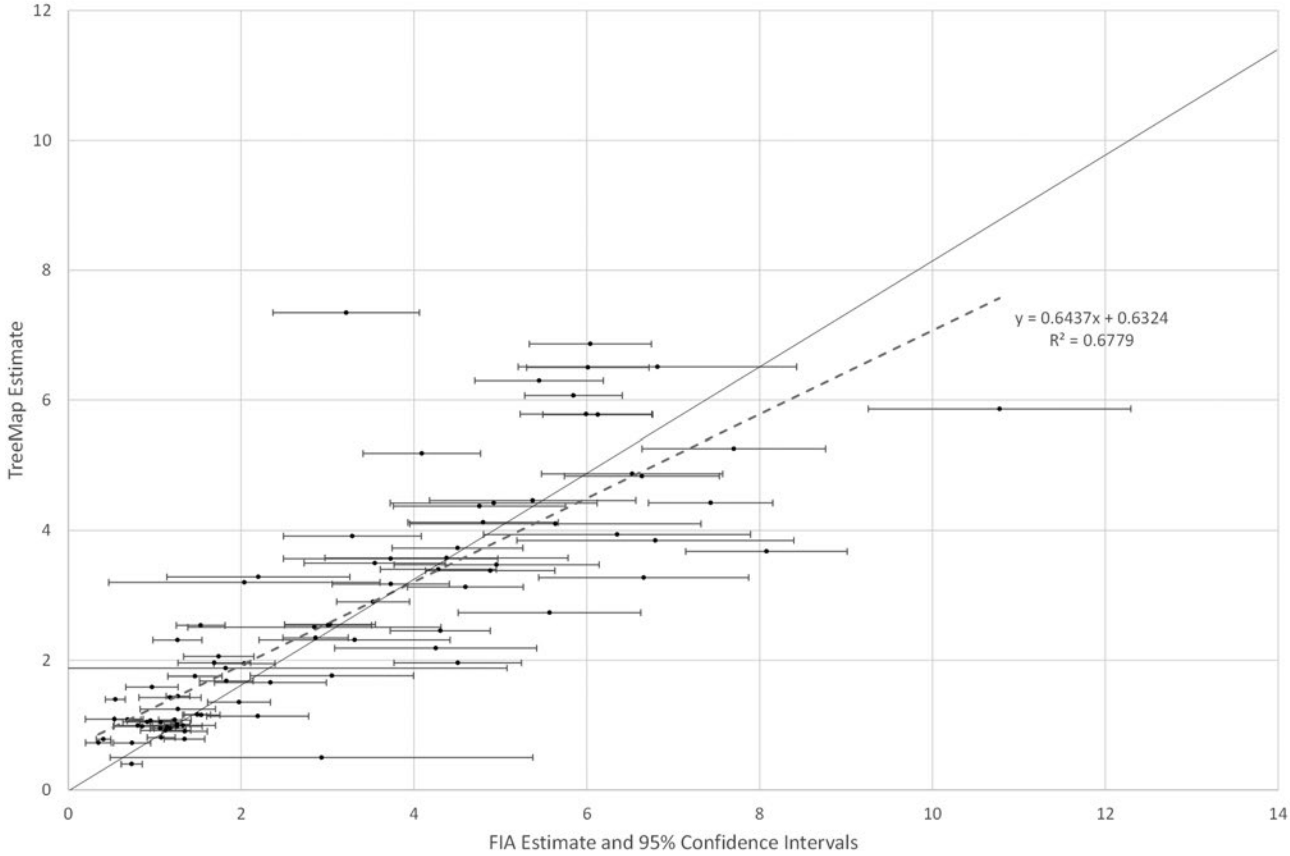

Figure 10 Mean aboveground dead tree carbon in tons per acre for 86 National Forest units.

disturbance code) than for other predictor variables (this includes location, topography, and biophysical variables). Thus, accuracy may be higher for EVC, EVH, EVG, and disturbance and for attributes that derive from these (e.g., number of live and dead trees, biomass, and carbon) than for many other attributes.

We showed maps of metrics such as forest type and live basal area to demonstrate the diverse capabilities of this dataset. Values for carbon and number of trees such as those included here show high-to-moderate correlations with those obtained using FIA's methods at the state level. For areas on the scale of 100,000 acres or smaller, FIA's sampling intensity (~1 plot per 6,000 acres) is too sparse to reliably make these types of estimates, especially where conditions are heterogenous, and as a seamless national-scale tree-level dataset, the TreeMap presents a method by which estimates can readily be drawn, providing potentially valuable estimates in areas where the accuracy of the modeled variables is high.

Conclusions

The TreeMap 2016 improved accuracy of imputing disturbed plots to disturbed pixels compared to TreeMap 2014 from 90.3% to 99.98% while retaining accuracy greater than 94% for forest cover, height, and vegetation group. The dataset was also validated at the site of approximately 2,749 forest plots not used in the model: accuracy ranged from 51.8% to 87.4% at these locations, depending on the variable assessed. The TreeMap 2016 was summarized at the state level and compared favorably with estimates from FIA (R^2 = 91.6–93.0%). A new feature of the TreeMap 2016 is that it provides plot-level attributes to facilitate mapping and geographical summaries of forest attributes for users, meaning that attributes such as aboveground live and dead carbon can be displayed on-the-fly in a GIS. Because validation indicated moderate to high accuracy for forest cover, height, vegetation group, tree species, and disturbance at both the pixel and continental scales, the dataset is likely suitable for use in many applications ranging from fuel treatment planning to estimation of terrestrial carbon resources.

Supplementary Materials

Supplementary data are available at *Journal of Forestry* online.

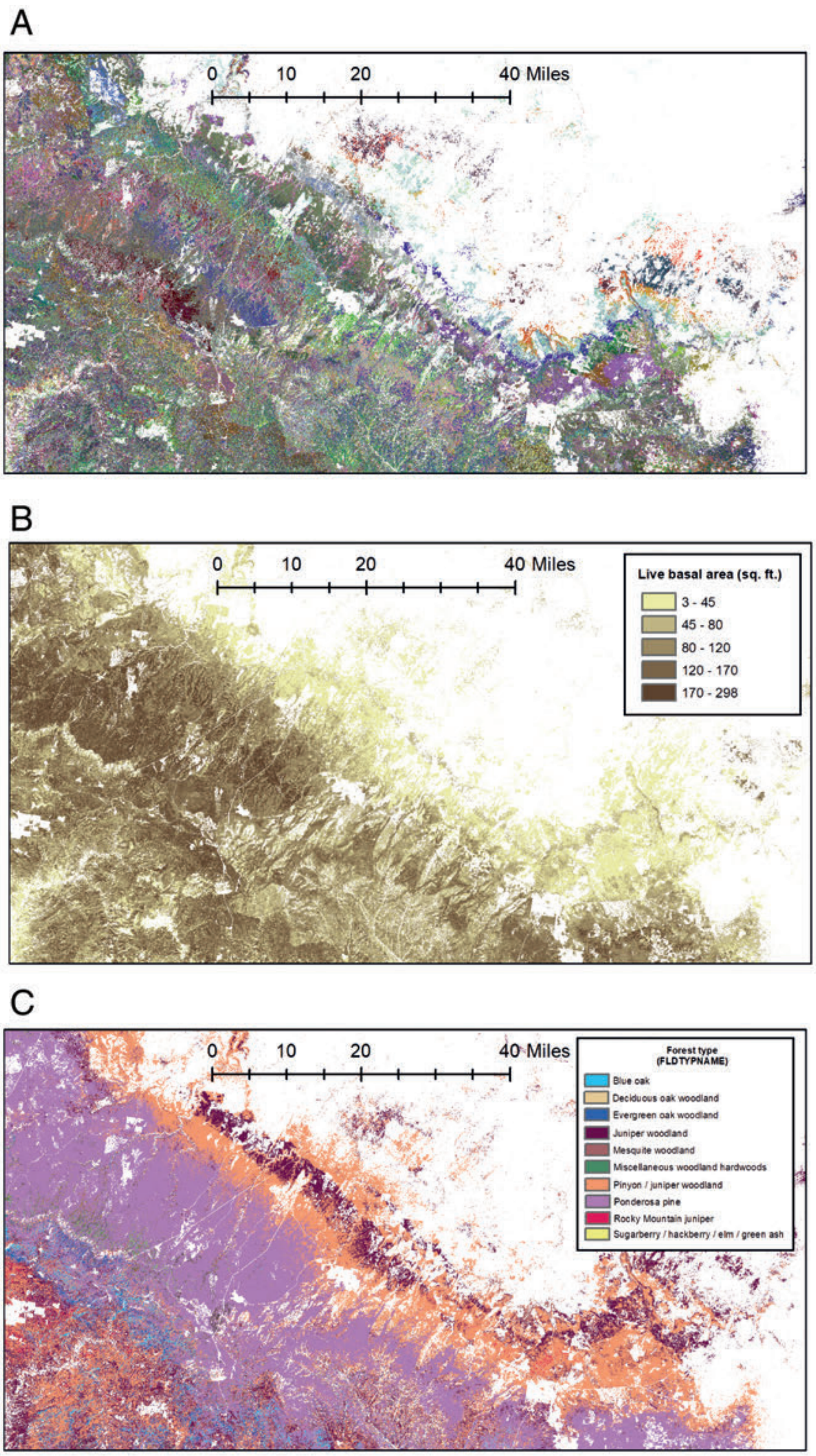

Figure 11 A) Plot identifiers (CNs) for a subset of the Mogollon Rim of Arizona. Each unique color corresponds to a different plot. B) Live basal area in square feet for the same subset of the Mogollon Rim of Arizona. C) Forest type for the same subset of the Mogollon Rim of Arizona, using the forest type assigned by FIA field crews or FLDTYPNAME. Fifty different forest types appeared in this portion of the Mogollon Rim; only those with 0.05% or greater share of the landscape are shown in the legend for simplicity. D) Number of live trees per acre for the same subset of the Mogollon Rim. E) Number of dead trees per acre for the same subset of the Mogollon Rim. F) Live tree carbon for the same subset of the Mogollon Rim.

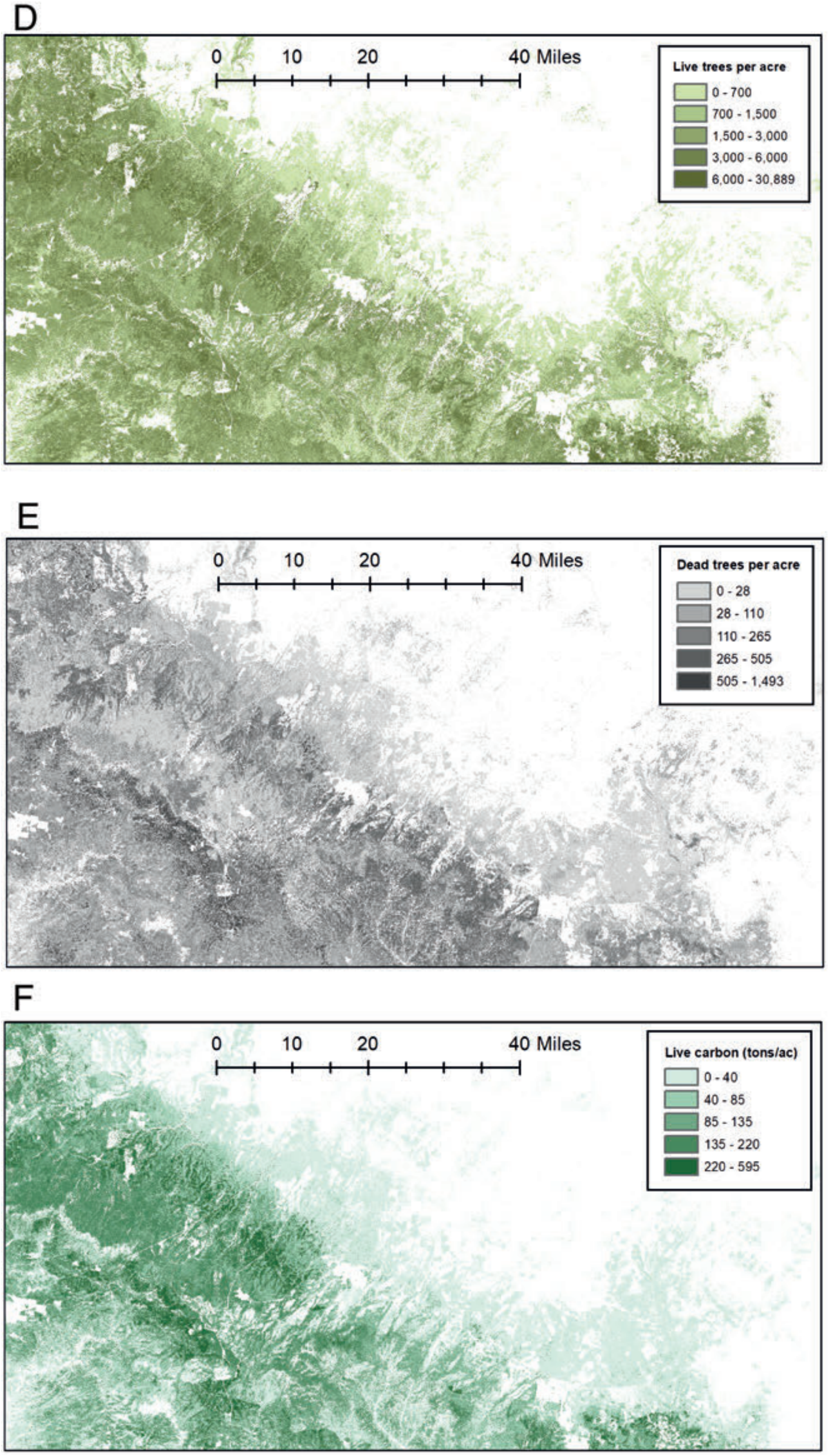

Figure 11 Continued.

Acknowledgements

We appreciate the assistance of Justin Holgerson at FIA in providing plot spatial locations for single-condition, 100% forested plots. We thank the two reviewers for insightful comments that greatly strengthened the article. Several of these review comments were incorporated into the article. This research was supported by the USDA Forest Service Fire Sciences Lab.

Literature Cited

Arner, Stanford L., Sharon Woudenberg, Shirley Waters, John Vissage, Colin MacLean, Mike Thompson, and Mark Hansen. 2001. *National Algorithms for Determining Stocking Class, Stand Size Class, and Forest Type for Forest and Inventory Analysis Plots*. USDA Forest Service Internal Rep., Northeastern Research Station, Newtown Square, PA.

Bechtold, William A, and Paul L Patterson. 2005. *The Enhanced Forest Inventory and Analysis Program — National Sampling Design and Estimation Procedures*. USDA Forest Service Gen. Tech. Rep. GTR-SRS-80, Southern Research Station, Fort Collins, CO, 85 p.

Burrill, Elizabeth A., Andrea M. Wilson, Jeffrey A. Turner, Scott A. Pugh, James Menlove, Glenn Christiansen, Barbara L. Conkling, and Winne David. 2018. "The Forest Inventory and Analysis Database: Database Description and User Guide Version 8.0 for Phase 2." [Online]. http://www.fia.fs.fed.us/library/database-documentation/.

Crookston, Nicholas L, and Andrew O Finley. 2008. "YaImpute: An R Package for KNN Imputation." *J. Stat. Softw.* 23(10):1–16.

Crookston, Nicholas L, and Albert R Stage. 1999. "Percent Canopy Cover and Stand Structure Statistics from the Forest Vegetation Simulator." *General Technical Report RMRS-GTR-24*. Ogden, UT.

Forest Inventory Analysis. 2022a. "Forest Inventory Analysis DataMart." Forest Inventory Analysis DataMart FIADB_1.9.0. 2022. https://apps.fs.usda.gov/fia/datamart/.

———. 2022b. "Forest Inventory and Analysis." 2022. https://www.fia.fs.fed.us/.

Healey, Sean P., Warren B. Cohen, Yang Zhiqiang, Ken Brewer, Evan Brooks, Noel Gorelick, Matthew Gregory, et al. 2016. "Next-Generation Forest Change Mapping across the United States: The Landscape Change Monitoring System (LCMS)." In *Pushing Boundaries: New Directions in Inventory Techniques and Applications: Forest Inventory and Analysis (FIA) Symposium 2015.*, edited by Sharon M. Stanton and Glenn A. Christensen. 2015 December 8-10; Portland, Oregon: General Technical Report PNW-GTR-931. USDA Forest Service, Pacific Northwest Research Station.

LANDFIRE. 2020a. "Download Mosaic Data Products." 2020. www.landfire.gov/version_download.php.

———. 2020b. "LANDFIRE Biophysical Gradient Raster Datasets." 2020. https://springernature.figshare.com/collections/LANDFIRE_Biophysical_Gradient_Raster_Datasets/5142572.

Moisen, Gretchen, J.A. Blackard, and M.V. Finco. 2004. "Small Area Estimation in Forests Affected by Wildfire in the Interior West." In *Remote Sensing for Field Users; Proceedings of the Tenth Forest Service Remote Sensing Applications Conference*, edited by Jerry Dean Greer. 2004 April 5-9. Salt Lake City, UT: American Society of Photogrammetry and Remote Sensing.

Ohmann, Janet L and Matthew J Gregory. 2002. "Predictive Mapping of Forest Composition and Structure with Direct Gradient Analysis and Nearest- Neighbor Imputation in Coastal Oregon, USA." *Can. J. For. Res.* 32:725–741. doi: 10.1139/X02-011.

Pierce, Kenneth B Jr, Janet L Ohmann, Michael C Wimberly, Matthew J Gregory, and Jeremy S Fried. 2009. "Mapping Wildland Fuels and Forest Structure for Land Management: A Comparison of Nearest Neighbor Imputation and Other Methods." *Can. J. For. Res.* 39: 1901–1916. doi:10.1139/X09-102.

Riley, Karin L., Christopher D O'Connor, Christopher J. Dunn, Jessica R. Haas, Richard D. Stratton, and Benjamin Gannon. 2022. "A National Map of Snag Hazard to Reduce Risk to Wildland Fire Responders." *Forests*. 13(8):1160. doi: 10.3390/f13081160.

Riley, Karin L., Isaac C. Grenfell, and Mark A. Finney. 2016. "Mapping Forest Vegetation for the Western United States Using Modified Random Forests Imputation of FIA Forest Plots." *Ecosphere*. 7(10):e01472. doi: 10.1002/ecs2.1472.

Riley, Karin L., Isaac C. Grenfell, Mark A. Finney, and John D. Shaw. 2021. "TreeMap 2016: A Tree-Level Model of the Forests of the Conterminous United States circa 2016." 2021. doi:10.2737/RDS-2021-0074.

Riley, Karin L., Isaac C. Grenfell, Mark A. Finney, and Rachel M. Wiener, Jason M. Houtman. 2019. "Fire Lab Tree List: A Tree-Level Model of the Conterminous United States Landscape circa 2014." 2019. https://www.fs.usda.gov/rds/archive/catalog/RDS-2019-0026.

Riley, Karin L., Isaac C. Grenfell, Mark A. Finney, and Jason M. Wiener. 2018. "Fire Lab Tree List: A Tree-Level Model of the Western US circa 2009 V1." 2018. doi:10.2737/RDS-2018-0003.

———. 2021. "TreeMap, a Tree-Level Model of Conterminous US Forests circa 2014 Produced by Imputation of FIA Plot Data." *Nature Scientific Data* 8(11):1–14. doi: 10.1038/s41597-020-00782-x.

Risk Management Assistance. 2021. "Risk Management Assistance Dashboard." 2021. https://nifc.maps.arcgis.com/apps/MapSeries/index.html?appid=c5bc811ee22e4da0bde8abec7c20b8b4.

Rollins, Matthew G. 2009. "LANDFIRE: A Nationally Consistent Vegetation, Wildland Fire, and Fuel Assessment." *Int. J. Wildland Fire* 18:235–249. doi: 10.1071/WF08088.

Shaw, John D, and Aaron Gagnon. 2020. "Field Note: A New Conversion of Forest Inventory and Analysis Data for Use in the Forest Vegetation Simulator." *Journal of Forestry*. doi: 10.1093/JOFORE/FVZ050.

USDA Forest Service. 2022. "EVALIDator." 2022. https://www.fia.fs.fed.us/tools-data/.

U.S. Geological Survey (USGS) Gap Analysis Project. 2020. "Protected Areas Database of the United States (PAD-US) 2.1: U.S. Geological Survey Data Release." 2020. doi:10.5066/P92QM3NT.

Wilson, B Tyler, Andrew J Lister, and Rachel I Riemann. 2012. "A Nearest-Neighbor Imputation Approach to Mapping Tree Species over Large Areas Using Forest Inventory Plots and Moderate Resolution Raster Data." *Forest Ecol. Manag.* 271:182–198. doi: 10.1016/j.foreco.2012.02.002.

Wilson, Barry T, Joseph F Knight, and Ronald E McRoberts. 2018. "Harmonic Regression of Landsat Time Series for Modeling Attributes from National Forest Inventory Data." *ISPRS J. Photogramm. Remote Sens.* 137:29–46. doi: 10.1016/j.isprsjprs.2018.01.006.

Wilson, Barry Tyler, Christopher W Woodall, and Douglas M Griffith. 2013. "Imputing Forest Carbon Stock Estimates from Inventory Plots to a Nationally Continuous Coverage." *Carbon Balance Manage.* 8(1):1–15. doi: 10.1186/1750-0680-8-1.