

Not-so-random forests: Comparing voting and decision tree ensembles for characterizing partial harvest events

Valerie J. Pasquarella^{a,b,*}, Luca L. Morreale^{a,b}, Christopher F. Brown^c, John B. Kilbride^d,
Jonathan R. Thompson^b

^a Department of Earth & Environment, Boston University, 685 Commonwealth Avenue, Boston, MA 02215, USA

^b Harvard Forest, Harvard University, 324 N Main St, Petersham, MA 01366, USA

^c Google LLC, 1600 Amphitheatre Pkwy., Mountain View, CA 94043, USA

^d College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR 97331, USA

ARTICLE INFO

Keywords:

Change detection

Forest harvest

Temporal segmentation

LandTrendr

Forest Inventory and Analysis

Ensemble methods

ABSTRACT

Ensemble-based change detection can improve map accuracies by combining information from multiple datasets. There is a growing literature investigating ensemble inputs and applications for forest disturbance detection and mapping. However, few studies have evaluated ensemble methods other than Random Forest classifiers, which rely on uninterpretable “black box” algorithms with hundreds of parameters. Additionally, most ensemble-based disturbance maps do not utilize independently and systematically collected field-based forest inventory measurements. Here, we compared three approaches for combining change detection results generated from multi-spectral Landsat time series with forest inventory measurements to map forest harvest events at an annual time step. We found that seven-parameter degenerate decision tree ensembles performed at least as well as 500-tree Random Forest ensembles trained and tested on the same LandTrendr segmentation results and both supervised decision tree methods consistently outperformed the top-performing voting approach (majority). Comparisons with an existing national forest disturbance dataset indicated notable improvements in accuracy that demonstrate the value of developing locally calibrated, process-specific disturbance datasets like the harvest event maps developed in this study. Furthermore, by using multi-date forest inventory measurements, we are able to establish a lower bound of 30% basal area removal on detectable harvests, providing biophysical context for our harvest event maps. Our results suggest that simple interpretable decision trees applied to multi-spectral temporal segmentation outputs can be as effective as more complex machine learning approaches for characterizing forest harvest events ranging from partial clearing to clear cuts, with important implications for locally accurate mapping of forest harvests and other types of disturbances.

1. Introduction

Whether for resource management, climate mitigation, or ecosystem management, policy makers and forest managers require accurate information describing patterns and rates of forest disturbances (Pickett and White, 2013). In most mesic temperate forests, including the forests that dominate the northeastern United States, timber harvesting is the dominant disturbance (Canham et al., 2013; Masek et al., 2011). Harvest regimes are strongly influenced by biophysical factors, including forest type and productivity (Canham et al., 2013) in combination with socioeconomic factors, such as markets, policies, ownership, and population density (Kittredge et al., 2017; Thompson et al., 2017). Thus, mapping

the spatial and temporal variation in forest harvest patterns and rates at local to regional scales is essential for understanding the impacts of changing harvest regimes on forest structure, composition, and productivity.

Open access to satellite imagery from the Landsat and Copernicus programs has led to significant advances in automated change detection approaches (Hansen and Loveland, 2012; Hemati et al., 2021; Kennedy et al., 2014; Woodcock et al., 2020; Zhu, 2017). Forest disturbance detection has been at the forefront of many of these advances (Banskota et al., 2014; Wulder et al., 2012) and remains the most common application of Landsat time series analysis (Hemati et al., 2021). However, forest disturbance maps can vary significantly in their ability to

* Corresponding author.

E-mail address: valpasq@bu.edu (V.J. Pasquarella).

<https://doi.org/10.1016/j.jag.2023.103561>

Received 30 June 2023; Received in revised form 9 November 2023; Accepted 10 November 2023

Available online 18 November 2023

1569-8432/© 2023 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

characterize different disturbance processes (Cohen et al., 2017; Hansen et al., 2013) and best practices for detecting and mapping complex non-stand-replacing disturbances like partial harvests, thinning, and degradation continue to evolve (e.g., Chen et al., 2021; Coops et al., 2020; Koltunov et al., 2020; Thomas et al., 2021; Tortini et al., 2019; Ye et al., 2021).

With greater access to imagery, algorithms, and computing resources, there has been an increased emphasis on using ensemble approaches to characterize forest disturbances including forest harvest events (e.g., Cohen et al., 2018; Healey et al., 2018). Ensemble approaches combine the outputs of different classifiers to improve the performance of a single classifier (Polikar, 2006). These approaches can vary in complexity from simple rules like voting strategies (e.g., Friedl and Brodley, 1997) to secondary classification, also known as “stacking” or “stacked generalization,” using machine learning methods (Healey et al., 2018; Wolpert, 1992).

Random Forest (RF) approaches are widely used for a variety of remote sensing applications (Belgiu and Drăguț, 2016) and have become a standard choice for forest disturbance mapping ensembles (e.g., Cohen et al., 2020, 2018; De Marzo et al., 2021; Healey et al., 2018; Schultz et al., 2016; Wang et al., 2019). While RF algorithms have proven effective, their complexity can render them an uninterpretable “black box” that can be difficult to scale across datasets and processing platforms. Furthermore, most studies default to RF classification with 500 trees as recommended by Belgiu and Drăguț (2016), focusing primarily on refinements to feature inputs and training datasets for improving model performance.

Few studies have evaluated more parsimonious methods for generating disturbance ensembles and their findings have been inconclusive. For example, in comparing RF with voting approaches, Hislop et al., (2019) found that lower error rates could be achieved through refinements to RF training data and feature inputs, however, majority voting strategies achieved comparable performance in some cases. Healey et al. (2018) considered simple logistic regression as an alternative to RF and found that a regression-based multi-algorithm change detection ensemble outperformed individual algorithms, though RF consistently achieved lower balanced omission and commission. These results suggest that machine learning approaches like RF may not necessarily be the best choice for building a change detection ensemble, and rigorous comparisons of simpler alternatives can yield important insights for applied use cases.

In this study, we evaluate the choice of change detection ensemble methods for mapping forest harvest events and address the tradeoffs between model complexity and accuracy. We specifically compared three different approaches for generating maps of potential harvest events: (1) voting strategies, (2) a standard 500-tree RF classifier, and (3) a degenerate decision tree (DDT) ensemble. These approaches were selected to represent tradeoffs between expressiveness (i.e., ability to effectively characterize complex phenomena), scalability, and interpretability. We apply these approaches to multi-spectral Landsat temporal segmentation results produced using the LandTrendr algorithm (Kennedy et al., 2018, 2010), a well-established temporal segmentation approach (Kennedy et al., 2010; Pasquarella et al., 2022). We used field-based forest inventory plot measurements to train supervised decision tree models, cross-validated the relative performance of different ensemble approaches, and estimated a lower bound on the level of partial harvesting (in terms of both absolute and percent basal area removal) that can be most accurately detected. We also compared ensemble results with an external reference dataset collected using the TimeSync photo-interpretation protocol (Cohen et al., 2010) and an existing disturbance detection ensemble product being generated as part of the Landscape Change Monitoring System (LCMS) project (Housman et al., 2021). Although our use case focuses on improved mapping of timber harvests in the industrial woodlands of the northeastern United States, we expect our findings will be more broadly relevant to future work on ensemble-based change detection, and the methods considered

here can be adapted to other landscapes, change processes, satellites, and temporal segmentation methods.

2. Study area

Our study area is the state of Maine in the northeastern United States (Fig. 1). Maine is the largest state in New England at 91,600 km² (comparable in size to the country of Portugal) and is the most forested state in the country by proportion of land area (89%). The forests of Maine span an ecological transition from spruce- and fir-dominated boreal forests in the north and west to northern hardwood forests dominated by beech, maple, birch, in the south and east (Duveneck et al., 2015). Forest composition is largely dictated by previous land use and the regional climate gradient, which transitions from cold and snowy in the north (min/max average temperature: Jan −17/−6, July 13/25 °C with 90 cm/yr precipitation) to comparatively mild in the south (min/max average temperature: Jan −10/0, July 15/26 °C with 140 cm/yr precipitation).

Maine is unique among forested regions in that it is largely owned by private entities, including corporations (59%) and family ownerships (32%) (Butler, 2017; Oswalt et al., 2019). Prior to the 1990 s, Maine's private corporate timberlands were largely owned by vertically integrated firms and harvest regimes were dominated by clearcutting (Sader et al., 2003). Ownerships have since transitioned to investor-driven financial owners, who often have shorter term investor horizons and altered harvest regimes (Chudy and Cubbage, 2020). These ownership transitions combined with legacies of widespread spruce-budworm outbreaks (Fraver et al., 2007), a policy-mandated shift to partial harvesting (Belair and Ducey, 2018; Canham et al., 2013), and a natural regime characterized by small-scale gap dynamics (Lorimer, 1977; Seymour et al., 2002) has resulted in fragmented and frequently disturbed forests. Thus, understanding impacts of changing ownership and policy on industrial forest landscapes requires accurate harvest mapping on decadal time scales. There is a long history of Landsat-based harvest mapping for the state of Maine (Jin and Sader, 2006, 2005; Sader et al., 2003; Sader and Legaard, 2008; Wilson and Sader, 2002). However, the majority of this work was conducted prior to widespread availability of imagery and cloud-based computing resources, resulting in new opportunities to use maturing time series analysis approaches for detection of harvest events at an annual time step across the entire state.

3. Methods

3.1. LandTrendr temporal segmentation

We used the LandTrendr temporal segmentation approach (Kennedy et al., 2018, 2010) to generate inputs for our harvest event detection ensembles. LandTrendr is a time series analysis algorithm that characterizes the per-pixel spectral trajectories using piece-wise linear models (Kennedy et al., 2010; Pasquarella et al., 2022). The algorithm has been implemented natively in Google Earth Engine (GEE) (Kennedy et al., 2018), making it a common choice for cloud-based change detection workflows (Pasquarella et al., 2022). We applied LandTrendr to annual medoid composites of all high-quality Landsat 5, 7 and 8 Collection 1 Surface Reflectance observations acquired between June 20 and September 20 (Northern Hemisphere growing season) for the years 1985–2020 using the parameters shown in Table 1.

Initial segmentation results were generated separately for three SWIR-based indices, (1) the Normalized Burn Ratio (NBR), (2) the Normalized Difference Moisture Index (NDMI), and (3) Tasseled Cap Wetness (TCW). These indices are sensitive to removal of forest cover and are often used for forest disturbance detection (e.g., Cohen et al., 2010; Collins and Woodcock, 1996; Franklin et al., 2000; Healey et al., 2006; Wilson and Sader, 2002). Though we expect NBR, NDMI, and TCW values to be highly correlated (Fiorella and Ripple, 1995; Jin and Sader, 2005), we also expect differences in their calculation to

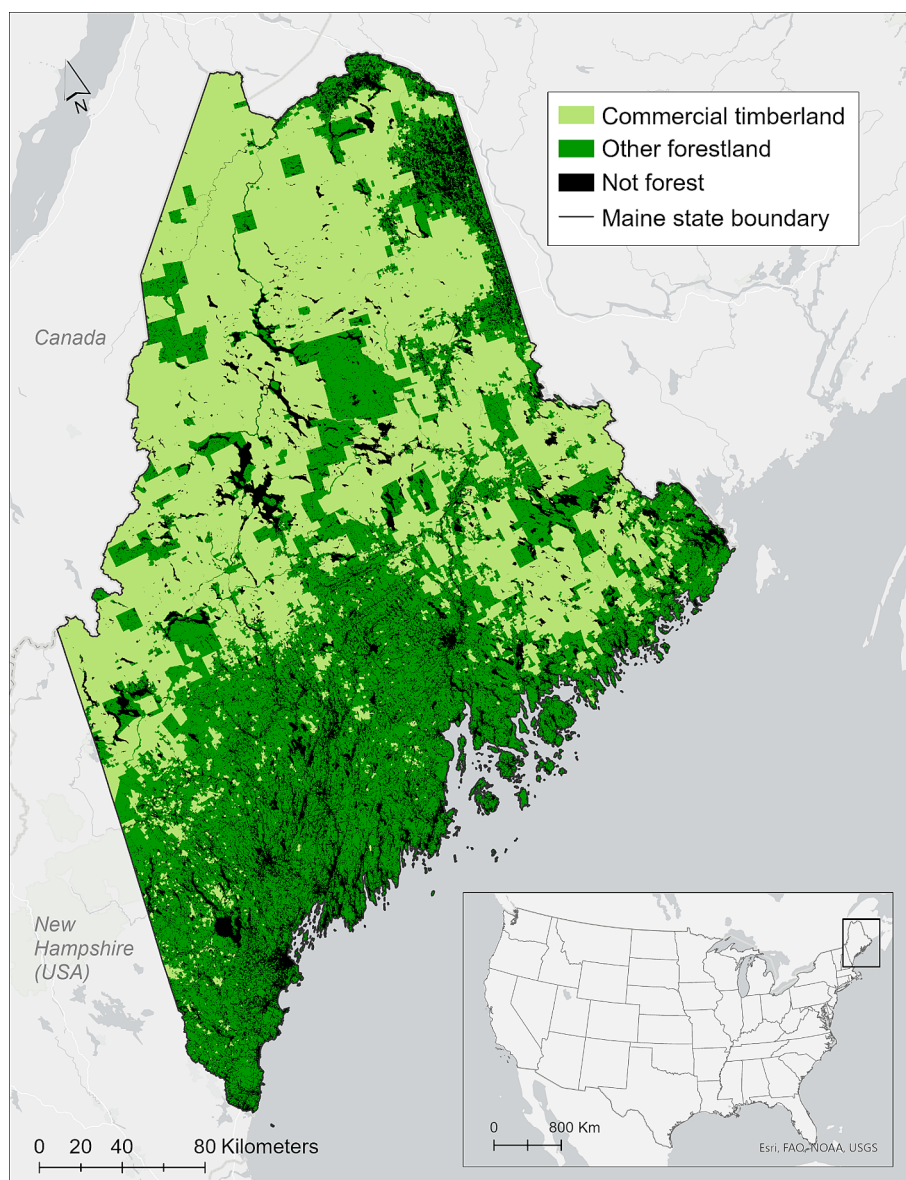


Fig. 1. Forestlands in the state of Maine, northeastern United States. Forests of different ownership types shown in shades of green, non-forest land cover in black (sources: US National Land Cover Dataset, 2019; James W. Sewall Company, 2018). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1

LandTrendr parameters used in this study.

Parameter	Value
maxSegments	8
spikeThreshold	0.9
vertexCountOvershoot	3
preventOneYearRecovery	true
recoveryThreshold	0.75
pvalThreshold	0.05
bestModelProportion	0.75
minObservationsNeeded	6

complement each other and improve harvest detection performance when combined in an ensemble approach (Cohen et al., 2018).

LandTrendr outputs include a series of segments, which correspond to relatively stable periods, and vertices, which were identified as inflection points along a spectral trajectory and are indicative of potential changes in surface conditions (Pasquarella et al., 2022). LandTrendr

segments can be processed in a number of ways to produce maps of potential change events, i.e., extracting only the segment with the greatest magnitude of change over a specified time period (e.g., Senf and Seidl, 2020) or using year-to-year changes in fitted values to indicate cover or condition change (e.g., Cohen et al., 2020, 2018). We considered all loss segments, i.e., those with spectral changes in the direction of decreased vegetation cover, as potential disturbance events. To differentiate harvests from longer-duration disturbances such as those related to drought or forest insect damage, we removed segments greater than two years in duration, leaving only short-term events (less than or equal to two years in duration) that are more likely associated with harvesting.

Because harvest events tend to be larger patch-based disturbances, a minimum mapping unit (mmu) is typically applied to LandTrendr results as a post-processing step (e.g., Cohen et al., 2018; Kennedy et al., 2018). We instead treated the mmu as a feature so that mmu thresholds could be learned rather than set a priori. We used the number of adjacent pixels with segments that began in the same year to estimate harvest patch size, hereafter referred to as mmu. We also extracted the spectral magnitude of change (mag), which provides an indication of harvest

intensity. To better match Landsat-based inputs with the scale of FIA plots and account for uncertainty in GPS coordinates, which can be on the order of 5 to 10 m (McRoberts et al., 2018; Strunk et al., 2019), mag and mmu features for each spectral index (NBR, NDMI and TCW) were smoothed to produce mean mag and max mmu for a 3×3 pixel (90×90 m) kernel (see Pasquarella, 2022 for archived GEE JavaScript workflow).

3.2. Forest inventory and analysis (FIA) measurements

We used all available FIA field plot measurements collected in the state of Maine between 1999 and 2019 to train and cross-validate our harvest detection ensembles. FIA plots in the northeastern U.S. are typically measured every 5 to 7 years (Gillespie, 1999; McRoberts et al., 2005; Tinkham et al., 2018) and we had access to true plot locations through a memorandum of understanding between the USFS and Harvard University (MOU #09MU11242305123). Our Maine FIA dataset consisted of 13,299 measurements (i.e., unique space–time coordinates) recorded for 3,265 plots (i.e., unique spatial locations), and of these, we analyzed the 3,220 FIA plots that had been remeasured at least once and our final dataset included 10,034 pairs of sequential FIA measurements.

FIA surveys record individual trees as being alive, dead, or removed in a given measurement cycle. We aggregated tree-level measurements of diameter at breast height and mortality and removal designations to plot-scale estimates of total basal area removed (m2) as a proxy for harvest intensity. We also calculate percent basal area removal by dividing the total basal area removed between measurements by the basal area of living trees at the time of the first measurement as a relative measure of change (Healey et al., 2006; Tao et al., 2019). Of the 10,034 FIA measurement pairs, 1,711 recorded basal area removal (harvest).

To integrate the Landsat-based and FIA datasets, we queried the LandTrendr results for all years between the first and second FIA measurement years to determine if a potential harvest event was detected between measurements (Fig. 2). The resulting dataset included a record for each FIA remeasurement with plot information from each

measurement pair (m_a and m_b) as well as the LandTrendr mag and mmu features for each of the three spectral indices we considered. This dataset served as the basis for all ensemble experiments and for the FIA-based cross-validation.

3.3. Ensemble approaches

We compared three ensemble approaches for producing forest harvest maps from multi-spectral LandTrendr segmentation results and FIA-based basal area removal estimates, specifically (1) voting schemes, (2) a 500-tree RF classification, and (3) a seven-parameter DDT ensemble.

3.3.1. Voting strategies

Voting schemes target a specific level of agreement across inputs, and the voting strategies used in this study did not consider the spectral or spatial properties of disturbance, only whether or not a disturbance event was detected. Of the methods tested in this study, voting has the distinct advantage of not requiring training data and has a high degree of interpretability but low expressive power. We tested three voting schemes: (1) one of three (*any*), where a detected change in any of the spectral indices during the FIA remeasurement period was considered a change, (2) two of three (*majority*), where detected changes in two of the three indices was required, and (3) three of three (*all*), where a change must be detected in all three indices. Though more complex weighting schemes could be applied, we assigned equal weights to all spectral indices.

3.3.2. Random forest (RF)

Random Forest approaches rely on an ensemble of many decision trees to estimate threshold-based splits for various subsets of feature inputs. For classification tasks, these splits are typically chosen to minimize the Gini impurity, a metric that quantifies class separability. The number of fitted parameters scales with both the number of trees in the ensemble as well as the number of input features, making RF classifiers less interpretable and more difficult to apply over large spatial

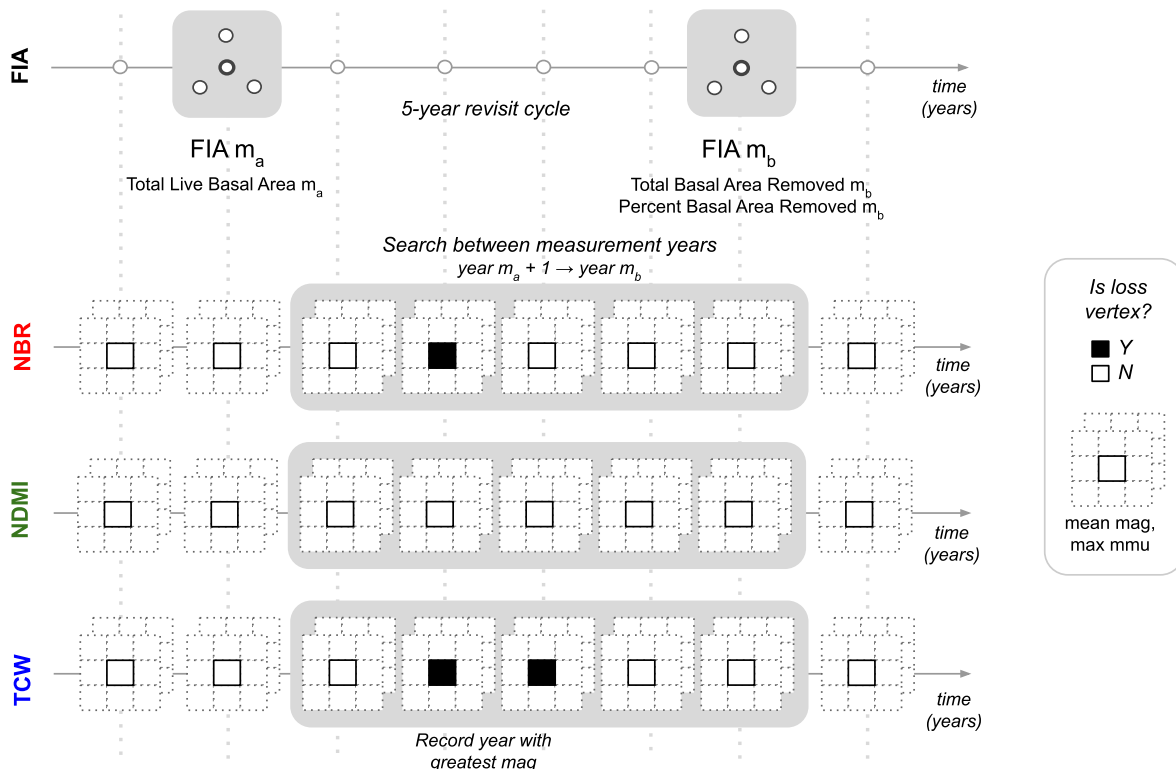


Fig. 2. Matching FIA plot re-measurements with annual LandTrendr results. The same process was used with annual LCMS products.

extents where hundreds of trees are used to classify each pixel.

We used the scikit-learn (version 0.24.1) implementation of RF (Pedregosa, 2011) for our comparisons and set the number of trees to 500 following numerous other studies (Cohen et al., 2018; De Marzo et al., 2021; Healey et al., 2018; Hislop et al., 2019; Wang et al., 2019). Because we wanted to test a generic use case, we used default settings with no further hyperparameter tuning. Models were trained using binary detectable harvest labels based on FIA measurements with the same LandTrendr features used for voting and DDT approaches.

3.3.3. Degenerate decision trees (DDT) ensemble

We also evaluated a degenerate decision trees (DDT) ensemble as an alternative to RF classification. Degenerate trees are a subclass of binary trees where each decision node has only a single parent node. Like RF classifiers, the DDT ensemble is a supervised approach; however, DDT ensembles can be optimized using any number of metrics, are less sensitive to class imbalances, and trained models have fewer parameters. Our DDT models include two decision nodes (mag and mmu) for each of the three spectral indices considered plus an additional threshold on the number of votes across indices for a total of seven fitted parameters (c_1 – c_7 ; Fig. 3).

Although there are a number of ways decision tree parameters could be optimized including Bayesian inference (i.e., multi-armed bandit problem models), simulated annealing, and genetic strategies (Brady, 1985; Katehakis and Veinott, 1987; Kirkpatrick et al., 1983), we implemented the DDT ensemble as a grid search using standard Python and Numpy operations (Harris et al., 2020). This implementation

requires explicit specification of parameter values for the grid search. We selected a step size of 0.01 in NBR and NDMI ratios and 0.001 in TCW reflectance. For mmu thresholds, we used a step size of five connected Landsat pixels, which translates to about a one-acre change in patch size, a standard minimum area for commercial timber harvest. The final vote count threshold (c_7 ; Fig. 3) can vary between one and three votes at the final decision node.

The F1 score, which is the harmonic mean of precision and recall, is commonly used in binary classification (Chinchor, 1992; Lipton et al., 2014). We selected the F1 score as the accuracy metric for optimizing the DDT classification since it is invariant to changes in true negative count and therefore useful for characterizing performance for positive labels in highly imbalanced datasets (Sokolova and Lapalme, 2009). All possible combinations of mag, mmu, and voting thresholds were tested to determine the set of parameters that gave the highest harvest classification accuracy. We provide an example of our DDT implementation at github.com/valpasq/lt-ensemble, including a Python notebook with example functions for running a sweep over series of thresholds and determining optimal thresholds for each feature as well as a Google Earth Engine script for applying thresholds to LandTrendr results (Pasquarella, 2022).

3.4. Assessment

3.4.1. Identifying a harvest detection threshold

To characterize performance over a range of harvest intensities, we trained and tested ensemble models using a series of different basal area

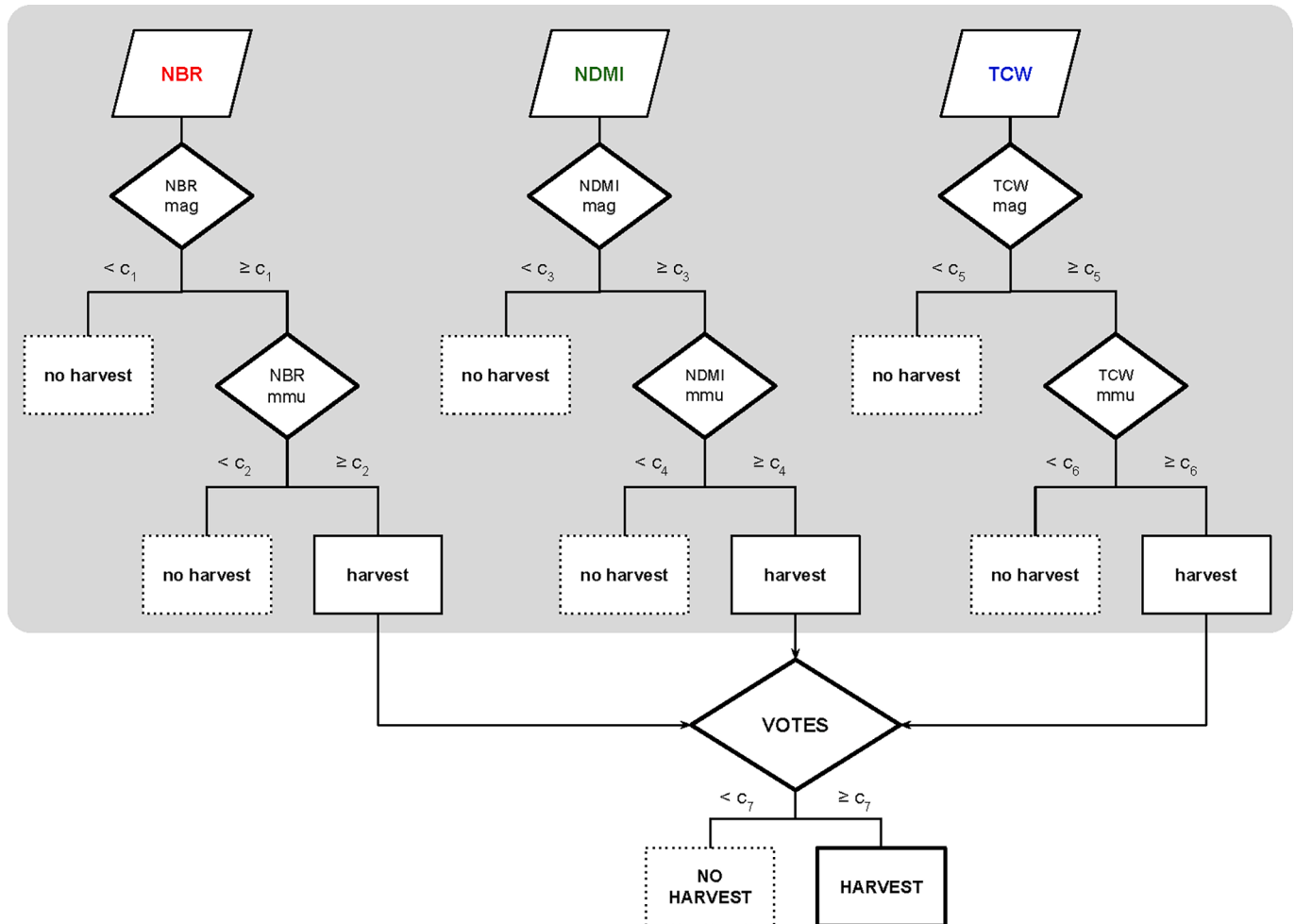


Fig. 3. Degenerate decision trees ensemble architecture consisting of seven trainable thresholds (constants, c_1 – c_7) on spectral (mag) and spatial (mmu) parameters for individual spectral indices and across indices (votes).

removal thresholds for determining what is considered a detectable harvest event. If an FIA remeasurement indicated basal area removal greater than the specified threshold, it was considered a detectable harvest in the training/test set for that threshold; otherwise, it was considered a non-detectable harvest (even though some basal area may have been removed). This allowed us to estimate the influence of varying biophysically-based definitions of harvest on our results and determine the threshold that best represents a detectable harvest in terms of percent basal area removal.

3.4.2. K-fold Cross-validation using FIA reference data

We used our FIA remeasurement dataset to perform a three-fold cross validation replicated ten times for a total of 30 folds. Train-test splits were kept consistent across models to facilitate direct comparison of cross-validation results. For supervised decision tree approaches, we present both training and testing results. For DDT, the training score is the best F1 score achieved across all possible combinations of parameter values, while for RF, this score is calculated by applying a trained model to its own training set. Large differences between these training and testing scores provide an indicator of potential overfitting.

3.4.3. External validation using TimeSync reference data

A second validation dataset was used to assess how well trained models generalized to interpretations based on using a different disturbance labeling protocol. We used TimeSync interpretations collected as part of another effort to model change processes in the northeastern US and Canada (Kilbride, 2018). Points were initially selected using a simple random sample. The TimeSync approach for reference data collection (Cohen et al., 2010) was used to identify segments and vertices and label change processes based on time series visualizations and assisted with historical high-resolution imagery available in Google Earth.

From the full reference dataset of 3,436 TimeSync pixels and interpretations, we extracted 1,294 unique spatial locations within our Maine study area. Interpretations for these locations included 634 events labeled harvest. We also combined TimeSync vertices labeled harvest, mechanical, hydrology, wind, debris, and “other” into a more general fast loss class following (Housman et al., 2021) for comparisons with LCMS products, which do not distinguish between harvests and other fast loss events.

3.4.4. Comparisons with LCMS products

As a final point of comparison, we used an existing national-scale forest disturbance dataset from the Landscape Change Monitoring System (LCMS) project (Lister et al., 2020) as a benchmark, with gains in performance relative to this readily available dataset suggesting development of local-scale ensembles is worthwhile. In contrast to the approaches tested in this study, which are all single-algorithm, single-sensor, multi-spectral ensembles, LCMS products employ a multi-sensor, multi-algorithm stacked generalization approach (Cohen et al., 2018; Healey et al., 2018; Housman et al., 2021). We acquired the full time series of annual LCMS change maps (v2020-5) from 1985 to 2020 from the FSGeodata Clearinghouse (USDA Forest Service, 2021). The LCMS annual change products include several change categories, specifically fast loss, slow loss, and gain (Housman et al., 2021). We focus on the fast loss results, which includes changes attributed to fire, harvest, mechanical, wind/ice, hydrology, debris, and other processes. We assessed the LCMS dataset using the same FIA and TimeSync datasets used to assess other ensemble methods tested in this study. Because *harvest* is a subset of the *fast loss* category mapped by LCMS, we expected to see higher rates of commission than omission due to detection of other types of disturbances for sites where no detectable harvest was observed in the reference datasets.

4. Results

4.1. Cross-validated performance as a function of basal area removal

The highest F1 scores were generally achieved when harvest was defined using a 30% basal area removal threshold, with the exception of the any votes model, which had the highest F1 score at 20% removal. The consistency in optimal basal area removal threshold for defining a detectable harvest event across approaches suggests that the 30% removal threshold is a physically meaningful definition of change. Repeating this analysis using total instead of percent basal area removed, we found 5 m² to be the optimal threshold for defining a detectable harvest. This correlates well with percentage-based findings, as usually about one-third of stand volume is removed in commercial and pre-commercial thinning operations (Sader et al., 2003) and the average plot-level live basal area for our dataset was about 16 m².

In the case of the degenerate trees approach, we are able to associate these lower bounds on detectable forest harvest with a set of thresholds on input feature values and vote counts. For the full model at a 30% basal area removal thresholds were estimated as follows: spectral magnitude (mag) thresholds of 0.10, 0.00, and 0.03 and minimum mapping unit (mmu) thresholds of 5 pixels, 20 pixels, and 10 pixels for NBR, NDMI, and TCW, respectively, and a one-vote threshold across the three indices considered. These thresholds can be directly applied to temporal segmentation results and compared across studies, making the DDT approach inherently more interpretable than an RF that relies on threshold estimates across hundreds of trees (see Fig. 3 for DDT structure).

4.2. Cross-validated performance at 30% removal threshold

The RF and DDT models consistently outperformed voting approaches when validated against the FIA remeasurement dataset (Table 2). Of the three voting ensembles, the majority strategy had the best performance in terms of F1 score ($M = 0.68$, $SD = 0.02$; Table 2) and achieved the most balanced omission/commission. The any index strategy had the lowest F1 scores ($M = 0.54$, $SD = 0.01$), largest errors of commission, and lowest errors of omission across all approaches. The all indices strategy was unsurprisingly the most conservative of the voting strategies, with relatively low errors of commission but the high rates of omission (Table 2).

Of the two supervised decision tree methods, the F1 scores for the cross-validated RF model ($M = 0.71$, $SD = 0.02$) were within one percentage point (0.01) of DDT models trained and tested on the same splits

Table 2

Model comparisons for FIA dataset using a 30% basal area removal threshold to define a detectable harvest event. Mean and standard deviation are reported for cross-validated results. The mean number of FIA measurements at this threshold was 708 ($SD = 18$) for harvest and 5980 ($SD = 39$) for no harvest.

Ensemble approach		F1	Omission Error	Commission Error
Voting	Any index	0.54 (0.01)	0.22 (0.02)	0.59 (0.01)
	Two indices	0.68 (0.02)	0.33 (0.02)	0.31 (0.02)
	All indices	0.65 (0.02)	0.47 (0.02)	0.18 (0.02)
Degenerate decision trees (DDT) ensemble	Full model	0.73	0.30	0.24
	Testing ($1/3$)	0.72 (0.02)	0.32 (0.02)	0.23 (0.03)
	Training ($2/3$)	0.74 (0.01)	0.31 (0.01)	0.21 (0.01)
Random Forest (RF)	Testing ($1/3$)	0.71 (0.02)	0.37 (0.02)	0.21 (0.02)
	Training ($2/3$)	0.87 (0.01)	0.23 (0.01)	0.00 (0.00)
LCMS		Fast loss	0.60	0.54
				0.13

($M = 0.72$, $SD = 0.02$), with the DDT approach exhibiting slightly better performance particularly in terms of lower omission. The RF models tended to have higher omission but lower commission. The difference in the performance of the RF between testing ($M = 0.71$, $SD = 0.02$) and training ($M = 0.87$, $SD = 0.01$) plus complete lack of commission indicate the model is likely overfitting to the training dataset. The DDT approach achieved more balanced training and testing F1 scores (Table 2), suggesting that the reduced number of fitted parameters enables comparable performance without overfitting. Additionally, the mean and median F1 scores across the DDT testing subsets were equivalent (0.72) and very comparable to the training F1 score for the full model (0.73; Table 2).

As would be expected, harvests removing a greater percentage of basal area are more detectable and omission generally decreased with

increasing percent basal area removal (Fig. 4). Commission tended to be highest at basal area removal percentages between 0% and the 30% threshold used to define harvest events during training, which can be attributed to events that are correctly identified as removals but labeled non-harvests at this threshold. The LCMS and all indices voting approaches tended to be the most conservative, with higher omission but lower commission.

4.3. Assessment using TimeSync reference dataset

Performance for all ensemble approaches was poorer when validated against the TimeSync dataset. The highest F1 score among voting strategies for the *harvest* category was achieved by the three-index voting strategy (0.54), though this all-index strategy only slightly outperforms

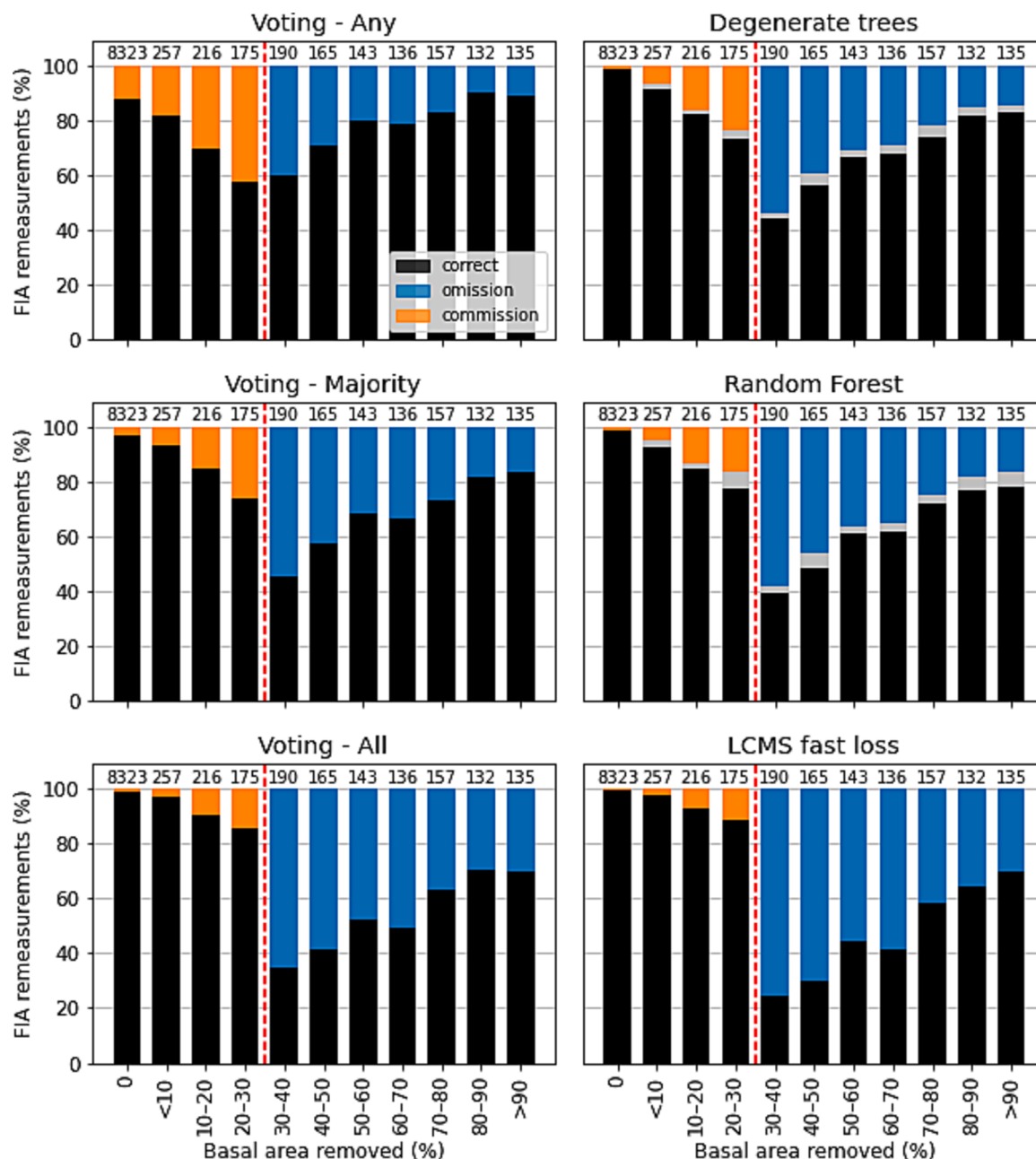


Fig. 4. Cross-validation using FIA remeasurements by percent basal area removal bins using a 30% threshold to define a detectable harvest (red dotted line). Correct labels (true positives and true negatives) are shown in black, omission errors in blue, and commission errors in orange, with gray areas indicating the range of variability over cross-validation splits. Total number of FIA remeasurements in each basal area removal bin is shown at the top of each bar. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the majority strategy (0.53). Using the more general *fast loss* category, F1 scores improved for the any and two-index strategies but decreased slightly for the three indices strategy due to an increase in omission error. The two indices combination had the highest F1 score across all voting approach and label combinations (0.56), though omission and commission errors for voting ensembles were generally much higher and less balanced than for the supervised decision tree approaches (Table 6).

The RF and DDT approaches performed similarly on the TimeSync *harvest* labels (F1 scores of 0.64 and 0.63, respectively), with the RF producing lower errors of commission but higher errors of omission (Table 6). The difference in performance among the decision trees methods was more apparent using the *fast loss* category, with the DDT achieving an F1 score of 0.67 and the RF again achieving a score of 0.63. The LCMS products had the lowest errors of commission on the TimeSync reference dataset; however, omission was greater than 50% for both *harvest* and the aggregated *fast loss* categories (0.54 and 0.57, respectively). In all cases, the *fast loss* aggregation resulted in a slightly higher rate of omission and lower rate of commission, which is logical given the broader number of categories that constitute a change under this broader definition.

4.4. Mapped results

To provide spatial context for our results, we scaled the majority voting and DDT ensembles using GEE (Pasquarella, 2022). Visual comparisons of these mapped results indicated that these methods produce similar spatial patterns to the LCMS stacked generalization model (Fig. 6). However, fast loss predictions from the LCMS dataset tended to be more conservative than the ensemble approaches tested here, with relatively low commission errors, but very high (>50%) errors of omission (Tables 2 and 3). Although it might be assumed that national LCMS products would be better suited for detecting higher-intensity stand-replacing disturbances, higher rates of omission were observed across all percent basal area removal categories in the FIA-based analysis (Fig. 4). Comparisons with the categorical harvest interpretations in the TimeSync assessments also indicate greater errors of omission (Fig. 5), suggesting that using fast loss classifications from the LCMS dataset to represent potential harvest events would have resulted in underestimation of affected areas with important implications for management and policy assessments.

5. Discussion

Accurate mapping of forest disturbance events is essential for assessing the past, present, and future impacts of forest management. In the industrial timberlands of Maine, harvest tends to be the most common type of forest disturbance, as evidenced by the dominant portion of *harvest* and *mechanical* disturbances in the TimeSync dataset (Fig. 5). The ensemble approaches tested here effectively apply a secondary classification to attribute harvesting events to a disturbance product. By running LandTrendr multiple times using different spectral indices, we initially detect a large set of possible disturbance events, decreasing

potential for omission errors while increasing commission errors. We then use ensemble methods to further refine LandTrendr commission errors by performing a secondary classification.

By combining multi-spectral LandTrendr temporal segmentation results, we were able to map potential harvest events with F1 scores of up to 0.73 based on forest inventory measurements and 0.64 based on visual interpretations (Tables 2 and 3). Though majority voting tended to exhibit the best performance of the three voting approaches, all voting schemes were consistently outperformed by supervised decision tree approaches (Tables 2 and 3), confirming the value of training data for generating more accurate harvest maps. Using forest inventory data on basal area removals for training and validation, we are able to define a lower bound of 30% on a detectable harvest in terms of percent basal area removed, which is comparable to a lower bound of 20% basal area removal characterized by Tortini et al. (2015) in the Michigan Upper Peninsula. Harvest detection accuracy tended to increase with higher proportions of basal area removed (Fig. 4) suggesting that fine-tuning spectral and spatial thresholds may be less critical in regions and time periods where production forestry systems are dominated by clear-cutting regimes (e.g., Cohen et al., 2002; Liu et al., 2004) compared with mixed-ownership landscapes dominated by partial harvests (e.g., Jarron et al., 2016). We also note that mean omission error for the all indices voting ensemble was 0.22 for the FIA cross-validation, 0.21 for the TimeSync *harvest* category, and 0.20 for the TimeSync *fast loss* category, suggesting an ~ 80% upper bound on LandTrendr's ability to characterize harvest events from annual time series of Landsat observations. These baseline omission errors could be further reduced by including results from other temporal segmentation approaches that utilize a higher frequency of observations and may be more sensitive to low-magnitude changes not detected in annual growing season composites (e.g., Verbesselt et al., 2010; Zhu and Woodcock, 2014), though at the cost of additional computational overhead. Detection of partial harvest events could also be improved using higher-resolution imagery, e.g., Sentinel-2 time series; however, the Landsat record is uniquely suited for mapping harvest events on the sorts of decadal time scales of interest in this study.

In comparing supervised ensembling approaches, we found that a three-tree, seven-parameter DDT model achieved comparable harvest detection performance to a 500-tree RF with more consistent performance between training and testing datasets, indicating the simpler model does not overfit and is better able to generalize to unseen examples (Table 5). ML algorithms like RF have become the norm in disturbance mapping. In contrast to this conventional wisdom, our analysis shows simpler decision trees can be just as accurate, more interpretable, and straightforward to apply. A 500-tree RF may take only seconds to train, but applying hundreds of decision trees at scale can become a very memory- and storage-intensive operation, requiring additional resources beyond those initially required to generate temporal segmentation or other change detection results. As an intermediate option, a grid search optimization strategy allowed us to exhaustively investigate decision tree parameter spaces and output optimized thresholds for individual input features as a single human-readable list.

Table 3

Comparisons of ensemble approach results for 30% basal area removal threshold and TimeSync interpretations. TimeSync labels were grouped to produce two binary comparisons for assessment: *harvest* (harvest versus all other categories) and *fast loss* (fire, harvest, mechanical, wind/ice, hydrology, debris, and "other" versus all other categories).

Ensemble approach		TimeSync <i>harvest</i>			TimeSync <i>fast loss</i>		
		F1	Omission	Commission	F1	Omission	Commission
Voting	Any index	0.47	0.20	0.66	0.52	0.21	0.62
	Two indices	0.53	0.33	0.55	0.56	0.36	0.50
	All indices	0.54	0.51	0.40	0.53	0.55	0.35
Decision trees	Degenerate decision trees (DDT)	0.64	0.26	0.43	0.67	0.29	0.37
	Random Forest (RF)	0.63	0.33	0.39	0.63	0.39	0.35
LCMS		0.57	0.54	0.23	0.57	0.57	0.16

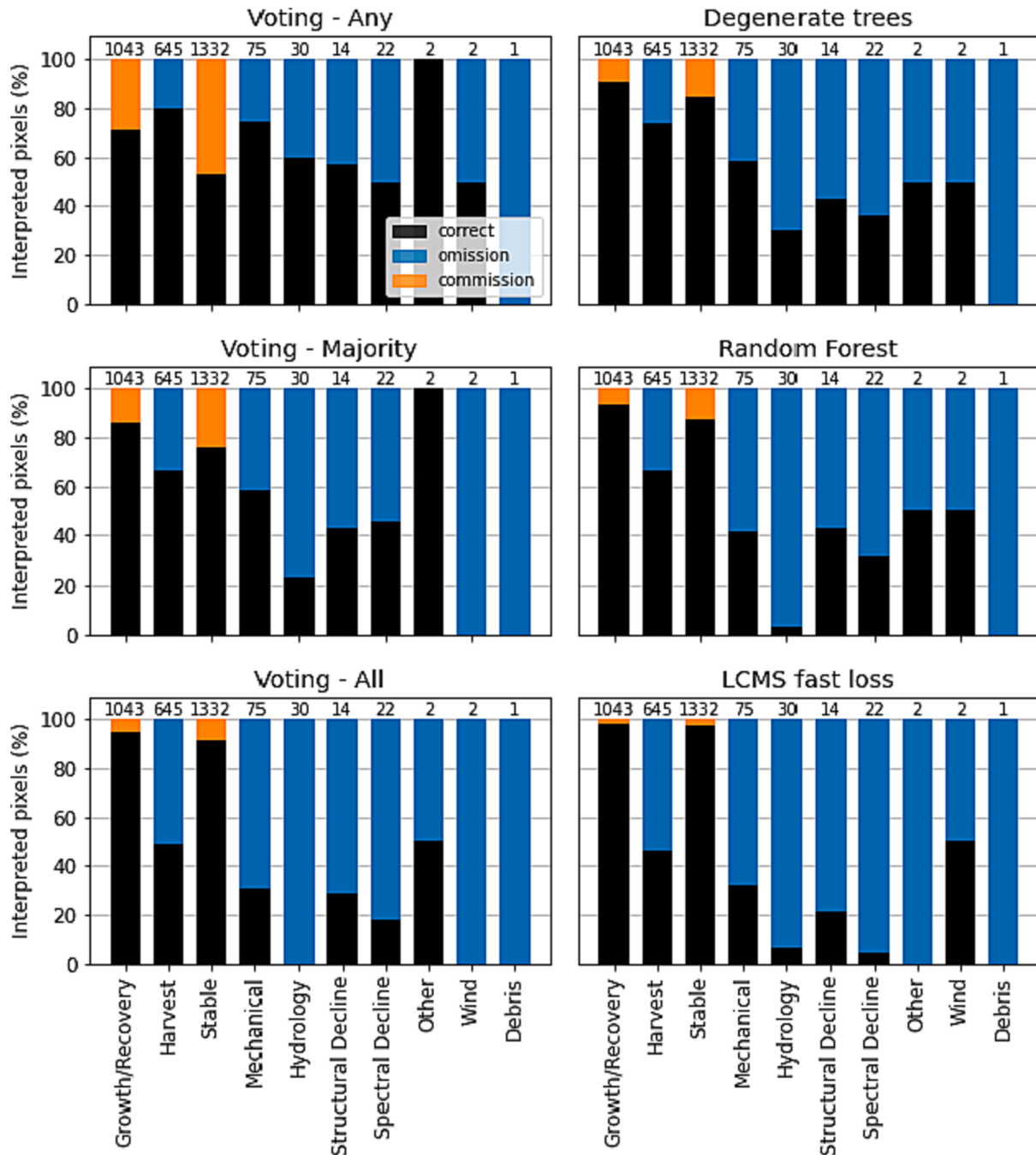


Fig. 5. Comparison of errors by TimeSync category for voting, degenerate decision trees, Random Forest, and an existing fast loss change product (LCMS) using a 30% threshold to define a detectable harvest. Fire, harvest, mechanical, wind/ice, hydrology, debris, and “other” are considered *fast loss* and growth/recovery, spectral decline, stable, and structural decline are considered *not fast loss*. Total number of interpretations for each category is shown at the top of each bar.

These thresholds can then be applied using Boolean logic, meaning small custom decision tree models can be trivially re-implemented in different programming languages and software environments with minimal computational overhead. In contrast, transferring saved models produced by different implementations of RF (e.g., the *RandomForest* and *ranger* packages for the R programming language, Scikit-learn, TensorFlow, GEE) to other platforms can be difficult or impossible (Abadi et al., 2015; Wright and Ziegler, 2015) and our findings indicate this extra effort may not be justified by a notable difference in performance.

With increasing availability of national and global change detection datasets, researchers interested in analyzing disturbance patterns and processes must choose between working with or adapting existing datasets or developing custom datasets more tailored to their needs like

those generated in this study. Our results suggest that improvements in detection accuracy relative to the more general-purpose national-scale LCMS disturbance detection ensemble justifies the development of custom ensemble models for detecting harvest events. Furthermore, given other precedents for implementing a series of regionally-trained RF models to improve sensitivity to local conditions over large extents (e.g., Hermosilla et al., 2022), it should similarly be feasible to take a regional or other locally-gridded approach to tuning and scaling DDT models to new areas. Simple decision tree approaches can be extended to characterize multiple disturbance agents by building binary classifications for each agent. Rather than rely on large models to generalize from large feature sets, the DDT approach facilitates development of smaller, more interpretable ensemble models that can easily be tuned and

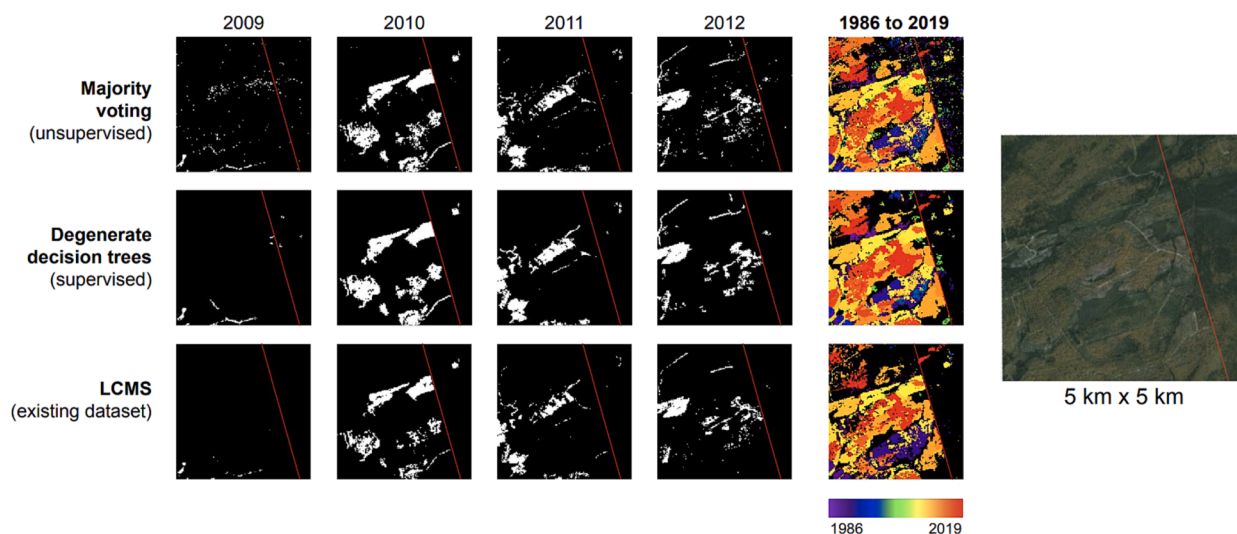


Fig. 6. Annual harvest detection results for 2009 through 2012 and year of most recent detection for the full time series (1986–2019) for voting, degenerate decision trees and existing LCMS dataset (stacked generalization approach). Results shown for a 5 km × 5 km site centered on a site along the northwestern boundary of Baxter State Park, Maine, USA (46.1171° N, -69.10205° W).

compared across disturbance types.

6. Conclusions

Counter to conventional wisdom, this study demonstrated that accurately mapping forest harvest events does not necessarily require advanced machine learning approaches like RF. We found that a seven-parameter degenerate decision tree (DDT) ensemble exhibited comparable performance to a 500-tree RF for ensemble-based harvest classification. In the common situation where high-quality training data like the forest inventory measurements used in this study are not available, our results suggest majority voting can also produce acceptable results based on multi-spectral change detection outputs alone. Given that models with fewer parameters are more interpretable and easier to apply at scale, we conclude that more parsimonious approaches can be preferable. As an additional benefit of our approach, we establish a biophysical interpretation of the forest harvest maps generated in this study, determining that using Landsat time series and a LandTrendr change detection approach, we most accurately detected harvests where at least 30% or around 5 m² of total basal area was removed. Overall, our findings indicate that simple ensemble models can be suitable alternatives to commonly used RF approaches, supporting their continued use as well as further exploration of best practices for ensemble-based mapping of forest disturbances including forest harvest events.

CRedit authorship contribution statement

Valerie J. Pasquarella: Conceptualization, Data curation, Methodology, Formal analysis, Investigation, Funding acquisition, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Luca L. Morreale:** Conceptualization, Data curation, Validation, Writing – review & editing. **Christopher F. Brown:** Conceptualization, Data curation, Methodology, Software, Visualization, Writing – review & editing. **John B. Kilbride:** Conceptualization, Data curation, Validation, Writing – review & editing. **Jonathan R. Thompson:** Conceptualization, Funding acquisition, Resources, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

the work reported in this paper.

Data availability

Earth Engine scripts and degenerate trees implementation example are available at github.com/valpasq/lt-ensemble. Authors do not have permission to share FIA measurement dataset.

Acknowledgements

We thank Lucy Lee for designing the Maine study area figure. We would also like to gratefully acknowledge our anonymous reviewers for their feedback on this manuscript.

Funding

This work was supported by the United States Department of Agriculture (USDA) National Institute of Food and Agriculture Award #2021-67023-34491, the National Science Foundation (NSF) Dynamics of Integrated Socio-Environmental Systems (DISES) Grant 22-05705, and the Harvard Forest Long Term Ecological Research Program (NSF DEB #1832210).

References

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Others, 2015. TensorFlow: Large-scale machine learning on heterogeneous systems.
- Banskota, A., Kayastha, N., Falkowski, M.J., Wulder, M.A., Froese, R.E., White, J.C., 2014. Forest Monitoring Using Landsat Time Series Data: A Review. *Can. J. Remote Sens.* 40, 362–384. <https://doi.org/10.1080/07038992.2014.987376>.
- Belair, E.P., Ducey, M.J., 2018. Patterns in Forest Harvesting in New England and New York: Using FIA Data to Evaluate Silvicultural Outcomes. *J. For.* 116, 273–282. <https://doi.org/10.1093/jofore/fvx019>.
- Belgiu, M., Drăguț, L., 2016. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* 114, 24–31. <https://doi.org/10.1016/j.isprsjprs.2016.01.011>.
- Brady, R.M., 1985. Optimization strategies gleaned from biological evolution. *Nature* 317, 804–806. <https://doi.org/10.1038/317804a0>.
- Butler, B.J., 2017. Forests of Maine, 2016. U.S. Department of Agriculture, Forest Service, Northern Research Station, Newtown Square, PA. Doi: .
- Canham, C.D., Rogers, N., Buchholz, T., 2013. Regional variation in forest harvest regimes in the northeastern United States. *Ecol. Appl.* 23, 515–522. <https://doi.org/10.1890/12-0180.1>.
- Chen, S., Woodcock, C.E., Bullock, E.L., Arévalo, P., Torchinava, P., Peng, S., Olofsson, P., 2021. Monitoring temperate forest degradation on Google Earth Engine

- using Landsat time series analysis. *Remote Sens. Environ.* 265, 112648 <https://doi.org/10.1016/j.rse.2021.112648>.
- Chinchor, N., 1992. MUC-4 evaluation metrics, in: Proceedings of the 4th Conference on Message Understanding, MUC4 '92. Association for Computational Linguistics, USA, pp. 22–29. Doi: 10.3115/1072064.1072067.
- Chudy, R.P., Cabbage, F.W., 2020. Research trends: Forest investments as a financial asset class. *For. Policy Econ.* 119, 102273 <https://doi.org/10.1016/j.forpol.2020.102273>.
- Cohen, W.B., Spies, T.A., Alig, R.J., Oetter, D.R., Maierberger, T.K., Fiorella, M., 2002. Characterizing 23 Years (1972–95) of Stand Replacement Disturbance in Western Oregon Forests with Landsat Imagery. *Ecosystems* 5, 122–137. <https://doi.org/10.1007/s10021-001-0060-X>.
- Cohen, W.B., Yang, Z., Kennedy, R., 2010. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 2. TimeSync — Tools for calibration and validation. *Remote Sens. Environ.* 114, 2911–2924. <https://doi.org/10.1016/j.rse.2010.07.010>.
- Cohen, W.B., Healey, S.P., Yang, Z., Stehman, S.V., Brewer, C.K., Brooks, E.B., Gorelick, N., Huang, C., Hughes, M.J., Kennedy, R.E., Loveland, T.R., Moisen, G.G., Schroeder, T.A., Vogelmann, J.E., Woodcock, C.E., Yang, L., Zhu, Z., 2017. How Similar Are Forest Disturbance Maps Derived from Different Landsat Time Series Algorithms? *Forests* 8, 98. <https://doi.org/10.3390/f8040098>.
- Cohen, W.B., Yang, Z., Healey, S.P., Kennedy, R.E., Gorelick, N., 2018. A LandTrendr multispectral ensemble for forest disturbance detection. *Remote Sens. Environ.* 205, 131–140. <https://doi.org/10.1016/j.rse.2017.11.015>.
- Cohen, W.B., Healey, S.P., Yang, Z., Zhu, Z., Gorelick, N., 2020. Diversity of Algorithm and Spectral Band Inputs Improves Landsat Monitoring of Forest Disturbance. *Remote Sensing* 12, 1673. <https://doi.org/10.3390/rs12101673>.
- Collins, J.B., Woodcock, C.E., 1996. An assessment of several linear change detection techniques for mapping forest mortality using multitemporal landsat TM data. *Remote Sens. Environ.* 56, 66–77. [https://doi.org/10.1016/0034-4257\(95\)00233-2](https://doi.org/10.1016/0034-4257(95)00233-2).
- Coops, N.C., Shang, C., Wulder, M.A., White, J.C., Hermosilla, T., 2020. Change in forest condition: Characterizing non-stand replacing disturbances using time series satellite imagery. *For. Ecol. Manage.* 474, 118370 <https://doi.org/10.1016/j.foreco.2020.118370>.
- De Marzo, T., Pflugmacher, D., Baumann, M., Lambin, E.F., Gasparri, I., Kummerle, T., 2021. Characterizing forest disturbances across the Argentine Dry Chaco based on Landsat time series. *Int. J. Appl. Earth Obs. Geoinf.* 98, 102310 <https://doi.org/10.1016/j.jag.2021.102310>.
- Duvenek, M.J., Thompson, J.R., Wilson, B.T., 2015. An imputed forest composition map for New England screened by species range boundaries. *For. Ecol. Manage.* 347, 107–115. <https://doi.org/10.1016/j.foreco.2015.03.016>.
- Fiorella, M., Ripple, W.J., 1995. Determining successional stage of temperate coniferous forests with Landsat satellite data. *Geographic Information Analysis: An Ecological Approach for the Management of Wildlife on the Forest Landscape*.
- Franklin, S.E., Moskal, L.M., Lavigne, M.B., Pugh, K., 2000. Interpretation and Classification of Partially Harvested Forest Stands in the Fundy Model Forest Using Multitemporal Landsat TM Digital Data. *Can. J. Remote Sens.* 26, 318–333. <https://doi.org/10.1080/07038992.2000.10874783>.
- Fraver, S., Seymour, R.S., Speer, J.H., White, A.S., 2007. Dendrochronological reconstruction of spruce budworm outbreaks in northern Maine, USA. *Can. J. Forest Res.* <https://doi.org/10.1139/x06-251>.
- Friedl, M.A., Brodley, C.E., 1997. Decision tree classification of land cover from remotely sensed data. *Remote Sens. Environ.* 61, 399–409. [https://doi.org/10.1016/S0034-4257\(97\)00049-7](https://doi.org/10.1016/S0034-4257(97)00049-7).
- Gillespie, A.J.R., 1999. Rationale for a National Annual Forest Inventory Program. *J. For.* 97, 16–20. <https://doi.org/10.1093/jof/97.12.16>.
- Hansen, M.C., Loveland, T.R., 2012. A review of large area monitoring of land cover change using Landsat data. *Remote Sens. Environ.* 122, 66–74. <https://doi.org/10.1016/j.rse.2011.08.024>.
- Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G., 2013. High-resolution global maps of 21st-century forest cover change. *Science* 342, 850–853. <https://doi.org/10.1126/science.1244693>.
- Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., Del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. *Nature* 585, 357–362. <https://doi.org/10.1038/s41586-020-2649-2>.
- Healey, S.P., Yang, Z., Cohen, W.B., Pierce, D.J., 2006. Application of two regression-based methods to estimate the effects of partial harvest on forest structure using Landsat data. *Remote Sens. Environ.* 101, 115–126. <https://doi.org/10.1016/j.rse.2005.12.006>.
- Healey, S.P., Cohen, W.B., Yang, Z., Kenneth Brewer, C., Brooks, E.B., Gorelick, N., Hernandez, A.J., Huang, C., Joseph Hughes, M., Kennedy, R.E., Loveland, T.R., Moisen, G.G., Schroeder, T.A., Stehman, S.V., Vogelmann, J.E., Woodcock, C.E., Yang, L., Zhu, Z., 2018. Mapping forest change using stacked generalization: An ensemble approach. *Remote Sens. Environ.* 204, 717–728. <https://doi.org/10.1016/j.rse.2017.09.029>.
- Hemati, M., Hasanlou, M., Mahdianpari, M., Mohammadianesh, F., 2021. A Systematic Review of Landsat Data for Change Detection Applications: 50 Years of Monitoring the Earth. *Remote Sensing* 13, 2869. <https://doi.org/10.3390/rs13152869>.
- Hermosilla, T., Bastyr, A., Coops, N.C., White, J.C., Wulder, M.A., 2022. Mapping the presence and distribution of tree species in Canada's forested ecosystems. *Remote Sens. Environ.* 282, 113276 <https://doi.org/10.1016/j.rse.2022.113276>.
- Hislop, S., Jones, S., Soto-Berelov, M., Skidmore, A., Haywood, A., Nguyen, T.H., 2019. A fusion approach to forest disturbance mapping using time series ensemble techniques. *Remote Sens. Environ.* 221, 188–197. <https://doi.org/10.1016/j.rse.2018.11.025>.
- Housman, I., Campbell, L., Goetz, W., Finco, M., Pugh, N., Megown, K., 2021. US Forest Service Landscape Change Monitoring System Methods (No. GTAC-10225-Brief1). USDA Forest Service.
- Jarron, L.R., Hermosilla, T., Coops, N.C., Wulder, M.A., White, J.C., Hobart, G.W., Leckie, D.G., 2016. Differentiation of Alternate Harvesting Practices Using Annual Time Series of Landsat Data. *For. Trees Livelihoods* 8, 15. <https://doi.org/10.3390/f8010015>.
- Jin, S., Sader, S.A., 2005. Comparison of time series tasseled cap wetness and the normalized difference moisture index in detecting forest disturbances. *Remote Sens. Environ.* 94, 364–372. <https://doi.org/10.1016/j.rse.2004.10.012>.
- Jin, S., Sader, S.A., 2006. Effects of forest ownership and change on forest harvest rates, types and trends in northern Maine. *For. Ecol. Manage.* 228, 177–186. <https://doi.org/10.1016/j.foreco.2006.03.009>.
- Katehakis, M.N., Veinott, A.F., 1987. The Multi-Armed Bandit Problem: Decomposition and Computation. *Math. Oper. Res.* 12, 262–268.
- Kennedy, R.E., Yang, Z., Cohen, W.B., 2010. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr — Temporal segmentation algorithms. *Remote Sens. Environ.* 114, 2897–2910. <https://doi.org/10.1016/j.rse.2010.07.008>.
- Kennedy, R.E., Andréfouët, S., Cohen, W.B., Gómez, C., Griffiths, P., Hais, M., Healey, S.P., Helmer, E.H., Hostert, P., Lyons, M.B., Meigs, G.W., Pflugmacher, D., Phinn, S.R., Powell, S.L., Scarth, P., Sen, S., Schroeder, T.A., Schneider, A., Sonnenschein, R., Vogelmann, J.E., Wulder, M.A., Zhu, Z., 2014. Bringing an ecological view of change to Landsat-based remote sensing. *Front. Ecol. Environ.* 12, 339–346. <https://doi.org/10.1890/1530066>.
- Kennedy, R.E., Yang, Z., Gorelick, N., Braaten, J., Cavalcante, L., Cohen, W.B., Healey, S., 2018. Implementation of the LandTrendr Algorithm on Google Earth Engine. *Remote Sensing* 10, 691. <https://doi.org/10.3390/rs10050691>.
- Kilbride, J.B., 2018. Forest Disturbance Detection and Aboveground Biomass Modeling Using Moderate-Resolution, Time-Series Satellite Imagery. The University of Maine.
- Kirkpatrick, S., Gelatt Jr, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. *Science* 220, 671–680. <https://doi.org/10.1126/science.220.4598.671>.
- Kittredge, D.B., Thompson, J.R., Morreale, L.L., Short Gianotti, A.G., Hutrya, L.R., 2017. Three decades of forest harvesting along a suburban–rural continuum. *Ecosphere* 8, e01882.
- Koltunov, A., Ramirez, C.M., Ustin, S.L., Slaton, M., Haunreiter, E., 2020. eDART: The Ecosystem Disturbance and Recovery Tracker system for monitoring landscape disturbances and their cumulative effects. *Remote Sens. Environ.* 238, 111482. <https://doi.org/10.1016/j.rse.2019.111482>.
- Lipton, Z.C., Elkan, C., Narayanaswamy, B., 2014. Thresholding Classifiers to Maximize F1 Score. *arXiv [stat.ML]*. Doi: 10.48550/arXiv.1402.1892.
- Lister, A.J., Andersen, H., Frescino, T., Gatziolis, D., Healey, S., Heath, L.S., Liknes, G.C., McRoberts, R., Moisen, G.G., Nelson, M., Riemann, R., Schleeuwis, K., Schroeder, T.A., Westfall, J., Wilson, B.T., 2020. Use of Remote Sensing Data to Improve the Efficiency of National Forest Inventories: A Case Study from the United States National Forest Inventory. *Forests* 11, 1364. <https://doi.org/10.3390/f11121364>.
- Liu, S., Loveland, T.R., Kurtz, R.M., 2004. Contemporary carbon dynamics in terrestrial ecosystems in the southeastern plains of the United States. *Environ. Manage.* 33. <https://doi.org/10.1007/s00267-003-9152-z>.
- Lorimer, C.G., 1977. The presettlement forest and natural disturbance cycle of northeastern Maine. *Ecology* 58, 139–148. <https://doi.org/10.2307/1935115>.
- Masek, J.G., Cohen, W.B., Leckie, D., Wulder, M.A., Vargas, R., de Jong, B., Healey, S., Law, B., Birdsey, R., Houghton, R.A., Mildrexler, D., Goward, S., Smith, W.B., 2011. Recent rates of forest harvest and conversion in North America. *J. Geophys. Res.* 116 <https://doi.org/10.1029/2010jg001471>.
- McRoberts, R.E., Holden, G.R., Nelson, M.D., Liknes, G.C., Gormanson, D.D., 2005. Using satellite imagery as ancillary data for increasing the precision of estimates for the Forest Inventory and Analysis program of the USDA Forest Service. *Can. J. For. Res.* 35, 2968–2980. <https://doi.org/10.1139/x05-222>.
- McRoberts, R.E., Chen, Q., Walters, B.F., Kaisershot, D.J., 2018. The effects of global positioning system receiver accuracy on airborne laser scanning-assisted estimates of aboveground biomass. *Remote Sens. Environ.* 207, 42–49. <https://doi.org/10.1016/j.rse.2017.09.036>.
- Oswalt, S.N., Smith, W.B., Miles, P.D., Pugh, S.A., 2019. Forest resources of the United States, 2017. U.S. Department of Agriculture, Forest Service, Washington, DC. Doi: 10.2737/wo-gtr-97.
- Pasquarella, V.J., Arévalo, P., Bratley, K.H., Bullock, E.L., Gorelick, N., Yang, Z., Kennedy, R.E., 2022. Demystifying LandTrendr and CCDC temporal segmentation. *Int. J. Appl. Earth Obs. Geoinf.* 110, 102806 <https://doi.org/10.1016/j.jag.2022.102806>.
- Pasquarella, V., 2022. valpasq/lt-ensemble: v0.0.0-release. Doi: 10.5281/zenodo.7301752.
- Pedregosa, F., 2011. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.
- Pickett, S.T.A., White, P.S., 2013. The Ecology of Natural Disturbance and Patch Dynamics. Elsevier.
- Polikar, R., 2006. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine* 6, 21–45. <https://doi.org/10.1109/MCAS.2006.1688199>.
- Sader, S.A., Legaard, K.R., 2008. Inclusion of forest harvest legacies, forest type, and regeneration spatial patterns in updated forest maps: A comparison of mapping results. *For. Ecol. Manage.* 255, 3846–3856. <https://doi.org/10.1016/j.foreco.2008.03.047>.

- Sader, S.A., Bertrand, M., Wilson, E.H., 2003. Satellite Change Detection of Forest Harvest Patterns on an Industrial Forest Landscape. *For. Sci.* 49, 341–353. <https://doi.org/10.1093/forestscience/49.3.341>.
- Schultz, M., Clevers, J.G.P.W., Carter, S., Verbesselt, J., Avitabile, V., Quang, H.V., Herold, M., 2016. Performance of vegetation indices from Landsat time series in deforestation monitoring. *Int. J. Appl. Earth Obs. Geoinf.* 52, 318–327. <https://doi.org/10.1016/j.jag.2016.06.020>.
- Senf, C., Seidl, R., 2020. Mapping the forest disturbance regimes of Europe. *Nature Sustainability* 4, 63–70. <https://doi.org/10.1038/s41893-020-00609-y>.
- Seymour, R.S., White, A.S., deMaynadier, P.G., 2002. Natural disturbance regimes in northeastern North America—evaluating silvicultural systems using natural scales and frequencies. *For. Ecol. Manage.* 155, 357–367. [https://doi.org/10.1016/S0378-1127\(01\)00572-2](https://doi.org/10.1016/S0378-1127(01)00572-2).
- Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* 45, 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>.
- Strunk, J., Packalen, P., Gould, P., Gatzliolis, D., Maki, C., Andersen, H.-E., McGaughey, R.J., 2019. Large Area Forest Yield Estimation with Pushbroom Digital Aerial Photogrammetry. *For. Trees Livelihoods* 10, 397. <https://doi.org/10.3390/f10050397>.
- Tao, X., Huang, C., Zhao, F., Schleeweis, K., Masek, J., Liang, S., 2019. Mapping forest disturbance intensity in North and South Carolina using annual Landsat observations and field inventory data. *Remote Sens. Environ.* 221, 351–362. <https://doi.org/10.1016/j.rse.2018.11.029>.
- Thomas, V.A., Wynne, R.H., Kauffman, J., McCurdy, W., Brooks, E.B., Thomas, R.Q., Rakestraw, J., 2021. Mapping thins to identify active forest management in southern pine plantations using Landsat time series stacks. *Remote Sens. Environ.* 252, 112127. <https://doi.org/10.1016/j.rse.2020.112127>.
- Thompson, J.R., Canham, C.D., Morreale, L., Kittredge, D.B., Butler, B., 2017. Social and biophysical variation in regional timber harvest regimes. *Ecol. Appl.* 27, 942–955. <https://doi.org/10.1002/eap.1497>.
- Tinkham, W.T., Mahoney, P.R., Hudak, A.T., Domke, G.M., Falkowski, M.J., Woodall, C. W., Smith, A.M.S., 2018. Applications of the United States Forest Inventory and Analysis dataset: a review and future directions. *Can. J. For. Res.* 48, 1251–1268. <https://doi.org/10.1139/cjfr-2018-0196>.
- Tortini, R., Mayer, A.L., Maiani, P., 2015. Using an OBCD Approach and Landsat TM Data to Detect Harvesting on Nonindustrial Private Property in Upper Michigan. *Remote Sensing* 7, 7809–7825. <https://doi.org/10.3390/rs70607809>.
- Tortini, R., Mayer, A.L., Hermosilla, T., Coops, N.C., Wulder, M.A., 2019. Using annual Landsat imagery to identify harvesting over a range of intensities for non-industrial family forests. *Landsc. Urban Plan.* 188, 143–150. <https://doi.org/10.1016/j.landurbplan.2018.04.012>.
- USDA Forest Service, 2021. USFS Landscape Change Monitoring System Southeastern Alaska version 2020-5.
- Verbesselt, J., Hyndman, R., Newnham, G., Culvenor, D., 2010. Detecting trend and seasonal changes in satellite image time series. *Remote Sens. Environ.* 114, 106–115. <https://doi.org/10.1016/j.rse.2009.08.014>.
- Wang, Y., Ziv, G., Adami, M., Mitchard, E., Batterman, S.A., Buermann, W., Schwantes Marimon, B., Marimon Junior, B.H., Matias Reis, S., Rodrigues, D., Galbraith, D., 2019. Mapping tropical disturbed forests using multi-decadal 30 m optical satellite imagery. *Remote Sens. Environ.* 221, 474–488. <https://doi.org/10.1016/j.rse.2018.11.028>.
- Wilson, E.H., Sader, S.A., 2002. Detection of forest harvest type using multiple dates of Landsat TM imagery. *Remote Sens. Environ.* 80, 385–396. [https://doi.org/10.1016/S0034-4257\(01\)00318-2](https://doi.org/10.1016/S0034-4257(01)00318-2).
- Wolpert, D.H., 1992. Stacked generalization. *Neural Netw.* 5, 241–259. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1).
- Woodcock, C.E., Loveland, T.R., Herold, M., Bauer, M.E., 2020. Transitioning from change detection to monitoring with remote sensing: A paradigm shift. *Remote Sens. Environ.* 238, 111558. <https://doi.org/10.1016/j.rse.2019.111558>.
- Wright, M.N., Ziegler, A., 2015. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *arXiv [stat.ML]*.
- Wulder, M.A., Masek, J.G., Cohen, W.B., Loveland, T.R., Woodcock, C.E., 2012. Opening the archive: How free data has enabled the science and monitoring promise of Landsat. *Remote Sens. Environ.* 122, 2–10. <https://doi.org/10.1016/j.rse.2012.01.010>.
- Ye, S., Rogan, J., Zhu, Z., Eastman, J.R., 2021. A near-real-time approach for monitoring forest disturbance using Landsat time series: stochastic continuous change detection. *Remote Sens. Environ.* 252, 112167. <https://doi.org/10.1016/j.rse.2020.112167>.
- Zhu, Z., 2017. Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications. *ISPRS J. Photogramm. Remote Sens.* 130, 370–384. <https://doi.org/10.1016/j.isprsjprs.2017.06.013>.
- Zhu, Z., Woodcock, C.E., 2014. Continuous change detection and classification of land cover using all available Landsat data. *Remote Sens. Environ.* 144, 152–171. <https://doi.org/10.1016/j.rse.2014.01.011>.