Forest Ecosystems 11 (2024) 100236

Contents lists available at ScienceDirect

Forest Ecosystems

journal homepage: www.keaipublishing.com/cn/journals/forest-ecosystems

Predicting time-to-harvest in mixed-species forests using a random survival forest algorithm

Dinuka Madhushan Senevirathne [a,*], Sheng-I Yang [a], Consuelo Brandeis [b], Donald G. Hodges [c]

[a] Warnell School of Forestry and Natural Resources, University of Georgia, 180 E Green Street, Athens, GA, 30602-2152, USA
[b] USDA Forest Service, Southern Research Station, Forest Inventory and Analysis, 4700 Old Kingston Pike, Knoxville, TN, 37919, USA
[c] University of Tennessee School of Natural Resources, 401 Agriculture and Natural Resources Building, Knoxville, TN, 37996, USA

ARTICLE INFO

Keywords:
Survival analysis
Southern Appalachian region
Network analysis
Harvesting
Time-to-event
Censored data

ABSTRACT

Survival analysis is composed of a group of analytical approaches that can be used to predict the occurrence of harvest activities, which provides insightful information about the dynamics of natural resources utilization in a region. Recently, random survival forest (RSF) has been proposed in survival analysis to capture the complex relationships among variables. The main objective of this study was to employ the RSF algorithm to examine the temporal evolution of tree harvest, accounting for stand and environmental variables. Specifically, the predictability of the RSF model was compared with the Cox proportional hazard (Cox) model, a popular model in survival analysis. Important variables in explaining the variation of harvest time were identified. Data collected by the USDA Forest Service, Forest Inventory and Analysis (FIA) program from permanent plots in the southern Appalachian region were utilized in the analysis. Results showed that the RSF model consistently outperformed the Cox model based on prediction accuracy. Among 14 variables examined, ownership, forest type, elevation, state, and slope emerged as most important. Utilizing only these five variables in a reduced model produced satisfactory prediction accuracy compared to the full model (i.e., the models with all variables included). The findings of this work provide insights for forest managers and policy makers to utilize survival analysis methods in understanding harvest activities at the regional scale.

1. Introduction

Survival analysis is composed of a group of analytical approaches that can be used to predict the occurrence of harvest activities, which provides insightful information about the dynamics of natural resources utilization in a region (Griess and Knoke, 2013; Melo et al., 2017; Paul et al., 2019). Survival analysis aims at analyzing the expected duration of time until the event of interest occurs (i.e., time-to-event analysis) (Lawless, 2003; Weathers and Cutler, 2017). The event of interest is tree harvest in this study, but it can also mean death, infection, and disturbance in different applications. Survival analysis has been used in the forestry field, mostly for predicting tree mortality in forest modeling (Knight et al., 2012; Reis et al., 2018; Solarik et al., 2012). Melo et al. (2017) provides an example of survival methodology applied to predicting harvest probabilities for permanent plots dominated by broadleaved, mixed, and coniferous forests in Quebec, Canada, using a Cox proportional hazards model. Due to the limited existing research, there is a need for further investigation into the application of survival analysis in predicting harvest occurrence.

Harvesting in forestry is influenced by various site factors, such as stand characteristics, site history, management practices, and timber markets. These factors are interconnected and may contribute to the variation observed in harvesting (Elad and Pertot, 2014; McCarthy-Neumann and Ibanez, 2012; Roberts et al., 2020). Melo et al. (2017) found that the probability of a stand being harvested is positively correlated with stand basal area but has a negative relationship with stem density or slope. In survival analysis, Cox (1972) proposed a proportional hazards model, denoted as Cox model, to incorporate covariates via a multiple linear regression. The Cox model has gained popularity since then because it can easily quantify the effect of different exploratory variables in predicting the occurrence of an event (Wang and Li, 2017). However, it has also been observed to yield inconsistent results and lower performance, primarily due to the complexity of solving the model, especially when dealing with numerous variables and nonlinear effects

* Corresponding author.
E-mail addresses: dms32073@uga.edu (D.M. Senevirathne), syang23@uga.edu (S. Yang), consuelo.brandeis@usda.gov (C. Brandeis), dhodges2@utk.edu (D.G. Hodges).

https://doi.org/10.1016/j.fecs.2024.100236
Received 15 February 2024; Received in revised form 29 July 2024; Accepted 2 August 2024

(Brieman, 1996). With increased computing power, random survival forest (RSF), an ensemble tree method for analysis of survival data was introduced by Ishwaran et al. (2008). As an extension of the original random forest algorithm, RSF captures the complex relationships among variables by constructing multiple decision trees. It has also proven advantageous for highly-dimensional data where a large number of predictor variables are included in model construction, addressing the above mentioned challenges (Wang and Li, 2017). Previous research has examined RSF and compared the performance with other survival analysis models in different fields, especially in medical research (Chen et al., 2013; Pickett et al., 2021; Qiu et al., 2020). However, to our knowledge, the predictability of the models for harvest occurrence has not been explicitly studied and compared in forestry, especially for mixed-species forests with various management objectives.

Therefore, the main objective of this study was to employ a random survival forest algorithm to examine the temporal evolution of tree harvest, accounting for stand and environmental variables (e.g., basal area, elevation). Specific objectives were to:

1) compare the prediction accuracy of the RSF model with the Cox model;
2) identify important variables in explaining the variation of harvest time;
3) construct a simplified RSF model with a reduced number of predictors, and evaluate the variation of harvest probabilities among regions, ownerships, and forest types.

As vital components of the global carbon cycle, the southern Appalachian region conserves critical and unique forest ecosystems. Forests in the region are primarily dominated by hardwood species, such as oaks (*Quercus* spp.), hickories (*Carya* spp.) and maples (*Acer* spp.). Data used in the analysis comes from the United States Department of Agriculture (USDA) Forest Service, Forest Inventory and Analysis (FIA) program's net of permanent sample forest inventory plots along the natural growing range of the mixed-species forests. The results of this study not only provide quantitative information about harvest activities in the region, but also improve current methodology in harvest prediction, which can inform decision-making in sustainable forest management.

2. Materials and methods

2.1. Study population and data

2.1.1. Dataset and sampling unit

The dataset used in this study was queried from the FIA database where measurements were collected from permanent plots installed in the southern Appalachian region. A total of six states were included: Alabama (AL), Georgia (GA), Kentucky (KY), North Carolina (NC), Tennessee (TN), and Virginia (VA). The geographical boundaries of the study area are shown in Fig. 1. This study area covers a diverse range of geologic formations, topographic features, soil compositions, climatic conditions, and associated vegetation.

To classify the forest types within this dataset, we employed the forest typing algorithm utilized by FIA (Burrill et al., 2021). This algorithm identified 13 primary mixed-species forest types (see Table 1). However, to ensure adequate sample sizes, forest types with fewer than 250 samples in each were aggregated into a single category as “other types”. Consequently, the dataset included seven forest type groups (Table 1).

Measurements taken from FIA permanent plots during 2000–2020 were selected. Each FIA plot includes four 7.3 m (24 feet) fixed-radius subplots which are remeasured every five to ten years. All trees with diameter at breast height (DBH) $\geq$ 12.7 cm (5 in.) are measured on a subplot. Due to constraints in data availability, to ensure an adequate sample size this study included plots located only on privately-owned lands and national forest lands.

2.1.2. Harvest events

To better capture the forest management at each installation, a subplot was treated as the basic sampling unit. Harvest in this study was defined as the removal of one or more trees on a subplot. The time to event is the duration between the year of the first inventory and the year at the last remeasurement (i.e., censor year) or the year where harvest occurs. The first measurement of a subplot was considered as the initial status of the sample. Since harvesting may have occurred any time within the time interval between two consecutive remeasurements, the middle year of the remeasurements was considered as the time of harvest in this study. For example, measurements were taken on subplot A in 2006 and 2014. No harvest was found in 2006, but trees were cut when the inventory crew re-visited the plot in 2014. The tree harvest could have occurred at any point of time during 2006 and 2014. In this case, the midpoint year (i.e., 2010) was taken as the harvest year, corresponding to a survival time of 4 years (i.e., 2010–2006 = 4).

In a situation where a subplot encounters multiple harvest events within the total study period (i.e., 2000–2020), each harvest cycle was treated as a separate interval. The remeasurements taken after the first harvest event served as the initial status record of the second interval. For instance, the inventory crew visited subplot A again in 2019, and another harvest event was found. The years of 2006 and 2014 serve as the initial status for the first and second intervals, respectively. A total of 23,025 unique intervals were used in analysis.

2.1.3. Predictor variables

Past studies indicated that including stand or site variables in the models can improve predictability. For example, Solarik et al. (2012) and Melo et al. (2017) highlighted the importance of DBH, height, basal area per unit area, forest types, and slope in harvest models. Based on literature and our knowledge, 14 stand and site variables were selected in model construction for each sampling unit (i.e., at the subplot level), which include: (1) state, (2) ownership, (3) slope, (4) forest type, (5) aspect, (6) elevation, (7) Simpson's diversity index, (8) stand density index (SDI), (9) quadratic mean diameter (QMD), (10) total tree volume per hectare, (11) basal area per hectare, (12) trees per hectare, (13) the shortest road distance between a plot and its closest mill, and (14) number of mills in the county. Variables (1)–(6) (i.e., from state to elevation) were obtained from the FIA database, and variables (7)–(12) were computed from the inventoried trees using the FIA data. The shortest distance from a plot to a mill (i.e., variable 13) was calculated using the network analysis tool in ArcGIS Pro (version 3.2) (ESRI, 2023). The unpaved roads, prohibited through traffic, private roads and gates were avoided in the used road networks. The information of plot and mill locations were sourced from the FIA database and the FIA National Resource Use Monitoring (NRUM) website. A summary of variables across ownerships, states, and forest types is given in Table 2, 3, 4 and 5.

2.2. Models for survival analysis

Harvest events were predicted by two models: random survival forest (RSF) and Cox proportional hazard model (Cox model). Detailed description of each model is given below.

2.2.1. Random survival forest (RSF)

Random survival forest is an ensemble tree-based algorithm that can be applied for survival data and is an extension of the random forest method (Ishwaran et al., 2014). Predictions are made by aggregating survival trees containing cumulative hazard function (CHF), indicating the cumulative risk/probability of a harvesting event occurrence over given time periods. The construction of survival trees using RSF involves recursive partitioning of the covariate space using binary splits, leading to the formation of clusters with similar survival outcomes. Binary categorical variables are directly handled with straightforward binary splits. For multi-level categorical variables, the RSF model considers up to a user-defined maximum number of possible binary splits (nsplit = 5 set in

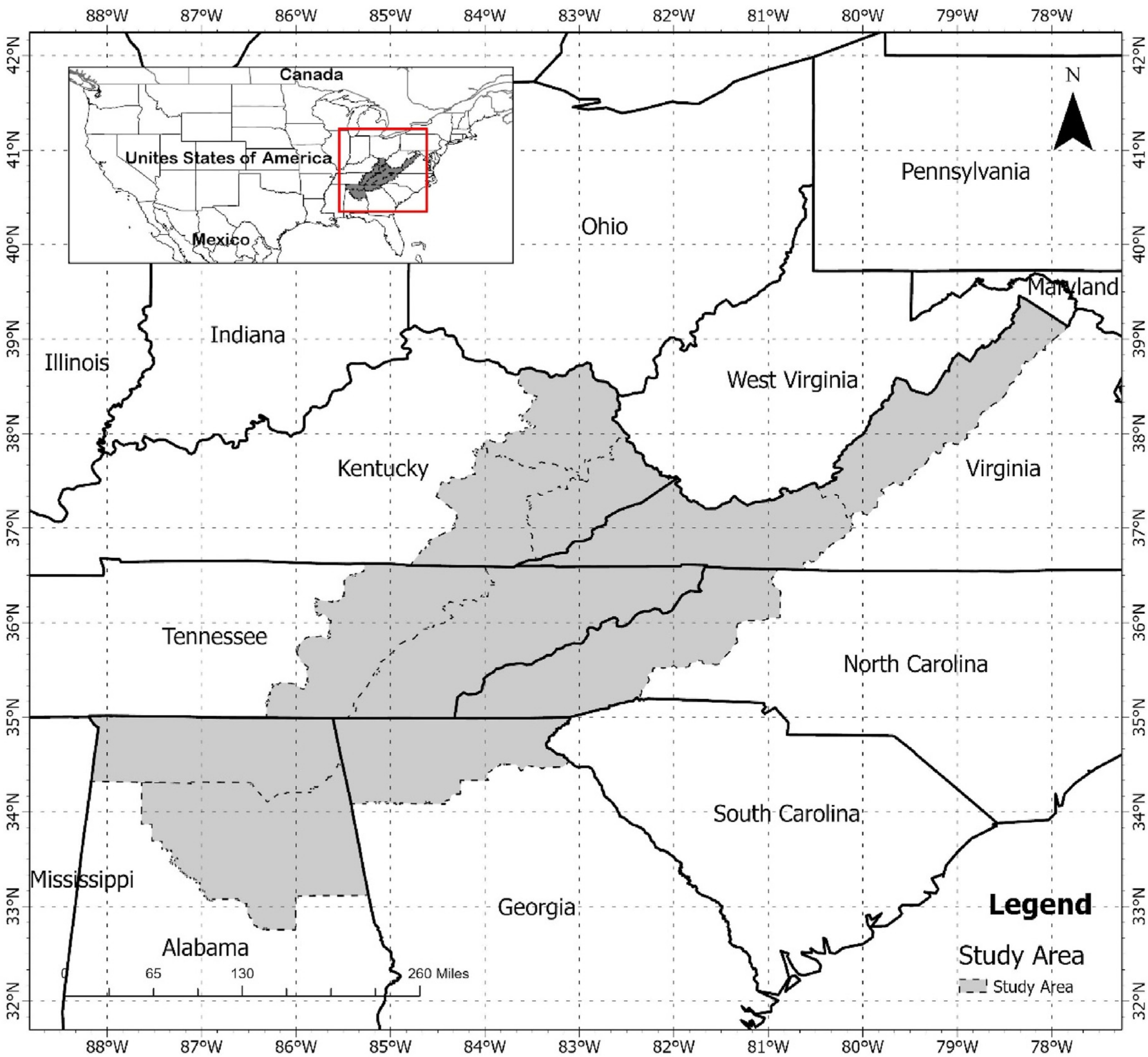

Fig. 1. Map of the study area.

this study) for each variable within a single node. It then selects the optimal split that separates the variable levels into two groups with the most similar survival outcomes within each group. Each survival tree is built using a bootstrap sample randomly selected with replacement, as well as a subset of features sampled at each split (Nasejje et al., 2017). The remaining unselected samples (referred to as out-of-bag data) are preserved for validation purposes. At each bootstrap sample predictions are made for out-of-bag data to calculate the prediction error for the aggregated CHF.

The RSF model was implemented using the rfsrc function in the "randomForestSRC" package in R (Ishwaran and Kogalur, 2023). To improve the model predictive outcomes, a proper selection of hyper-parameters is important in building RSF models. Hyperparameters in a RSF model include number of survival trees (ntrees), number of features randomly selected at a given split (mtry), number of random potential split points considered at each node (nsplit), and minimum number of observations that are required at each node to split further (minsplit). Based on multiple preliminary runs and guidelines from literature (Dauda, 2022; Hao et al., 2021), the hyperparameters were determined as ntree = 600, mtry = 4 (default $\approx \sqrt{\text{number of variables}}$), nsplit = 5, samptype = "swr" (sample without replacement), minsplit = 1% of data and nodesize = 50. The default values were used for the remaining parameters.

Table 1
Number of events among primary forest types in the study.

Primary forest type	Number of events
Oak-hickory	**16,317**
Loblolly-shortleaf pine	**2626**
Oak-pine	**2210**
Maple-beech-birch	**764**
White-red-jack pine	**394**
Elm-ash-cottonwood	**272**
Other	**442**
Oak-gum-cypress	144
Other hardwoods	120
Other eastern softwoods	81
Longleaf-slash pine	53
Exotic hardwoods	26
Spruce-fir	13
Aspen-birch	5

2.2.2. *Cox proportional hazard model (Cox model)*

The Cox model is a semi-parametric regression method for analyzing survival data. It computes the ratio of the event occurrence rates between treated and control groups, indicating the probability of event occurrence, given that it still has not occurred divided by the time interval. This computed ratio is generally known as hazard ratio (HR) in survival analysis. Hazard ratio provides the information about the influence of

Table 2

Summary of variables in national forests across states and harvest status. From columns "Time to event (years)" to "Distance to the closest mill (km)", mean is given followed by standard deviation in parentheses.

State	Harvested?	Number of events	Time to event[a] (years)	Slope (degrees)	Elevation (m)	Aspect (degrees)	Total volume per ha ($m^3 \cdot ha^{-1}$)	Basal area per hectare ($m^2 \cdot ha^{-1}$)	Trees per hectare (ha^{-1})	QMD (cm)	**SDI (trees·ha^{-1})**	Diversity	Distance to the closest mill (km)	Number of mills in the county
AL	Yes	28	8.5 (4.2)	14.6 (8.1)	266.5 (68.6)	180.1 (77.8)	30.2 (26.7)	5.3 (2.7)	559.5 (486.2)	18.0 (7.3)	244.8 (140.1)	0.7 (0.2)	33.0 (11.1)	3.5 (0.9)
	No	263	13.1 (3.6)	22.7 (12.2)	283.7 (73.2)	204.2 (94.6)	38.5 (27.7)	5.7 (2.9)	384.3 (356.0)	23.3 (7.4)	263.4 (189.2)	0.8 (0.1)	31.9 (12.3)	2.7 (1.1)
GA	Yes	7	6.6 (3.0)	15.0 (9.5)	347.9 (52.4)	146.0 (101.7)	26.9 (11.3)	5.7 (2.8)	590.8 (389.8)	21.0 (5.0)	383.2 (212.0)	0.6 (0.2)	14.7 (1.8)	2.9 (0.4)
	No	1308	8.9 (4.0)	37.6 (17.0)	692.8 (221.9)	192.3 (100.1)	52.6 (34.6)	7.5 (3.4)	326.9 (303.9)	26.1 (7.5)	287.5 (202.1)	0.8 (0.2)	25.3 (11.2)	1.7 (0.9)
KY	Yes	19	9.3 (3.7)	34.0 (15.4)	370.3 (63.8)	179.0 (78.2)	37.4 (26.6)	6.2 (2.5)	528.2 (330.2)	23.1 (6.0)	380.9 (154.9)	0.8 (0.1)	22.1 (12.3)	3.2 (1.2)
	No	398	13.0 (4.3)	36.1 (16.5)	365.3 (63.8)	183.4 (101.4)	44.0 (31.2)	6.4 (3.1)	404.8 (420.7)	24.8 (8.3)	300.6 (214.1)	0.8 (0.2)	24.0 (10.5)	2.8 (1.3)
NC	Yes	16	4.3 (4.3)	42.8 (12.9)	1099.8 (179.2)	174.9 (138.9)	14.0 (18.6)	5.0 (3.0)	1072.4 (1004.2)	13.8 (5.0)	301.3 (182.0)	0.7 (0.1)	20.2 (6.5)	5.8 (4.1)
	No	697	15.9 (3.4)	43.8 (18.4)	954.1 (288.3)	194.8 (92.8)	47.0 (35.5)	7.0 (3.7)	334.2 (349.9)	25.9 (9.1)	270.2 (224.9)	0.8 (0.2)	29.1 (16.0)	3.4 (2.5)
TN	Yes	17	8.5 (4.7)	39.2 (10.7)	833.0 (368.8)	166.5 (65.9)	32.9 (21.4)	6.5 (2.7)	542.0 (529.2)	20.8 (5.7)	317.2 (227.3)	0.8 (0.1)	27.8 (8.5)	2.0 (0.9)
	No	1192	9.3 (3.7)	46.2 (20.1)	732.3 (268.8)	188.7 (106.2)	50.6 (37.9)	7.3 (3.7)	369.6 (375.9)	25.2 (8.5)	290.5 (216.3)	0.8 (0.2)	25.0 (13.4)	1.9 (0.6)
VA	Yes	33	9.4 (5.3)	20.2 (8.6)	699.1 (279.3)	170.0 (64.1)	43.2 (27.0)	6.8 (3.2)	403.1 (321.0)	24.0 (7.2)	308.1 (229.9)	0.7 (0.2)	22.8 (9.4)	2.3 (0.9)
	No	1838	10.8 (5.0)	36.9 (17.5)	745.4 (209.0)	180.5 (104.3)	40.3 (28.1)	6.8 (3.2)	349.1 (319.3)	25.2 (8.1)	283.3 (208.9)	0.7 (0.2)	33.6 (15.9)	2.1 (1.1)

[a] or censor year for non-harvested plots.

Table 3

Summary of variables in private lands across states and harvest status. From columns "Time to event (years)" to "Distance to the closest mill (km)", mean is given followed by standard deviation in parentheses.

State	Harvested?	Number of events	Time to event[a] (years)	Slope (degrees)	Elevation (m)	Aspect (degrees)	Total volume per ha($m^3 \cdot ha^{-1}$)	Basal area per hectare ($m^2 \cdot ha^{-1}$)	Trees per hectare (ha^{-1})	QMD (cm)	**SDI (trees·ha^{-1})**	Diversity	Distance to the closest mill (km)	Number of mills in the county
AL	Yes	1224	6.1 (3.7)	15.3 (10.0)	224.7 (69.7)	181.4 (100.1)	29.9 (24.5)	5.6 (3.2)	548.4 (498.7)	19.1 (7.4)	283.0 (217.1)	0.6 (0.3)	28.8 (14.3)	2.3 (1.3)
	No	3083	12.6 (3.8)	18.4 (13.1)	234.1 (81.6)	192.8 (101.7)	28.3 (25.0)	4.9 (3.0)	501.9 (501.1)	19.9 (8.5)	255.2 (207.8)	0.7 (0.2)	28.3 (14.2)	2.3 (1.5)
GA	Yes	269	5.5 (3.4)	17.7 (14.0)	312.4 (104.4)	181.2 (92.5)	30.5 (27.5)	5.5 (3.3)	462.0 (373.0)	19.9 (7.4)	265.0 (199.8)	0.6 (0.2)	24.9 (14.9)	2.2 (1.1)
	No	1297	12.8 (3.3)	23.3 (16.3)	388.7 (154.1)	180.8 (103.9)	35.5 (28.4)	5.6 (3.2)	475.3 (547.2)	21.2 (8.0)	272.4 (214.2)	0.8 (0.2)	20.2 (12.1)	2.2 (1.1)
KY	Yes	578	6.3 (4.2)	41.5 (16.0)	361.3 (139.2)	195.6 (100.4)	40.9 (26.3)	6.0 (3.0)	357.3 (319.3)	25.1 (7.6)	294.6 (222.2)	0.8 (0.2)	24.4 (13.1)	2.2 (1.7)
	No	2804	12.2 (4.0)	42.8 (18.4)	356.8 (117.2)	190.4 (100.4)	32.1 (25.3)	5.1 (3.0)	395.9 (370.7)	22.6 (8.7)	263.2 (214.3)	0.8 (0.2)	24.9 (13.0)	2.2 (1.7)
NC	Yes	244	7.4 (4.1)	31.8 (17.6)	728.2 (303.3)	187.5 (102.4)	52.3 (39.2)	7.5 (3.6)	385.5 (419.4)	25.5 (8.0)	303.0 (218.4)	0.7 (0.2)	19.1 (9.3)	3.3 (2.7)
	No	1715	15.1 (3.7)	37.6 (18.3)	762.8 (285.3)	176.9 (104.3)	43.0 (35.5)	6.6 (3.8)	390.3 (395.6)	24.2 (8.6)	281.2 (214.2)	0.8 (0.2)	19.2 (11.9)	3.5 (2.4)
TN	Yes	507	5.9 (3.6)	25.5 (16.7)	448.3 (141.9)	194.7 (108.5)	42.6 (31.4)	5.9 (3.2)	334.9 (308.8)	24.7 (8.3)	268.4 (207.9)	0.8 (0.2)	23.9 (11.2)	2.0 (1.1)
	No	2963	12.2 (3.6)	27.2 (18.5)	439.7 (146.4)	188.7 (102.3)	35.8 (28.6)	5.4 (3.2)	386.5 (389.0)	22.6 (8.7)	252.9 (204.3)	0.8 (0.2)	23.6 (12.4)	2.0 (1.0)
VA	Yes	331	6.3 (4.0)	30.2 (19.7)	621.8 (209.5)	188.9 (105.5)	46.1 (33.4)	6.7 (3.5)	293.1 (261.3)	25.8 (8.1)	258.8 (193.5)	0.8 (0.2)	29.1 (14.2)	2.0 (1.1)
	No	2405	13.5 (3.7)	36.7 (21.0)	618.5 (199.1)	183.6 (106.0)	35.5 (28.5)	5.7 (3.4)	368.3 (414.2)	23.3 (8.7)	250.6 (211.4)	0.7 (0.2)	29.2 (13.7)	1.8 (1.0)

[a] or censor year for non-harvested plots.

Table 4
Summary of variables in national forests across ownerships, forest type, and harvest status. From columns “Time to event (years)” to “Distance to the closest mill (km)”, mean is given followed by standard deviation in parentheses.

Forest type[b]	Harvested?	Number of events	Time to event[a] (years)	Slope (degrees)	Elevation (m)	Aspect (degrees)	Total volume per ha ($m^3 \cdot ha^{-1}$)	Basal area per hectare ($m^2 \cdot ha^{-1}$)	Trees per hectare (ha^{-1})	QMD (cm)	**SDI (trees·ha^{-1})**	Diversity	Distance to the closest mill (km)	Number of mills in the county
WRJP	No	169	10.0 (4.2)	36.1 (20.8)	603.5 (188.5)	190.7 (106.7)	71.1 (46.4)	9.2 (4.2)	370.9 (301.2)	27.2 (7.4)	358.9 (249.6)	0.7 (0.2)	27.7 (16.4)	1.8 (0.9)
LSP	Yes	12	7.5 (3.7)	13.5 (6.2)	365.0 (139.2)	193.6 (57.1)	29.1 (14.0)	6.0 (2.0)	522.6 (270.4)	20.6 (3.8)	341.0 (111.7)	0.7 (0.1)	29.4 (10.7)	2.8 (0.8)
	No	325	9.8 (4.4)	32.3 (17.0)	491.2 (169.5)	193.7 (85.2)	43.8 (30.7)	7.2 (3.2)	513.5 (419.0)	21.3 (6.2)	325.1 (208.4)	0.7 (0.2)	28.8 (12.7)	1.8 (0.9)
OP	Yes	18	9.4 (4.2)	17.2 (9.0)	267.7 (67.4)	179.6 (92.9)	31.1 (30.1)	5.9 (2.8)	596.2 (578.2)	18.8 (6.9)	250.7 (147.4)	0.7 (0.2)	32.7 (14.3)	4.0 (0.0)
	No	622	10.6 (4.5)	35.8 (19.0)	583.5 (185.8)	207.3 (86.5)	42.3 (33.5)	6.7 (3.3)	406.5 (383.6)	23.5 (7.4)	293.0 (207.2)	0.8 (0.2)	28.9 (14.8)	2.0 (1.1)
OH	Yes	82	8.6 (4.8)	32.3 (14.3)	686.1 (323.5)	164.8 (86.5)	32.9 (24.7)	6.1 (2.8)	586.4 (606.6)	20.9 (7.1)	313.3 (181.3)	0.7 (0.2)	21.6 (8.8)	3.3 (2.4)
	No	4312	11.1 (4.9)	40.4 (18.2)	726.2 (262.1)	185.7 (103.1)	45.7 (32.2)	6.9 (3.3)	332.4 (324.0)	25.8 (8.3)	276.1 (204.5)	0.8 (0.2)	28.3 (14.4)	2.2 (1.4)
EAC	No	4	13.5 (1.7)	25.0 (0.0)	466.3 (130.1)	117.0 (0.0)	18.4 (18.8)	3.4 (3.2)	457.5 (505.3)	22.0 (11.0)	235.9 (339.0)	0.8 (0.2)	26.7 (13.9)	2.2 (1.5)
MBB	No	194	11.1 (5.2)	39.7 (20.4)	1095.6 (339.5)	166.0 (128.9)	55.5 (40.0)	7.9 (4.1)	322.6 (405.7)	28.8 (9.0)	316.1 (284.3)	0.7 (0.2)	33.9 (14.8)	2.0 (0.8)
Other	Yes	8	2.5 (0.5)	7.5 (2.7)	787.9 (640.3)	217.5 (13.4)	44.6 (42.3)	5.1 (5.0)	576.1 (388.3)	18.7 (12.1)	358.0 (417.0)	0.7 (0.2)	34.8 (1.7)	2.0 (0.0)
	No	66	12.5 (4.6)	32.3 (18.2)	950.0 (464.4)	177.7 (112.8)	48.5 (29.1)	6.8 (3.1)	346.7 (308.2)	24.0 (7.6)	268.5 (209.1)	0.7 (0.2)	31.1 (14.1)	3.9 (3.7)

[a] or censor year for non-harvested plots.
[b] WRJP - white/red/jack pine, LSP - loblolly/shortleaf pine, OP - oak/pine, OH - oak-hickory, EAC - elm/ash/cottonwood, MBB - maple/beech/birch.

Table 5
Summary of variables in private lands across ownerships, forest type, and harvest status. From columns “Time to event (years)” to “Distance to the closest mill (km)”, mean is given followed by standard deviation in parentheses.

Forest type[b]	Harvested?	Number of events	Time to event[a] (years)	Slope (degrees)	Elevation (m)	Aspect (degrees)	Total volume per ha ($m^3 \cdot ha^{-1}$)	Basal area per hectare ($m^2 \cdot ha^{-1}$)	Trees per hectare (ha^{-1})	QMD (cm)	**SDI (trees·ha^{-1})**	Diversity	Distance to the closest mill (km)	Number of mills in the county
WRJP	Yes	45	6.3 (3.8)	20.9 (12.1)	691.6 (251.2)	215.3 (105.7)	57.0 (45.6)	8.5 (3.9)	440.6 (648.3)	24.8 (9.4)	287.2 (198.3)	0.6 (0.3)	19.6 (9.2)	2.7 (1.7)
	No	180	12.6 (4.1)	28.8 (15.3)	654.7 (205.3)	180.7 (96.5)	45.7 (38.6)	6.8 (4.2)	333.8 (359.8)	24.5 (8.3)	267.4 (220.7)	0.7 (0.2)	22.6 (12.5)	2.1 (1.2)
LSP	Yes	945	5.5 (3.4)	14.5 (10.1)	251.4 (107.3)	185.2 (96.6)	28.6 (22.8)	5.8 (3.2)	570.4 (508.0)	18.2 (6.7)	282.4 (220.8)	0.5 (0.2)	28.3 (14.7)	2.0 (1.0)
	No	1344	11.7 (4.0)	18.4 (14.6)	300.0 (141.9)	195.3 (99.5)	29.6 (26.7)	5.2 (3.3)	584.0 (650.6)	18.3 (7.6)	270.5 (224.7)	0.6 (0.2)	25.9 (13.0)	2.2 (1.2)
OP	Yes	243	6.7 (3.9)	17.4 (12.2)	331.4 (163.9)	174.1 (99.9)	31.1 (26.0)	5.3 (3.1)	535.1 (427.0)	19.9 (7.8)	306.0 (215.0)	0.8 (0.2)	26.1 (13.8)	2.5 (1.9)
	No	1327	12.8 (3.9)	24.3 (16.6)	387.0 (198.8)	189.5 (95.6)	30.2 (27.2)	5.1 (3.2)	503.8 (489.9)	19.7 (8.1)	265.5 (214.7)	0.8 (0.2)	25.1 (13.6)	2.2 (1.3)
OH	Yes	1725	6.5 (4.0)	31.1 (18.2)	436.1 (229.9)	189.6 (101.4)	42.1 (31.0)	6.1 (3.3)	355.4 (325.0)	24.6 (8.0)	279.7 (211.1)	0.8 (0.2)	25.1 (12.8)	2.4 (1.7)
	No	10198	13.1 (3.8)	33.5 (19.7)	466.5 (238.6)	185.4 (103.1)	35.3 (28.4)	5.5 (3.3)	398.2 (395.1)	22.8 (8.6)	262.1 (207.6)	0.8 (0.2)	24.7 (13.4)	2.4 (1.7)
EAC	Yes	23	5.2 (3.4)	6.5 (5.5)	227.8 (73.1)	127.3 (94.7)	43.4 (30.3)	6.3 (3.5)	248.8 (222.3)	27.1 (8.2)	221.3 (154.5)	0.7 (0.2)	21.5 (13.1)	1.6 (0.5)
	No	245	13.0 (3.6)	15.7 (18.0)	311.0 (148.8)	172.4 (93.4)	35.0 (36.6)	5.5 (4.1)	317.6 (346.1)	23.8 (11.3)	221.6 (227.9)	0.7 (0.2)	24.1 (14.6)	2.3 (1.6)
MBB	Yes	97	5.5 (3.8)	42.1 (18.6)	516.1 (258.2)	206.8 (145.1)	46.3 (31.4)	6.2 (3.3)	219.4 (199.9)	28.8 (8.3)	237.3 (198.7)	0.7 (0.2)	25.3 (17.0)	2.2 (1.3)
	No	473	12.8 (4.1)	46.4 (21.1)	520.7 (273.1)	195.6 (124.9)	35.7 (26.1)	5.4 (3.3)	298.2 (297.5)	25.4 (9.4)	245.4 (216.7)	0.7 (0.2)	28.5 (14.9)	1.9 (1.5)
Other	Yes	53	5.6 (3.4)	21.6 (14.3)	303.2 (197.0)	170.1 (87.6)	40.4 (40.6)	6.2 (4.3)	452.1 (458.6)	22.2 (8.9)	257.0 (179.2)	0.7 (0.2)	32.2 (11.3)	2.7 (2.0)
	No	315	13.2 (3.9)	25.7 (19.4)	442.3 (307.5)	175.1 (108.3)	29.8 (29.8)	4.7 (3.5)	354.1 (361.8)	20.7 (9.2)	204.1 (185.5)	0.7 (0.2)	26.7 (13.0)	2.2 (1.7)

[a] or censor year for non-harvested plots.
[b] WRJP - white/red/jack pine, LSP - loblolly/shortleaf pine, OP - oak/pine, OH - oak-hickory, EAC - elm/ash/cottonwood, MBB - maple/beech/birch.

Table 6
List of predictor variables included in the full and simplified models.

Models	Variables[a]
Model 1 (Full model)	ST, OWN, SLP, FT, ASP, ELV, SD, SDI, QMD, AV, BAPH, TPH, DCM, NM
Model 2	ST, OWN, SLP, FT, ASP, ELV, AV, DCM, NM
Model 3	ST, OWN, SLP, FT, ELV, NM
Model 4	ST, OWN, SLP, FT, ELV
Model 5	ST, OWN, FT, ELV
Model 6	OWN, FT

[a] Variables: State - ST, ownership - OWN, slope - SLP, forest type - FT, aspect – ASP, elevation - ELV, Simpsons diversity - SD, Reineke's stand diversity index – SDI, quadratic mean diameter - QMD, average volume - AV, basal area per hectare - BAPH, trees per hectare - TPH, distance to the closest mill - DCM, number of mills in the county - NM.

different levels of variables on event occurrence (Spruance et al., 2004). The Cox model assumes that the HR remains proportional over time for different strata or groups, such as forest types, indicating that the probability of a plot being harvested changes at a consistent rate along time across these groups. In the current study, the Cox model was employed to explore the feasibility of predicting occurrences of harvesting events. These analyses determine the effects of predictor variables on event probabilities by estimating hazard ratios, thus enabling the estimation of the harvesting event occurrence probability. The model was constructed using the "survival" package in R (Therneau et al., 2023).

2.3. Model construction and evaluation

2.3.1. Procedure

To account for the potential spatial correlation of subplots at each plot installation (i.e., a whole FIA sample plot), a cluster bootstrap technique (Davidson and Hinkley, 1997; Field and Welsh, 2007; Ren et al., 2010) was used in model construction and evaluation. Detailed steps are given as.

Step 1: All FIA sample plots were randomly selected with replacement. All subplots within the selected FIA plots were used to construct a bootstrap sample.
Step 2: Models (RSF and Cox) were constructed using the bootstrap sample generated in step 1. An identical set of predictor variables listed in "Predictor variables (section 2.1.3)" was used to build both models. Statistics listed in section 2.3.2 "Evaluation statistics" were computed.
Step 3: Steps 1–2 were repeated 1,000 times (i.e., 1,000 bootstrap samples).
Step 4: For each statistic, point estimate, upper and lower boundaries of the 95% confidence intervals were computed from the median, 2.5% and 97.5% quantiles from 1,000 bootstrap samples.

2.3.2. Evaluation statistics

The predictability of the models was evaluated based on discrimination and calibration. Discrimination refers to the capability of predictive models in categorizing data into distinct classes, while calibration measures the agreement between the model's anticipated probabilities and the inherent probabilities within the studied population. Such

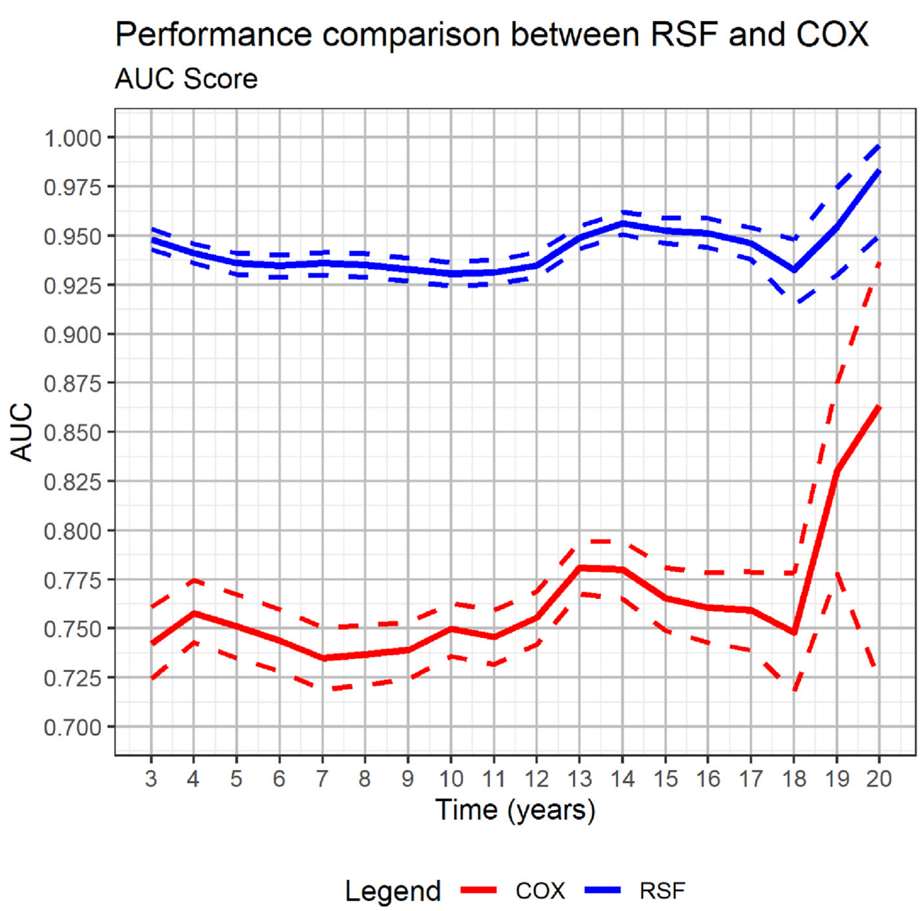

Fig. 2. Median (solid lines) and 95% confidence intervals (dashed lines) of AUC curves (area under the receiver operating characteristic curve) between random survival forest (RSF) and Cox proportional hazard (Cox) models.

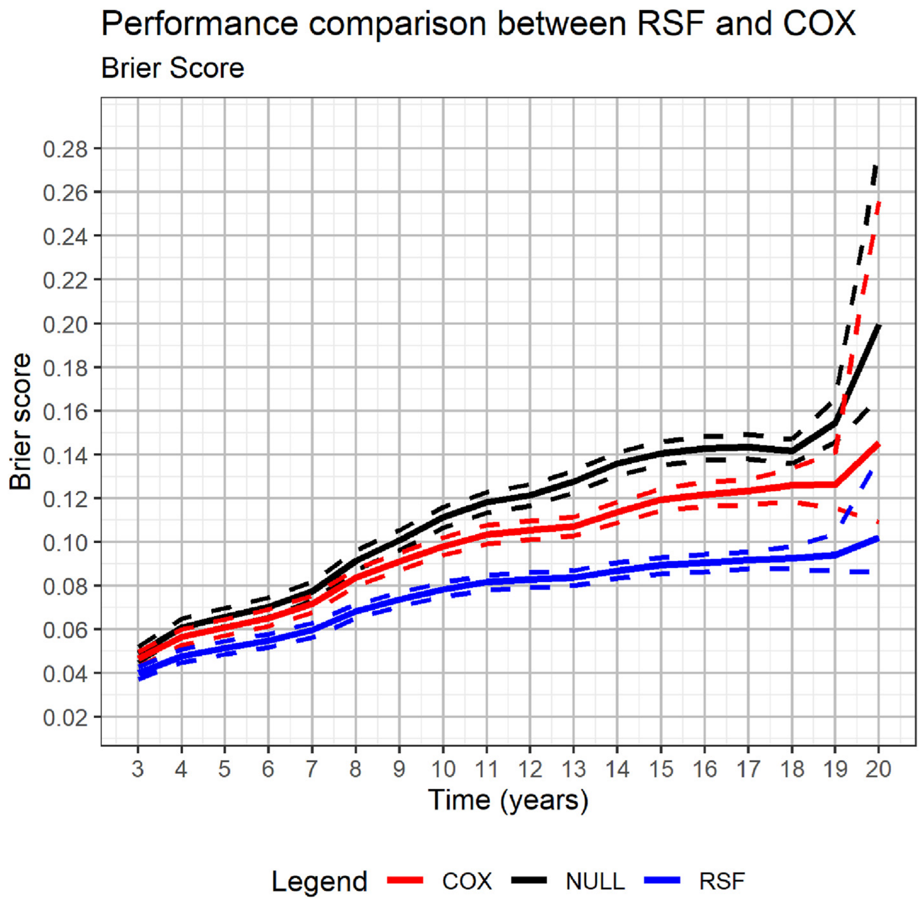

Fig. 3. Median (solid lines) and 95% confidence intervals (dashed lines) of Brier score curves among random survival forest (RSF), Cox proportional hazard (Cox) and null models, where null model was constructed with no predictor variables included.

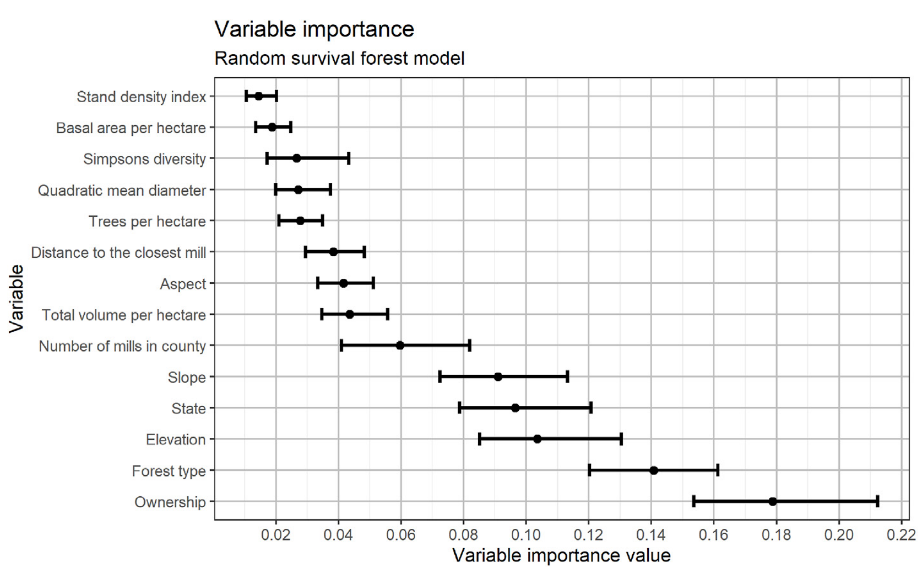

Fig. 4. Median (point) and 95% confidence limit (whiskers) of variable importance (VIMP) values generated using the random survival forest (RSF) model.

assessments can be facilitated by the utility of prediction error curves, a practice demonstrated in comparative model performance evaluations (Hu et al., 2023; Luck et al., 2017; Pickett et al., 2021; Wilstrup and Cave, 2022). In survival analysis, discrimination and calibration are commonly quantified using two statistics: area under the receiver operating characteristic curve (AUC, also known as AUROC) and Brier score. The AUC quantifies a model's proficiency in selected individuals with varying probabilities to experience an event within a designated prediction time window. This assessment depends upon the condition that these individuals remain event-free until the specific time point. Occupying a scale from 0 to 1, a higher AUC value signifies an enhanced ability to discriminate between event occurrences and non-occurrences within the

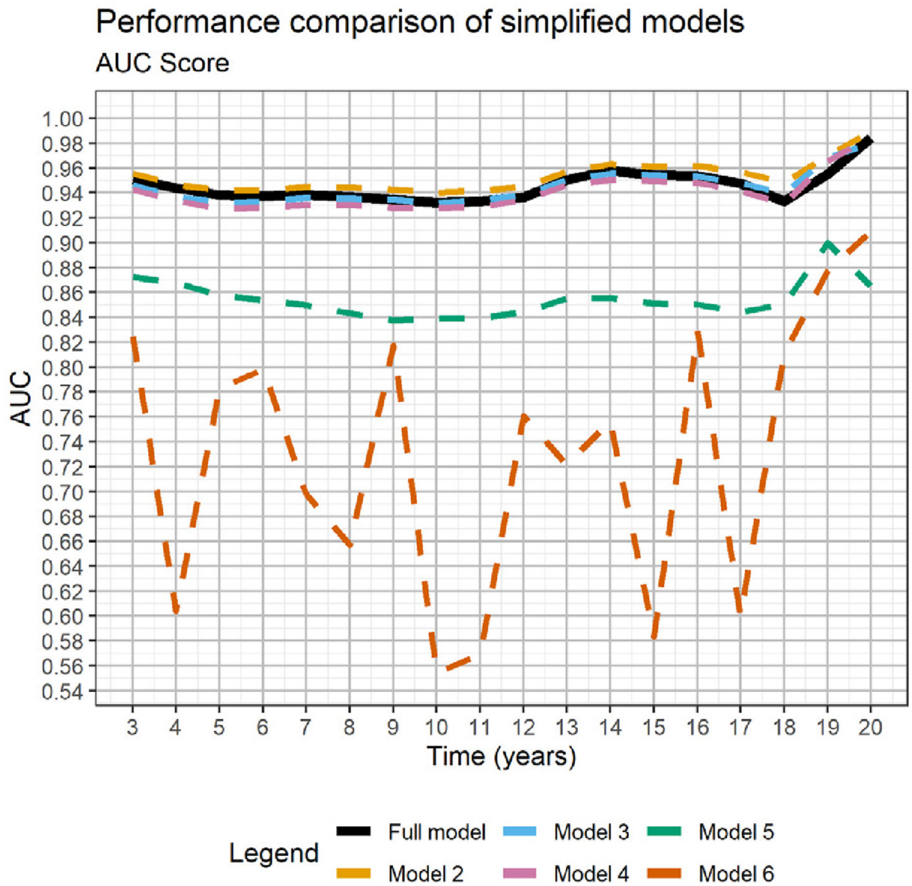

Fig. 5. Predictive performance comparison of simplified (based on variable importance) vs full random survival forest (RSF) models using area under the ROC curve (AUC) values.

stipulated time frame. The Brier score evaluates model's predictive precision by capturing the mean square error between the model's projected event probability and the actual event indicator. Concentrating on individuals who remain event-free within the prediction window, a lower dynamic Brier score signals better predictive accuracy. To better understand the predictive accuracy of our models, we incorporated a null model into the analysis. The null model, acting as a baseline for comparison, represents a model constructed without any predictor variables (i.e., Kaplan-Meier estimator). In this context, the null model assumes a constant harvesting probability across all observations, regardless of ownership, forest type, or other variables. The statistics of AUC and Brier score were computed using "riskRegression" (Gerds et al., 2023) in R. The 95% confidence intervals were used to compare the difference in prediction accuracy over time between both RSF and Cox models, incorporating complete set of variables.

In addition, a further investigation was conducted to evaluate the importance of variables in RSF model, which was measured using variable importance (VIMP) function in the "randomForestSRC" package in R (Ishwaran and Kogalur, 2023). VIMP measures the decrease in prediction accuracy for out-of-bag cases when a variable is excluded from the model. Variable importance values serve as a relative measure of the significance of each variable, with a higher value indicative of greater importance.

2.3.3. Simplifying RSF models with a reduced number of variables

With the relatively important variables identified, we constructed simpler models to investigate the possibility of achieving accurate predictions with limited information. The least important variables were excluded from the full model, resulting in 5 alternative, simpler models. These simplified models were evaluated for accuracy reduction compared to the full model, by constructing and comparing AUC and Brier score graphs. Table 6 details the variables included in each simplified model.

2.3.4. Evaluating the variation of harvest probabilities over time

With the optimal models identified, the variation of harvest probabilities across ownership, forest type, elevation, and state were evaluated. The optimal-simplified RSF model was used to generate survival curves (depicting changing survival probabilities over time).

3. Results

3.1. Comparing predictability between the RSF and cox models

Based on AUC and Brier score, the predictability of random survival forest model is consistently better than the Cox proportional hazards model (see Figs. 2 and 3). For a given model, similar AUC values over time imply that the model maintains a consistent level of discrimination capability over time. Notably, RSF has a significantly higher AUC than Cox model over 20 years (see 95% confidence intervals in Fig. 2), which indicates that the prediction produced by RSF is more reliable than the Cox model. Furthermore, an obvious increase of AUC was observed after 18 years ($t > 18$ yrs) for both models. The increase of prediction accuracy may be due to the absence of harvest events observed over 18 years.

In terms of prediction precision, a gradual increase of Brier score over time was found for both models, indicating that the predicted survival probabilities diverge from the actual status with an increase of prediction time. RSF provided significantly more precise predictions than Cox

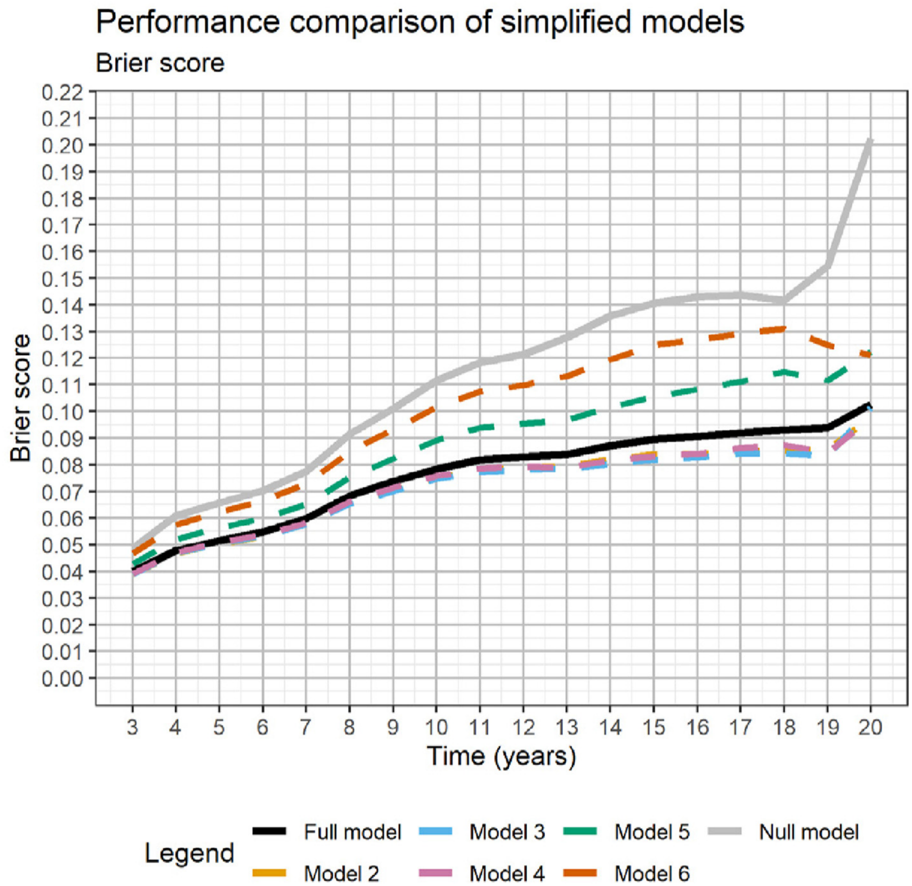

Fig. 6. Predictive performance comparison of simplified (based on variable importance) vs full random survival forest (RSF) models using Brier score.

model except when $t > 19$ (see 95% confidence intervals in Fig. 3). Notably, the inclusion of predictor variables in building RSF models led to a substantial improvement in prediction accuracy, as evidenced by a significantly lower Brier score compared to the null model (Fig. 3). It is important to note that the null model is not graphically displayed for AUC, but it is significantly lower as the AUC calculations treats a score of 0.5 (representing a 50% probability) as the null value.

The difference in Brier score becomes even more pronounced over time, suggesting that RSF models effectively capture patterns in harvesting probabilities. A similar trend was found for the Cox model. However, the 95% confidence intervals of the Cox and null models overlap when $t < 9$ and $t > 19$, suggesting that the Cox model's prediction accuracy is not significantly superior to the null model at these specific time frames considered.

3.2. Simplifying RSF models with a reduced number of predictors

Among the 14 variables, ownership and forest type stand out as the first and second significantly higher importance variables, followed by elevation, state, and slope (see Fig. 4). Notably, the number of mills, despite its lower overall importance, did not show significantly lower impact compared to slope and state. All the remaining variables show considerably lower importance and significant differences are not observed amongst each other.

Based on AUC and Brier score (see Figs. 5 and 6), the prediction accuracy remains at a relatively similar level even when 8 variables are removed (model 4), utilizing only the 5 most important ones (ownership, forest type, elevation, state, and slope). However, further simplification to just the top 4 variables (ownership, forest type, elevation, and state) results in both lower AUC and Brier score values, indicating lower prediction accuracy compared to the full model and other models. This signifies the insufficiency of relying on less than 5 important variables.

3.3. Comparison of harvesting probabilities across ownership, forest type, state, and elevation

The estimated survival curves (see Fig. 7, 8, 9 and 10) reveal variations in harvesting probability predictions across different land ownerships, forest types, elevations, and states. Notably, privately owned forests are consistently harvested earlier than national forests. In Virginia's oak-pine forests, for example, the probability of harvesting privately owned land exceeds 30%, while national forests have less than a 10% probability. In Alabama, the difference narrows down to 5% for oak-hickory forests, whereas in oak-pine forests, harvesting probabilities for the two ownership types converge after 14 years. However, the overall survival probability pattern holds: Private lands exhibit a higher likelihood of harvest within shorter timeframes compared to national forests, where plots are expected to be harvested at longer intervals.

Within privately-owned lands, the influence of forest type and state on harvesting probabilities reveal diverse patterns across these categories. For example, elm-ash-cottonwood forests show the highest harvesting probability in Alabama, Kentucky, and Tennessee, while Virginia claims higher harvesting probability among oak-pine forests. Conversely, in Georgia and North Carolina white-red-jack pine forests show the highest. National forests, in contrast, exhibit less variability in harvesting probabilities across states and forest types.

Figs. 9 and 10 suggest that, in most cases, a clear distinction in harvesting probabilities between higher and lower elevations is not evident. However, notable differences are observed across forest types within Alabama and Georgia. Additionally, North Carolina exhibits slightly

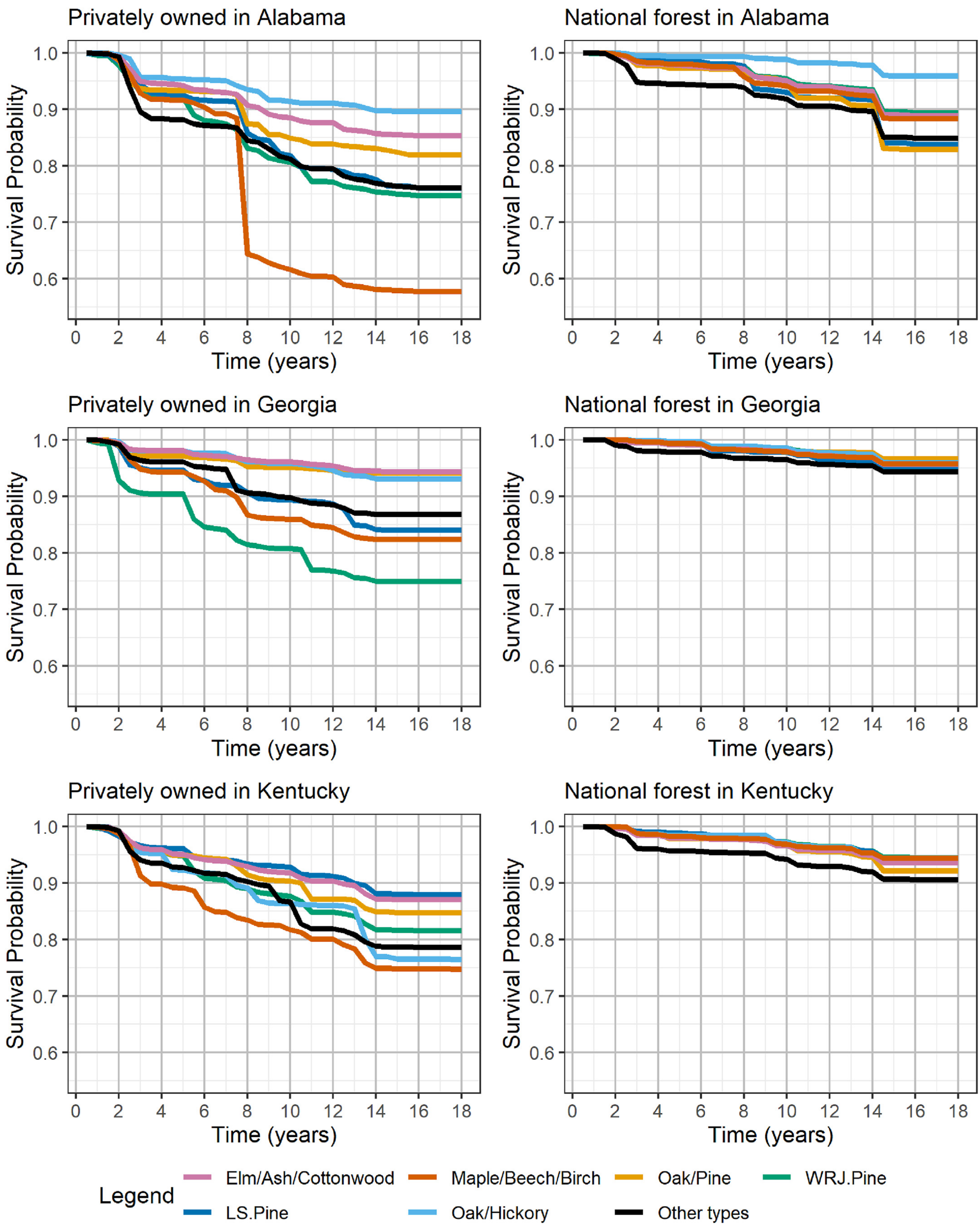

Fig. 7. Survival probability change over time across different forest types. Column 1: privately owned, column 2: national forest; row 1: Alabama, row 2: Georgia, row 3: Kentucky.

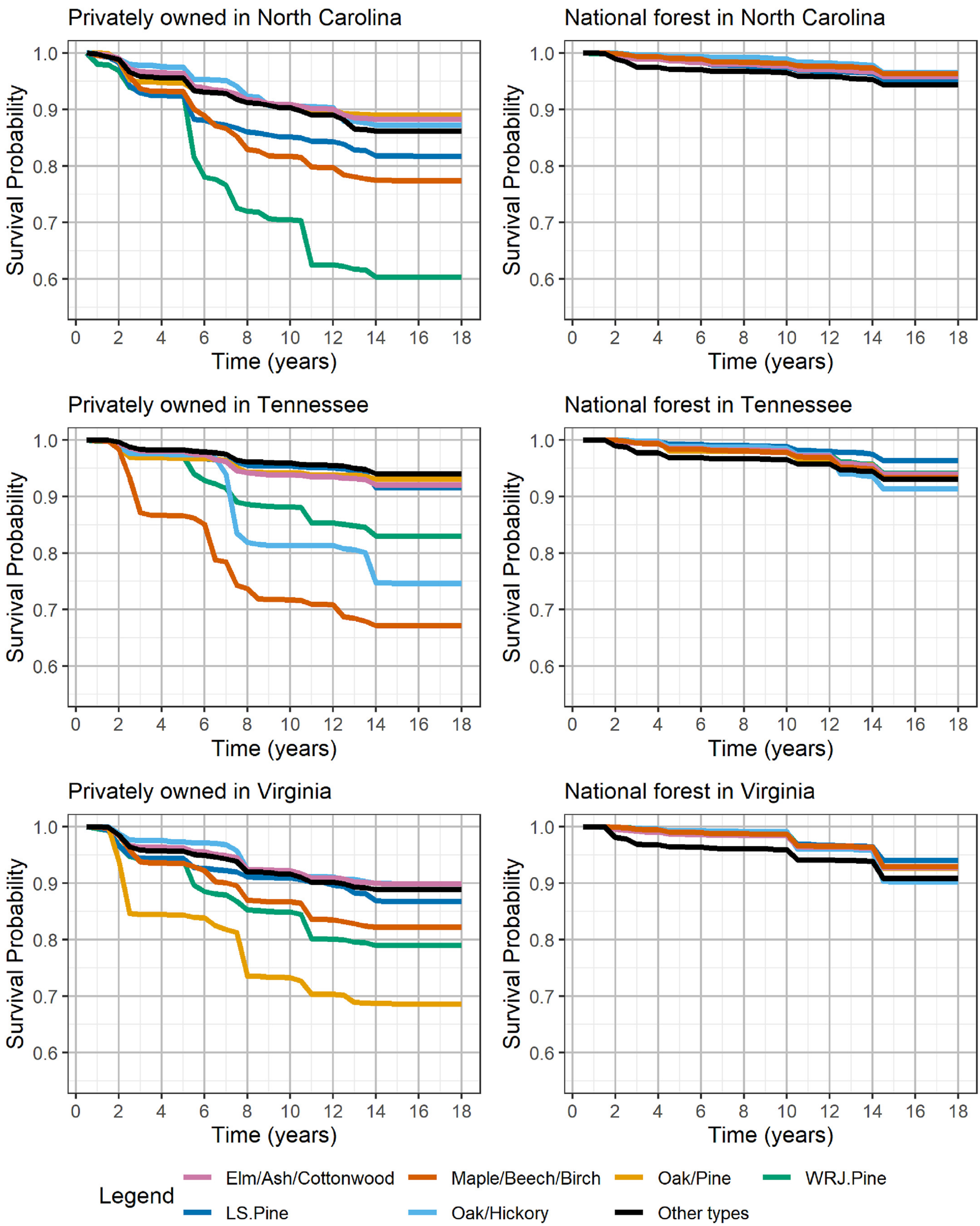

Fig. 8. Survival probability change over time across different forest types. Column 1: privately owned, column 2: national forest;row 1: North Carolina, row 2: Tennessee, row 3: Virginia.

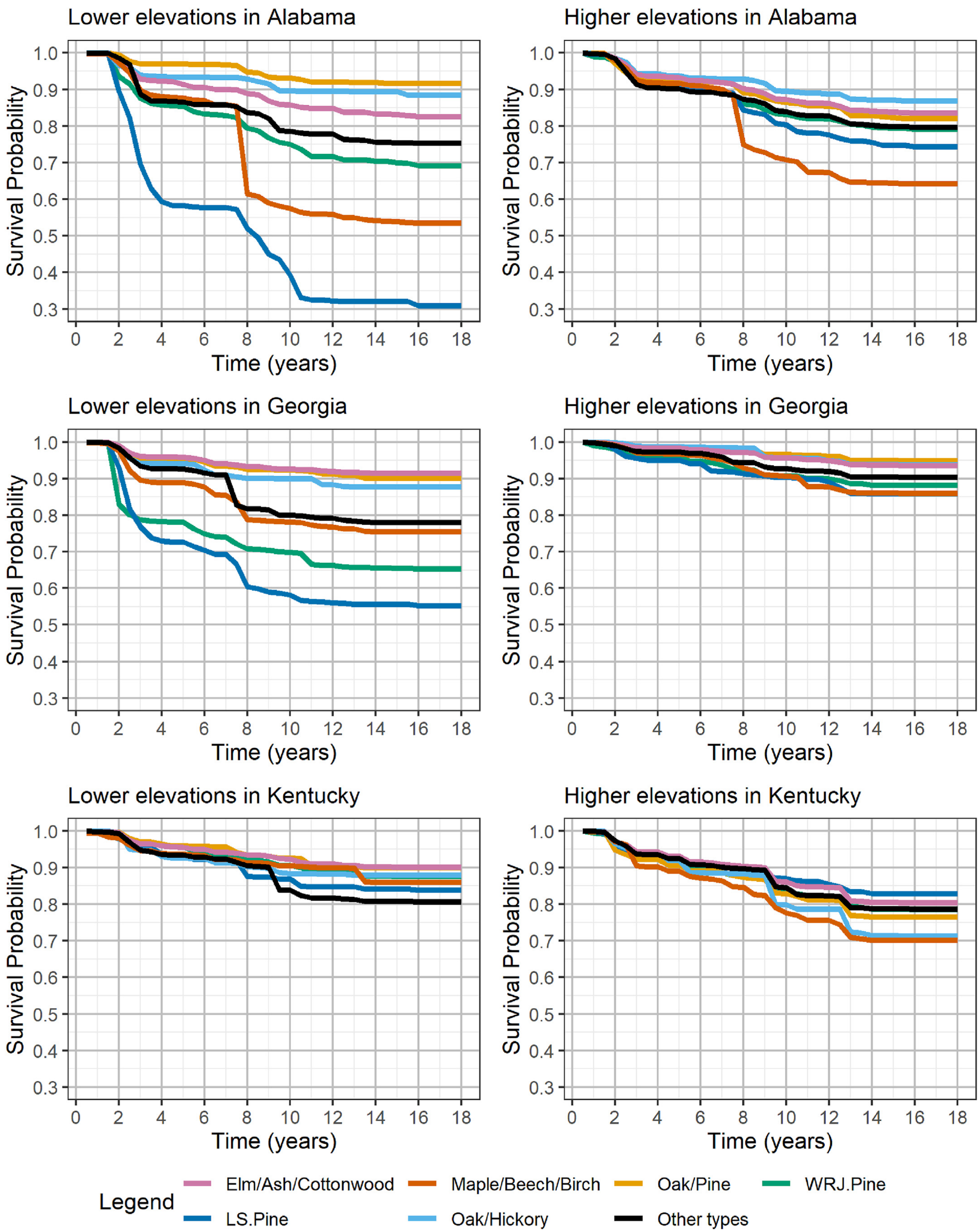

Fig. 9. Survival probability change over time across different forest types. Column 1: lower elevations, column 2: higher elevations; row 1: Alabama, row 2: Georgia, row 3: Kentucky.

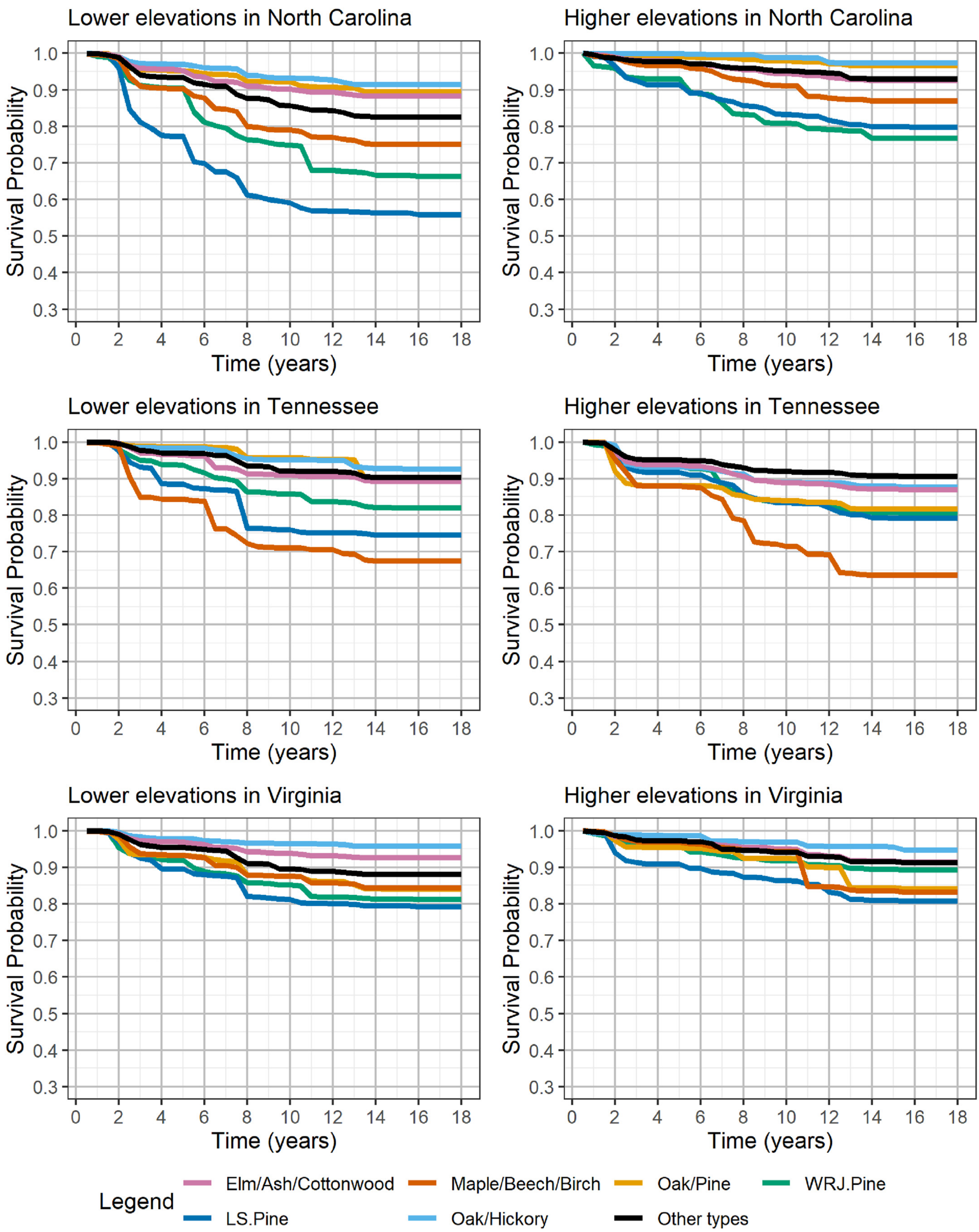

Fig. 10. Survival probability change over time across different forest types. Column 1: lower elevations, column 2: higher elevations; row 1: North Carolina, row 2: Tennessee, row 3: Virginia.

higher harvesting probabilities at lower elevations. While other states (Kentucky and Virginia) lack clear elevation-based trends, closer examination reveals variations in harvesting probabilities across forest types, with specific types showing higher probabilities at either elevation.

4. Discussion

4.1. Comparing model performance

In our study, the predictability of the RSF model was better than the Cox model according to both AUC and Brier scores. Consistent AUC scores across all time frames suggest that the RSF model can accurately classify harvest and non-harvest events with a similar accuracy level over time. Decreasing prediction precision with increasing prediction time was reported in other studies (Gerds and Schumacher, 2006; Heller, 2021). The Brier score increased in higher time frames, but the magnitude of increase is minor for the RSF model (from 0.04 to 0.10). It is worth noting that both the models' prediction accuracies were better than the null: RSF in all time frames and Cox in the time frames from nine to seventeen years.

The reason that a model can show better prediction accuracy in one metric but not in the other can be explained by the nature of the metrics themselves. The AUC measures the prediction accuracy of the model based on true positive rate and false positive rate. This indicates accuracy of the model in classifying harvested vs non-harvested. The Brier score measures the prediction accuracy of the model based on absolute values indicating the difference between the predicted and actual harvesting/survival probabilities. A better prediction accuracy in terms of AUC but not in terms of Brier score indicates that the classification is accurate, but the predicted probabilities are over- or under-estimated. Nevertheless, in terms of our study, the RSF model demonstrated consistently higher prediction accuracy compared to the Cox in both AUC and Brier score. However, the prediction accuracy, as measured by Brier score, declined with time. This suggests that the prediction may be over- or under-estimated in higher time frames considered.

Our findings align with previous studies in different fields of study. Many studies demonstrate that the RSF model consistently provides more reliable predictions than the Cox model for survival. For example, Weathers and Cutler (2017) reported that the RSF model exhibited superior predictive accuracy. Pickett et al. (2021) also found that the prediction accuracy of RSF model is superior to Cox model in dynamic survival prediction in two clinical datasets. Furthermore, Dietrich et al. (2016) found that the RSF model was predicting and agreeing with the Cox model predictions and mentioned that the RSF model effectively tackles issues related to multicollinearity and is well-suited for datasets with high dimensionality. On the contrary, Qiu et al. (2020) compared the Cox model and RSF model in predicting progression in high-grade glioma (HGG) after proton and carbon ion therapy (PBRT), concluding that the Cox model had higher prediction accuracy compared to RSF in smaller datasets with HGG patients treated with PBRT. Here, it should be noted that the observation was for smaller datasets. Another study by Miao et al. (2015) examined the use of RSF in predicting cardiovascular disease and found that while RSF can identify non-linear effects of variables, it cannot entirely replace the Cox model as it is less effective in identifying predictors with lower population ratios and is sensitive to noise.

In our study, we found the RSF produced more accurate predictions than the Cox model even at lower sampling time frames. Our results also highlighted the importance of considering different evaluation metrics to gain a more detailed understanding of model performance. However, further research is needed to explore the models' performance in different forest settings and to identify potential sources of bias or limitations. Additionally, Luck et al. (2017) explored the application of deep learning models in modeling patient-specific kidney survival distribution, revealing that deep learning models demonstrated superior accuracy in prediction. These findings suggest that further exploration of deep learning techniques in survival analysis may prove beneficial and warrant future investigation.

4.2. Important variables

Ownership (privately owned or national forest) was identified as the most influential factor is likely because management objectives and methods differ between private lands and national forests, where forests on the majority of public lands are generally overseen with a focus on both conservation (or non-consumptive) and economic objectives while most private lands are managed with the intention of profit maximization (Schulze et al., 2008; Young et al., 2015). These findings are further supported by Figs. 7 and 8, which illustrate a higher probability of harvesting over time on private lands compared to national forest lands across all studied states.

Indication of forest type as a crucial variable is coherent with the understanding that growth rate and harvest cycles vary between different species, leading to differing harvesting timings. Product preferences (e.g., sawlog, pulpwood) across species also influence optimal harvesting times. Prominence of state as an important variable may be due to various reasons (e.g., geographical features). The variation across different states, and forest types can also be due to differing demand and policies (Canham et al., 2013).

Elevation and slope were found to be important factors, clearly indicating that harvesting frequencies differ substantially depending on the terrain features. This is further supported by Figs. 9 and 10, which illustrate that in Alabama, Georgia, and North Carolina, lower elevations exhibit a higher probability of harvesting over time, indicating a negative relationship between elevation and harvesting probability. This trend, however, is not observed in the other states studied. In line with our findings, Melo et al. (2017) have identified slope as a significant variable in predicting harvest events at plot level in Quebec, Canada. On the contrary they also identified basal area, and stem density as significant variables. In our findings, these variables were less significant maybe as we had additional information included such as elevation, and ownership which were playing a higher impacting role in harvest prediction.

While the impacts of other variables are relatively lower compared to the above mentioned, they still may play a role in determining the forest's resilience to harvesting and tree growth. Previous studies have found variables indicating average tree size and stand density (ex: diameter, volume and basal area) to be very influential variables in forest survival prediction (Antón-Fernández and Astrup, 2012; Kuehne et al., 2019; Melo et al., 2017).

4.3. Model simplification

Reducing the number of predictor variables in the RSF model did not considerably reduce its prediction accuracy. By utilizing only 5 variables: ownership, forest type, elevation, state, and slope, the model 4 achieved a similar level of prediction accuracy compared to the full model. The result implies that applying RSF can possibly reduce the cost of data acquirement and processing in practice while maintaining a high level of prediction accuracy. Notably, further reducing the number of predictors in the model led to a decline in prediction accuracy. Models relying on less than 5 variables (models 5 and 6) exhibited obviously lower predictive capabilities, which implies that the top five most important variables hold critical information for accurately predicting harvest activities.

4.4. Further discussion

Logistic regression has been extensively used to quantify the probability of binary outcomes in forestry literature (e.g., Eastaugh and Hasenauer, 2012; Thurnher et al., 2011). In this study, we further explored an alternative method in modeling harvest occurrence, which can account for changing risk over time and handle censored subjects (e.g., the event has not occurred for some subjects by the end of the study period). When understanding dynamic risk and the timing of event occurrence are of interest, survival analysis is advantageous as time is explicitly modeled via a hazard function, unlike logistic regression. By adding random forests in the survival model, it can handle complex and nonlinear relationships among predictor variables, which is not easily done by conventional regression methods.

Lastly, the prediction accuracy was not considerably improved by adding the market-related variables (i.e., the number of mills in the county, and distance to the closest mill from each plot) in the models. It may be because the occurrence of harvest in the region may not be solely for timber production purposes. Alternatively, other variables might have captured market influences indirectly through their connections to forest characteristics or management practices. However, given that many studies have indicated that market plays an important factor in driving harvest activities (Zhao et al., 2020; Steele et al., 2015), exploring additional market metrics to examine the market influences on harvest probabilities is warranted.

5. Conclusion

In summary, this study demonstrates that random survival forest (RSF) is a useful quantitative tool for predicting harvest activities on specific site characteristics, with considerably higher prediction accuracy compared to the existing survival model (i.e., the Cox model). Forest managers and policy makers can use RSF to develop an effective strategic plan that considers the potential changes in future harvest probabilities over time in a region. Such information can be used to assist in optimizing resource allocation, prioritizing management interventions, and mitigating potential risks associated with forest management practices especially in biodiversity-rich settings like the southern Appalachian region. Five key variables (ownership, forest type, elevation, state, and slope) identified in this study are recommended for consideration when building the model. The reduced model with the five key variables yields a similar accuracy level compared to the full model, offering a practical solution for situations with resource constraints or limited data availability. While this study focused on six states, exploring wider applications using the proposed methodology is recommended for the next step of this study.

Funding

USDA Forest Service (projects # 21CR11330145063 and 21CR11330145069) and The Nature Conservancy (project # 030122-1).

CRediT authorship contribution statement

Dinuka Madhushan Senevirathne: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sheng-I Yang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Consuelo Brandeis:** Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Data curation, Conceptualization. **Donald G. Hodges:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Funding acquisition.

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.

Acknowledgements

Support from the USDA Forest Service (projects # 21CR11330145063 and 21CR11330145069) and The Nature Conservancy (project # 030122-1) is greatly appreciated.

References

Antón-Fernández, C., Astrup, R., 2012. Empirical harvest models and their use in regional business-as-usual scenarios of timber supply and carbon stock development. Scand. J. Forest Res. 27 (4), 379–392. https://doi.org/10.1080/02827581.2011.644576.
Brieman, L., 1996. Heuristics of instability and stabilization in model selection. Ann. Stat. 24 (6), 2350–2383.
Burrill, E.A., DiTommaso, A.M., Turner, J.A., Pugh, S.A., Menlove, J., Christiansen, G., Perry, C.J., Conkling, B.L., 2021. The Forest Inventory and Analysis Database: Database Description and User Guide Version 9.0.1 for Phase 2. US Department of Agriculture, Forest Services, USA.
Canham, C.D., Rogers, N., Buchholz, T., 2013. Regional variation in forest harvest regimes in the northeastern United States. Ecol. Appl. 23 (3), 515–522. https://doi.org/10.1890/12-0180.1.
Chen, Y., Jia, Z., Mercola, D., Xie, X., 2013. A gradient boosting algorithm for survival analysis via direct optimization of concordance index. Comput. Math. Methods Med., 873595 https://doi.org/10.1155/2013/873595.
Cox, D.R., 1972. Regrssion models and life-tables. J. Royal Stat. Soc. 34, 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.
Dauda, K.A., 2022. Optimal tuning of random survival forest hyperparameter with an application to liver disease. Malays. J. Med. Sci. 29 (6), 67–76. https://doi.org/10.21315/mjms2022.29.6.7.
Davidson, A.C., Hinkley, D.V., 1997. Bootstrapping Methods and Their Application. Cambridge University Press. https://doi.org/10.1017/CBO9780511802843.
Dietrich, S., Floegel, A., Troll, M., Kuhn, T., Rathmann, W., Peters, A., Sookthai, D., von Bergen, M., Kaaks, R., Adamski, J., Prehn, C., Boeing, H., Schulze, M.B., Illig, T., Pischon, T., Knuppel, S., Wang-Sattler, R., Drogan, D., 2016. Random survival forest in practice: a method for modelling complex metabolomics data in time to event analysis. Int. J. Epidemiol. 45 (5), 1406–1420. https://doi.org/10.1093/ije/dyw145.
Eastaugh, C.S., Hasenauer, H., 2012. A statistical thinning model for initialising large-scale ecosystem models. Scand. J. Forest Res. 27 (6), 567–577. https://doi.org/10.1080/02827581.2012.679679.
Elad, Y., Pertot, I., 2014. Climate change impacts on plant pathogens and plant diseases. J. Crop Improv. 28 (1), 99–139. https://doi.org/10.1080/15427528.2014.865412.
ESRI, 2023. ArcGIS Pro. 3.2. ESRI Inc.
Field, C.A., Welsh, A.H., 2007. Bootstrapping clustered data. J. R. Statist. Soc. 69, 369–390.
Gerds, T.A., Schumacher, M., 2006. Consistent estimation of the expected Brier score in general survival models with right-censored event times. Biom. J. 48 (6), 1029–1040. https://doi.org/10.1002/bimj.200610301.
Griess, V.C., Knoke, T., 2013. Bioeconomic modeling of mixed Norway spruce—European beech stands: economic consequences of considering ecological effects. Eur. J. Forest Res. 132 (3), 511–522. https://doi.org/10.1007/s10342-013-0692-3.
Hao, L., Kim, J., Kwon, S., Ha, I.D., 2021. Deep learning-based survival analysis for high-dimensional survival data. Mathematics 9 (11), 1244. https://doi.org/10.3390/math9111244.
Heller, G., 2021. The added value of new covariates to the brier score in cox survival models. Lifetime Data Anal. 27 (1), 1–14. https://doi.org/10.1007/s10985-020-09509-x.
Hu, X., Yin, S., Zhang, X., Menon, C., Fang, C., Chen, Z., Elgendi, M., Liang, Y., 2023. Blood pressure stratification using photoplethysmography and light gradient boosting machine. Front. Physiol. 14, 1072273. https://doi.org/10.3389/fphys.2023.1072273.
Ishwaran, H., Kogalur, U.B., 2023. Package 'randomForestSRC. CRAN. https://www.randomforestsrc.org/. (Accessed 15 February 2024).
Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S., 2008. Random survival forests. Ann. Appl. Stat. 2 (3), 841–860. https://doi.org/10.1214/08-aoas169.
Ishwaran, H., Gerds, T.A., Kogalur, U.B., Moore, R.D., Gange, S.J., Lau, B.M., 2014. Random survival forests for competing risks. Biostatistics 15 (4), 757–773. https://doi.org/10.1093/biostatistics/kxu010.
Knight, K.S., Brown, J.P., Long, R.P., 2012. Factors affecting the survival of ash (*Fraxinus* spp.) trees infested by emerald ash borer (*Agrilus planipennis*). Biol. Invasions 15 (2), 371–383. https://doi.org/10.1007/s10530-012-0292-z.
Kuehne, C., Weiskittel, A.R., Legaard, K.R., Simons-Legaard, E.M., 2019. Development and comparison of various stand- and tree-level modeling approaches to predict harvest occurrence and intensity across the mixed forests in Maine, northeastern US. Scand. J. Forest Res. 34 (8), 739–750. https://doi.org/10.1080/02827581.2019.1694975.
Lawless, J.F., 2003. Statistical Models and Methods for Lifetime Data, second ed. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118033005.
Luck, M., Sylvain, T., Cardinal, H., Lodi, A., Bengio, Y., 2017. Deep learning for patient-specific Kidney graft survival analysis. https://doi.org/10.48550/arXiv.1705.10245.
McCarthy-Neumann, S., Ibanez, I., 2012. Tree range expansion may be enhanced by escape from negative plant-soil feedbacks. Ecology 93 (12), 2637–2649. https://doi.org/10.1890/11-2281.1.
Melo, L.C., Schneider, R., Manso, R., Saucier, J.P., Fortin, M., 2017. Using survival analysis to predict the harvesting of forest stands in Quebec, Canada. Can. J. Forest Res. 47 (8), 1066–1074. https://doi.org/10.1139/cjfr-2016-0498.

Miao, F., Cai, Y.P., Zhang, Y.T., Li, C.Y., 2015. Is random survival forest an alternative to Cox proportional model on predicting cardiovascular disease?. In: Lacković, I., Vasic, D. (Eds.), 6th European Conference of the International Federation for Medical and Biological Engineering. IFMBE Proceedings, vol. 45. Springer, Cham. https://doi.org/10.1007/978-3-319-11128-5_184.

Nasejje, J.B., Mwambi, H., Dheda, K., Lesosky, M., 2017. A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. BMC Med. Res. Methodol. 17 (1), 115. https://doi.org/10.1186/s12874-017-0383-8.

Paul, C., Brandl, S., Friedrich, S., Falk, W., Härtl, F., Knoke, T., 2019. Climate change and mixed forests: how do altered survival probabilities impact economically desirable species proportions of Norway spruce and European beech? Ann. Forest Sci. 76 (14). https://doi.org/10.1007/s13595-018-0793-8.

Pickett, K.L., Suresh, K., Campbell, K.R., Davis, S., Juarez-Colunga, E., 2021. Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker. BMC Med. Res. Methodol. 21 (1), 216. https://doi.org/10.1186/s12874-021-01375-x.

Qiu, X., Gao, J., Yang, J., Hu, J., Hu, W., Kong, L., Lu, J.J., 2020. A comparison study of machine learning (random survival forest) and classic statistic (Cox proportional hazards) for predicting progression in high-grade glioma after proton and carbon ion radiotherapy. Front. Oncol. 10, 551420. https://doi.org/10.3389/fonc.2020.551420.

Reis, L.P., de Souza, A.L., dos Reis, P.C.M., Mazzei, L., Soares, C.P.B., Miquelino Eleto Torres, C.M., da Silva, L.F., Ruschel, A.R., Rêgo, L.J.S., Leite, H.G., 2018. Estimation of mortality and survival of individual trees after harvesting wood using artificial neural networks in the amazon rain forest. Ecol. Eng. 112, 140–147. https://doi.org/10.1016/j.ecoleng.2017.12.014.

Ren, S., Lai, H., Tong, W., Aminzadeh, M., Hou, X., Lai, S., 2010. Nonparametric bootstrapping for hierarchical data. J. Appl. Stat. 37, 1487–1498.

Roberts, M., Gilligan, C.A., Kleczkowski, A., Hanley, N., Whalley, A.E., Healey, J.R., 2020. The effect of forest management options on forest resilience to pathogens. Front. For. Glob. Change 3. https://doi.org/10.3389/ffgc.2020.00007.

Schulze, M., Grogan, J., Vidal, E., 2008. Technical challenges to sustainable forest management in concessions on public lands in the Brazilian Amazon. J. Sustain. For. 26 (1), 61–76. https://doi.org/10.1300/j091v26n01_03.

Solarik, K.A., Volney, W.J.A., Lieffers, V.J., Spence, J.R., Hamann, A., 2012. Factors affecting white spruce and aspen survival after partial harvest. J. Appl. Ecol. 49 (1), 145–154. https://doi.org/10.1111/j.1365-2664.2011.02089.x.

Spruance, S.L., Reid, J.E., Grace, M., Samore, M., 2004. Hazard ratio in clinical trials. Antimicrob. Agents Chemother. 48 (8), 2787–2792. https://doi.org/10.1128/AAC.48.8.2787-2792.2004.

Steele, M.Z., Shackleton, C.M., Uma Shaanker, R., Ganeshaiah, K.N., Radloff, S., 2015. The influence of livelihood dependency, local ecological knowledge and market proximity on the ecological impacts of harvesting non-timber forest products. Forest Policy Econ. 50, 285–291. https://doi.org/10.1016/j.forpol.2014.07.011.

Therneau, T.M., Lumley, T., Elizabeth, A., Cynthia, C., 2023. Package 'survival'. https://github.com/therneau/survival. (Accessed 15 February 2024).

Thurnher, C., Klopf, M., Hasenauer, H., 2011. Forests in transition: a harvesting model for uneven-aged mixed species forests in Austria. Forestry 84 (5), 517–526. https://doi.org/10.1093/forestry/cpr021.

Wang, H., Li, G., 2017. A selective review on random survival forests for high dimensional data. Quant. Biosci. 36 (2), 85–96. https://doi.org/10.22283/qbs.2017.36.2.85.

Gerds, T.A., Ohlendorff, J., Ozenne, B., 2023. riskRegression: Risk Regression Models and Prediction Scores for Survival Analysis with Competing Risks (Version R Package Version 2023, 12.21. CRAN. https://CRAN.R-project.org/package=riskRegression (Accessed 15 February 2024).

Weathers, B., Cutler, D.R., 2017. Comparison of Survival Curves between Cox Proportional Hazards, Random Survival Forests, and Conditional Inference Forests in. Survival Analysis, 927. Utah State University, All Graduate Plan B and other Reports (Accessed 15 February 2024). https://digitalcommons.usu.edu/gradreports/927.

Wilstrup, C., Cave, C., 2022. Combining symbolic regression with the Cox proportional hazards model improves prediction of heart failure deaths. BMC Med. Inform Decis. Mak. 22 (1), 196. https://doi.org/10.1186/s12911-022-01943-1.

Young, T., Wang, Y., Guess, F., Fly, M., Hodges, D., Poudyal, N., 2015. Understanding the characteristics of non-industrial private forest landowners who harvest trees. Small-scale For. 14 (3), 273–285. https://doi.org/10.1007/s11842-015-9287-9.

Zhao, J.H., Daigneault, A., Weiskittel, A., 2020. Forest landowner harvest decisions in a new era of conservation stewardship and changing markets in Maine, USA. Forest Policy Econ, 118, 102251. https://doi.org/10.1016/j.forpol.2020.102251.