Forest Ecology and Management 285 (2012) 101–122

Contents lists available at SciVerse ScienceDirect

Forest Ecology and Management

journal homepage: www.elsevier.com/locate/foreco

Applying survival analysis to managed even-aged stands of ponderosa pine for assessment of tree mortality in the western United States

Fabian C.C. Uzoh [a,*], Sylvia R. Mori [b]

[a] Ecology and Management of Western Forests Influenced by Mediterranean Climates, United States Department of Agriculture, Forest Service, Pacific Southwest Research Station, Redding, CA 96002, United States
[b] Statistics Unit, United States Department of Agriculture, Forest Service, Pacific Southwest Research Station, Albany, CA 94710, United States

ARTICLE INFO

Article history:
Received 14 February 2012
Received in revised form 1 August 2012
Accepted 3 August 2012
Available online 12 September 2012

Keywords:
Cox Proportional Hazard model
Probability of survival
Tree mortality
Pinus ponderosa
Logistic regression
Generalized Additive Model (GAM)

ABSTRACT

A critical component of a growth and yield simulator is an estimate of mortality rates. The mortality models presented here are developed from long-term permanent plots in provinces from throughout the geographic range of ponderosa pine in the United States extending from the Black Hills of South Dakota to the Pacific Coast. The study had two objectives: estimation of the probability of a tree survival for the next 5 years and the probability of a tree surviving longer than a given time period (survival trend) for a given set of covariates. The probability of a tree surviving for the next 5 years was estimated using a logistic model regressed on 18 covariates measured 5 years before the last measurement period with 15 smoothing variables (S1–S15) for spatial effects of latitude and longitude surface. The fitted model showed that the probability of survival increased with increasing diameter at breast height (DBH), DBH periodic annual increment (PAIDBH) and increasing plot basal area/number of trees per hectare (PBAH/TPH), and decreased with increasing average of the 5 tallest trees in the plot (AVGHT5) when other selected covariates were included in the model. The probability of a tree surviving longer than a given time period was estimated by fitting the Cox Proportional Hazard model to the last observed survival period regressed on 13 covariates measured at the first measurement period. This probability also increased with increasing DBH and PAIDBH, and decreased with increasing AVGHT5. The Akaike's Information Criterion (AIC) and graphs of partial residuals were used in the selection of covariates included in the models.

Published by Elsevier B.V.

1. Introduction

Competition among trees is one of the main factors determining their growth and mortality (Oliver and Uzoh, 1997; Zeide, 2004). The competition stressors can be long-term or short-term (van Mantgem et al., 2003). For a model to adequately characterize tree growth it must include estimates of mortality rates (or survival), because mortality is an integral part of stand dynamics (Monserud, 1976; Hamilton, 1986; Hann and Wang, 1990). The literature on modeling tree mortality is voluminous; nevertheless, mortality estimates remain the weakest link in growth and yield simulators because of estimation difficulties (Hamilton, 1986). There are two main causes of tree mortality: external and internal factors. Mortality resulting from external factors tends to be episodic and often even catastrophic, especially if mortality is a result of factors such as bark beetles, root disease, or wind. Mortality resulting from internal factors arises from inter-tree competition and it tends to be more uniform and constant (Oliver and Uzoh, 1997).

The first generation of statistical mortality models was at the stand level, predicting the future number of trees per unit area (Lee, 1971; Moser, 1972; Ek, 1974; Somers et al., 1980; Clutter et al., 1983; Harms, 1983). Subsequently, Hamilton (1974) and Monserud (1976) introduced the use of logistic regression models for individual tree mortality response.

Survival analysis is a recent improvement in assessing mortality trends and dynamics of individual trees. Woodall et al. (2005) provided a superb history and reason for the use of survival analysis in modeling tree mortality. In general, survival analysis is a collection of statistical procedures for data analysis used for studying the occurrence and timing of events for which the outcome variable of interest is most often death, which was the purpose for their original designs (Kleinbaum and Klein, 2005; Woodall et al., 2005; Allison, 2010). The uniqueness of survival analysis stems from the fact that it allows for censoring of observations (lack of exact time of death) and inclusion of time-dependent covariates, and dealing with non-normal distributions (Woodall et al., 2005). These features of survival data make it difficult to handle with commonly-used conventional statistical methods, but ignoring them will reduce the precision of the estimates (Allison, 2010).

* Corresponding author. Tel.: +1 530 226 2549; fax: +1 530 226 5091.
E-mail address: fuzoh@fs.fed.us (F.C.C. Uzoh).

0378-1127/$ - see front matter Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.foreco.2012.08.006

Consequently, conventional approaches such as Logistic regression are inadequate for dealing with either censoring or time-dependent covariates, something at which survival analysis excels (Allison, 2010).

Kleinbaum and Klein (2005) defined two quantitative terms that should be considered in any survival analysis. These are the survivor function, denoted by $S(t)$ and the hazard function denoted by $h(t)$ for the survival time t. The survivor function $S(t)$ gives the probability that an individual (in our case, a tree) survives longer than some specified time t: that is, $S(t)$ gives the probability that the random variable T exceeds the specified time t (Kleinbaum and Klein, 2005; Woodall et al., 2005; Allison 2010). The hazard function focuses on failing, that is, on the event occurring because the hazard function $h(t)$ gives the instantaneous potential per unit time for death to occur, given that the individual has survived up to time t. In contrast the survivor function focuses on not failing, that is, on the event not occurring. In a unique sense, both functions can be considered as giving the opposite side of the information given by the other (Kleinbaum and Klein, 2005).

We used two survival modeling approaches in our investigation: the logistic regression (parametric approach, McCullagh and Nelder, 1991) for 5-year responses, and the Cox Proportional Hazard, regression (semi-parametric approach, Cox, 1972, 1975; Lee and Wang, 2003) for censored responses.

The objective of this study is to develop an individual tree mortality model applicable for even-aged pure stands of ponderosa pine throughout its geographic range in the United States. The objective is divided into two phase: (1) to build a survival model for predicting the probability of a tree surviving to the next 5 years based on plot/tree information measured 5 years before the last measurement period and (2) to build a survival model to estimate the probability of a tree surviving longer than a given time period based on plot/tree information at the first measurement period. The Cox PH model here is used as an explanatory model but not as a predictive model, because of its non-parametric distribution assumption for the survival time. For phase (1), we fitted the Logistic model, and for phase (2), we fitted the Cox Proportional Hazard (PH) model. For both models, we considered the following groups of measured candidate explanatory (predictors) variables:

1. *Plot spatial information*: latitude, longitude, elevation, slope and aspect.
2. *Stand structure information at measurement time*: plot basal area, number of trees per hectare, stand age, average of the five tallest trees in the plot (however, if less than five, then we just calculated their average.), site index, site density index, and basal area per hectare in larger diameter.
3. *Tree information at measurement time*: DBH, DBH periodic annual increment, and tree basal area.

2. Methods

2.1. Data

The measurements were made in years ranging from 1938 to 1998. Some plots were measured repeatedly (for example, 1938–1943–1948–1953; 1963–1967–1971–1979–1984–1989–1994), and others only once or twice. Several datasets used in this study were from long-term permanent plots consisting of: (a) levels-of-growing-stock studies established in the 1960s using a similar study design with five or six stand density levels replicated three times (Myers, 1967) and (b) initial spacing and permanent-plot thinning studies. Individual-tree data were from plots initiated from both artificial stands and natural stands located in the five provinces of ponderosa pine in the western United States (Fig. 1) and initially covering a wide range of size classes. Stands were free, or mostly free, of competing shrubs that reduce growth of young ponderosa pine especially in central Oregon and California (Oliver, 1984; Oliver and Ryker, 1990; Cochran and Barrett, 1999). Results from individual installations of the levels-of-growing-stock studies have been previously reported (Tables 1 and 2), as were growth models based on five installations (Oliver and Edminster, 1988; Uzoh and Oliver, 2006, 2008).

Trees were tagged and repeatedly measured on periods of different length (ranging from 2 years to 18 years lag), but about 68% of the measurement were done every 5 years. Seventy-eight percent of the plots were measured for more than 10 years. The data resulting from this study consisted of 305 plots with a total of 29,449 trees. Of those trees, 28,901 trees were used for fitting the Cox PH model. Some trees had to be removed from the analysis because some plots were measured only once, while at others plots, measurements were done in 2-year interval but not in the same years. Of these, 20,118 trees were used for the logistic model because they were measured in 5-year intervals. Table A1 in the Appendix shows the summary of the original dataset and Table A2 shows a summary of the last 5-year measurement periods. Fig. 2 shows the distribution of mortality within the study area.

Basic records for each plot included latitude, longitude, elevation, aspect, slope percent, and plot size. Individual tree measurements included diameter at breast height (DBH) and total height. Different methods were used at different locations for sampling tree height. At some locations, every tree height in each plot was measured; at others, a systematic sample of tree heights were measured; yet at other locations, height sample trees were randomly selected within 2 in. diameter classes across the range of tree sizes. Height measurements were repeated on the same trees (Uzoh and Oliver, 2006). Mortality was noted and the causal agent investigated. The data for this analysis consists only of the initial, and the last two measurement periods, called from now on "Initial", "Prior" and "Current". The two models (the Logit model and the Cox Proportional Hazard model (Cox PH model)) aim to predict/explain the current survival response with the information from the prior or initial period. Therefore, only the values of the explanatory variables from the initial measurement period, from the 5-year period prior to the last measurement period and the current survival status, are used in this analysis.

Many trees in a number of plots suffered competition-induced mortality. For the Cox PH model, since a tree either died during an interval or is alive at the end of the study and the year of death is unknown, we have a case of interval censoring (Allison, 2010). Some of the measurement periods were of different length, therefore, it is possible that the estimated survival probability has some bias or added imprecision. It is possible that a greater than 10-year time lag between measurements increases the bias or decreases the precision of the estimate, however, only 4.3% of all the 28,901 trees had a time-lag greater than 10 years and of these only 1.2% died in those intervals. We fitted the COX model with these trees removed and the trees and the slopes' trend of the coefficients did not change and the deviance residual plots showed the same pattern as that shown for the whole dataset. As a result, we chose not to remove the trees from the analysis.

2.2. Statistical analyses

2.2.1. Logistic model for predicting 5-year survival probability

We used the Logistic model from the family of the Generalized Additive Models (GAMs) (Hastie and Tibshirani, 1990).

2.2.1.1. Logit model.

$$\log\left(\frac{p}{1-p}\right) = g(LAT, LONG) + \sum_{j=0}^{m} c_j * x_j$$