Forest Ecology and Management 276 (2012) 139–145

Contents lists available at SciVerse ScienceDirect

Forest Ecology and Management

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

Evaluation of mixed-effects models for predicting Douglas-fir mortality

Jeremiah D. Groom *, David W. Hann, Hailemariam Temesgen

Department of Forest Engineering Resources & Management, Oregon State University, Corvallis, OR 97331, United States

ARTICLE INFO

Article history:
Received 16 January 2012
Received in revised form 23 March 2012
Accepted 27 March 2012
Available online 26 April 2012

Keywords:
Generalized linear model
Mixed model
Douglas-fir
Mortality

ABSTRACT

We examined the performance of several generalized linear fixed- and mixed-effects individual-tree mortality models for Douglas-fir stands in the Pacific Northwest. The mixed-effects models accounted for sampling and study design overdispersion. Inclusion of a random intercept term reduced model bias by 88% relative to the fixed-effects model; however, model discrimination did not substantially differ. An uninformed version of the mixed model that used only its fixed effects parameters produced predicted mortality values that exceeded the fixed-effects model bias by 31%. Overall, we did not find compelling evidence to suggest that the mixed models fit our data better than the fixed-effects model. In particular, the mixed models produced fixed-effects parameter estimates that predicted unreasonably high mortality rates for trees approaching 1 m in diameter at breast height.

1. Introduction

Tree mortality is a critical component of stand growth and yield models. It is also highly variable and difficult to predict (Lee, 1971; Dobbertin and Biging, 1998). The nature of data collected to model and quantify mortality, however, may challenge the assumptions inherent in statistical tools used to estimate mortality. In this study we examine a generalized linear mixed-effects method to account for data structure and lack of independence.

Lee (1971) and Staebler (1953) described tree mortality as either regular or irregular. Irregular mortality includes death occurring from insects, disease, fire, snow damage, and wind. This type of mortality typically is episodic, brief, and difficult to predict. Regular mortality is more predictable, and includes influences such as competition for light, moisture, and nutrients. As stands become more crowded, a degree of mortality usually occurs. Trees may die for several possibly co-occurring reasons: suppression where stands are differentiating, weakening due to insects and disease, and buckling where stems become tall and thin (Oliver and Larson, 1996). Trees in stands characterized by regular mortality exhibit a preponderance of mortality amongst smaller-diameter individuals that are over-topped by neighbors (Peet and Christensen, 1987). Mortality rates become low for established trees until larger diameters are reached and the mortality rate increases again (Buchman et al., 1983; Harcombe, 1987; Monserud and Sterba, 1999). Although both classes of mortality may affect stands, only single-tree regular mortality models are routinely incorporated in most growth and yield simulators such as FVS (Dixon, 2011) and ORGANON (Hann, 2011).

Single-tree mortality models have been developed using a variety of data and approaches. Logistic models are common for data sets where revisit frequency consists of equal-length time periods (Hamilton, 1986; Bigler and Bugmann, 2003; Jutras et al., 2003; Moore et al., 2004; Adame et al., 2010). However, if the time periods differ, a common solution is to use the logistic model but insert time as a power upon survival probabilities or use a complimentary log–log link function (e.g., Monserud, 1976; Eid and Tuhus, 2001; Moore et al., 2004; Temesgen and Mitchell, 2005; Fortin et al., 2008). For stands where remeasurement occurred multiple times, researchers either avoid pseudoreplication at the level of the tree by omitting all but the last remeasurement for each tree (Hamilton, 1986) or include the remeasurement information (Temesgen and Mitchell, 2005; Fortin et al., 2008).

Data used in these analyses are from nested samples, with the highest level referred to as installations. Each installation contains one or more plots; each plot contains many trees with repeated measurements. Analyses performed on individual tree mortality data has recently begun to account for the structured nature and non-independence by using generalized linear mixed-effects models. Logistic models by Adame et al. (2010) and Jutras et al. (2003) include random intercepts for study plots or stands. A complimentary log–log model by Fortin et al. (2008) included an adjusted intercept with random effects for study plot and specific time interval nested within plot.

Prediction performance for nonlinear mixed-effects models may be improved (less bias and greater precision) when compared to corresponding fixed-effects models conditional on the availability

* Corresponding author. Present address: Oregon Department of Forestry, 2600 State Street, Salem, OR 97310, United States. Tel.: +1 503 945 7394; fax: +1 503 945 7490.

E-mail addresses: jgroom@odf.state.or.us, jeremygroom@gmail.com (J.D. Groom), david.hann@oregonstate.edu (D.W. Hann), hailemariam.temesgen@oregonstate.edu (H. Temesgen).

0378-1127/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.foreco.2012.03.029

of previous information on the subject; however, in absence of random-effects information, predictions using just the fixed portions of the parameterization from the nonlinear mixed-effects model exhibit greater bias and less precision than even the original fixed-effects model (Monleon, 2003; Temesgen et al., 2008; Garber et al., 2009). Setting the random effect to zero follows from prediction theory only for linear mixed models, but it has a different meaning for nonlinear models. Consider a linear mixed model where X is a $(n \times p)$ design matrix where n is the number of observations and p is the number of fixed-effects parameters, β is a vector of linear slope values, Z is a $(n \times r)$ design matrix where r is the number of random effects parameters, γ represents G-sided random effects parameterization, and ε is the random error:

$$y = X\beta + Z\gamma + \varepsilon, \text{ where } E(\gamma) = E(\varepsilon) = 0$$

Then, conditional on the random effect, and because the expectation is a linear operator,

$$E(y|\gamma) = X\beta + Z\gamma$$

Unconditionally,

$$E(y) = E(X\beta + Z\gamma + \varepsilon) = X\beta + ZE(\gamma) = X\beta$$

Thus, in a linear model, the unconditional expectation can be calculated from the conditional expectation by setting the random effect to zero:

$$E(y) = E(y|\gamma = 0)$$

For a nonlinear model, this is not the case. The nonlinear mixed model can be written as:

$$Y = f(X,\ \beta,\ Z,\ \gamma) + \varepsilon, \text{ where } E(\gamma) = E(\varepsilon) = 0$$

Conditional on installation:

$$E(y|\gamma) = f(X,\ \beta,\ Z,\ \gamma)$$

Unconditionally:

$$E(y) = E[E(y|\gamma)] = E[f(X,\ \beta,\ Z,\ \gamma]$$

Unlike linear models, for nonlinear models, the unconditional model is not the same as the conditional model with the random effects set to zero:

$E(y) \neq E(y|\gamma = 0)$ because $E[f(X, \beta, Z, \gamma)] = \int f(X, \beta, Z, \gamma) d\mu(\gamma) \neq f(X, \beta, Z, \gamma = 0)$, where $\mu(\gamma)$ is the distribution function of γ.

The model for $E(y)$ is known as the population-average model and the model for $E(y|\gamma)$ is known as the subject-specific model. For nonlinear mixed models, those versions are different. Choosing which type of model and inference is appropriate for each objective is fundamental when dealing with nonlinear mixed models. For a tree from a completely new stand that does not have information to estimate the random effects and, therefore, condition on the stand effect, the proper model is a population average model. When using the subject-specific model with $\gamma = 0$ (i.e., the subject-specific model for the average stand), prediction performance is expected to decline. Again, in linear mixed models this is not an issue, because setting $\gamma = 0$ yields the population-average model.

Forest management requires models that are useful beyond their study areas. Generalized or nonlinear mixed-effects models can increase bias when applied to novel data (e.g., Robinson and Wykoff, 2004). Mixed models require estimated information about a hierarchical level that may be unknown for novel data sets. One technique to extend generalized linear or nonlinear mixed-effect model applicability is to utilize minimal data from new stands for estimating the random effects parameters. This allows the application of nonlinear mixed effects models beyond their original data frames (Monleon, 2003; Temesgen et al., 2008; Garber et al., 2009). However, this technique may be limited by the response variable type. In those studies it worked for tree height, a continuous static variable. Our study's response variable, individual tree mortality, is rare, binomial, dynamic, and requires several years of data collection to observe. Thus, incorporating subsample information from new plots to inform mixed-effects model predictions is generally unfeasible.

The objectives of this study are to (1) determine whether a generalized linear mixed model fit to repeatedly remeasured Douglas-fir (*Pseudotsuga menziesii* [Mirb.]) trees can improve mortality estimation over a previous nonlinear estimation approach (Hann et al., 2003, 2006), and (2) compare the predictive abilities of mixed-effects models to nonlinear least squares estimation in the presence and absence of random effects information. We expect biased predictions from the mixed model that lacks random effects information, but examine the degree by which those results are useful relative to the nonlinear least squares predictions. Taken together, our goal is to examine how well models met our objectives and whether we produce a model that is useful for current Douglas-fir growth and yield simulators.

2. Methods

2.1. Study area and data acquisition

Data used in this analysis were obtained from randomly located installations on nine land ownerships and represent a subset of data described in Hann et al. (2003, 2006). One of the uses of the overall data collection effort was to calibrate the ORGANON stand development model (Hann, 2011) for intensively managed Douglas-fir in the Pacific Northwest region of the USA and Canada. What follows is a description of the subsetted data. The data were from 304 permanent sample installations from Southwest British Columbia, Western Washington, and Northwestern Oregon. The 820 plots within those installations contained 195,795 revisit data collected from 70,720 Douglas-fir trees. Trees were revisited one to 18 times over the course of data collection. Time between revisits was not equal among trees or plots, and varied from 3 to 7 years (median = 5 years). The fixed-area plots varied in size from 0.041 to 0.486 ha (mean = 0.069). The average breast height age was 27.8 years and ranged from 3 to 108 years. Plots included in this study were not subject to thinning or fertilization experimental treatments.

We further reduced the data set according to two criteria. The first criterion only permitted data from installations that had two or more plots. This criterion was necessary for creating mixed-effects mortality predictions (described below), and it removed 12,616 trees, 38,314 observations, and 67 single-plot installations from the data set. The second criterion was that we retained only trees with DBH <101.6 cm. We removed larger-DBH trees to allay model convergence issues likely arising from a paucity of mortality information leading to a lack of fit at that extreme. This removed eight observations and five trees (<0.01% of data) and permitted model convergence. The resulting data set included 157,473 revisits of 58,099 trees in 753 plots located within 201 installations.

2.2. Mortality estimation

We based this analysis on a general equation of mortality given differing plot revisit schedules as described by Hann et al. (2006):

$$PM = 1.0 - [1.0 + e^{-(X\beta)}]^{-\text{PLEN}} + \varepsilon_{\text{PM}} \quad (1)$$

where PLEN is the length of the growth period in 5-year increments (i.e., length of a growth period in years divided by 5), PM is the 5-year mortality rate, and ε_{PM} is the random error on PM. The response variable distribution is $y \sim$ Bernoulli (π) where the observed response was y and π is the corresponding response probability.