Modeling Mortality of Loblolly Pine (*Pinus taeda* L.) Plantations

Ram Thapa

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Forestry

Harold E. Burkhart, Chair
Philip J. Radtke
Marion R. Reynolds Jr.
Inyoung Kim

February 7, 2014
Blacksburg, Virginia

Keywords: Lobolly pine plantations, Mortality, Climate and soil, Difference mortality equation, Multilevel logistic regression, Cox proportional hazards model, Shared frailty

Modeling Mortality of Loblolly Pine (*Pinus taeda* L.) Plantations

Ram Thapa

(ABSTRACT)

Accurate prediction of mortality is an important component of forest growth and yield prediction systems, yet mortality remains one of the least understood components of the system. Whole-stand and individual-tree mortality models were developed for loblolly pine plantations throughout its geographic range in the United States. The model for predicting stand mortality were developed using stand characteristics and biophysical variables. The models were constructed using two modeling approaches. In the first approach, mortality functions for directly predicting tree number reduction were developed using algebraic difference equation method. In the second approach, a two-step modeling strategy was used where a model predicting the probability of tree death occurring over a period was developed in the first step and a function that estimates the reduction in tree number was developed in the second step. Individual-tree mortality models were developed using multilevel logistic regression and survival analysis techniques. Multilevel data structure inherent in permanent sample plots data i.e. measurement occasions nested within trees (e.g., repeated measurements) and trees nested within plots, is often ignored in modeling tree mortality in forestry applications. Multilevel mixed-effects logistic regression takes into account the full hierarchical structure of the data. Multilevel mixed-effects models gave better predictions than the fixed effects model; however, the model fits and predictions were further improved by taking into account the full hierarchical structure of the data. Semiparametric proportional hazards regression was also used to develop model for individual-tree mortality. Shared frailty model, mixed model extension of Cox proportional hazards model, was used to account for unobserved heterogeneity not explained by the observed covariates in the Cox model.

Dedication

To my late father Jagat Bahadur Thapa.

His words of inspiration and encouragement

in pursuit of excellence, still linger on.

Acknowledgments

I would like to express my deep appreciation and gratitude to my advisor, Dr. Harold E. Burkhart, for the patient guidance and mentorship he provided me. I could not have imagined having a better advisor and mentor for my PhD study. Without his guidance and persistent help this dissertation would not have been possible.

I would like to thank my committee members, Drs. Philip J. Radtke, Marion R. Reynolds Jr. and Inyoung Kim for the friendly guidance, insightful comments and suggestions. Special thanks goes to Dr. Yili Hong, who provided valuable statistical advice with Chapter 4.

I would like to thank the Forest Modeling Research Cooperative (FMRC) at Virginia Tech, the Center for Advanced Forestry System (CAFS) and Pine Integrated Network: Education, Mitigation, and Adaptation project (PINEMAP) for the financial support towards my graduate program. I am thankful to the Department of Forest Resources and Environmental Conservation (FREC) at Virginia Tech for the administrative support and cooperation. Ralph Amateis was helpful in providing the data and gave me valuable experience in the field.

I greatly appreciate the support my friends gave me during my time in Blacksburg. Thanks to my past and present fellow graduate students Nabin Gyawali, Charles Obuya Sabatia, Mickey Allen, Gavin Corral and Santosh Subedi for their encouragement and friendship.

Most importantly, none of this would have been possible without the love and patience of my family. Thank you mother and Suresh. They were always supporting me and encouraging me with their best wishes.

Finally, I would be remiss if I didn't acknowledge the support, encouragement, understanding and unwavering love of my wife, Reena. She was always there cheering me up and stood by me through the good times and bad.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Loblolly pine (*Pinus taeda*) is the most widely planted timber species in the southeastern United States. There were over 30 million acres of pine plantations in the southern US in early 2000's and they were mainly comprised of loblolly pine which was projected to increase to 55 million acres by 2040 (Wear and Greis, 2002). Southern forest production has doubled in the last 50 years and intensive management practices such as planting genetically improved seedlings, site preparation, and fertilization are attributed to this increased productivity. Loblolly pine plantations account for about 80% of 2 millinon acres of forest plantation established each year in the southern US (McKeand et al., 2003).

Growth and mortality are the basic components of forest stand dynamics and accurate prediction of mortality is crucial to forest growth and yield prediction system (Monserud, 1976). Modeling of forest survival is usually done at two resolutions namely whole stand and individual tree level. Whole stand survival models predict the future stems per unit area when an initial number of trees and corresponding age are given. Predicting survival at young age is generally difficult and hence the whole stand survival models for early stand age have also been modeled separately from the rest of the continuum of survival in the

stand (e.g. Amateis et al., 1997). Forest survival modeling at individual tree level is more common and the descriptive level is the individual tree. Individual tree level survival model uses some flexible function that is bounded between 0 and 1.

Mortality models can be used either deterministically or stochastically like many statistical models. Both at the tree and the stand level, a deterministic model yields the mathematical expectations of the growth and a stochastic model explains natural variability of the growth by including random components. Although variability of the growth may be of interest sometimes, rather than the mean response, stochastic growth models are rarely used (Fortin and Langevin, 2012). Most growth and yield models are used for providing deterministic predictions or may include few stochastic components. Weber et al. (1986) made 100 years projections using the deterministic and stochastic mortality algorithms of the STEMS individual-tree projection model. They found that there was no practical differences in mean stand values for number of trees, basal area, volume, or diameter distributions and confirmed the effectiveness of deterministic mortality estimation approach. However, Vanclay (1991) and Fortin and Langevin (2012) found that the predictions of some response variables were different from stochastic and deterministic simulations.

1.1 Stand-level mortality

Many approaches have been taken to model stand-level mortality in forest stands. In 1970's and before, linear and polynomial functions were used to model mortality, e.g., Staebler (1953) used linear functions to predict percent survival from age, site index, and mean diameter and Lee (1971) applied linear regression analysis to data from existing yield tables to predict mortality rates of lodgepole pine. Whole-stand mortality models have been commonly developed using derivative of the generalized Gamma distribution (Weibull or exponential)

or the difference equation approach. Somers et al. (1980) used the Weibull distribution to predict mortality of young natural loblolly pine stands in South Carolina. Algebraic difference approach generally produces satisfactory results by predicting number of live trees at some future point in time from the current number of trees, age and other site variables (Lemin and Burkhart, 1983). The remeasurement data from permanent sample plots that are used to model tree mortality consists of number of trees (N_1) at an initial age (A_1), number of trees (N_2) at subsequent age (A_2) and stand site index. An equation is fitted that predicts (N_2) as a function of (A_1), (A_2) and (N_1), often with some function of site index included. Stand-level mortality equations that predict changes in the number of trees per unit area, developed using an algebraic difference approach, can be represented as in Equation (1.1):

$$N_2 = f(N_1, A_2, A_1) \tag{1.1}$$

where N_i is the number of trees per unit area at age A_i such that $A_1 \leq A_2$.

Mortality functions derived from differential equations should possess certain desirable properties such as consistency, path-invariance, asymptotic limit of stocking approaching zero as age goes to infinity, negligible in-growth assumption for even-aged stands (Clutter et al., 1983; Diéguez-Aranda et al., 2005). Site index is often an important factor in modeling mortality and mortality generally increases with higher site index (Bailey et al., 1985; Diéguez-Aranda et al., 2005; Zhao et al., 2007). Hence, some function of site index also enters into Equation (1.1). Silvicultural factors like thinning also affect stand-level mortality (Bailey et al., 1985; Eid and Øyen, 2003). Stand-level mortality models based on the difference equations approach assume ingrowth to be negligible (Weiskittel et al., 2011) but it may not be relevant for planted stands where ingrowth is generally not a consideration.

There has been increased interest in predicting growth responses of forest trees to climate changes and environmental effects like carbon sequestration. The number of models that account for the effects of human-induced changes, particularly climate change, in forest productivity have increased lately (Waring et al., 2006). Several studies have investigated the effect of site, topographic, climatic, soil and genetic factors on growth and yield. Climate and soil data have often been used in forest growth and yield models to improve empirical estimation. Past work in forest and growth yield modeling has focused largely on improving dominant height and site index model prediction. Climate and soil information were used in this study to improve the estimation of stand mortality.

1.2 Tree-level mortality

At the individual tree-level, forest mortality analyses have largely focused on logistic regression modeling. Hamilton (1974) introduced the logistic function as an individual tree mortality model and it has been widely used since then for modeling mortality of many tree species (e.g. Monserud, 1976; Buchman, 1979; Hamilton, 1986; Avila and Burkhart, 1992; Vanclay, 1995; Yao et al., 2001; Zhao et al., 2007). A generalized logistic model formulation proposed by Monserud (1976) is

$$\pi = \left(1 + e^{-\mathbf{X}\boldsymbol{\beta}}\right)^{-t} \tag{1.2}$$

where π is probability of survival for all trees over a remeasurement interval t, $\mathbf{x}$ is a vector of explanatory variables, and $\boldsymbol{\beta}$ is the vector of parameters.

Limitations of stand-level mortality models based on difference equation approach are that ingrowth must be assumed to be negligible or predicted with another function which is rel-

evant to natural stands but for planted stand ingrowth is generally not a consideration. These whole-stand models predict some level of mortality even when no mortality occurs (Weiskittel et al., 2011). Nearly all individual tree mortality models use logistic regression to model tree level mortality, however, some recent models have been developed that use different approaches. Guan and Gertner (1991a,b) developed a mortality model that was based on artificial neural network. Dobbertin and Biging (1998) used the recursive partitioning (also known as CART) method to predict tree mortality. However, these methods have not led to significant improvement in modeling tree mortality as compared to classical statistical methods (Monserud and Sterba, 1999). Kiernan et al. (2009); Ma et al. (2013) used marginal generalized estimating equations (GEE) proposed by Liang and Zeger (1986). GEE model is an extension of the generalized linear model to longitudinal data. Multiple sources of heterogeneity occur naturally in data from permanent plot systems due to the multilevel data structure inherent in the design. Measurement incidents are nested within trees and trees are nested within plots. Multilevel mixed effects logistic models are more appropriate for such hierarchical data than the traditional logistic regression that does not account for multiple sources of heterogeneity present in the data.

Survival analysis technique in assessing mortality and dynamics of individual trees is a novel approach in forest growth and yield systems. Woodall et al. (2005) provided reasons for using survival analysis in modeling tree. Antón-Fernández (2008) listed lack of data on time-to-event and intrinsic peculiarities of the mortality data as two reasons for the lack of interest in using survival analysis in forestry. The intrinsic peculiarities include the time-dependent nature of the explanatory variables used in predicting mortality and the censoring present in most mortality data. Data in permanent plots are often collected at regular interval of time (e.g. 3 or 5 years) and hence the mortality data from permanent plots are often interval censored. Most of the tree- and stand-level covariates used in logistic regression for modeling

individual tree mortality are time-dependent (e.g. diameter at breast height, crown class, stand age, stand density etc.). The traditional logistic regression does not allow the inclusion of the time-dependent nature of those covariates. Traditional logistic regression uses the values of covariates at the beginning of the measurement period and assumes them to be constant through that period. Leffondŕe et al. (2003) evaluated Cox's model and logistic regression for matched case-control data with time-dependent covariates. They reported that logistic regression estimates were less accurate in some simulation scenarios that involved time-dependent covariates and when the covariates were correlated. They also observed that the logistic regression tended to over-estimate effects of some time-dependent covariates while under-estimating others. Hence, using survival analysis for data with time-dependent covariates would more accurately assess the effects of the covariates on tree mortality. Survival analysis techniques have potential to handle time-dependent covariates, censored observations and allow testing the assumption of a constant hazard function.

Survival analysis techniques have been used sporadically in forestry to model mortality (e.g. Volney, 1998; Rose et al., 2004; Woodall et al., 2005; Rose et al., 2006; Antón-Fernández, 2008). Burgman et al. (1994) developed mortality models for mountain ash and alpine ash using Cox proportional hazards function. Volney (1998) used semiparametric proportional hazards function to study tree mortality in older jack pine stands in a central Saskatchewan forest. Using survival analysis approach, Rose et al. (2004) presented a method for deriving whole-stand survival models that are capable of modeling complex hazard functions. Rose et al. (2006) used Cox's proportional hazards function for modeling interval-censored individual-tree survival and they incorporated random effects to account for multiple source of variability inherent in most permanent plot data. In addition, their model incorporated silvicultural treatment effects on tree survival. Woodall et al. (2005) applied survival analysis for analyzing tree mortality in Minnesota. However, they used diameter at breast height as

a surrogate for time-to-event, a traditionally used time variable in survival analysis. They concluded that the survival analysis approach may lead to more efficient and statistically defensible evaluation of tree mortality. Magnussen et al. (2005) modeled relationship between spruce budworm defoliation and survival-times using Cox proportional hazards model in Prince George Forest Region of British Columbia, Canada. Antón-Fernández (2008) compared survival regression models that incorporated time-dependent covariates with logistic regression approach for modeling individual tree mortality. She concluded that the survival regression models outperformed the logistic. Uzoh and Mori (2012) developed mortality models for even-aged stands of ponderosa pine in the western US using logistic regression and Cox proportional hazards model. Survival analysis technique was used in this study to handle time-dependent covariates and allow for censoring of observations in addition to testing the assumption of a constant hazard function and dealing with non-normal distribution (Klein and Moeschberger, 2003). Shared frailty model was used to account for unobserved heterogeneity or variation not explained by observed covariates in the survival mode.

1.3 Objectives

This research attempts to model both whole-stand and individual-tree mortality of loblolly pine plantations throughout its geographic range in the United States. To meet the research goal, the following specific objectives were outlined:

I. Fit plot level functions ($N_2 = f(N_1, A_1, A_2)$) using algebraic difference approach.

II. Use two-step regression approach to mortality modeling.

III. Incorporate climate and soil information in the mortality models to improve predictions.

IV. Model individual-tree level mortality using multilevel mixed-effects logistic regression.

V. Model individual-tree level survival (or mortality) using Cox proportional hazards model.

VI. Fit shared frailty model to account for the unobserved heterogeneity not explained by the covariates considered in the Cox model.

The first three objectives served the purpose of modeling stand-level mortality of loblolly pine. Climate and soil information were used to explore the potential for improving mortality estimates from the models. The last three objectives served the purpose of modeling tree-level mortality. Multilevel logistic regression model was fitted to account for the full hierarchical structure inherent in data obtained from permanent sample plots. Both plot and tree information from permanent plots were used to develop a survival model to estimate the probability of a tree surviving to a given time period. Shared frailty model was used to account for unobserved heterogeneity. The research methods, analyses and results are organized in three chapters that follow. Chapter 2 addresses those first three specific objectives. Chapter 3 addresses the fourth specific objective. Chapter 4 addresses the last two specific objectives. The final chapter summarizes the research and provides future directions for research on modeling stand and tree mortality.

References

Amateis, R. L., Burkhart, H. E., and Liu, J. (1997). Modeling survival in juvenile and mature loblolly pine plantations. *Forest Ecology and Management*, 90(1):51–58.

Antón-Fernández, C. (2008). *Towards greater accuracy in individual-tree mortality regression*. PhD thesis, Michigan Technological University.

Avila, O. B. and Burkhart, H. E. (1992). Modeling survival of loblolly pine trees in thinned and unthinned plantations. *Canadian Journal of Forest Research*, 22(12):1878–1882.

Bailey, R. L., B. E. Borders, K. D. W., and Jones, Jr, E. P. (1985). A compatible model relating slash pine plantation survival to density, age, site index, and type and intensity of thinning. *Forest Science*, 31(1):180–189.

Buchman, R. G. (1979). Mortality functions. In *A generalized forest growth projection system applied to the Lake States region*, number NC-49, pages 47–55. USDA For. Ser.

Burgman, M. A., Incoll, W., Ades, P. K., Ferguson, I., Fletcher, T. D., and Wohlers, A. (1994). Mortality models for mountain and alpine ash. *Forest Ecology and Management*, 67(1-3):319–327.

Clutter, J. L., Fortson, J. C., Pienaar, L. V., Brister, G. H., and Bailey, R. L. (1983). *Timber management: a quantitative approach*. John Wiley & Sons Inc, New York, NY.

Diéguez-Aranda, U., Castedo-Dorado, F., Álvarez-González, J. G., and Rodŕiguez-Soalleiro, R. (2005). Modelling mortality of Scots pine (*Pinus sylvestris* L.) plantations in the northwest of Spain. *European Journal of Forest Research*, 124(2):143–153.

Dobbertin, M. and Biging, G. S. (1998). Using the non-parametric classifier CART to model forest tree mortality. *Forest Science*, 44(4):507–516.

Eid, T. and Øyen, B. H. (2003). Models for prediction of mortality in even-aged forest. *Scandinavian Journal of Forest Research*, 18(1):64–77.

Fortin, M. and Langevin, L. (2012). Stochastic or deterministic single-tree models: is there any difference in growth predictions? *Annals of Forest Science*, 69(2):271–282.

Guan, B. T. and Gertner, G. Z. (1991a). Modeling red pine tree survival with an artificial neural network. *Forest Science*, 37(5):1429–1440.

Guan, B. T. and Gertner, G. Z. (1991b). Using a parallel distributed processing system to model individual tree mortality. *Forest Science*, 37(3):871–885.

Hamilton, D. A. (1974). Event probabilities estimated by regression. Res. Pap. INT-152, USDA For. Ser.

Hamilton, Jr., D. A. (1986). A logistic model of mortality in thinned and unthinned mixed conifer stands of Northern Idaho. *Forest Science*, 32(4):989–1000.

Kiernan, D., Bevilacqua, E., Nyland, R., and Zhang, L. (2009). Modeling tree mortality in low- to medium-density uneven-aged hardwood stands under a selection system using generalized estimating equations. *Forest Science*, 55(4):343–351.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data*. Springer-Verlag, New York, NY.

Lee, Y. (1971). Predicting mortality for even-aged stands of lodgepole pine. *Forestry Chornicle*, 47(1):29–32.

Leffondŕe, K., Abrahamowicz, M., and Siemiatycki, J. (2003). Evaluation of Cox's model and logistic regression for matched case-control data with time-dependent covariates: a simulation study. *Statistics in Medicine*, 22(24):3781–3794.

Lemin, R. C. and Burkhart, H. E. (1983). Predicting mortality after thinning in old-field loblolly pine plantations. *Southern Journal of Applied Forestry*, 7(1):20–23.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Ma, Z., Peng, C., Li, W., Zhu, Q., Wang, W., Song, X., and Liu, J. (2013). Modeling individual tree mortality rates using marginal and random effects regression models. *Natural Resource Modeling*, 26(2):131–153.

Magnussen, S., Alfaro, R. I., and Boudewyn, P. (2005). Survival-time analysis of white spruce during spruce budworm defoliation. *Silva Fennica*, 39(2):177–189.

McKeand, S. E., Mullin, T. J., Byram, T. D., and White, T. L. (2003). Deplyoment of genetically improved loblolly and slash pine in the south. *Journal of Forestry*, 101(3):32–37.

Monserud, R. A. (1976). Simulation of forest tree mortality. *Forest Science*, 22(4):438–444.

Monserud, R. A. and Sterba, H. (1999). Modeling individual tree mortality for Austrian forest species. *Forest Ecology and Management*, 113(2-3):109–123.

Rose, E. C., Clutter, M. L., Shiver, B. D., Hall, D. B., and Borders, B. (2004). A generalized methodology for developing whole-stand survival models. *Forest Science*, 50(5):686–695.

Rose, E. C., Hall, D. B., Shiver, B. D., Clutter, M. L., and Borders, B. (2006). A multilevel approach to individual tree survival prediction. *Forest Science*, 52(1):31–43.

Somers, G. L., Oderwald, R. G., Harms, W. R., and Langdon, G. O. (1980). Predicting mortality with a Weibull distribution. *Forest Science*, 26(2):291–300.

Staebler, G. R. (1953). Mortality estimation in fully stocked stands of young-growth Douglas-fir. Res. Pap. 4, USDA For. Serv., Pacific Northwest Forest and Range Experiment Station.

Uzoh, F. C. and Mori, S. R. (2012). Applying survival analysis to managed even-aged stands of ponderosa pine for assessment of tree mortality in the western United States. *Forest Ecology and Management*, 285(0):101 – 122.

Vanclay, J. K. (1991). Compatible deterministic and stochastic predictions by probabilistic modeling of individual trees. *Forest Science*, 37(6):1656–1663.

Vanclay, J. K. (1995). Synthesis: growth models for tropical forests: a synthesis of models and methods. *Forest Science*, 41(1):7–42.

Volney, W. J. A. (1998). Ten-year tree mortality following a jack pine budworm outbreak in Saskatchewan. *Canadian Journal of Forest Research*, 28(12):1784–1793.

Waring, R. H., Milner, K. S., Jolly, W. M., Phillips, L., and McWethy, D. (2006). Assessment of site index and forest growth capacity across the Pacific and Inland Northwest U.S.A. with a {MODIS} satellite-derived vegetation index. *Forest Ecology and Management*, 228(13):285–291.

Wear, D. N. and Greis, J. G. (2002). Southern forest resource assessment: summary of findigs. *Journal of Forestry*, 100(7):6–14.

Weber, L. R., Ek, A. R., and Droessler, T. D. (1986). Comparison of stochastic and deterministic mortality equation in an individual tree based stand growth model. *Canadian Journal of Forest Research*, 16(5):1139–1141.

Weiskittel, A. R., Hann, D. W., Kershaw, J. A., and Vanclay, J. K. (2011). *Forest growth and yield modeling*. Wiley-Blackwell, Chicester, UK, 2 edition.

Woodall, C. W., Grambschb, P. L., and Thomas, W. (2005). Applying survival analysis to a large-scale forest inventory for assessment of tree mortality in Minnesota. *Ecological Modelling*, 189(1-2):199–208.

Yao, X., Titus, S. J., and MacDonald, S. E. (2001). A generalized logistic model of individual tree mortality for aspen, white spruce, and lodgepole pine in Alberta mixedwood forests. *Canadian Journal of Forest Research*, 31(2):283–291.

Zhao, D., Borders, B., Wang, M., and Kane, M. (2007). Modeling mortality of second-rotation loblolly pine plantations in the Piedmont/Upper Coastal Plain and Lower Coastal Plain of the southern United States. *Forest Ecology and Management*, 252(1-3):132–143.

Chapter 2

Modeling stand-level mortality of loblolly pine plantation using stand, climate and soil variables

Abstract

Accurate prediction of mortality is an important component of forest growth and yield prediction systems, yet mortality remains one of the least understood components of the system. Remeasurement data collected from permanent sample plots established in 1980/81 across the natural range of loblolly pine in the Atlantic Coastal Plain, Gulf Coastal Plain and Piedmont were used. The biophysical data for plot locations for the years 1980 to 2003 were obtained from the Oak Ridge National Laboratory, Distributed Active Archive Center. The main objective of this study was to develop a model for predicting stand mortality of loblolly pine plantations using the stand characteristics and biophysical variables. The potential of using biophysical variables to refine stand mortality estimates was explored. Models were constructed using two modeling ap-

proaches. In the first approach, mortality functions for directly predicting tree number reduction were developed using an algebraic difference equation method called direct prediction approach here. In the second approach, a two-step modeling strategy was used. In the first step, a model predicting the probability of tree death occurring over a measurement period was developed; in the second step, a function that estimates the reduction in tree number was developed. Performance of all the modeling approaches was compared using leave-one-cluster-out cross-validation procedure. When the biophysical variables were used in the model, the direct prediction approach performed the best. The effects of biophysical variables were not significant in predicting probability of stand mortality at the physiographic level (i.e. for the Coastal Plain and Piedmont) in the two-step approach.

2.1 Introduction

Mortality in forest growth and yield systems can be classified as regular (or non-catastrophic) that results from competition for scarce resources within a stand and irregular (or catastrophic) that results from random disturbances such as fire, wind, snow or insect outbreaks (Vanclay, 1995). Many growth models of even-aged forests account for regular mortality when modeling tree mortality or survival (Monserud, 1976; Monserud and Sterba, 1999). Accurate prediction of mortality is an important part of forest growth and yield prediction systems (Monserud, 1976). However, mortality is one of the least understood components of the system (Hamilton, 1986) due to complex interactions among different factors such as environmental, physiological, pathological and random events. In the early period of forest growth and yield modeling, forest mortality or survival models were overlooked mainly due to difficulty in modeling or predicting mortality based on insufficient long-term data from permanent plots.

Forest survival is usually modeled at the whole stand or individual tree level. Whole stand survival models predict the future stems for a given initial number of trees and corresponding age. Stand-level mortality has been modeled using many approaches and strategies. The algebraic difference equation approach produces satisfactory results by predicting number of live trees at some future point in time from the current number of trees, age and other site variables (Lemin and Burkhart, 1983). Clutter and Jones (1980) developed a survival function based on a difference equation that implied mortality in a stand represented a change in tree per unit area with a change in time. Bailey et al. (1985) and Clutter et al. (1984) used this approach to develop survival models for slash pine and loblolly pine, respectively, in the southeastern United States. The stand-level mortality equations that are developed using an algebraic difference equation approach are given in Equation (1.1) in Chapter 1.

A limitation of stand-level mortality models based on the difference equation approach is that these models predict some level of mortality even when no mortality occurs (Weiskittel et al., 2011b). Permanent plot records often contain data that show no mortality even over several years (Woollons, 1998). Discarding data from the plots where no mortality occurred introduces a bias of significant magnitude resulting in overestimation of mortality. If all the data are retained, difficulties arise in model fitting and mortality rate may be underestimated (Woollons, 1998; Eid and Øyen, 2003). Woollons (1998) suggested a two-step approach to counter the problem above. In the first step, a logistic model predicting the probability of mortality in the subsequent measurement interval, given current stand conditions, is developed. In the second step, an equation that estimates the tree-number reduction is developed and these estimates are modified using deterministic or stochastic approaches (Woollons, 1998; Monserud and Sterba, 1999).

Climate and soil data have been used in forest growth and yield models in order to improve empirical estimation. Past work has focused largely on improving dominant height and site index model prediction (e.g. Woollons et al., 1997; Monserud et al., 2006; Wang et al., 2007; Bravo-Oviedo et al., 2008, 2010; Nunes et al., 2011; Weiskittel et al., 2011a). Monserud et al. (2006) found that different climatic variables had strong impact on site productivity of lodgepole pine for the province of Alberta and accounted for about one quarter of the variation in the species site index. Wang et al. (2007) developed dominant height and site index models for *Eucalyptus globulus* plantations in southeastern Australia and tested effects of fertilizer application and various environmental variables on the dominant height. They reported that climate variables greatly improved model fit and reduced the residual variability among plots. Inclusion of climate and soil variables in developing dominant height growth models has improved their applicability by reducing the bias and improving the model efficiency (Bravo-Oviedo et al., 2008). However, Woollons et al. (1997) did not

find significant improvement in prediction of dominant height growth by including climatic and edaphic variables in conjunction with traditional plot measures in the model. Crookston et al. (2010) adjusted predictors in the Forest Vegetation Simulator (FVS) to account for expected climate effects on mortality, growth and regeneration; the modified model Climate-FVS allows incorporation of climate change impacts in forest plans.

Little or no work has been done towards using climate and soil data in improving estimation of stand-level mortality. In this study, climate and soil data were used in estimating mortality of loblolly pine plantations. Studies based on algebraic difference approach, the most common approach for predicting stand mortality, have not used climate and soil data in predicting stand mortality (e.g. Pienaar et al., 1990; Woollons, 1998; Diéguez-Aranda et al., 2005). Similarly, climate and soil data have not been utilized in any study that used two-step regression approach for modeling stand mortality (e.g. Woollons, 1998; Álvarez González et al., 2004; Diéguez-Aranda et al., 2005; Zhao et al., 2007).

The main objective of this chapter was to model stand-level mortality of loblolly pine plantations using stand and biophysical variables. In this study climate and soil information were used to improve the estimation of stand mortality in both the algebraic difference and two-step regression approach for modeling stand-level mortality.

2.2 Data

2.2.1 Region wide thinning study data

Data sets that are managed by the Forest Modeling Research Cooperative at Virginia Tech were used for this study. The data were collected from permanent sample plots established across the natural range of loblolly pine in the Piedmont, Atlantic Coastal Plain, and Gulf

Coastal Plain regions (Figure 2.1). Piedmont is a relatively flat and topographically featureless region located in the eastern US between the Atlantic Coastal Plain and the Appalachian mountains. Atlantic Coastal Plain is a low relief region along the east coast of the US and the Gulf Coastal Plain is the region that extends around the Gulf of Mexico in the southern United States.

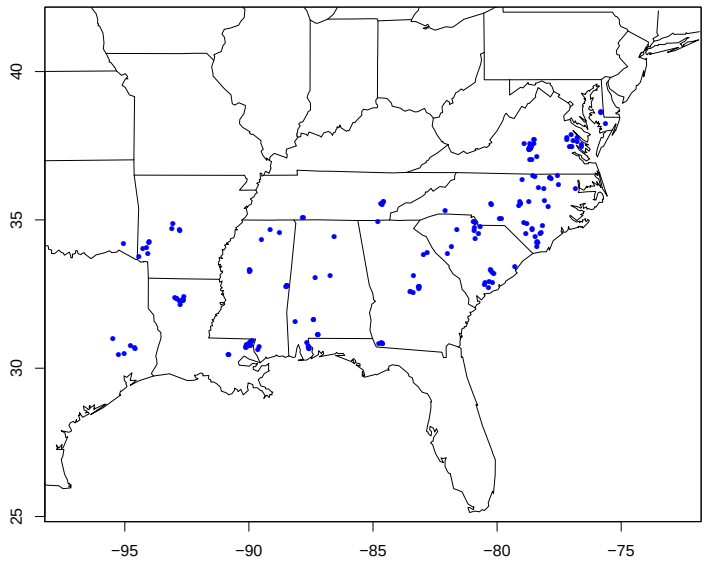


Figure 2.1: Location of 186 permanent plots over the loblolly pine range

A region wide thinning study with permanent plots was established in the dormant seasons of 1980-1981 and 1981-1982 in loblolly pine plantations on cutover, site-prepared lands across the natural range of loblolly pine (Burkhart et al., 1985). At each of 186 sites, located throughout the Piedmont, Atlantic Coastal Plain, and Gulf Coastal Plain physiographic regions of the southeastern US, three treatment plots with similar initial spacing, basal area, and site index were established: an unthinned control, lightly thinned (approximately 1/3 basal area removed), and heavily thinned (approximately 1/2 basal area removed) plot.

Information on diameter at breast height (DBH), total height, crown class (dominant, codom-

inant, intermediate and suppressed) and height to live crown was collected on all trees. Plots were remeasured at 3-year intervals after installation. Data from eight measurements (an initial measurement and seven remeasurements) were gathered in this study. Number of trees per ha at different ages in unthinned plots at all 186 locations are given in Figure 2.2.

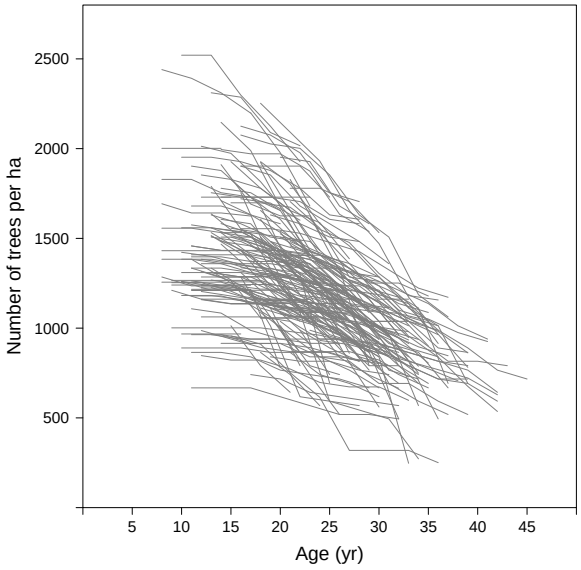


Figure 2.2: Number of trees per ha in unthinned plots

Only the measurements from unthinned plots that were measured at least four times (i.e. measurement at plot establishment and at least three remeasurements afterward) were considered in model fitting. Since modeling mortality due to insect attack was not an objective of the study, measurements from plots that had incidence of insect attacks were dropped from the analyses. There were 151 plots in total and 822 intervals of 3-year in the data used.

Summary statistics at the time of plot establishment and seven remeasurements for 186 permanent plots are given in Table 4.1. Number of observations (N), mean, minimum and maximum for the stand characteristics for all inventories (measurement at plot establishment and seven remeasurements) are given.

Table 2.1: Summary statistics of stand characteristics

Measurements	**N**	**Age (years)**			**Dominant height (m)** [1]			**Site index (m)** [2]			**Basal area (m^2 ha^{-1})**			**Number of trees (ha^{-1})**		
		Mean	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**
Plot establishment	557	15.2	8.0	25.0	11.9	4.2	22.6	17.8	10.8	25.8	18.5	2.9	53.5	961.3	226.3	2520.5
1^{st} remeasurement	557	18.2	11.0	28.0	14.0	6.2	24.5	17.8	10.8	25.8	22.3	4.7	53.7	931.6	226.3	2520.5
2^{nd} remeasurement	535	21.1	14.0	31.0	15.5	7.8	24.4	17.7	10.8	24.8	25.2	4.3	47.1	895.7	109.8	2311.1
3^{rd} remeasurement	464	23.9	17.0	34.0	16.8	9.1	25.5	17.5	10.8	24.8	27.6	9.7	47.3	866.0	226.3	2197.9
4^{th} remeasurement	385	26.6	20.0	37.0	18.3	10.4	25.9	17.5	10.8	24.8	30.3	10.7	48.7	824.9	226.3	1927.4
5^{th} remeasurement	218	29.6	23.0	40.0	20.0	11.2	27.9	17.6	11.5	23.6	33.7	11.4	50.3	833.7	226.3	1630.9
6^{th} remeasurement	182	32.3	26.0	43.0	21.1	12.5	27.9	17.6	11.5	23.6	34.1	12.2	52.2	771.9	285.5	1606.2
7^{th} remeasurement	115	34.9	29.0	45.0	22.1	13.5	30.0	17.7	12.5	23.6	35.1	11.7	50.3	724.1	247.1	1171.8

[1]Calculated following protocol developed by PINEMAP

[2]Calculated using the equation of Diéguez-Aranda et al. (2005) under PINEMAP protocol

2.2.2 Biophysical data

Daily precipitation and temperature data were obtained for each of the 186 locations for the years 1980 to 2003 from the Oak Ridge National Laboratory, Distributed Active Archive Center (Thornton et al., 2012). These data are an interpolation and extrapolation of daily weather data from a network of weather stations, accomplished using the models of Thornton et al. (1997). The following information on annual and seasonal climate condition for each location was obtained after processing the daily weather data:

i. Annual Growing Degree Days (5^0C baseline)

ii. Annual Precipitation (mm)

iii. Annual Minimum Temperature (^{0}C)

iv. Annual Maximum Temperature (^{0}C)

v. Mean Annual Temperature (^{0}C)

vi. Growing Season Precipitation (mm)

vii. Growing Season Growing Degree Days (5^0C baseline)

viii. Growing Season Dryness Index

ix. Mean Growing Season Temperature (^{0}C)

x. Length of Growing Season (days)

xi. Number of Days in Growing Season with Precipitation $\geq$ 13 mm

xii. January Mean Maximum Temperature (^{0}C)

xiii. July Mean Maximum Temperature (^{0}C)

xiv. Summer Dryness Index

xv. Summer Growing Degree Days (5^0C baseline)

xvi. Summer Precipitation (mm)

xvii. Summer Mean Maximum Temperature (^{0}C)

xviii. January-July Temperature Differential (^{0}C)

The growing season was defined as the number of days between the last spring frost (minimum daily temperature $\leq 0\,^0$C) and the first fall frost. When the first fall frost did not occur by the end of the year, 365 was used as the day of the first fall frost. Summer was defined as the time period from the first day of June to the last day of August.

The climate event variables were not used in their original form but as lagged explanatory variables in fitting models. Simple moving average of order 3 of each variable as in Equation (2.1) was used in the analysis,

$$m_{it} = \frac{\sum_{j=0}^{k} x_{i(t-j)}}{k+1} \tag{2.1}$$

where $i = 1, \ldots, N$, $k = 2$, N is a number of plots, t is year, x is a biophysical variable and m_{it} is moving average of biophysical variable for plot i at year t.

The estimates of climate variables at year t for a plot were obtained by averaging the last three values of the time series ($k = 0, 1, 2$), including the value of present time series t. Since the climate data obtained from the Oak Ridge National Laboratory had measurements starting 1980, estimate of the average for year 1981 was a moving average of two values of the time series (i.e. 1980 and 1981). The estimates of all averages after year 1981 were moving average

of three values of the time series. Averaging eliminates some of the randomness in the data. The use of average effect of last three years of climate variable observations in a model for predicting probability of stand mortality was more plausible than the use of single year effect of the variable.

The soil data were obtained from the USDA Natural Resource Conservation Service SSURGO soil database (Soil Survey Staff, 2012). The following soil characteristics were downloaded from the Soil Data Mart:

i. Soil available water storage capacity for the depth 0 to 150 cm

ii. Percent sand

iii. Percent silt

iv. Percent clay

v. Percent organic matter

vi. Soil depth (to the 2-meter USDA observation maximum)

The soil characteristics for each study location were extracted from the map unit data using GIS point-polygon data extraction techniques.

2.3 Methods

A two-step regression approach was used to model observed mortality in loblolly pine plantations. In the first step, a model predicting probability of tree death occurring over a measurement period was developed using data from all plots, with and without occurrence

of mortality. The climate and soil data were also used in addition to stand variables to estimate the probability in the first step. In the second step, a mortality function that estimates the reduction in tree number due to natural mortality was developed using data from plots with occurrence of mortality only.

2.3.1 Model for predicting mortality probability

The logistic function has been extensively used in modeling mortality of stand or individual trees (e.g. Monserud, 1976; Buchman, 1979; Hamilton, 1986; Avila and Burkhart, 1992; Vanclay, 1995; Woollons, 1998; Yao et al., 2001; Diéguez-Aranda et al., 2005; Zhao et al., 2007). Let the stand mortality occurring over an interval be represented by binary outcomes (i.e. $y_i = 1$ when mortality occurs in the stand and $y_i = 0$ when mortality does not occur). The traditional logistic model is given by:

$$\pi_i = \left(1 + e^{-\mathbf{x}_i'\boldsymbol{\beta}}\right)^{-1} \tag{2.2}$$

where π_i is probability of stand mortality occurring in plot i over an interval (i.e. $\Pr(y_i = 1)$), $\mathbf{x}_i$ is a vector of explanatory variables, and $\boldsymbol{\beta}$ is a vector of parameters.

Permanent sample plot systems in forestry observe the response for each plot repeatedly at several time points. This hierarchical structure of data from permanent sample plots is accounted for by using the two-level model formulation given by Equation (2.3). Let i $(i = 1, \ldots, N)$ where N denotes number of plots (clusters) and t $(t = 1, \ldots, n_i)$ where n_i denotes measurement occasions on plot i (nested observations). The total number of observations is $\sum_{i=1}^{N} n_i$.

$$\pi_{it} = \left(1 + e^{-\mathbf{x}_{it}^{'}\boldsymbol{\beta}}\right)^{-1} \tag{2.3}$$

where π_{it} represents the probability of mortality occurring in plot i at measurement occasion t (i.e. $\Pr(y_{it} = 1)$), $\mathbf{x}_{it}$ is the vector of explanatory variables of plot i at measurement occasion t and $\boldsymbol{\beta}$ is the vector of parameters to be estimated.

The parameter estimates for both fixed-effects and mixed-effects logistic regression models were obtained using the SAS procedure NLMIXED (SAS Institute Inc., 2011).

2.3.1.1 Generalized estimating equations

Generalized linear models (GLM) are fixed effects models that assumes independence among all observations. Hence, straightforward application of GLM to longitudinal data from permanent sample plots is not appropriate (Hedeker and Gibbons, 2006). However, they can be extended to account for the correlation inherent in longitudinal data using Generalized Estimating Equations (GEE) proposed by Liang and Zeger (1986).

In GEE models the joint distribution of a subject's response vector $\boldsymbol{y}_i$ does not need to be specified but only the marginal distribution of y_{it} at each time point is specified. As in GLM the mean response μ_{it} is linearly related to covariates via an appropriate link function

$$g(\mu_{it}) = \mathbf{x}_{it}^{'}\boldsymbol{\beta} \tag{2.4}$$

where $g(.)$ is a known link function such as logit link for binary data and $\mathbf{x}_{it}$ is the covariate vector for subject i at time t. The variance is given as a function of the mean

$$V(y_{it}) = \phi\, \upsilon(\mu_{it}) \tag{2.5}$$

where $\upsilon(\mu_{it})$ is a known variance function and ϕ is a scale parameter. For binary response $\upsilon(\mu_{it}) = \mu_{it}(1-\mu_{it})$ and $\phi = 1$.

The "working" correlation matrix $\boldsymbol{R}_i(\boldsymbol{\alpha})$ of size $n_i \times n_i$ is specified for each subject i. $\boldsymbol{\alpha}$ represents a vector of association parameters and is assumed to be the same for all subjects. Note that when $\boldsymbol{R}_i(\boldsymbol{\alpha}) = \boldsymbol{I}$, the $n_i \times n_i$ identity matrix, the GEE reduces to the quasi-likelihood estimating equations for a generalized linear model which assume the repeated measures are independent. The working variance-covariance matrix for $\boldsymbol{y}_i$ can be specified as

$$V(\boldsymbol{\alpha}) = \phi\, \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{A}_i^{1/2} \tag{2.6}$$

where $\boldsymbol{A}_i = \text{diag}\{\upsilon(\mu_{it})\}$ is a diagonal matrix with jth diagonal elements $\upsilon(\mu_{it})$.

The GEE estimator of $\boldsymbol{\beta}$ is the solution of

$$\sum_{i=1}^{N} \boldsymbol{D}_i^{'} [\boldsymbol{V}(\hat{\boldsymbol{\alpha}})]^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = 0 \tag{2.7}$$

where $\hat{\boldsymbol{\alpha}}$ is a consistent estimate of $\boldsymbol{\alpha}$ and $\boldsymbol{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$.

Obtaining GEE estimates requires an iteration between the quasi-likelihood solution for $\boldsymbol{\alpha}$ and a robust method for estimating $\boldsymbol{\alpha}$ as a function of $\boldsymbol{\beta}$ until convergence. The GEE approach has many appealing properties. In many longitudinal designs the GEE estimator $\hat{\boldsymbol{\beta}}$ is almost as efficient as the maximum likelihood estimator and it yields a consistent estimator of $\boldsymbol{\beta}$ even if the with-in subjects correlation structure among the repeated measures is mis-

specified (Lipsitz and Fitzmaurice, 2008). The SAS procedures GENMOD and GLIMMIX (SAS Institute Inc., 2011) were used to fit GEE based on linearization.

2.3.2 Model for predicting tree number reduction

Mortality functions for predicting tree number reduction were developed using algebraic difference equation approach. Clutter et al. (1983) stated the following logical properties a whole stand mortality models should possess:

1. If $A_2 = A_1$, then $N_2 = N_1$.
2. For even-aged stands, if $A_2 > A_1$, then $N_2 \leq N_1$.
3. For even-aged stands, as $A_2 \to \infty$ then $N_2 \to 0$.
4. If N_2 is predicted at age A_2 and A_2 and N_2 are then used to predict N_3 at age A_3 (such that $A_3 > A_2 > A_1$), the result should be the same as that obtained by a single projection from A_1 to A_3. This is path invariance property of the model.

Here N_i represents the number of trees per unit area at age A_i.

Integration of mortality rate equations is often an effective approach for developing a difference equation model for stand-level mortality. In the simplest form constant mortality is modeled by an exponential population decline:

$$\frac{1}{N}\frac{dN}{dA} = \alpha \tag{2.8}$$

where N is number of trees per unit area at age A, $\frac{dN}{dA}$ is instantaneous mortality rate at age A and α is constant.

Or, in the integrated form

$$N_2 = N_1 \exp(\alpha(A_2 - A_1)) \tag{2.9}$$

The assumption of constant mortality rate for all age, site indexes and stand densities is too simplistic for forest stands since mortality rate is related to stand variables (Burkhart and Tomé, 2012). The relative rate of instantaneous mortality could be related to stand age, site index and stand densities as in Equations (2.10) - (2.12).

$$\frac{1}{N}\frac{dN}{dA} = \alpha N^{\beta} f(S) A^{\delta} \tag{2.10}$$

or

$$\frac{1}{N}\frac{dN}{dA} = \alpha N^{\beta} \left(f(S) + \frac{\delta}{A} \right) \tag{2.11}$$

or

$$\frac{1}{N}\frac{dN}{dA} = \alpha N^{\beta} f(S) \delta^{A} \tag{2.12}$$

where $f(S)$ is a function of site index and α, β and δ are parameters.

Different functions have been used to represent the effects of site index on mortality, which can be written as the following general functional form

$$f(S) = \gamma_0 + \gamma_1 S^{\gamma_2} \tag{2.13}$$

The combination of $\beta = 0$ or $\beta \neq 0$ with different functions of age and site index yields many differential equations and integrating them with the initial conditions gives the corresponding difference equation models. Integrating differential equation (2.10) over the initial condition

that $\delta \neq -1$ gives the following models:

when $\beta \neq 0$

$$N_2 = \left(N_1^{b_1} + f(S)\left(A_2^{b_2} - A_1^{b_2}\right)\right)^{\frac{1}{b_1}} \tag{2.14}$$

with $b_1 = -\beta$ and $b_2 = \delta + 1$

when $\beta = 0$

$$N_2 = N_1 \exp\left(f(S)\left(A_2^{b_1} - A_1^{b_1}\right)\right) \tag{2.15}$$

with $b_1 = \delta + 1$

Clutter and Jones (1980), Woollons (1998) and Diéguez-Aranda et al. (2005) used survival functions similar to equation (2.14) and Pienaar and Shiver (1981) and Pienaar et al. (1990) used functions similar to Equation (2.15).

Integrating differential equation (2.11) gives the following models:

when $\beta \neq 0$

$$N_2 = \left(N_1^{b_1} + f(S)(A_2 - A_1) + \ln\left(\frac{A_2}{A_1}\right)^{b_2}\right)^{\frac{1}{b_1}} \tag{2.16}$$

with $b_1 = -\beta$ and $b_2 = -\alpha\beta\delta$

when $\beta = 0$

$$N_2 = N_1 \left(\frac{A_2}{A_1}\right)^{b_1} \exp(f(S)(A_2 - A_1)) \tag{2.17}$$

with $b_1 = -\alpha\beta\delta$

Bailey et al. (1985) used functions similar to Equation (2.17).

Integrating differential Equation (2.12) with the initial condition $\delta > 1$ gives the following models:

when $\beta \neq 0$

$$N_2 = \left(N_1^{b_1} + f(S)\left(b_2^{A_2} - b_2^{A_1}\right)\right)^{\frac{1}{b_1}} \tag{2.18}$$

with $b_1 = -\beta$ and $b_2 = \delta$

when $\beta = 0$

$$N_2 = N_1 \exp\left(f(S)\left(b_1^{A_2} - b_1^{A_1}\right)\right) \tag{2.19}$$

with $b_1 = \delta$

Zhao et al. (2007) used functions similar to Equation (2.18)

The equations (2.9),(2.14), (2.15), (2.16), (2.17), (2.18) and (2.19) were fitted to the data collected from permanent plot systems to predict the mortality in loblolly pine plantations. Estimates of model parameters and the measures of goodness-of-fit were obtained using the SAS procedure MODEL (SAS Institute Inc., 2010) and `nls2` function in **nls2** package in R (Grothendieck, 2013).

2.3.3 Exploratory factor analysis

Climate and soil variables were used in estimation of stand mortality probability. Eleven climate event and six soil variables listed in Section 2.2.2 were included. The essence of factor analysis is its ability to uncover the relationships between the assumed latent variables and the manifest variables. Exploratory factor analysis (EFA) was used to explore the possible underlying factor structure of the set of climate and soil variables without making any

assumptions about which manifest variables were related to which factors.

Let a set of observed or manifest variables, $\mathbf{x}' = (x_1, x_2, \ldots, x_q)$, assumed to be linked to k unobserved common factors $f_1, f_2, \ldots, f_k$, where $k < q$, by a regression model of the form

$$\begin{aligned} x_1 &= \lambda_{11}\, f_1 + \lambda_{12}\, f_2 + \cdots + \lambda_{1k}\, f_k + u_1 \\ x_2 &= \lambda_{21}\, f_1 + \lambda_{22}\, f_2 + \cdots + \lambda_{2k}\, f_k + u_2 \\ &\vdots \\ x_q &= \lambda_{q1}\, f_1 + \lambda_{q2}\, f_2 + \cdots + \lambda_{qk}\, f_k + u_q \end{aligned}$$

where λ_{ij}'s are factor loadings and u_i are random disturbances. Factor loadings are used in the interpretation of the factors i.e. larger values relate a factor to the corresponding observed variables.

The above regression equations can be written as

$$\mathbf{x} = \mathbf{\Lambda f} + \mathbf{u} \tag{2.20}$$

where

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1k} \\ \vdots & \ddots & \vdots \\ \lambda_{q1} & \cdots & \lambda_{qk} \end{pmatrix}, \ \mathbf{f} = \begin{pmatrix} f_1 \\ \vdots \\ f_q \end{pmatrix}, \ \mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_q \end{pmatrix}$$

The random disturbance terms $u_1, \ldots, u_q$ are called specific variates and elements of $\mathbf{u}$ are specific to each x_i. It is further assumed that

$$
\begin{aligned}
E(\mathbf{f}) &= \mathbf{0}, \\
\mathrm{Cov}(\mathbf{f}) &= \mathbf{I}, \\
E(\mathbf{u}) &= \mathbf{0}, \\
\mathrm{Cov}(\mathbf{u}) = \mathbf{\Psi} &= \begin{pmatrix} \Psi_1 & 0 & \cdots & 0 \\ 0 & \Psi_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Psi_p \end{pmatrix}, \\
\mathrm{Cov}(\mathbf{u}, \mathbf{f}) &= \mathbf{0}
\end{aligned}
$$

Then the covariance matrix $\mathrm{Cov}(\mathbf{x})$ is given by

$$
\mathrm{Cov}(\mathbf{x}) = \mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi} \tag{2.21}
$$

The factor analysis model implies that

$$
\begin{aligned}
\mathrm{Var}(x_i) = \sigma_i^2 &= \sum_{j=1}^{k} \lambda_{ij}^2 + \Psi_i \\
\mathrm{Cov}(x_i, x_j) = \sigma_{ij} &= \sum_{l=1}^{k} \lambda_{il}\lambda_{jl}, \qquad i \neq j
\end{aligned}
$$

In practice, $\mathbf{\Sigma}$ is estimated by the sample covariance matrix $\mathbf{S}$. The estimates $\hat{\mathbf{\Lambda}}$ of factor loadings and estimates $\hat{\mathbf{\Psi}}$ of the specific variances are obtained such that

$$\mathbf{S} \approx \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^T + \hat{\mathbf{\Psi}}$$

Determination of an adequate number of factors, k, is critical in fitting an exploratory factor analysis model. There are too many high loadings if too few factors are selected and factors may be fragmented and difficult to interpret with too many factors (Jolliffee, 2002). The number of factors, k, can be selected subjectively by examining results corresponding to different values of k or the scree diagram. The factor loadings are not unique and the problem of non-uniqueness of factor loadings can be solved by introducing some constraints in the original model. The constraint is imposed where

$$\mathbf{G} = \mathbf{\Lambda}^T\mathbf{\Psi}^{-1}\mathbf{\Lambda} \qquad \text{is diagonal}$$

The factor rotation makes interpretation of the loadings simple by splitting the variables into disjoint sets, each associated with one factor. Two types of rotation can be applied: orthogonal rotation that restricts the rotated factors to be uncorrelated and oblique rotation that allows rotated factors to be correlated. Various methods of rotation were assessed before selecting the one that provided a simple factor structure with a meaningful and easily interpretable solution.

2.3.4 Approaches for projecting number of trees

Estimated number of live trees per unit area $\hat{N_2}$ at age A_2 was calculated deterministically using the following approaches

I. Direct prediction approach

This approach estimates the number of live trees $\hat{N}_2$ at age A_2 directly using the algebraic difference form mortality function given by Equation (2.22). The function is obtained by fitting it to all plots data (with and without occurrence of mortality) and it does not consider the probability of stand mortality when estimating N_2.

$$N_2 = f(N_1, A_2, A_1, S) \tag{2.22}$$

where N_i is the number of trees per ha at age A_i and S is the site index.

II. Threshold probability based approach

This approach involves simulating stand mortality using a cutoff probability. The probability of stand mortality ($\hat{\pi}_{it}$) occurring over a period is obtained by Equation (2.3) in Section 2.3.1. If the predicted probability ($\hat{\pi}_{it}$) is greather than or equal to the cutoff, natural mortality occurs in the stand and the trees number $\hat{N}_2$ at age A_2 is determined using the algebraic difference form mortality function selected from Section 2.3.2. If the predicted probability ($\hat{\pi}_{it}$) is less than the threshold, then the $\hat{N}_2 = N_1$ at age A_2. The threshold was selected as a cutoff that maximizes the sensitivity and specificity simultaneously in Receiver-Operating Characteristic (ROC) curves analysis.

III. Decision theory based approach

This approach is based on the decision theory and the predicted number of trees N_{adj2} at age A_2 is determined as (Woollons, 1998; Diéguez-Aranda et al., 2005):

$$N_{adj2} = N_1 - \hat{\pi}_{it}(N_1 - \hat{N}_2) \tag{2.23}$$

where N_1 is the number of trees per ha at the beginning of the period and $\hat{N}_2$ is the number of trees at age A_2 estimated by difference form survival model.

2.3.5 Evaluation and validation of the model for tree number reduction

Since the data used in this study were repeated observations (i.e. longitudinal), leave-one-cluster-out cross-validation was performed for each model as a measure of their predictive performance. It is more generalized version of leave-one-out cross-validation where a whole cluster forms a validation set as opposed to single observation in leave-one-out cross-validation. Cross-validation measures the predictive ability of model on a set of data not used in parameter estimation of the model. Leave-one-cluster-out cross-validation is a more sophisticated version of training/test sets where the predictive accuracy measures were obtained as follows:

Let y_{it} denotes observation from plot i $(i = 1, \ldots, N)$ at measurement occasion t $(t = 1, \ldots, n_i)$. Vectors $\mathbf{y}_1, \ldots, \mathbf{y}_N$ are assumed to be independent.

1. Let observations n_i from plot i form the validation set and the model is fitted using the observations from remaining plots $N - 1$ from training set.

2. Compute the n_i errors for plot i as $(\mathbf{e}_{i,-i} = \mathbf{y}_i - \hat{\mathbf{y}}_{i,-i})$ for the omitted n_i observations. $\hat{\mathbf{y}}_{i,-i}$ is a vector of predicted values for the observations in validation set.

3. Repeat Step 1 and 2 for $i = 1, \ldots, N$ plots to obtain $\mathbf{e}_{1,-1}, \mathbf{e}_{2,-2}, \ldots, \mathbf{e}_{N,-N}$.

4. Compute evaluation statistics such as RMSE, MAE and AIC.

Leave-one-cluster-out cross-validation makes use of available data much more efficiently as observations from a single cluster are omitted at each step. Model fits and cross-validation results of difference form candidate models were evaluated using the following three model evaluation measures:

- Root Mean Square Error (RMSE)

  RMSE is a criterion for measuring the quality of estimation of a model and is given by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\sum_{t=1}^{n_i}(y_{it} - \hat{y}_{it})^2}{n - p}}$$

  where y_{it} and $\hat{y}_{it}$ are observed and predicted values, respectively, p is number of model parameters and $n = \sum_i^N n_i$ is the total number of observations.

- Mean Absolute Error (MAE)

  MAE is also a criterion for assessing average model performance and is given by

$$MAE = \frac{\sum_{i=1}^{N}\sum_{t=1}^{n_i}|y_{it} - \hat{y}_{it}|}{n}$$

- Akaike's Information Criterion (AIC)

  AIC is model selection criterion based on the relationship between information theory and likelihood theory. It has the following form when applied to Gaussian or normal models

$$AIC = n\left[\log\left(\frac{\sum_{i=1}^{N}\sum_{t=1}^{n_i}(y_{it} - \hat{y}_{it}}{n}\right) + \log(2\pi) + 1\right] + 2p$$

  where the first part $\log\left(\frac{\sum_{i=1}^{N}\sum_{t=1}^{n_i}(y_{it} - \hat{y}_{it}}{n}\right)$ measures the goodness-of-fit of the model, which is penalized by model complexity in the second part $2p$.

2.4 Results and discussion

2.4.1 Factor analysis on biophysical data

Explanatory factor analysis was performed on the climate and soil data (simple moving average of order 3) to extract underlying factors with fewer dimension than the original data so that those factors could be used as explanatory variables in a model for estimating stand mortality probability.

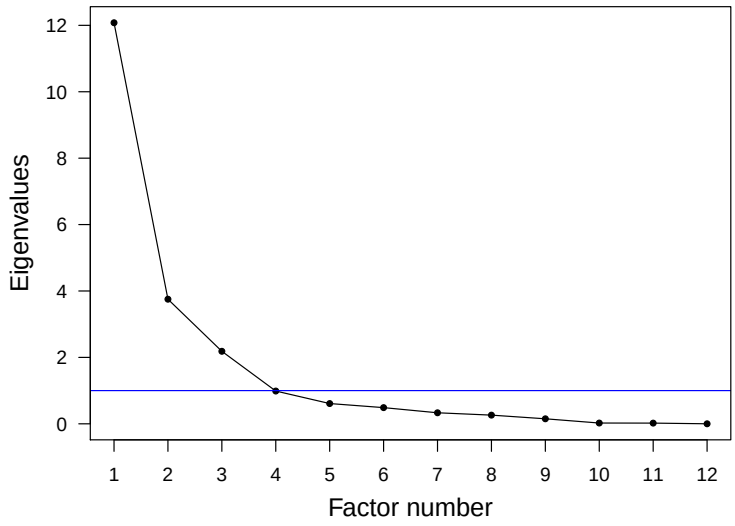


Figure 2.3: Scree plot for determining the number of factors

Three factors were selected based on the scree plot in Figure 2.3 and the Kaiser's criterion which suggests retaining only factors with eigenvalues greater than 1. In the four factors model there were variables with split loadings and the variables loaded in a factors were characteristically different making the interpretation of factor illogical. With the three factors, variables loaded in each factor were similar making the interpretation of factor meaningful and easy.

The variables loaded in each factor with their corresponding loadings are given in Table 2.2.

Table 2.2: Obliquely rotated factor loadings for climate and soil data

Variable	Factor1	Factor2	Factor3
Summer Mean Maximum Temperature (^{0}C)	**0.9864**	-0.1033	-0.0204
July Mean Maximum Temperature (^{0}C)	**0.9267**	-0.2437	-0.0391
Summer Growing Degree Days (5^{0}C baseline)	**0.9162**	0.1211	-0.1541
Mean Growing Season Temperature (^{0}C)	**0.8591**	0.0980	-0.0632
Annual maximum temperature (^{0}C)	**0.8311**	0.3167	-0.1198
Annual Growing Degree Days (5^{0}C baseline)	**0.7844**	0.3819	-0.1709
Mean Annual Temperature (^{0}C)	**0.7780**	0.3868	-0.1804
Growing Season Growing Degree Days (5^{0}C baseline)	**0.7599**	0.3807	-0.1926
Annual minimum temperature (^{0}C)	**0.7046**	0.4417	-0.2321
January Mean Maximum Temperature (^{0}C)	**0.6509**	0.4495	-0.1421
Length of Growing Season (days)	**0.5963**	0.4555	-0.2333
Growing Season Precipitation (mm)	0.2231	**0.8719**	-0.0789
Number of Days in Growing Season with Precipitation $\geq$ 13mm	0.2579	**0.8399**	-0.1014
Summer Precipitation (mm)	-0.2035	**0.8309**	-0.2446
Annual Precipitation (mm)	0.1697	**0.8304**	0.1243
Summer Dryness Index	0.5197	**-0.7256**	0.2128
Growing Season Dryness Index	0.3428	**-0.8617**	-0.0469
Percent Silt	0.1357	0.1168	**0.8639**
Percent Clay	0.0018	-0.0036	**0.7536**
Soil Available Water Storage Capacity 0 to 150 cm	0.1640	0.2545	**0.4758**
Percent Sand	0.0742	0.0707	**-0.7916**
Eigenvalues	12.079	3.7523	2.1839
Variance explained by each factor	7.3085	5.0221	2.4904
Number of variables	11	6	4

Two variables Percent Organic Matter and Soil Depth (cm) did not load significantly (loading <0.30) into any factor and the variable January-July Temperature Differential (^{0}C) had split loading across three factors (loading >0.50 in two factors) and hence those variables were removed from the analysis. The factor pattern matrix in Table 2.2 was used to interpret the meaning of the factors.

The variables significantly loaded in the first factor were Summer Mean Maximum Temperature (^{0}C), July Mean Maximum Temperature (^{0}C), Summer Growing Degree Days (5^{0}C baseline), Mean Growing Season Temperature (^{0}C), Annual maximum temperature (^{0}C), Annual Growing Degree Days (5^{0}C baseline), Mean Annual Temperature (^{0}C), Growing Season

Growing Degree Days (5^0C baseline), Annual minimum temperature (^{0}C), January Mean Maximum Temperature (^{0}C) and Length of Growing Season (days). All of these variables relate to temperature or heat and hence, this factor was named "heat index". The second factor was identified by Growing Season Precipitation (mm), Number of Days in Growing Season with Precipitation $\geq$ 13mm, Summer Precipitation (mm), Annual Precipitation (mm), Summer Dryness Index, Growing Season Dryness Index. Variables in the second factor relate to moisture and dryness of the climate and this factor was named "drought index". The variables loaded in the third factors are Percent Silt, Percent Clay, Percent Sand and Soil Available Water Storage Capacity 0-150 cm. These variables mainly define soil texture and therefore the third factor was named "soil texture index". The names "heat index" and "drought index" used in this study do not have any relevance to the terms used in meteorological science. The terms were used for convenience of interpretation of factors.

The Kernel density plots of all three factors are given Figure 2.4, which show the distribution of factor scores. Each factor scores were calculated as linear combination of factor scoring coefficients and the raw data values.

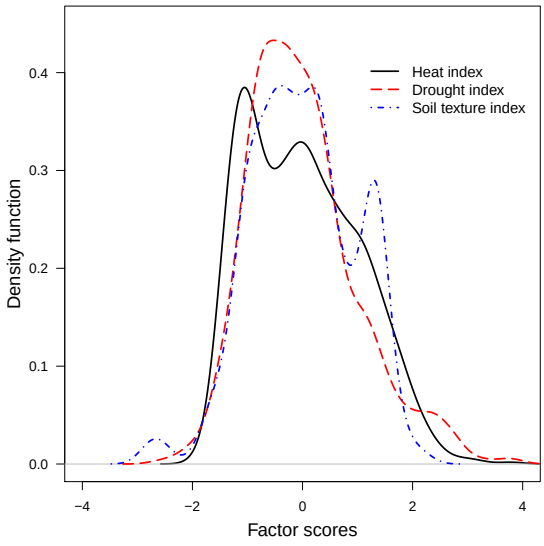


Figure 2.4: Kernel density plots of three factors

2.4.2 Model for predicting tree number reduction

Various algebraic difference form equations were fitted to two sets of data. First the equations were fitted to plot data with and without occurrence of mortality (Table 2.3) and then to a subset of plot data with occurrence of mortality over the remeasurement interval (Table 2.6).

2.4.2.1 Model for direct prediction of tree number reduction

Data from all plots (with and without occurrence of mortality) were used to obtain models for directly predicting reduction in tree numbers per unit area. Sixteen candidate models were fitted to the data to obtain the mortality functions (Table 2.3). The parameter estimates (standard error), $p-$value and the statistics such as RMSE, MAE and AIC are given for both model fit and leave-one-out cross-validation. Model M14 (Equation 2.24) had the smallest RMSE, MAE and AIC for both model fit and cross-validation and hence it was selected as the equation for direct estimation of tree number reduction.

$$N_2 = \left[N_1^{-0.7503} + 0.0213 \left(\frac{S}{10000} \right)^{2.0017} \left(A_2^{2.9607} - A_1^{2.9607} \right) \right]^{-\frac{1}{0.7503}} \tag{2.24}$$

Diéguez-Aranda et al. (2005) proposed an equation of the same form for estimating the reduction in tree number in even-aged stands of Scots pine in Galicia, Spain. The standard error of estimate of b_1 of Model M14 was quite large relative to $\hat{b}_1$ (Table 2.3). This model satisfies logical properties of whole stand mortality models described in Section 2.3.2 and setting parameter $b_1 = 0$ forces Equation 2.24 to collapse. Statistical fits are important in model selection, however, better approach is fitting growth model based on ecologically important variables (Yang et al., 2003). A model with good ecological behaviors may be preferred over a model with entirely good statistical fit. Hence, the parameter b_1 was retained

in the equation even the standard error of $\hat{b}_1$ was large relative to $\hat{b}_1$.

The RMSE for fit and cross-validation in Table 2.3 were calculated from 822 observations with a 3-year time-step of prediction. The RMSE of model M14 was 72.93 and 73.59 trees ha^{-1} for fit and cross-validation, respectively. For the same base model, Álvarez González et al. (2004) reported model fit RMSE of 86.77 trees ha^{-1} and cross-validation RMSE of 91.81 trees ha^{-1} for 3-year time-step prediction for *Pinus radiata* in northwestern Spain. Similarly, Diéguez-Aranda et al. (2005) reported RMSE of 87.5 and 93.4 trees ha^{-1} for fit and cross-validation, respectively, for 6-year time-step prediction for Scots pine in the same region. The RMSEs reported in this study for loblolly pine were less than or comparable to the RMSEs reported in those two studies for different pine species.

Table 2.3: Parameter estimates, fit statistics and leave-one-out cross-validation statistics of models fitted with all plots data for direct prediction of the tree number reduction

Model	Equation	Parameter	Estimate (SE)	Pr > |t|	Fit			Cross-validation		
					RMSE	MAE	AIC	RMSE	MAE	AIC
M1	$N_2 = N_1 \exp\left(b_1(A_2 - A_1)\right)$	b_1	-0.0228 (0.0008)	< .0001	82.67	59.66	9593.78	82.77	59.79	9597.01
M2	$N_2 = N_1 \exp\left(b_1\left(A_2^{b_2} - A_1^{b_2}\right)\right)$	b_1	-0.0001 (0.0001)	0.0293	75.12	52.83	9437.38	75.34	53.05	9443.65
		b_2	2.4180 (0.1247)	< .0001						
M3	$N_2 = N_1 \exp\left(b_1(A_2 - A_1)\right)\left(\frac{A_2}{A_1}\right)^{b_2}$	b_1	-0.0526 (0.0024)	< .0001	75.21	52.63	9439.25	75.40	52.82	9444.48
		b_2	0.5892 (0.0444)	< .0001						
M4	$N_2 = N_1 \exp\left(b_1\left(b_2^{A_2} - b_2^{A_1}\right)\right)$	b_1	-0.1168 (0.0243)	< .0001	75.94	53.86	9455.15	76.18	54.12	9461.80
		b_2	1.0573 (0.0050)	< .0001						
M5	$N_2 = \left[N_1^{b_0} + b_1(A_2 - A_1)\right]^{\frac{1}{b_0}}$	b_0	0.5739 (0.1264)	< .0001	81.80	59.99	9577.44	82.07	60.26	9583.70
		b_1	-0.8050 (0.9018)	0.3723						
M6	$N_2 = \left[N_1^{b_0} + b_1\left(A_2^{b_2} - A_1^{b_2}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.3098 (0.1477)	0.0363	74.96	52.04	9434.85	75.31	52.41	9444.48
		b_1	2.1×10^{-6} (2×10^{-6})	0.3358						
		b_2	2.6046 (0.1602)	< .0001						
M7	$N_2 = \left[N_1^{b_0} + b_1(A_2 - A_1) + b_2 \ln\left(\frac{A_2}{A_1}\right)\right]^{\frac{1}{b_0}}$	b_0	0.0660 (0.1206)	0.5845	75.24	52.71	9440.96	75.53	53.02	9449.34
		b_1	-0.0055 (0.0148)	0.7088						
		b_2	0.0618 (0.1649)	0.7079						
M8	$N_2 = \left[N_1^{b_0} + b_1\left(b_2^{A_2} - b_2^{A_1}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.3362 (0.1533)	0.0286	75.76	53.00	9452.24	76.12	53.38	9462.15
		b_1	0.0025 (0.0021)	0.2325						
		b_2	1.0665 (0.0069)	< .0001						
M9	$N_2 = N_1 \exp\left(b_1 S^{b_2}(A_2 - A_1)\right)$	b_1	-0.0015 (0.0012)	0.2229	82.24	59.54	9586.116	82.59	59.86	9595.09
		b_2	0.9561 (0.2860)	0.0009						

Model	Equation	Parameter	Estimate (SE)	Pr > |t|	Fit			Cross-validation		
					RMSE	MAE	AIC	RMSE	MAE	AIC
M10	$N_2 = N_1 \exp\left(b_1 \left(\frac{S}{10000}\right)^{b_2} \left(A_2^{b_3} - A_1^{b_3}\right)\right)$	b_1	-0.5080 (0.7744)	0.5120	73.88	52.35	9411.02	74.33	52.74	9424.25
		b_2	1.3408 (0.2384)	$< .0001$						
		b_3	2.4694 (0.1201)	$< .0001$						
M11	$N_2 = N_1 \exp\left(b_1 S^{b_2} (A_2 - A_1)\right) \left(\frac{A_2}{A_1}\right)^{b_3}$	b_1	-0.0104 (0.0033)	0.0015	74.02	52.22	9414.13	74.39	52.54	9424.47
		b_2	0.5817 (0.1079)	$< .0001$						
		b_3	0.6215 (0.0442)	$< .0001$						
M12	$N_2 = N_1 \exp\left(b_1 S^{b_2} \left(b_3^{A_2} - b_3^{A_1}\right)\right)$	b_1	-0.0026 (0.0019)	0.1833	74.79	53.43	9431.12	75.26	53.85	9444.65
		b_2	1.3089 (0.2430)	$< .0001$						
		b_3	1.0600 (0.0049)	$< .0001$						
M13	$N_2 = \left[N_1^{b_0} + b_1 S^{b_2} (A_2 - A_1)\right]^{\frac{1}{b_0}}$	b_0	0.5013 (0.1302)	0.0001	81.60	59.92	9574.44	82.10	60.38	9586.43
		b_1	-0.0631 (0.1015)	0.5348						
		b_2	0.6643 (0.2824)	0.0189						
M14	$N_2 = \left[N_1^{b_0} + b_1 \left(\frac{S}{10000}\right)^{b_2} \left(A_2^{b_3} - A_1^{b_3}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.7503 (0.1655)	$< .0001$	72.93	50.34	9390.70	73.59	50.87	9408.52
		b_1	0.0213 (0.0357)	0.5499						
		b_2	2.0017 (0.2866)	$< .0001$						
		b_3	2.9607 (0.1766)	$< .0001$						
M15	$N_2 = \left[N_1^{b_0} + b_1 \left(\frac{S}{10000}\right)^{b_2} (A_2 - A_1) + b_3 \ln\left(\frac{A_2}{A_1}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.1218 (0.1240)	0.3262	74.02	52.07	9415.14	74.50	52.50	9428.68
		b_1	0.1344 (0.1004)	0.1807						
		b_2	0.6082 (0.1101)	$< .0001$						
		b_3	-0.0322 (0.0055)	$< .0001$						
M16	$N_2 = \left[N_1^{b_0} + b_1 \left(\frac{S}{10000}\right)^{b_2} \left(b_3^{A_2} - b_3^{A_1}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.7564 (0.1705)	$< .0001$	73.87	51.40	9411.73	74.56	51.97	9430.20
		b_1	36.665 (59.427)	0.5387						
		b_2	1.9402 (0.2905)	$< .0001$						
		b_3	1.0827 (0.0078)	$< .0001$						

In Table 2.3, the models where $\frac{1}{N}\frac{dN}{dA}$ was not proportional to a function of site index $f(S)$ (models M1-M8) performed poorly in terms of model fit and leave-one-out cross-validation relative to the models where $\frac{1}{N}\frac{dN}{dA}$ was proportional to a function of site index (models M9 - M16). Inclusion of a function of site index as an explanatory variable reduced the RMSE, MAE and AIC of fit and cross-validation in general. Álvarez González et al. (2004) and Diéguez-Aranda et al. (2005) observed improvement in the estimates by including site index in mortality models. Zhao et al. (2007) found that the models in which $\frac{1}{N}\frac{dN}{dA}$ was proportional to site index provided accurate estimates and performed better in both Lower Coastal Plain and Piedmont/Upper Coastal Plain regions. Among the models that did not include the function of site index, model M6 performed the best. The selected model M14 for direct prediction of tree number reduction was in fact obtained by including the function of site index in model M6. Inclusion of the function of site index in the model caused RMSE of model fit to drop from 74.96 to 72.93, MAE from 52.04 to 50.34 and AIC from 9434.85 to 9390.70. In cross-validation results, there was a drop in RMSE, MAE and AIC of similar magnitude.

The models for direct prediction of tree number reduction were obtained for two physiographic regions, namely Coastal Plain and Piedmont, as well. The same base model that was selected for the whole physiographic region was selected for the Coastal Plain and Piedmont based on fit statistics. The corresponding models for direct prediction of tree number reduction for the Coastal Plain and Piedmont regions are given by Equation (2.25) and Equation (2.26), respectively.

$$N_2 = \left[N_1^{-0.9166} + 0.0056 \left(\frac{S}{10000} \right)^{1.9465} \left(A_2^{2.9983} - A_1^{2.9983} \right) \right]^{-\frac{1}{0.9166}} \tag{2.25}$$

$$N_2 = \left[N_1^{-0.9183} + 1.8 \times 10^{-5} \left(\frac{S}{10000} \right)^{1.0345} \left(A_2^{2.8873} - A_1^{2.8873} \right) \right]^{-\frac{1}{0.9183}} \tag{2.26}$$

Model fits of mortality functions of both Coastal Plain and Piedmont areas for directly predicting the reduction in tree numbers per unit area are given in Appendix A.1.

2.4.3 Incorporating biophysical variables in a model for direct prediction of tree number reduction

Potential of using biophysical variables to refine stand mortality estimates was explored. An attempt was made to incorporate a measure of biophysical variables into the model for direct prediction of tree number reduction. The data were arbitrarily divided into three heat index (HI) classes and separate sets of coefficient for each of the classes were estimated (Table 2.4). Similarly, the data set was divided into three drought index (DI) classes and soil texture index (STI) classes separately and sets of coefficient for each class were estimated in a similar fashion as was done for HI classes. The potential contribution of biophysical variables to mortality prediction was assessed by fitting model M14 in Table 2.3 to the divided data separately and examining the relationship between the mean of each class of biophysical variables and the coefficients of the model. A similar approach has been implemented before in incorporating crown ratio and other variables into taper equations by Burkhart and Walton (1985); Muhairwe et al. (1994) and Jiang et al. (2007).

Table 2.4: Estimated coefficients in different classes of biophysical variables

Biophysical variable	**Parameter estimates**			
	b_0	b_1	b_2	b_3
Heat index classes				
-1.885 $\geq HI <$ -0.584	-0.5524	2.8 $\times 10^{-5}$	1.2134	3.6982
-0.584 $\geq HI <$ 0.482	-1.0349	2 $\times 10^{-5}$	1.1033	2.7849
0.482 $\geq HI <$ 3.182	-0.6585	0.3952	2.2630	2.8218
Drought index classes				
-2.641 $\geq DI <$ -0.574	-0.9328	0.0590	2.5077	3.2677
-0.574 $\geq DI <$ 0.233	-0.9426	0.0004	1.6367	3.0805
0.233 $\geq DI <$ 3.909	-0.6233	0.0097	1.5368	2.5862
Soil texture index classes				
-2.822 $\geq STI <$ -0.454	-0.7557	0.0029	1.6836	2.9464
-0.454 $\geq STI <$ 0.369	-0.9634	0.0032	1.8797	2.9559
0.369 $\geq STI <$ 2.225	-0.9477	1.2074	3.1147	3.4455

Results from separate, independent fittings showed that the estimate of parameter b_2 first decreased and then increased with increasing heat index classes, it decreased with increasing drought index classes and it increased with increasing soil texture index classes (Table 2.4). The results suggested that biophysical variables could potentially be added in the mortality model. The biophysical variables were included in model M14 by expressing parameter b_2 as a linear function of HI, DI and STI as

$$b_2 = c_0 + c_1\, HI + c_2\, DI + c_3\, STI \tag{2.27}$$

The model M14 became

$$N_{2\,cli} = \left[N_1^{b_0} + b_1 \left(\frac{S}{10000} \right)^{c_0 + c_1 HI + c_2 DI + c_3 STI} \left(A_2^{b_3} - A_1^{b_3} \right) \right]^{\frac{1}{b_0}} \tag{2.28}$$

Equation (2.28) was fitted to the whole data set by including each biophysical variable (HI, DI and STI) one at a time as a function of b_2. Parameters of HI and DI were highly significant (p-value <0.0009) and the parameter of STI was marginally significant (p-value = 0.0314) in the three respective fits. Then the model was fitted with two variables combinations followed by all three variables. Among all the seven combinations (HI, DI, STI, $HI + DI$, $HI + STI$, $DI + STI$ and $HI + DI + STI$), the RMSE, MAE and AIC of the leave-one-cluster-out cross-validation were smallest when b_2 was expressed as linear combination of HI and DI variables. The parameter of STI was not significant (p-value = 0.2815) when all three biophysical variables were included in Equation (2.27). When the quadratic form of HI, DI and STI was used in Equation (2.27), the values of RMSE, MAE and AIC were not better (i.e. not smaller) than when the variables were in first-degree. The parameter b_2 was expressed as a function of HI and DI and the final model was

$$N_{2\,cli} = \left[N_1^{b_0} + b_1 \left(\frac{S}{10000} \right)^{c_0 + c_1 HI + c_2 DI} \left(A_2^{b_3} - A_1^{b_3} \right) \right]^{\frac{1}{b_0}} \tag{2.29}$$

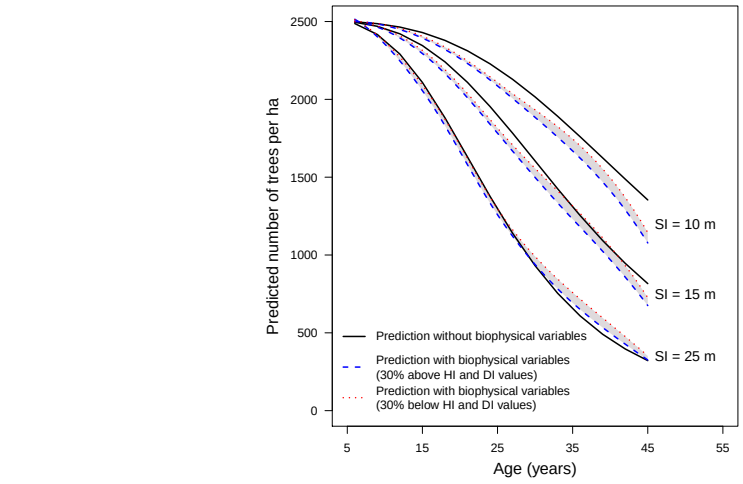
The parameter estimates of Equation (2.29) are given in Table 2.5.

Table 2.5: Parameter estimates for direct prediction model incorporating biophysical variables

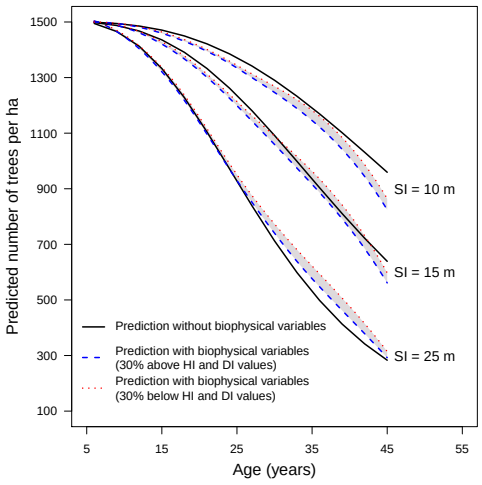
Parameter	**Estimate**	**SE**	**t-value**	**Pr** $> \|t\|$
b_0	-0.9103	0.1697	-5.36	$< .0001$
b_1	0.0012	0.0020	0.58	0.5591
b_3	3.0413	0.1767	17.21	$< .0001$
c_0	1.7405	0.2717	6.40	$< .0001$
c_1	-0.0223	0.0045	-4.93	$< .0001$
c_2	-0.0132	0.0052	-2.52	0.0120

Impact on predicted number of trees per ha was graphically assessed by holding all variables constant except HI and DI over three different values of initial number of trees per ha and site indexes. Figure 2.5 shows surviving number of trees per ha at different site indexes for

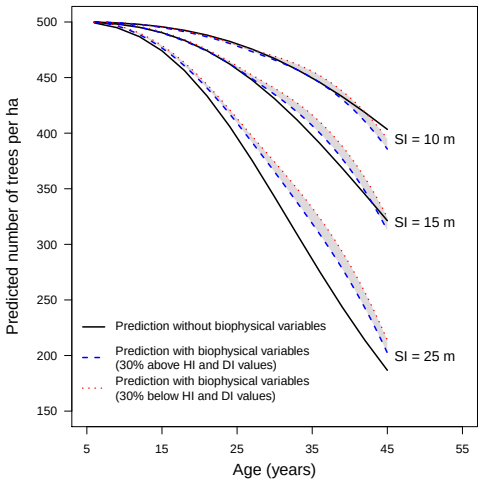
a particular initial stand density and Figure 2.6 shows the surviving number of trees per ha at different stand densities for a specified site index.



(a)

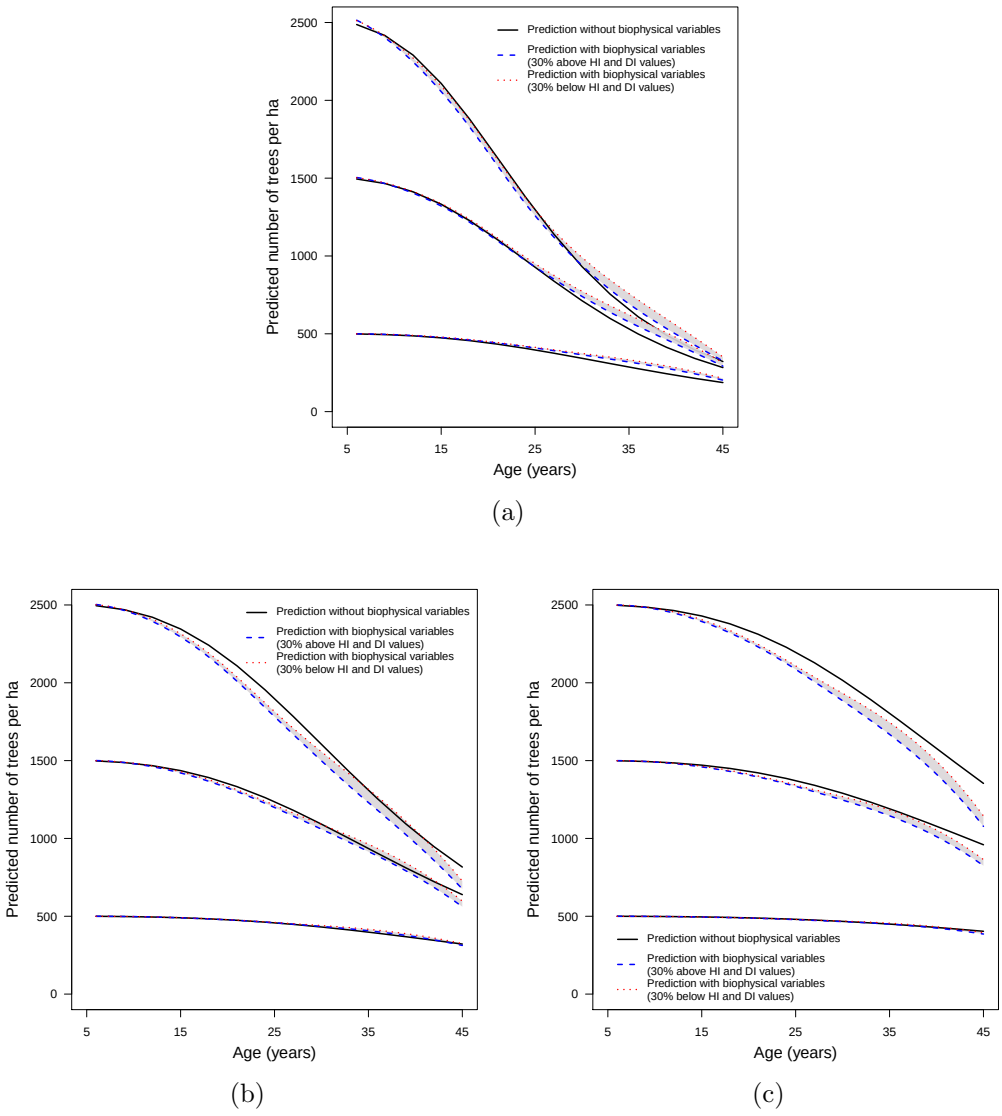


(b)



(c)

Figure 2.5: Predicted number of trees per ha at different site indexes, holding everything constant except HI and DI (a) high initial density (2500 tree per ha), (b) medium initial density (1500 tree per ha) and (c) low initial density (500 tree per ha)



(a)

(b)

(c)

Figure 2.6: Predicted number of trees per ha at different tree densities, holding everything constant except HI and DI (a) high site index (25 m), (b) medium site index (15 m) and (c) low site index (10 m)

Three initial densities 2500, 1500 and 500 tress per ha at age 5 years and three site indexes 25, 15 and 10 m at base age 25 years were considered. The HI and DI values were generated from normal distribution with mean and variance equal to the mean and variance of HI and DI, respectively in the data. The dash and dotted curves in Figure 2.5 and Figure 2.6

represent prediction using biophysical variables with HI and DI values 30% above and below the HI and DI values generated from normal distribution. The values were varied by plus and minus 30% in order to evaluate whether or not that magnitude of difference resulted in a substantial shift in the predicted outcome. The number of surviving trees from age 5 through 45 years, represented by the solid curve was predicted by model M14 in Table 2.3 (or Equation 2.24). The number of surviving trees represented by dash and dotted curves were predicted by model in Equation (2.29). Both models have the same base form, and the parameter b_2 in model M14 was expressed as a linear function of biophysical variables to obtain Equation (2.29).

In higher density stands prediction from the model that incorporated biophysical variables ($\hat{N}_{2\,cli}$) was close to the prediction from base model that did not incorporate biophysical variables ($\hat{N}_2$) at site indexes 25 m and 15 m (Figure 2.5 (a)). At lower site index (10 m) in that high density stand, the $\hat{N}_{2\,cli}$ curve was above $\hat{N}_2$ and the gap between them increased with increasing age. In medium density stands the trends were similar to high density stands at three site indexes (Figure 2.5 (b)). However, in lower density stands the gap between $\hat{N}_{2\,cli}$ and $\hat{N}_2$ curves was more at high site index of 25 m (Figure 2.5 (c)). From the plots, it appeared that the survival may not be affected by relatively large change (30% here) in HI and DI and thus there would be no detectable changes within the range of climate change that will be generated by the climate models. In both the base model prediction and prediction incorporating biophysical variables, number of surviving trees per ha decreased (i.e. mortality increased) with the increasing site index at all initial densities (Figure 2.6). Bailey et al. (1985) Álvarez González et al. (2004), Diéguez-Aranda et al. (2005) and Zhao et al. (2007) found similar results in their studies. Empirical evidence suggests that density-dependent mortality in plantations starts earlier in better sites and increases with site productivity (Diéguez-Aranda et al., 2005). However, mortality has been found to

decrease with increasing site index in few studies (e.g. Woollons, 1998; Zhao et al., 2007).

Potential of using biophysical variables in lieu of site index in model M14 in Table 2.3 was also explored. The base model considered for this purpose was model M6, the model that preformed best among alternatives that did not incorporate site index and has similar form as model M14:

$$N_2 = \left[N_1^{b_0} + b_1 \left(A_2^{b_2} - A_1^{b_2}\right)\right]^{\frac{1}{b_0}} \tag{2.30}$$

with the parameter b_1 expressed as a linear function of biophysical variables as

$$b_1 = c_0 + c_1\ HI + c_2\ DI + c_3\ STI \tag{2.31}$$

This is equivalent to adding $(c_0 + c_1\ HI + c_2\ DI + c_3\ STI)$ in place of $b_1 S^{b_2}$ in model M14. Model fit was assessed in a similar way as it was done for Equation (2.28). All the combinations of biophysical variables (HI, DI, STI, $HI + DI$, $HI + STI$, $DI + STI$ and $HI + DI + STI$) were non-significant. There was no improvement in model fit and cross-validation results as compared to the model M14 (with site index) in terms of RMSE, MAE and AIC. The results suggested that incorporating biophysical variables directly in lieu of site index was not feasible.

Similarly, an attempt was made to refine the stand mortality estimates at physiographic regions, Coastal Plain and Piedmont, by incorporating a measure of biophysical variables into the corresponding models. Final models (M14's) for direct prediction of tree number reductions for the Coastal Plain and Piedmont listed in Appendix A.1 and Appendix A.2, respectively, were fitted to each regional data set by including one biophysical variable at a time, two variables and all three variables combinations as a function of parameter b_2 in the

same way as was done for the whole region. For Coastal Plain, when b_2 was expressed as a function of DI only, the RMSE, MAE and AIC of the model fits and leave-one-cluster-out cross-validation were smallest. None of the other biophysical variables were significant. The final model for Coastal Plain was

$$N_{2\,cli} = \left[N_1^{-1.1259} + 1.7 \times 10^{-3} \left(\frac{S}{10000} \right)^{1.9926 - 0.0193 DI} \left(A_2^{3.0643} - A_1^{3.0643} \right) \right]^{-\frac{1}{1.1259}} \tag{2.32}$$

For Piedmont, the RMSE, MAE and AIC of the model fits and leave-one-cluster-out cross-validation were smallest when b_2 was expressed as a function of HI among all the combinations. The final model for Piedmont was

$$N_{2\,cli} = \left[N_1^{-0.7061} + 2.2 \times 10^{-4} \left(\frac{S}{10000} \right)^{1.0751 - 0.0322 HI} \left(A_2^{2.6044} - A_1^{2.6044} \right) \right]^{-\frac{1}{0.7061}} \tag{2.33}$$

2.4.4 Model for prediction of tree number reduction in two-step approach

A subset of data from plots with occurrence of mortality was used to obtain a model for predicting reduction in tree numbers per unit area in the two-step approach. Sixteen candidate models were fitted to data from plots with occurrence of mortality to obtain the mortality functions (Table 2.6). The parameter estimates (standard error), $p-$value and the statistics such as RMSE, MAE and AIC are given for both model fit and leave-one-out cross-validation. Model M14 (Equation 2.34) had the smallest RMSE, MAE and AIC for both model fit and cross-validation among all models and hence it was selected as the equation for estimating

tree number reduction in the two-step modeling approach.

$$N_2 = \left[N_1^{-0.4081} + 0.0773 \left(\frac{S}{10000} \right)^{1.6355} \left(A_2^{2.4838} - A_1^{2.4838} \right) \right]^{-\frac{1}{0.4081}} \tag{2.34}$$

This equation has the same form as the equation for direct estimation of tree number reduction (Equation 2.24). The standard error of $\hat{b}_1$ was quite larger relative to $\hat{b}_1$ (Table 2.6) here as well but the parameter b_1 was retained in the equation.

Table 2.6: Parameter estimates, fit statistics and leave-one-out cross-validation statistics of models fitted with plots data with occurrence of mortality only for prediction of tree number reduction in two-step approach

Model	Equation	Parameter	Estimate (SE)	Pr > |t|	Fit			Cross-validation		
					RMSE	MAE	AIC	RMSE	MAE	AIC
M1	$N_2 = N_1 \exp\left(b_1(A_2 - A_1)\right)$	b_1	-0.0268 (0.0008)	< .0001	81.31	57.11	8123.76	81.37	57.23	8126.63
M2	$N_2 = N_1 \exp\left(b_1(A_2^{b_2} - A_1^{b_2})\right)$	b_1	-0.0004 (0.0002)	0.0252	75.89	52.47	8028.59	76.14	52.71	8034.16
		b_2	2.1356 (0.1208)	< .0001						
M3	$N_2 = N_1 \exp\left(b_1(A_2 - A_1)\right)\left(\frac{A_2}{A_1}\right)^{b_2}$	b_1	-0.0544 (0.0029)	< .0001	76.03	52.71	8031.05	76.25	52.93	8036.13
		b_2	0.5753 (0.0567)	< .0001						
M4	$N_2 = N_1 \exp\left(b_1\left(b_2^{A_2} - b_2^{A_1}\right)\right)$	b_1	-0.2120 (0.0483)	< .0001	76.37	53.18	8037.35	76.64	53.44	8043.21
		b_2	1.0458 (0.0049)	< .0001						
M5	$N_2 = \left[N_1^{b_0} + b_1(A_2 - A_1)\right]^{\frac{1}{b_0}}$	b_0	0.6386 (0.1134)	< .0001	79.78	57.45	8098.34	80.05	57.74	8104.11
		b_1	-1.6679 (1.6412)	0.3099						
M6	$N_2 = \left[N_1^{b_0} + b_1\left(A_2^{b_2} - A_1^{b_2}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.0255 (0.1419)	0.8574	75.95	52.41	8030.56	76.31	52.79	8039.29
		b_1	-7.3×10^{-6} (3×10^{-5})	0.8130						
		b_2	2.1516 (0.1547)	< .0001						
M7	$N_2 = \left[N_1^{b_0} + b_1(A_2 - A_1) + b_2 \ln\left(\frac{A_2}{A_1}\right)\right]^{\frac{1}{b_0}}$	b_0	0.1653 (0.1218)	0.1752	75.98	52.19	8031.26	76.30	53.26	8039.03
		b_1	-0.0286 (0.0453)	0.5281						
		b_2	0.2956 (0.4623)	0.4623						
M8	$N_2 = \left[N_1^{b_0} + b_1\left(b_2^{A_2} - b_2^{A_1}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.0203 (0.1454)	0.8890	76.42	53.13	8039.34	76.81	53.53	8048.33
		b_1	0.0036 (0.0216)	0.8662						
		b_2	1.0464 (0.0066)	< .0001						
M9	$N_2 = N_1 \exp\left(b_1 S^{b_2}(A_2 - A_1)\right)$	b_1	-0.0017 (0.0013)	0.1766	80.65	57.23	8113.48	80.98	57.53	8121.55
		b_2	0.9624 (0.2579)	0.0002						

Model	Equation	Parameter	Estimate (SE)	Pr > |t|	Fit			Cross-validation		
					RMSE	MAE	AIC	RMSE	MAE	AIC
M10	$N_2 = N_1 \exp\left(b_1 \left(\frac{S}{10000}\right)^{b_2} \left(A_2^{b_3} - A_1^{b_3}\right)\right)$	b_1	-1.0606 (1.5691)	0.4993	74.47	52.34	8003.18	74.97	52.74	8014.75
		b_2	1.2871 (0.2317)	$< .0001$						
		b_3	2.1981 (0.1190)	$< .0001$						
M11	$N_2 = N_1 \exp\left(b_1 S^{b_2} (A_2 - A_1)\right) \left(\frac{A_2}{A_1}\right)^{b_3}$	b_1	-0.0104 (0.0035)	0.0029	74.71	52.24	8007.60	75.11	52.58	8017.12
		b_2	0.5942 (0.1139)	$< .0001$						
		b_3	0.6131 (0.0563)	$< .0001$						
M12	$N_2 = N_1 \exp\left(b_1 S^{b_2} \left(b_3^{A_2} - b_3^{A_1}\right)\right)$	b_1	-0.0052 (0.0038)	0.1735	75.02	53.00	8013.48	75.54	53.41	8025.24
		b_2	1.2584 (0.2338)	$< .0001$						
		b_3	1.0487 (0.0049)	$< .0001$						
M13	$N_2 = \left[N_1^{b_0} + b_1 S^{b_2} (A_2 - A_1)\right]^{\frac{1}{b_0}}$	b_0	0.5629 (0.1168)	$< .0001$	79.50	57.57	8094.43	80.00	58.04	8105.17
		b_1	-0.1391 (0.1977)	0.4819						
		b_2	0.6385 (0.2516)	0.0114						
M14	$N_2 = \left[N_1^{b_0} + b_1 \left(\frac{S}{10000}\right)^{b_2} \left(A_2^{b_3} - A_1^{b_3}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.4081 (0.1592)	0.0106	74.17	51.26	7998.52	74.83	51.83	8013.97
		b_1	0.0773 (0.1205)	0.5217						
		b_2	1.6355 (0.2730)	$< .0001$						
		b_3	2.4838 (0.1716)	$< .0001$						
M15	$N_2 = \left[N_1^{b_0} + b_1 S^{b_2} (A_2 - A_1) + b_3 \ln\left(\frac{A_2}{A_1}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.0322 (0.1268)	0.7998	74.76	52.18	8009.538	75.26	52.64	8021.98
		b_1	0.0003 (0.0008)	0.7388						
		b_2	0.6011 (0.1169)	$< .0001$						
		b_3	-0.0158 (0.0487)	0.7456						
M16	$N_2 = \left[N_1^{b_0} + b_1 S^{b_2} \left(b_3^{A_2} - b_3^{A_1}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.3782 (0.1619)	0.0198	74.78	52.01	8009.88	75.47	52.60	8025.89
		b_1	0.0459 (0.0384)	0.3001						
		b_2	1.5620 (0.2736)	$< .0001$						
		b_3	1.0605 (0.0074)	$< .0001$						

The models for prediction of tree number reduction in the two-step approach were obtained for the Coastal Plain and Piedmont regions. The corresponding models for Coastal Plain and Piedmont are given by Equation (2.35) and Equation (2.36), respectively

$$N_2 = \left[N_1^{-0.5236} + 0.0279 \left(\frac{S}{10000} \right)^{1.5934} \left(A_2^{2.5556} - A_1^{2.5556} \right) \right]^{-\frac{1}{0.5236}} \tag{2.35}$$

$$N_2 = \left[N_1^{-0.5797} + 5.4 \times 10^{-4} \left(\frac{S}{10000} \right)^{0.9312} \left(A_2^{2.3422} - A_1^{2.3422} \right) \right]^{-\frac{1}{0.5797}} \tag{2.36}$$

Model fits of mortality functions of both Coastal Plain and Piedmont for use in two-step prediction approach are given in Appendix A.2.

2.4.4.1 Model for predicting stand mortality probability

Three different logistic regression models were fitted to data from all plots. The first model was a fixed-effects logistic model that did not take into account the clustering effect present in this data due to repeated measurements of each permanent sample plot. This clustering effect, arising due to the nature of repeated data, was taken into account in the model by fitting Equation (2.37) with (i) mixed-effects logistic model and (ii) Generalized Estimating Equations (GEE), a marginal model proposed by Liang and Zeger (1986).

$$\log \left[\frac{\pi_{it}}{1 - \pi_{it}} \right] = \beta_0 + \beta_1 A_{it} + \beta_2 S_i + \beta_3 N_{it} + \beta_4 HI_i + \beta_5 DI_i + \beta_6 STI_i \tag{2.37}$$

where π_{it} is the probability of mortality occurring in plot i at measurement occasion t, A_{it} is age of plot i at time t, S_i is site index of plot i, N_{it} is number of trees per ha of plot i at time t, HI_i is heat index of plot i, DI_i is drought index of plot i, STI_i is soil texture index

of plot i and $\beta_1, \ldots, \beta_6$ are parameters.

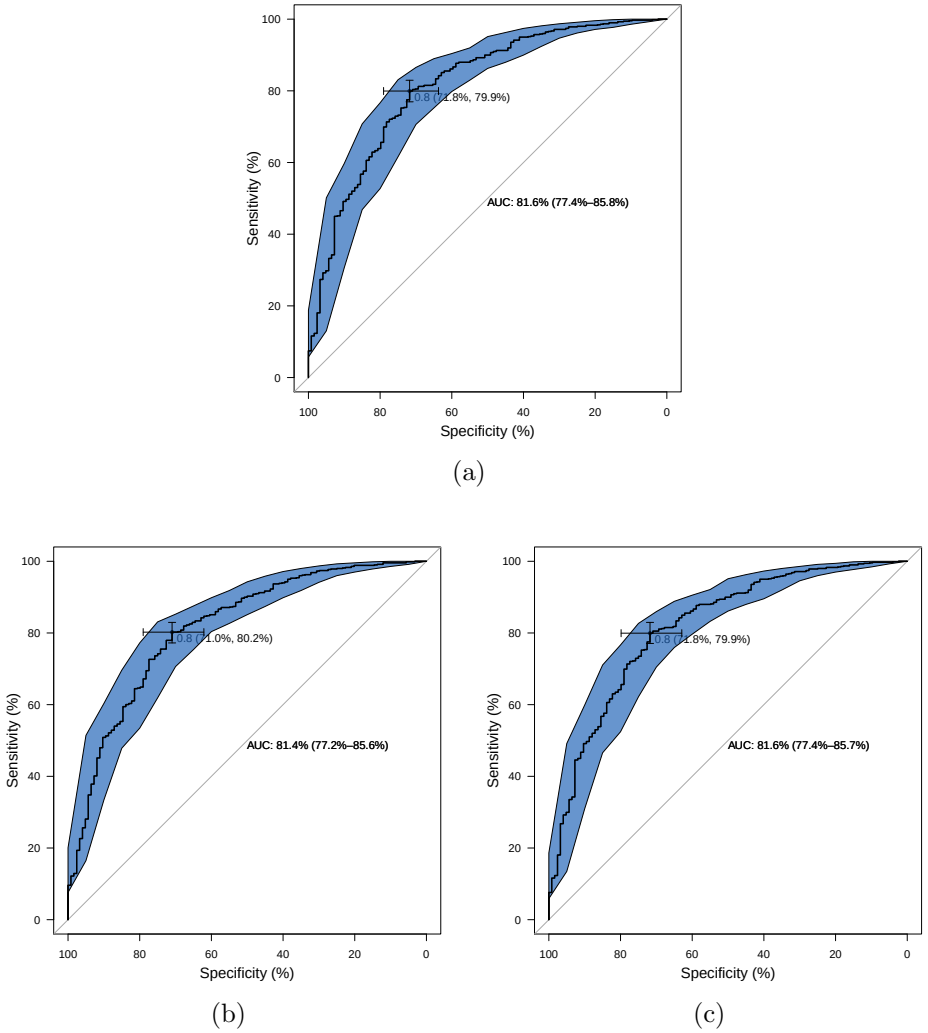
The parameter estimates, standard error, $p-$value, loglikelihood and AIC of three logistic models are given in Table 2.7.

Table 2.7: Parameter estimates of different logistic models for predicting mortality probability at regional level

Model	**Parameter**	**Estimate**	**SE**	**Pr** $>\|t\|$	**-2 Log** L	**AIC**
Marginal logistic	β_0	-8.5989	1.5400	<.0001	543.0	557.0
	β_1	0.2360	0.0233	<.0001		
	β_2	0.1215	0.0582	0.0371		
	β_3	0.0029	0.0005	<.0001		
	β_4	0.3244	0.1273	0.0110		
	β_5	0.0350	0.1280	0.7846		
	β_6	-0.2273	0.1195	0.0576		
Mixed-effects logistic	β_0	-8.5719	1.6260	<.0001	544.4	560.4
	β_1	0.2419	0.0251	<.0001		
	β_2	0.1091	0.0597	0.0695		
	β_3	0.0030	0.0005	<.0001		
	β_4	0.3595	0.1318	0.0071		
	β_5	-0.0295	0.1299	0.8207		
	β_6	-0.3515	0.1328	0.0090		
Marginal GEE logistic	β_0	-8.6314	1.4188	<.0001	-	-
	β_1	0.2362	0.0230	<.0001		
	β_2	0.1223	0.0562	0.0311		
	β_3	0.0029	0.0005	<.0001		
	β_4	0.3231	0.1105	0.0036		
	β_5	-0.0360	0.1145	0.7535		
	β_6	-0.2298	0.1056	0.0299		

Values for both the -2 Log L and AIC (Table 2.7) were lower for a fixed-effects model suggesting that the mixed-effects model may not be required for modeling mortality probability. All of the parameters were significant except for the drought index and soil texture index. With the marginal model GEE, all the parameters were significant except the drought index. Since clustering effects were present in the data, using the GEE logistic regression for modeling stand mortality probability was reasonable and provided more accurate specification of correlation structure of the data. Kiernan et al. (2009) used GEE to predict individual-tree mortality with repeated measure data from permanent plots and they observed that GEE model was better able to capture the change in the probability of mortality over time.

Receiver operating characteristics (ROC) graphs of model fits were generated to graphically visualize and select the best classifier among the three logistic models in Table 2.7, based on their performance in Figure 3.4. In ROC curves the true positive rate (sensitivity) is plotted against the false positive rate (1-specificity) for different cutoff points. The sensitivity-specificity pair describes the diagnostic accuracy.



(a)

(b)

(c)

Figure 2.7: ROC curve and AUC (a) marginal logistic, (b) mixed-effects logistic and (c) marginal GEE logistic

The Area Under Curve (AUC) of all three models was above 0.81 (Figure 3.4) suggesting a good discrimination accuracy (Harrell, 2001) and the AUCs were similar too. The model for predicting the probability of stands death occurring over a period of 3-year was fitted using GEE logistic regression. The parameter estimates of the fitted model are given in Table 2.7 and the fitted model was

$$\begin{aligned}\log\left[\frac{\pi_{it}}{1-\pi_{it}}\right] &= -8.6314 + 0.2362\,A_{it} + 0.1223\,S_i + 0.0029\,N_{it} \\ &\quad + 0.3231\,HI_i - 0.0360\,DI_i - 0.2298\,STI_i\end{aligned} \tag{2.38}$$

Stand characteristics namely stand age (A), site index (S) and trees per ha (N) were all significant in predicting the probability of mortality of stand over a 3-year period. The positive estimated coefficients of A, S, N and HI indicate expected increase in the log odds of stand mortality for higher age, site index, tree density and heat index over the period. Similarly, negative coefficients of DI and STI indicate decrease in the log odds of mortality for higher drought index and soil texture index. Age, site index and number of trees have been used in other stand mortality models (Woollons, 1998; Álvarez González et al., 2004; Diéguez-Aranda et al., 2005; Zhao et al., 2007) .

Parameter estimates for H, DI and STI were not significant (Table 2.8). The effects of climate and soil variables in predicting stand mortality of loblolly pine plantations at resolution smaller than regional i.e. at the level of physiographic regions (Coastal Plain and Piedmont) were not significant.

Table 2.8: Parameter estimates of different logistic models for predicting mortality probability for Coastal Plain and Piedmont

Model	Parameter	Coastal Plain					Piedmont				
		Estimate	**SE**	**Pr** $>\|t\|$	**-2 Log** L	**AIC**	**Estimate**	**SE**	**Pr** $>\|t\|$	**-2 Log** L	**AIC**
Marginal logistic	b_0	-12.3048	2.3212	<.0001	260.5	274.5	-5.6087	2.9008	0.0541	213.1	227.1
	b_1	0.2416	0.0328	<.0001			0.2013	0.0390	<.0001		
	b_2	0.2383	0.0791	0.0027			0.0449	0.1157	0.6986		
	b_3	0.0045	0.0008	<.0001			0.0018	0.0007	0.0093		
	b_4	0.1452	0.1959	0.4588			0.2997	0.2143	0.1631		
	b_5	-0.0101	0.1911	0.9577			-0.0742	0.1972	0.7069		
	b_6	-0.1266	0.1749	0.4697			-0.1443	0.2012	0.4739		
Mixed-effects logistic	b_0	-12.4755	2.7483	<.0001	262.9	277.9	-5.6123	2.9205	0.0594	213.1	229.1
	b_1	0.2669	0.0418	<.0001			0.2014	0.0399	<.0001		
	b_2	0.2212	0.0850	0.0111			0.0449	0.1160	0.7001		
	b_3	0.0046	0.0010	<.0001			0.0018	0.0007	0.0124		
	b_4	0.2695	0.2140	0.2116			0.3006	0.2277	0.1917		
	b_5	-0.1824	0.2182	0.4059			-0.0744	0.1978	0.7082		
	b_6	-0.0661	0.1794	0.7133			-0.1442	0.2017	0.4773		
Marginal GEE logistic	b_0	-12.3326	2.1330	<.0001	-	-	-5.6266	2.5065	0.0285	-	-
	b_1	0.2418	0.0320	<.0001			0.2016	0.0380	<.0001		
	b_2	0.2390	0.0791	0.0034			0.0451	0.1039	0.6658		
	b_3	0.0046	0.0007	<.0001			0.0018	0.0007	0.0114		
	b_4	0.1469	0.1782	0.4102			0.3038	0.1786	0.0903		
	b_5	-0.0114	0.1841	0.9508			-0.0753	0.2012	0.7084		
	b_6	-0.1285	0.1603	0.4235			-0.1444	0.1889	0.4452		

Parameter estimates, standard error, $p-$value, loglikelihood and AIC of three models for predicting probability of stand mortality for Coastal Plain and Piedmont physiographic regions are given in Table 2.8. Some results of exploratory factor analysis for the Coastal Plain and Piedmont regions are given in Appendix B.

The parameters of heat index (HI), drought index (DI) and soil texture index (STI) were not significant for both Coastal Plain and Piedmont (Table 2.8) suggesting that the effects of climate and soil variables on predicting stand mortality may not be pronounced at physiographic level in the range of loblolly pine. However, the effects of climate and soil were significant at the regional level (Table 2.7). Analyses done with two separate physiographic regions may have already accounted for the effects of climate and soil variables to some extent in predicting stand mortality and, hence, those biophysical variables must have been non significant in predicting the probability of mortality of stand at physiographical regions level.

Equation (2.38) predicts the probability of stand mortality occurrence; this probability was translated into discrete events (i.e. mortality occurred and mortality did not occur) deterministically by selecting a threshold probability. Cutpoint analysis was used to determine the optimal threshold. The cutoff on the predicted probability of stand mortality was chosen such that a response was classified as positive response (i.e. stand mortality occurs) if the predicted probability exceeded the cutoff. There are various methods of choosing the cutoff and in this study a cutoff that maximized both the sensitivity and the specificity was chosen (Figure 2.8). Hein and Weiskittel (2010) used the cutpoint analyses to determine optimal threshold in their study on branch mortality.

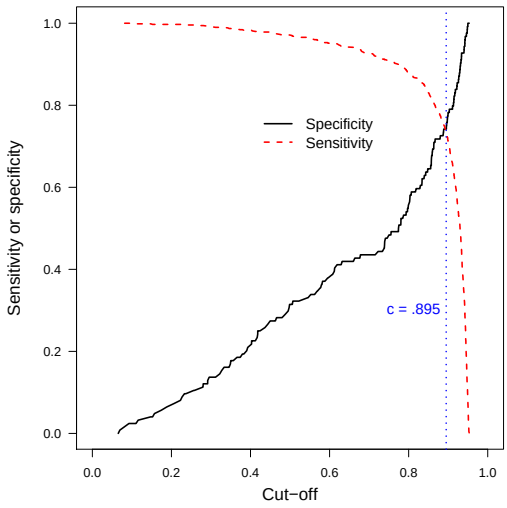


Figure 2.8: Cut-off probability that maximizes sensitivity and specificity

The cutoff probability for dichotomizing the predicted probability to 0 and 1 was 0.895. The chosen cutoff maximized the sum of sensitivity and specificity i.e. it maximized the percentage of well-classified plots in each possible outcome.

An alternative deterministic method used was based on decision theory (e.g. Woollons et al., 1997; Álvarez González et al., 2004; Diéguez-Aranda et al., 2005; Zhao et al., 2007) and the predicted number of trees at age A_2 was calculated using Equation (2.23) in Section 2.3.4.

2.4.5 Cross-validation of approaches for projection of tree numbers

Leave-one-cluster-out cross-validation was performed for all three prediction approaches as a measure of model performance. Root Mean Square Error (RMSE), mean absolute error (MAE) and AIC are given in Table 2.9 for each modeling strategy.

Table 2.9: Model performance measures from cross-validation for whole region

Modeling approach	**RMSE**	**MAE**	**AIC**
Direct prediction (without biophy. var.)	73.59	50.87	9408.52
Direct prediction (with biophy. var.)	72.49	49.69	9387.91
Threshold probability based	74.66	52.19	9435.87
Decision theory based	72.72	49.79	9388.87

When the biophysical variables were not incorporated in the direct prediction model, the decision theory based approach had the smallest RMSE, MAE and AIC. However, when those variables were incorporated in the direct prediction model by expressing parameter b_2 as a linear function of HI and DI, the direct prediction approach outperformed the decision theory based approach exhibiting the smallest RMSE, MAE and AIC (Table 2.9). Bravo-Oviedo et al. (2008) reported that parameters of their selected dynamic equation, when expressed as a function of climatic and soil attributes, led to an improvement in site index prediction of Mediterranean maritime pine in Spain. The authors found that the inclusion of climatic attributes improved the applicability of the inter-regional model in regions in regions where climate and soil type lead to intra-regional variability. Nunes et al. (2011) observed further increase in predictive ability of dominant height growth model of maritime pine in Portugal by expanding parameters of base models as sub-functions of climate and soil variables.

The threshold probability based approach had the largest RMSE, MAE and AIC among all

modeling strategies. Harrell (2001) listed some reasons why the cutoff approach should be avoided when dichotomizing the predicted probabilities in logistic regressions. There might be better strategies for dichotomizing the predicted probabilities but other possibilities were not explored in this study. The observed numbers of trees per ha at age A_2 for all sample plots were compared with the predicted values obtained by using the three approaches of tree number projection in Figure 2.9. The time-step of predictions was 3-year for all the approaches.

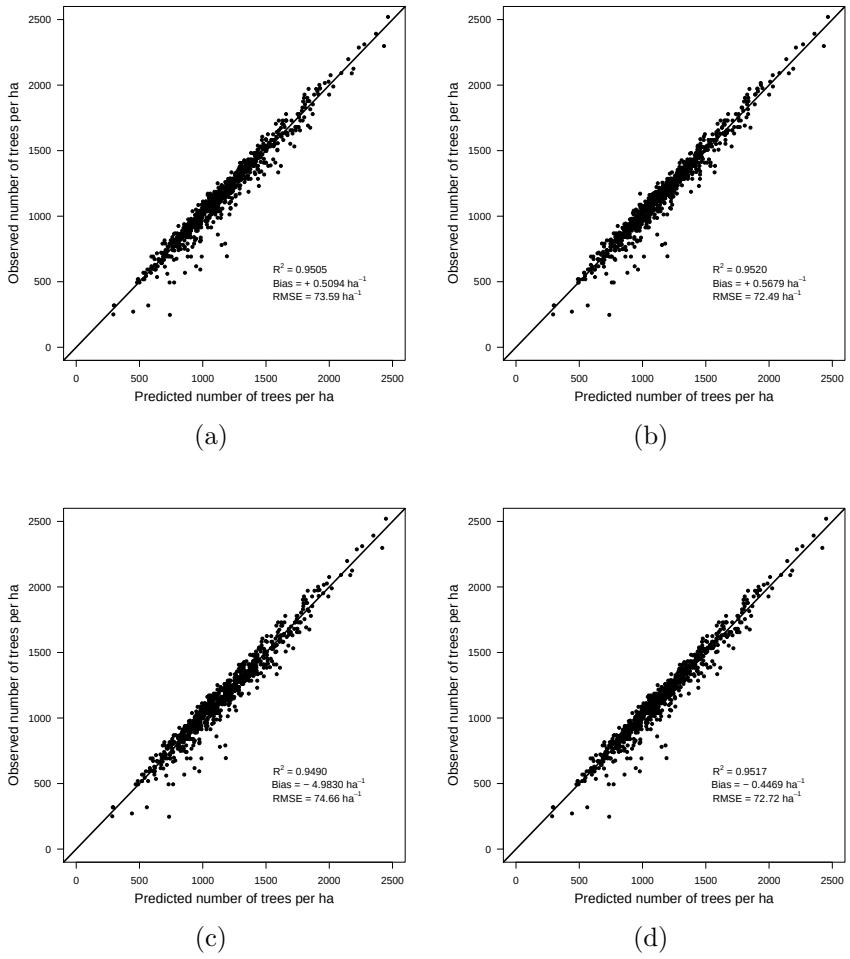


Figure 2.9: Observed against predicted number of trees per ha for (a) direct prediction (without biophysical variables), (b) direct prediction (with biophysical variables), (c) threshold probability based and (d) decision theory based

A linear model was fitted and the coefficient of determination (R^2), bias and RMSE were reported. The results from Figure 2.9 were similar to the conclusions from cross-validation measures in Table 2.9. R^2 values were similar for all three approaches. The decision theory based approach had the smallest bias which overestimated the number of trees per ha by 0.4469 ha^{-1} on average and the threshold approach had the largest bias which overestimated the number of trees per ha by 4.9830 ha^{-1} on average. The direct prediction (without biophysical variables) underestimated the number of trees per ha by 0.5094 ha^{-1} and the direct prediction (with biophysical variables) approach underestimated by 0.5679 ha^{-1} on average. Diéguez-Aranda et al. (2005) in their study reported that the decision theory based approach performed the best and gave the smallest critical error (a measure of prediction interval given in Reynolds (1984)), followed by the direct prediction approach and then the threshold approach.

In this study when the biophysical variables were incorporated in the direct prediction approach, it performed better than the decision theory based approach at least in reducing the RMSE, MAE and AIC in cross-validation process. This result suggests that the direct prediction approach could still be preferred over the two-step approach and this is the only approach that directly guarantees the path-invariance property of whole stand mortality models. Using an equation to directly predict the tree number reduction without using climate and soil information and without taking into account the probability of stand mortality may be a good alternative as well.

Leave-one-cluster-out cross-validation was performed for the three approaches for two physiograhic regions too (Table 2.10).

Table 2.10: Model performance measures from cross-validation for physiographic regions

Modeling approach	**Coastal Plain**			**Piedmont**		
	RMSE	**MAE**	**AIC**	**RMSE**	**MAE**	**AIC**
Direct prediction (without biophy. var.)	70.41	50.35	5205.75	72.84	50.65	3364.83
Direct prediction (with biophy. var.)	70.10	49.66	5203.67	70.63	49.55	3348.80
Threshold probability based	75.41	54.48	5288.20	79.02	58.26	3437.65
Decision theory based	70.12	49.96	5202.02	72.00	49.80	3358.04

The results for two physiographic regions were similar to the whole region results in Table 2.9. Direct prediction approach that incorporated DI variable in the model performed best in the Coastal Plain and had the smallest values for RMSE and MAE and its AIC was close to the decision theory based approach. In the Piedmont, the same approach that incorporated HI variable performed the best in terms of RMSE, MAE and AIC. The RMSE was reduced by 2 units and AIC dropped by about 16 units by incorporating the biophysical variables in the model for the Piedmont region.

2.5 Conclusion

Models for stand-level mortality were developed for loblolly pine plantations in the southeastern region and separately for the Coastal Plain and Piedmont regions. Number of trees per ha, stand age and site index were the stand characteristics used in the model for prediction of tree number reduction. Three factors namely heat index, drought index and soil texture index, extracted from explanatory factor analysis, were used as surrogates for the climate and soil information in the analysis. Based on the statistics RMSE, MAE and AIC from model fit and leave-one-cluster-out cross-validation results, the mortality function derived from a differential equation where the relative rate of instantaneous mortality was related

to a power function of stand density, stand age and site index performed the best for the whole region and both the Coastal Plain and Piedmont. The power of site index (parameter b_2) in the final model for prediction of tree number reduction was expressed as a function of heat index and drought index for the whole region, as a function of drought index for the Coastal Plain and as a function of heat index for the Piedmont in order to improve the estimation of stand mortality. When we used biophysical variables in lieu of site index in the final model M14, there was no improvement in model fit and cross-validation results as compared to the model M14 with site index term. For the whole region the probability of stand mortality occurring over a 3-year period was related to number of trees per ha, stand age, site index, heat index and soil texture index. The log odds of stand mortality increased for higher age, site index and tree density. The effects of climate and soil variables in prediction of stand mortality in the Coastal Plain and Piedmont were not significant. At the physiographic level the effects of biophysical variables in predicting the probability of stand mortality were not pronounced in the two-step regression approach probably due to the fact that the separate analyses by physiographic regions may have already accounted to some degree for the effects of biophysical variables in predicting stand mortality. Among the three prediction approaches, the decision theory based approach performed the best when the biophysical variables were not included in the direct prediction models. But the direct prediction approach outperformed the decision theory based approach when the biophysical variables were incorporated in the direct prediction model. The threshold probability based approach performed the poorest among all the modeling strategies. Similar results were obtained for the Coastal Plain and Piedmont when the three prediction strategies were compared.

References

Álvarez González, J. G., Dorado, F. C., González, A. D. R., Sánchez, C. A. L., and von Gadow, K. (2004). A two-step mortality model for even-aged stands of *Pinus radiata* D. Don in Galicia (Northwestern Spain) . *Annals of Forest Science*, 61(5):439–448.

Avila, O. B. and Burkhart, H. E. (1992). Modeling survival of loblolly pine trees in thinned and unthinned plantations. *Canadian Journal of Forest Research*, 22(12):1878–1882.

Bailey, R. L., B. E. Borders, K. D. W., and Jones, Jr, E. P. (1985). A compatible model relating slash pine plantation survival to density, age, site index, and type and intensity of thinning. *Forest Science*, 31(1):180–189.

Bravo-Oviedo, A., Gallardo-Andrés, C., del Río, M., and Montero, G. (2010). Regional changes of *Pinus pinaster* site index in Spain using a climate-based dominant height model. *Canadian Journal of Forest Research*, 40(10):2036–2048.

Bravo-Oviedo, A., Tomeé, M., Bravo, F., Montero, G., and del Río, M. (2008). Dominant height growth equations including site attributes in the generalized algebraic difference approach. *Canadian Journal of Forest Research*, 38(9):2348–2358.

Buchman, R. G. (1979). Mortality functions. In *A generalized forest growth projection system applied to the Lake States region*, number NC-49, pages 47–55. USDA For. Ser.

Burkhart, H. E., Cloeren, D. C., and Amateis, R. L. (1985). Yield relationships in unthinned loblolly pine plantations on cutover, site-prepared lands. *Southern Journal of Applied Forestry*, 9(2):84–91.

Burkhart, H. E. and Tomé, M. (2012). *Modeling Forest Trees and Stands*. Springer.

Burkhart, H. E. and Walton, S. B. (1985). Incorporating crown ratio into taper equations for loblolly pine trees. *Forest Science*, 31(2):478–484.

Clutter, J. L., Fortson, J. C., Pienaar, L. V., Brister, G. H., and Bailey, R. L. (1983). *Timber management: a quantitative approach*. John Wiley & Sons Inc, New York, NY.

Clutter, J. L., Harms, W. R., Brister, G. H., and Rheney, J. W. (1984). Stand structure and yields of site-prepared loblolly pine plantations in the lower coastal plain of the Carolinas, Georgia, and north Florida. Gen. Tech. Rep. SE-27, USDA For. Ser.

Clutter, J. L. and Jones, E. P. (1980). Prediction of growth after thinning in old field slash pine plantations. Gen. Tech. Rep SE-27, USDA For. Ser.

Crookston, N. L., Rahfeldt, G. E., Dixon, G. E., and Weiskittel, A. R. (2010). Addressing climate change in the forest vegetation simulator to assess impacts on landscape forest dynamics. *Forest Ecology and Management*, 260(7):1198–1211.

Diéguez-Aranda, U., Castedo-Dorado, F., Álvarez-González, J. G., and Rodŕiguez-Soalleiro, R. (2005). Modelling mortality of Scots pine (*Pinus sylvestris* L.) plantations in the northwest of Spain. *European Journal of Forest Research*, 124(2):143–153.

Eid, T. and Øyen, B. H. (2003). Models for prediction of mortality in even-aged forest. *Scandinavian Journal of Forest Research*, 18(1):64–77.

Grothendieck, G. (2013). **nls2***: Non-linear regression with brute force*. R package version 0.2.

Hamilton, Jr., D. A. (1986). A logistic model of mortality in thinned and unthinned mixed conifer stands of Northern Idaho. *Forest Science*, 32(4):989–1000.

Harrell, Jr., F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer-Verlag, New York.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal data analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Hein, S. and Weiskittel, A. (2010). Cutpoint analysis for models with binary outcomes: a case study on branch mortality. *European Journal of Forest Research*, 129(4):585–590.

Jiang, L., Brooks, J. R., and Hobbs, G. R. (2007). Using crown ratio in yellow-poplar compatible taper and volume equations. *Northern Journal of Applied Forestry*, 24(4):271–275.

Jolliffee, I. T. (2002). *Principal component analysis*. Springer-Verlag, New York.

Kiernan, D., Bevilacqua, E., Nyland, R., and Zhang, L. (2009). Modeling tree mortality in low- to medium-density uneven-aged hardwood stands under a selection system using generalized estimating equations. *Forest Science*, 55(4):343–351.

Lemin, R. C. and Burkhart, H. E. (1983). Predicting mortality after thinning in old-field loblolly pine plantations. *Southern Journal of Applied Forestry*, 7(1):20–23.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Lipsitz, S. and Fitzmaurice, G. (2008). Generalized estimating equations for longitudinal data analysis. In Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G., editors, *Longitudinal Data Analysis*, pages 43–78. Chapman and Hall/CRC, Boca Raton, FL.

Monserud, R. A. (1976). Simulation of forest tree mortality. *Forest Science*, 22(4):438–444.

Monserud, R. A., Haung, S., and Yang, Y. (2006). Predicting lodgepole pine site index from climatic parameters in Alberta. *Forestry Chronicle*, 82(4):562–571.

Monserud, R. A. and Sterba, H. (1999). Modeling individual tree mortality for Austrian forest species. *Forest Ecology and Management*, 113(2-3):109–123.

Muhairwe, C. K., Lemay, V. M., and Kozak, A. (1994). Effects of adding tree, stand, and site variables to Kozak's variable-exponent taper equation. *Canadian Journal of Forest Research*, 24(2):252–259.

Nunes, L., Patrício, M., and Tomeé, J. (2011). Modeling dominant height growth of maritime pine in Portugal using GADA methodology with parameters depending on soil and climate variables. *Annals of Forest Science*, 68(2):311–323.

Pienaar, L. V., Page, H. H., and Rheney, J. W. (1990). Yield prediction for mechanically site-prepared slash pine plantations. *Southern Journal of Applied Forestry*, 14(3):104–109.

Pienaar, L. V. and Shiver, B. D. (1981). Survival functions for site-prepared slash pine plantations in the flatwoods of Georgia and northern Florida. *Southern Journal of Applied Forestry*, 5(2):59–62.

Reynolds, M. R. (1984). Estimating the error in model predictions. *Forest Science*, 30(2):454–469.

SAS Institute Inc. (2010). *SAS/ETS® 9.22 User's Guide*. Cary, NC.

SAS Institute Inc. (2011). *SAS/STAT® 9.3 User's Guide*. Cary, NC.

Soil Survey Staff (2012). Soil survey geographic (SSURGO) database for the South Eastern US. Natural Resources Conservation Service, United States Department of Agriculture. Available online at `http://soildatamart.nrcs.usda.gov` . Accessed 25/4/2012.

Thornton, P. E., Running, S. W., and White, M. A. (1997). Generating surfaces of daily meteorological variables over large regions of complex terrain. *Journal of Hydrology*, 190(3):214–251.

Thornton, P. E., Thornton, M. M., Mayer, B. W., Wilhelmi, N., Wei, Y., and Cook, R. B. (2012). Daymet–Daily surface weather on a 1 km grid for North America, 1980 - 2011. Acquired online (`http://daymet.ornl.gov/`) on 12/11/2012 from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA.

Vanclay, J. K. (1995). Synthesis: growth models for tropical forests: a synthesis of models and methods. *Forest Science*, 41(1):7–42.

Wang, Y., LeMay, V. M., and Baker, T. G. (2007). Modelling and prediction of dominant height and site index of *Eucalyptus globulus* plantations using a nonlinear mixed-effects model approach. *Canadian Journal of Forest Research*, 37(8):1390–1403.

Weiskittel, A. R., Crookston, N. L., and Radtke, P. J. (2011a). Linking climate, gross primary productivity, and site index across forests of the western united states. *Canadian Journal of Forest Research*, 41(8):1710–1721.

Weiskittel, A. R., Hann, D. W., Kershaw, J. A., and Vanclay, J. K. (2011b). *Forest growth and yield modeling.* Wiley-Blackwell, Chicester, UK, 2 edition.

Woollons, R. C. (1998). Even-aged stand mortality estimation through a two-step regression process. *Forest Ecology and Management*, 105(1):189–195.

Woollons, R. C., Snowdon, P., and Mitchell, N. D. (1997). Augmenting empirical stand projection equations with edaphic and climatic variables. *Forest Ecology and Management*, 98(3):267–275.

Yang, Y., Titus, S. J., and Huang, S. (2003). Modeling individual tree mortality for white spruce in Alberta. *Ecological Modelling*, 163(3):209 – 222.

Yao, X., Titus, S. J., and MacDonald, S. E. (2001). A generalized logistic model of individual tree mortality for aspen, white spruce, and lodgepole pine in Alberta mixedwood forests. *Canadian Journal of Forest Research*, 31(2):283–291.

Zhao, D., Borders, B., Wang, M., and Kane, M. (2007). Modeling mortality of second-rotation loblolly pine plantations in the Piedmont/Upper Coastal Plain and Lower Coastal Plain of the southern United States. *Forest Ecology and Management*, 252(1-3):132–143.

Chapter 3

Modeling Loblolly Pine (*Pinus taeda* L.) Mortality Using Multilevel Mixed-effects Logistic Regression

Abstract

Tree mortality being an important component of forest growth and yield system, its accurate prediction is vital to the forest growth and yield system. Repeated measure data collected from permanent sample plots established in 1980/81 across the natural range of loblolly pine in the Atlantic Coastal Plain, Gulf Coastal Plain and Piedmont were used. In permanent plot system, repeated measurements on each tree are nested within the tree which in turn are nested within a plot. Such hierarchical data structure is often ignored in modeling tree mortality in forestry applications. The main objective of this study was to develop individual-tree mortality using multilevel mixed-effects logistic regression taking into account the full hierarchical structure of the data. Both the tree- and stand-level variables used used in the analyses. Three-level logistic re-

gression fitted the data significantly better than the two-level logistic or fixed-effects logistic models. Multilevel mixed-effects models gave better predictions than the fixed effects model; however, the model fits and predictions were further improved by taking into account the full hierarchical structure of the data. Area under the ROC curve was higher for the three-level logistic model and its performance was the best with validation data that was not used in model fitting.

3.1 Introduction

Tree mortality and tree growth are two basic processes of stand dynamics and accurate prediction of mortality is an important component of forest growth and yield systems. However, accurate prediction of mortality is difficult due to causes of tree mortality being highly variable. Mortality in forest growth and yield systems is classified as regular (or non-catastrophic) and irregular (or catastrophic). Regular mortality results from competition for scarce resources such as water, nutrients and lights within a stand and irregular mortality results from random disturbances such as fire, wind, snow or insect outbreak (Vanclay, 1995).

Estimates of regular or non-catastrophic mortality may be based on concepts of limiting stand density (SDI) for pure even-aged stands (e.g. Reineke (1933) stand density index, Yoda et al. (1963) self-thinning in fully stocked even-aged stands) or may use empirical relationships fitted to stand- or tree-level data. The concepts of limiting stand density were often used as the basis for modeling mortality in growth models for pure stands until the 1960's and 70's. Such theoretical approaches based on limiting conditions could not explain all regular mortality and hence whole stand and individual tree mortality models based on empirical functions were introduced to supplement those theoretical approaches. Stand level mortality models predict changes of stand density over time and often estimate mortality in terms of basal area or number of trees per unit area. Individual tree mortality models predict the probability of survival or mortality for an individual tree.

Individual tree mortality models are considerably different from stand level mortality models in that they predict the probability of survival for each individual tree involved in growth models (Clutter et al., 1983). Since Hamilton (1974) introduced the logistic function as an individual tree mortality model in forestry, it has been widely used in modeling mortality of

many tree species. However, many other functions have been used to model individual tree mortality, e.g., the Weibull function (Somers et al., 1980), Richards function (Buford and Hafley, 1985) and gamma function (Kobe and Coates, 1997).

Stand-level mortality models based on the difference equation approach have limitations. Ingrowth must be assumed to be negligible or predicted with another function which is relevant to natural stands. These models predict some level of mortality even when no mortality occurs in a stand (Weiskittel et al., 2011). Stand-level mortality functions are usually fitted using least-squares methods (Woollons, 1998; Eid and Øyen, 2003) that assume Gaussian distribution with constant variance for underlying data. However, these assumptions are rarely appropriate for repeated measure data that exhibit varying degree of dispersion and skewness with respect to the mean (Affleck, 2007) and contain a large proportion of zero counts i.e. showing no mortality even over several years (Woollons, 1998). Building on the Poisson regression model, Affleck (2007) compared a set of nonlinear models that considered stochastic structures appropriate for count data and accommodated additional heterogeneity such as a large zero fraction.

Permanent sample plot systems in forestry observe the response variable for each subject (plot or tree) repeatedly at several times. Repeated observations on a subject in longitudinal studies are typically correlated. The observations on trees are nested in nature, i.e. units at one level are contained within units of another level. In permanent plot systems, measurement occasions are nested within a tree (i.e. repeated measurements) and trees are nested within a plot. Such full hierarchical data structure is often ignored in modeling tree mortality in forestry applications, probably due to difficulty in modeling (e.g. Avila and Burkhart, 1992; Eid and Tuhus, 2001; Yao et al., 2001; Yang et al., 2003; Crecente-Campo et al., 2009). Multilevel logistic regression has become increasingly popular for data with hierarchical structure (Hedeker, 2003). Multilevel model provides a systematic analysis of

effects of covariates measured at various levels of hierarchical structure on the outcome variable and it corrects for the biases in parameter estimates resulting from clustering. A large and growing literature exists for multilevel logistic regression (Gibbons and Hedeker, 1997). Two-level logistic models commonly used in modeling individual-tree mortality take into account the hierarchical structure that occurs between the plots and trees but they ignore the hierarchical structure that occurs between the trees and measurement occasions (e.g. Alenius et al., 2003; Adame et al., 2010; Groom et al., 2012; Timilsina and Staudhammer, 2012).

The main objective of this chapter was to use multilevel mixed effect logistic regression to model individual tree mortality. The multilevel logistic regression that accounted for the full hierarchical structure of the data from permanent plots was of particular interest of this chapter.

3.2 Data

The region wide thinning study data used in Chapter 2 was used this study. The details of the plot establishment and data is described in Section 2.2 in Chapter 2. There were 171 plots in total with 105 plots in the Coastal Plain and 66 plots in the Piedmont region. There were 9,972 loblolly pine trees at those 171 plots. Information about diameter at breast height (DBH), total tree height, stand age, stand basal area per ha and height of dominant stand were used in the analyses. The stand-level variables age, stand basal area per ha and dominant height are the three main drivers of forest stand dynamics i.e. stand age, stand density and site quality, respectively (Burkhart and Tomé, 2012). For tree-level variables, DBH and total tree height were used. Summary statistics for permanent plots in the Coastal Plain and Piedmont region are given in Table 3.1. Number of dead and survived trees (δ),

survived proportion, and mean (standard deviation), minimum and maximum of tree and stand-level variables are given.

Table 3.1: Summary statistics of tree and stand-level characteristics in the Coastal Plain and Piedmont

Region	**No. of trees**		**Survived proportion**	**Variable**	**Mean (SD)**	**Min**	**Max**
	Survived (δ=0)	**Dead (δ=1)**					
Coastal Plain	4049	1912	0.6792	DBH (cm)	17.69 (5.66)	1.27	43.18
				Total height (m)	15.38 (4.54)	1.52	33.53
				Stand-level			
				Age (year)	22.59 (6.98)	8.00	45.00
				Dominant height (m)	16.53 (4.46)	4.17	30.02
				Basal area (m^2 ha^{-1})	25.79 (8.60)	2.96	53.70
Piedmont	3009	1002	0.7502	DBH (cm)	17.15 (5.19)	1.52	39.88
				Total height (m)	14.96 (4.05)	1.52	28.96
				Stand-level			
				Age (year)	23.15 (6.77)	9.00	43.00
				Dominant height (m)	15.79 (3.69)	5.96	26.50
				Basal area (m^2 ha^{-1})	26.14 (8.58)	4.57	49.65

3.3 Methods

3.3.1 Traditional logistic regression

Individual tree mortality is a binary event i.e. dependent variable is dichotomous that follows binomial distribution. The traditional logistic regression model with probability of mortality is expressed as

$$\log\left[\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right] = \boldsymbol{x}_i^{'}\boldsymbol{\beta} \tag{3.1}$$

which leads to

$$\pi(\boldsymbol{x}_i) = \frac{1}{1+\exp(\boldsymbol{x}_i^{'}\boldsymbol{\beta})} \tag{3.2}$$

where $\pi(\boldsymbol{x}_i)$ is a probability of mortality of tree i that depends on vector of explanatory variables, $\boldsymbol{x}_i$ is a vector of explanatory variables and $\boldsymbol{\beta}$ is a vector of parameters.

The general form of the logistic regression model is

$$y_i = \mathrm{E}(y_i) + \varepsilon_i \tag{3.3}$$

where the observations y_i are independent Bernoulli random variables with expected values

$$\mathrm{E}(y_i) = \frac{\exp(\boldsymbol{x}_i^{'}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^{'}\boldsymbol{\beta})} \tag{3.4}$$

The distribution of the errors (ε_i) is binomial with zero mean and variance $\pi(\boldsymbol{x}_i)[1 - \pi(\boldsymbol{x}_i)]$. The maximum likelihood method is used to estimate the parameters in the linear predictor $\boldsymbol{x}_i^{'}\boldsymbol{\beta}$. Numerical methods like Newton-Raphson algorithm are used to compute the maximum-likelihood estimates of model parameters.

3.3.2 Multilevel logistic regression with random effects

Multilevel logistic regression provides a systematic analysis of effects of covariates measured at various levels of hierarchical structure on the response variable and it takes into account the clustering effects present in the data.

3.3.2.1 Two-level logistic regression model

Let i $(i = 1, \ldots, N)$ denote the level-2 units (clusters) and j $(j = 1, \ldots, n_i)$ denote the level-1 units (nested observations). The total number of level-1 observations across level-2 units are $\sum_{i=1}^{N} n_i$. Let Y_{ij} be the value of the dichotomous variable associated with level-1

unit j nested within level-2 unit i.

The logistic regression model in the equation (3.1) can be extended to a two-level random intercept model in its logit form as (Hedeker and Gibbons, 2006)

$$\log\left[\frac{\pi(\boldsymbol{x}_{ij})}{1-\pi(\boldsymbol{x}_{ij})}\right] = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \upsilon_i \tag{3.5}$$

where $\pi(\boldsymbol{x}_{ij})$ is a probability of a response, $\boldsymbol{x}_{ij}$ is a $(p+1)\times 1$ vector of explanatory variables including 1 for the intercept, $\boldsymbol{\beta}$ is a $(p+1)\times 1$ vector of unknown parameters representing fixed effects, and υ_i is a level-2 random effect assumed to have a $\mathcal{N}(0,\sigma_\upsilon^2)$. The random effects are typically expressed in standardized form as $\upsilon_i = \sigma_\upsilon\theta_i$ and the model is given as

$$\log\left[\frac{\pi(\boldsymbol{x}_{ij})}{1-\pi(\boldsymbol{x}_{ij})}\right] = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \sigma_\upsilon\theta_i \tag{3.6}$$

The random-effects variance term (the population standard deviation σ_υ) is explicitly included in the model making it on the same scale as other regression coefficients.

The equation (3.1) can be extended to include multiple random effects. For this, let $\boldsymbol{z}_{ij}$ denote the $r\times 1$ vector of random-effect variables (that includes a column of ones for the random intercept). Let $\boldsymbol{\upsilon}_i$ is a vector of r random subject effects and is assumed to have a $\mathcal{MN}(\mathbf{0},\boldsymbol{\Sigma}_\upsilon)$. The random effects are expressed in standardized form as $\boldsymbol{\upsilon}_i = \boldsymbol{T}\boldsymbol{\theta}_i$ where $\boldsymbol{T}\boldsymbol{T}' = \boldsymbol{\Sigma}_\upsilon$ is the Cholesky factorization of $\boldsymbol{\Sigma}_\upsilon$. The model is

$$\log\left[\frac{\pi(\boldsymbol{x}_{ij})}{1-\pi(\boldsymbol{x}_{ij})}\right] = \boldsymbol{x}_{ij}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ij}^{'}\boldsymbol{T}\boldsymbol{\theta}_i \tag{3.7}$$

The Cholesky factor $\boldsymbol{T}$ is usually estimated instead of the variance-covariance matrix $\boldsymbol{\Sigma}_\upsilon$ allowing for more stable estimation of near-zero variance terms (Hedeker and Gibbons, 2006).

3.3.2.2 Three-level logistic regression model

The simple two-level model equation (3.5) may be inadequate for data from the permanent plots since the trees which are measured at several time points are also nested within the plots. Some correlation between observations of two trees from the same plot can be expected. This can be modeled using a three-level model. Let i ($i = 1, \ldots, N$) denote the level-3 units (sample plots), j ($j = 1, \ldots, n_i$) denote the level-2 units (trees), and t ($t = 1, \ldots, n_{ij}$) denote the level-1 unit (measurement occasions). The mixed-effects logistic regression model can be written as

$$\log\left[\frac{\pi(\boldsymbol{x}_{ijt})}{1-\pi(\boldsymbol{x}_{ijt})}\right] = \boldsymbol{x}_{ijt}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{ijt}^{(2)'}\boldsymbol{T}^{(2)}\boldsymbol{\theta}_{ij}^{(2)} + \boldsymbol{z}_{ijt}^{(3)'}\boldsymbol{T}^{(3)}\boldsymbol{\theta}_{i}^{(3)} \tag{3.8}$$

where $\boldsymbol{x}_{ijt}$ is a $p\times 1$ vector of explanatory variables, $\boldsymbol{\beta}$ is a $p\times 1$ vector of unknown parameters representing fixed effects, $\boldsymbol{z}_{ijt}^{(2)}$ is the vector for the r random effects at level-2, $\boldsymbol{z}_{ijt}^{(3)}$ is the vector for the m random effects at level-3, $\boldsymbol{T}^{(2)}\boldsymbol{T}^{(2)'} = \boldsymbol{\Sigma}_{\boldsymbol{v}}^{(2)}$ and $\boldsymbol{T}^{(3)}\boldsymbol{T}^{(3)'} = \boldsymbol{\Sigma}_{\boldsymbol{v}}^{(3)}$ are the Cholesky decomposition of $r \times r$ matrix $\boldsymbol{\Sigma}_{\boldsymbol{v}}^{(2)}$ and $m \times m$ matrix $\boldsymbol{\Sigma}_{\boldsymbol{v}}^{(3)}$ respectively, and $\boldsymbol{\theta}_{ij}^{(2)}$ and $\boldsymbol{\theta}_{i}^{(3)}$ are the vectors of standardized level-2 and level-3 random effects respectively. The superscripts indicate the level at which the random effects vary across cluster units. The random effects at each level have a multivariate normal distribution and random effects at different levels are mutually independent and independent of level-1 residuals.

3.3.3 Parameter estimation for three-level logistic regression model

3.3.3.1 Maximum likelihood estimation

Let the outcome variable $Y_{ijt} = 1$ for a positive response (i.e. when tree dies) and $Y_{ijt} = 0$ for a negative response from tree j ($j = 1, \ldots, n_i$) in plot i ($i = 1, \ldots, N$) at measurement occasion t ($t = 1, \ldots, n_{ij}$).

The conditional probability that $Y_{ijt} = 1$, conditional on the random effects $\boldsymbol{\theta}^*$ is given by

$$Pr(Y_{ijt} = 1 \mid \boldsymbol{\theta}^*) = \Psi(\zeta_{ijt}) \tag{3.9}$$

where $\zeta_{ijt} = \boldsymbol{x}'_{ijt}\boldsymbol{\beta} + \boldsymbol{z}^{(2)'}_{ijt}\boldsymbol{T}^{(2)}\boldsymbol{\theta}^{(2)}_{ij} + \boldsymbol{z}^{(3)'}_{ijt}\boldsymbol{T}^{(3)}\boldsymbol{\theta}^{(3)}_{i}$, $\boldsymbol{\theta}^*$ represents all the random effects in the model $\left(\boldsymbol{\theta}^{(2)}_{ij} \text{ and } \boldsymbol{\theta}^{(3)}_{i}\right)$, and $\Psi(\cdot)$ represents the standard logistic cumulative distribution function (cdf).

The probability that $Y_{ijt} = 0$ is simply

$$Pr(Y_{ijt} = 0 \mid \boldsymbol{\theta}^*) = 1 - \Psi(\zeta_{ijt}) \tag{3.10}$$

Assuming independence of the responses conditional on the random effects, the conditional likelihood of the $n_i \times 1$ response vector $\boldsymbol{Y}_i$, given $\boldsymbol{\theta}^*$, is given by

$$l(\boldsymbol{Y}_i \mid \boldsymbol{\theta}^*) = \prod_{j=1}^{n_i} \prod_{t=1}^{n_{ij}} \Psi(\zeta_{ijt})^{Y_{ijt}} \left[1 - \Psi(\zeta_{ijt})\right]^{1-Y_{ijt}} \tag{3.11}$$

Then the marginal probability of $\boldsymbol{Y}_i$ in the population of subjects is expressed as the following integral of the conditional likelihood $l(\cdot)$, weighted by the prior density $g(\cdot)$

$$h(\boldsymbol{Y}_i) = \int_{\boldsymbol{\theta}^*} l(\boldsymbol{Y}_i \mid \boldsymbol{\theta}^*) g(\boldsymbol{\theta}^*) d(\boldsymbol{\theta}^*) \tag{3.12}$$

where $g(\boldsymbol{\theta}^*)$ represents the population distribution of the random effects (joint distribution of $\boldsymbol{\theta}_{ij}^{(2)}$ and $\boldsymbol{\theta}_i^{(3)}$, each with a multivariate standard normal density).

The marginal likelihood of the response patterns $\boldsymbol{Y}_i$ from all subjects (i.e. total sample N) is given by

$$L = \prod_{i=1}^{N} h(\boldsymbol{Y}_i) \tag{3.13}$$

or

$$\log L = \sum_{i=1}^{N} \log h(\boldsymbol{Y}_i) \tag{3.14}$$

Let $\boldsymbol{\eta}$ be an arbitrary parameter vector, then for $\boldsymbol{\beta}$, and the unique elements $v\left(\boldsymbol{T}_{(2)}\right)$ and $v\left(\boldsymbol{T}_{(3)}\right)$ of the Cholesky factors $\boldsymbol{T}_{(2)}$ and $\boldsymbol{T}_{(3)}$ respectively, we get

$$\frac{\partial \log L}{\partial \boldsymbol{\eta}} = \sum_{i=1}^{N} h^{-1}(\boldsymbol{Y}_i) \frac{\partial \log h(\boldsymbol{Y}_i)}{\partial \eta} \tag{3.15}$$

Now expressing the marginal likelihood in the following form

$$
\begin{aligned}
h(\boldsymbol{Y}_i) &= \int_{\boldsymbol{\theta}^*} l\left(\boldsymbol{Y}_i \mid \boldsymbol{\theta}^*\right) g(\boldsymbol{\theta}^*) d\boldsymbol{\theta}^* \\
&= \int_{\boldsymbol{\theta}^*} \left(\prod_{j=1}^{n_i} \prod_{t=1}^{n_{ij}} \Psi(\zeta_{ijt})^{Y_{ijt}} \left[1 - \Psi(\zeta_{ijt})\right]^{1-Y_{ijt}} \right) g(\boldsymbol{\theta}^*) d\boldsymbol{\theta}^* \\
&= \int_{\boldsymbol{\theta}^*} \left[\exp\left(\log\left\{ \prod_{j=1}^{n_i} \prod_{t=1}^{n_{ij}} \Psi(\zeta_{ijk})^{Y_{ijt}} \left[1 - \Psi(\zeta_{ijt})\right]^{1-Y_{ijt}} \right\} \right) \right] g(\boldsymbol{\theta}^*) d\boldsymbol{\theta}^* \\
&= \int_{\boldsymbol{\theta}^*} \left[\exp\left(\sum_{j=1}^{n_i} \sum_{t=1}^{n_{ij}} Y_{ij} \log[\Psi(\zeta_{ijt})] + (1 - Y_{ijt}) \log[1 - \Psi(\zeta_{ijt})] \right) \right] g(\boldsymbol{\theta}^*) d\boldsymbol{\theta}^*
\end{aligned}
$$

Now,

$$
\begin{aligned}
\frac{\partial \log h(\boldsymbol{Y}_i)}{\partial \eta} &= \int_{\boldsymbol{\theta}^*} \sum_{j=1}^{n_i} \sum_{t=1}^{n_{ij}} \left[\frac{Y_{ijt}}{\Psi(\zeta_{ijt})} \partial\Psi(\zeta_{ijt}) + \frac{1 - Y_{ijt}}{1 - \Psi(\zeta_{ijt})} \left(-\partial\Psi(\zeta_{ijt})\right) \right] \frac{\partial \zeta_{ijt}}{\partial \eta} l(\boldsymbol{Y}_i \mid \boldsymbol{\theta}^*) g(\boldsymbol{\theta}^*) d(\boldsymbol{\theta}^*) \\
&= \int_{\boldsymbol{\theta}^*} \sum_{j=1}^{n_i} \sum_{t=1}^{n_{ij}} \left[\frac{Y_{ijt} - \Psi(\zeta_{ijt})}{\Psi(\zeta_{ijt})(1 - \Psi(\zeta_{ijt}))} \right] \partial\Psi(\zeta_{ijt}) \frac{\partial \zeta_{ijt}}{\partial \eta} l(\boldsymbol{Y}_i \mid \boldsymbol{\theta}^*) g(\boldsymbol{\theta}^*) d(\boldsymbol{\theta}^*)
\end{aligned}
$$

Yielding,

$$
\frac{\partial \log h(\boldsymbol{Y}_i)}{\partial \eta} = \sum_{i=1}^{N} h^{-1}(\boldsymbol{Y}_i) \int_{\boldsymbol{\theta}^*} \sum_{j=1}^{n_i} \sum_{t=1}^{n_{ij}} \left[Y_{ijt} - \Psi(\zeta_{ijt}) \right] \frac{\partial \zeta_{ijt}}{\partial \eta} l(\boldsymbol{Y}_i \mid \boldsymbol{\theta}^*) g(\boldsymbol{\theta}^*) d(\boldsymbol{\theta}^*)
$$

Since, $\partial\Psi(\zeta_{ijt})$ equals the pdf, which for the standard logistic is $\Psi(\zeta_{ijt})(1 - \Psi(\zeta_{ijt}))$

$$\frac{\partial \zeta_{ijt}}{\partial \boldsymbol{\beta}^{'}} = \boldsymbol{x}_{ij}^{'} \tag{3.16}$$

$$\frac{\partial \zeta_{ijt}}{\partial \left(v\left(\boldsymbol{T}^{(2)} \right) \right)^{'}} = \left(\boldsymbol{\theta}_{ij}^{(2)'} \otimes \boldsymbol{z}_{ijt}^{(2)'} \right) \boldsymbol{J}_{r}^{'} \tag{3.17}$$

$$\frac{\partial \zeta_{ijt}}{\partial \left(v\left(\boldsymbol{T}^{(3)} \right) \right)^{'}} = \left(\boldsymbol{\theta}_{i}^{(3)'} \otimes \boldsymbol{z}_{ijt}^{(3)'} \right) \boldsymbol{J}_{m}^{'} \tag{3.18}$$

where $\otimes$ is the Kronecker product and $\boldsymbol{J}_r$ is the transformation matrix that eliminates the elements above the main diagonal. Details of the marginal maximum likelihood estimation for three-level logistic regression can be found in Gibbons and Hedeker (1997) and Liu and Hedeker (2006). The integral in Equation (3.12) does not have a closed-form solution and hence it is approximated using a Laplace approximation.

3.3.4 Model evaluation and validation

Statistical significance of each model parameter was considered for model selection and their relationships with predicted mortality were also plotted. Akaike's Information Criteria (AIC) and log likelihood of the competing models were compared. Receiver operating characteristics (ROC) graphs generated to select the best classifier using both model fitting data and independent cross-validation data. 20% of the data was set aside for cross-validation. ROC plot shows the sensitivity (the proportion of correctly classified positive observations) and specificity (the proportion of correctly classified negative observations) as the output threshold is varied over the range of all possible values. The area under the curve (AUC) measures the performance of a classifier and is often used when a general measure of predictiveness is desired. A model with AUC greater than 0.8 has some utility in predicting the responses of

individual subjects (Harrell, 2001).

A generalized coefficient of determination R_N^2 index, suggested by Nagelkerke (1991), was used to compare models. This unitless index is based on log likelihood and its value ranges from 0 to 1. Let L is -2 log likelihood for the fitted model and L^0 is -2 log likelihood for a null model with no predictive information. R_N^2 is given by

$$R_N^2 = \frac{1 - \exp\left(-\frac{LR}{n}\right)}{1 - \exp\left(-\frac{L^0}{n}\right)} \tag{3.19}$$

where $LR = L^0 - L$, the log likelihood ratio statistics for testing the global null hypothesis that $\hat{\boldsymbol{\beta}} = 0$ and n is the total number of observations.

Rank correlation between predicted probability of response and actual response can be used as a measure of the fitted model's predictive discrimination (Harrell, 2001). Somers' D_{xy} rank correlation between predicted probabilities and observed response is given by

$$D_{xy} = 2(c - 0.5) \tag{3.20}$$

where c is the AUC. Model with $D_{xy} = 0$ makes random predictions and the model with $D_{xy} = 1$ makes perfect predictions.

Indexes derived from likelihood ratio tests were also used as a measure of discriminating power of the predictor such as indexes of discrimination D [(Logistic model likelihood ratio χ^2 -1)/n], unreliability U (lack of calibration) and overall quality ($Q = D - U$). Higher value of D and Q means better prediction.

Prediction bias from models across range of different variables were examined. The mean bias was calculated as

$$Bias = \frac{\sum(y_{ij} - \hat{\pi}_{ij})}{n} \tag{3.21}$$

where y_i is observed response (1 or 0), $\hat{\pi}_{ij}$ is the fitted value and n is the number of observations.

3.4 Results and discussion

A level plot representation of the data matrix of the data used in this study is given in Figure 3.1. It shows the hierarchical structure of the data where trees are nested within plots and repeated measurements on each tree are nested within trees.

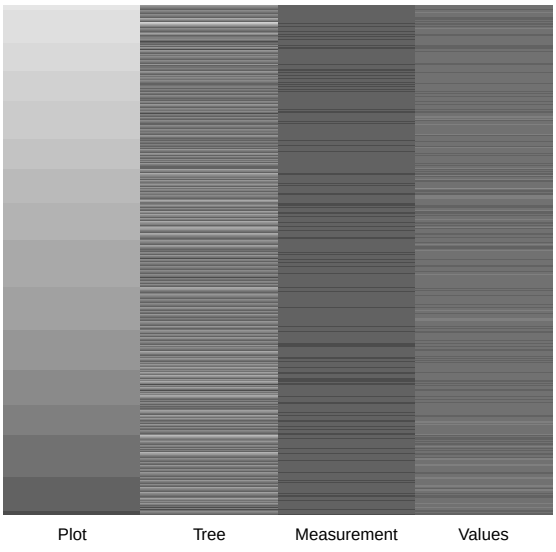


Figure 3.1: Level plot representation of data matrix of region wide thinning study data

Logistic regression models with fixed effects only (Equation 3.22) and with random effects added (Equations 3.23 and 3.24) were fitted including both the tree- and stand-level variables namely diameter at breast height (DBH), total tree height (TH), stand age (A) and stand

basal area per ha (BAP) and indicator variable representing two physiographic regions (PR). Although the period of study (1980/81 - 2002/03) is sufficient time period to represent the natural mortality of loblolly pine, the mortality rates may have been affected by other irregular environmental and climatic conditions during that time period.

The first model fitted was the fixed-effects logistic (Model 1) given by Equation (3.22).

$$\begin{aligned} \log\left[\frac{\pi(x_{ij})}{1-\pi(x_{ij})}\right] &= \beta_0 + \beta_1 PR + \beta_2 DBH_{ij} + \beta_3 TH_{ij} + \beta_4 A_i \\ &\quad + \beta_5 BAP_i + \beta_6 DBH_{ij} \times TH_{ij} + \beta_7 TH_{ij} \times A_i \end{aligned} \tag{3.22}$$

where $\pi(x_{ij})$ is a probability of 3-year mortality of jth tree in ith plot, PR is physiographic regions (Coastal Plains and Piedmont), DBH_{ij} is diameter at breast height of jth tree in ith plot, TH_{ij} is total height of jth tree in ith plot, A_i is age of plot i and BAP_i is basal area per ha of plot i.

Random effect parameters were added in the multilevel logistic regression models, one at a time, to the intercept and plot-level variable A_i. Two-level logistic regression (Model 2) was fitted to account for correlation in observations from trees (level-1 units) in a plot (level-2 units) given by Equation (3.23).

$$\log\left[\frac{\pi(x_{ij})}{1-\pi(x_{ij})}\right] = b_{0i} + b_{1i}PR + b_{2i}DBH_{ij} + b_{3i}TH_{ij} + b_{4i}A_i + b_{5i}BAP_i + b_{6i}DBH_{ij} \times TH_{ij} + b_{7i}TH_{ij} \times A_i \tag{3.23}$$

where

$$b_{0i} = \beta_0 + \upsilon_{0i}$$

$$b_{4i} = \beta_4 + \upsilon_{4i}$$

$$b_{1i} = \beta_1$$

$$b_{2i} = \beta_2$$

$$b_{3i} = \beta_3$$

$$b_{5i} = \beta_5$$

$$b_{6i} = \beta_6$$

$$b_{7i} = \beta_7$$

υ_{0i} and υ_{4i} are the random effects of intercept and stand age (A), respectively representing the effect of the ith plot and $\boldsymbol{\upsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\upsilon)$. Unstructured variance-covariance structure was used for random-effects.

Three-level logistic regression (Model 3) was fitted to account for correlation in repeated measurements (level-1 units) from a tree (level-2 units), which are nested in a plot (level-3 units) and is given in Equation (3.24).

$$\log\left[\frac{\pi(x_{ijt})}{1-\pi(x_{ijt})}\right] = b_{0ij} + b_{1ij}PR + b_{2ij}DBH_{ijt} + b_{3ij}TH_{ijt} + b_{4ij}A_{it} + b_{5ij}BAP_{it} + b_{6ij}DBH_{ijt} \times TH_{ijt} + b_{7ij}TH_{ijt} \times A_{it} \tag{3.24}$$

where

$$\begin{aligned} b_{0ij} &= \beta_0 + \upsilon_{0ij}^{(2)} + \upsilon_{0i}^{(3)} \\ b_{4ij} &= \beta_4 + \upsilon_{4ij}^{(2)} + \upsilon_{4i}^{(3)} \\ b_{1ij} &= \beta_1 \\ b_{2ij} &= \beta_2 \\ b_{3ij} &= \beta_3 \\ b_{5ij} &= \beta_5 \\ b_{6ij} &= \beta_6 \\ b_{7ij} &= \beta_7 \end{aligned}$$

$\upsilon_{0ij}^{(2)}$ and $\upsilon_{4ij}^{(2)}$ are the random effects of intercept and stand age (A), respectively representing the effect of the jth tree in ith plot and $\boldsymbol{\upsilon}_{ij}^{(2)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\upsilon}^{(2)})$. $\upsilon_{0i}^{(3)}$ and $\upsilon_{4i}^{(3)}$ are the random effects of intercept and stand age (A), respectively representing the effect of the ith plot and $\boldsymbol{\upsilon}_{i}^{(3)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\upsilon}^{(3)})$.

Parameter estimates (standard errors) of fixed effects and random-effect parameters in two- and three-level logistic regressions are given in Table 3.2. The Laplace approximation to the likelihood was implemented in parameter estimation.

Table 3.2: Parameter estimates (SE) of fixed effects and different random-effects parameters

Parameters	**Model 1** **Fixed effects**	**Model 2** **2-level random effects**	**Model 3** **3-level random effects**
β_0: (Intercept)	-2.4806 (0.2338)*	-4.2337 (0.7093)*	-4.1837 (0.5325)*
β_1: PR	-0.7577 (0.0450)*	-1.0826 (0.2820)*	-1.2589 (0.3025)*
β_2: DBH	-0.4309 (0.0197)*	-0.3194 (0.0222)*	-0.2656 (0.0294)*
β_3: TH	-0.0919 (0.0225)*	-0.6631 (0.0242)*	-0.7880 (0.0439)*
β_4: A	0.1817 (0.0142)*	0.0555 (0.0404)	0.0602 (0.0366)
β_5: BAP	0.1099 (0.0042)*	0.3966 (0.0227)*	0.4170 (0.0162)*
β_6: DBH × TH	0.0094 (0.0011)*	0.0085 (0.0013)*	0.0040 (0.0018)*
β_7: TH × A	-0.0056 (0.0010)*	-4.5×10^{-5} (0.0011)	0.0032 (0.0020)
Plot-level			
$\sigma^2_{\upsilon_{0i}}$: Int. variance		10.7626 (1.0204)	12.7128 (1.1832)
$\sigma^2_{\upsilon_{0i}\upsilon_{4i}}$: Int.-slope covar.		-0.5235 (0.0618)	-0.6227 (0.0704)
$\sigma^2_{\upsilon_{4i}}$: Slope variance		0.0341 (0.0044)	0.0405 (0.0051)
Tree-level			
$\sigma^2_{\upsilon_{0ji}}$: Int. variance			0.3162 (0.1487)
$\sigma^2_{\upsilon_{0ij}\upsilon_{4ij}}$: Int.-slope covar.			-0.0328 (0.0101)
$\sigma^2_{\upsilon_{4ij}}$: Slope variance			0.0034 (0.0006)
AIC	17185.02	14793.99	14678.49
Log Likelihood	-8584.51	-7385.99	-7325.25

*p-value < 0.05

All the fixed effects parameters were significant in Model 1 and the parameters for A and $TH \times A$ were not significant in Model 2 and 3. *DBH*, *total height* and *basal area* strongly influenced the predicted 3-year probability of mortality. Size of the fixed-effect parameter estimates in Model 2 and 3 were larger than in Model 1 and hence, ignoring the random effects underestimated most of the fixed-effect parameter sizes (Table 3.2). In Model 2 and Model 3 the random-effect parameters for intercept and slope were added one at a time. In Model 2, the likelihood ratio test for $H_0 : \sigma^2_{\upsilon_{4i}} = \sigma^2_{\upsilon_{0i}\upsilon_{4i}} = 0$ indicated that the model with intercept and slope random effects fitted the data better ($\chi^2_2 = 491.85$, p <0.0001). The likelihood ratio test for Model 3 ($H_0 : \sigma^2_{\upsilon_{4i}} = \sigma^2_{\upsilon_{0i}\upsilon_{4i}} = \sigma^2_{\upsilon_{4ij}} = \sigma^2_{\upsilon_{0ij}\upsilon_{4ij}} = 0$) gave the similar results suggesting the better fit obtained with intercept and slope random effects ($\chi^2_4 = 587.17$, p <0.0001). Likelihood ratio tests showed that the Model 3 fitted the data significantly better than both Model 1 ($\chi^2_6 = 2518.52$, p <0.0001) and Model 2 ($\chi^2_3 = 121.49$, p <0.0001). Model 3 had the smallest AIC among three models and this model had the highest log

likelihood value. When Model 3 was refitted by deleting the interaction term $TH \times A$, stand age (A) was highly significant (p <0.0001). In the study of individual-tree mortality models, Ma et al. (2013) reported that the mixed-effects logistic model incorporating both plot and time random effects performed the best as compared to standard logistic and marginal logistic model based on the GEE approach. Their model performance was based on the evaluation statistics such as mean prediction error, average absolute prediction error, the variance prediction error and the mean square error. A logistic model based on the GEE has been found to better capture the changes in probability of mortality over time than the standard logistic regression in a study of modeling tree mortality using longitudinal data from permanent plots (e.g Kiernan et al., 2009).

The negative estimated coefficient of PR indicate expected decrease in the log odds of tree mortality going from the Coastal Plane to the Piedmont. The increase in DBH and TH results in the expected decrease in the log odds of tree mortality. The positive and significant interaction effect between DBH and TH indicate that when DBH increases, the log odds of mortality increases with increase in TH. The significant positive coefficient of BAP is associated with increasing log odds of mortality for higher values of BAP. These results are consistent with other studies of individual tree mortality (e.g Adame et al., 2010; Groom et al., 2012; Timilsina and Staudhammer, 2012).

There was significant variation in both the individual intercept and slope of A_i for both Model 2 and Model 3. The estimates of random-effect parameters in Model 3 indicated that the conditional distribution of the random effects for *tree* had much less variability as compared to the conditional distribution of the random effects for the *plots* (the plot-to-plot variability had the greater contribution than the tree-to-tree variability). Standard errors of the random-effect parameters in Table 3.2 were obtained from 200 bootstrap samples. Bootstrap estimates of both the fixed- and random-effect parameters with 95% confidence

interval (percentile CIs) are given in Appendix C. Bootstrap estimates were similar to the ones obtained from the model fit with original data. Histograms of bootstrap estimates of random-effect for intercept and slope (Age) of Model 2 are given in Figure 3.2.

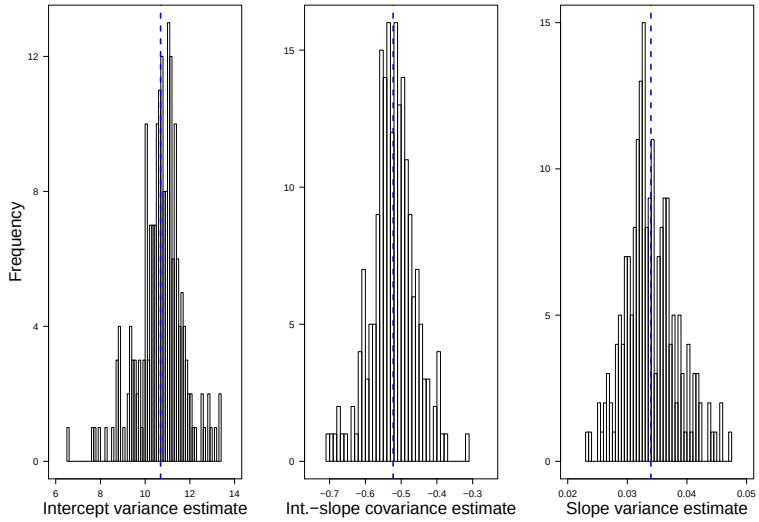
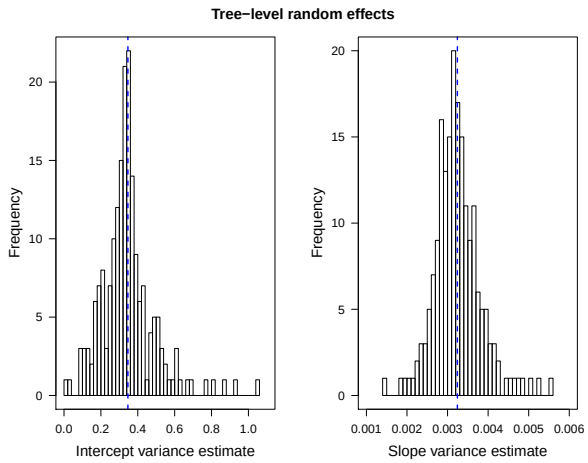
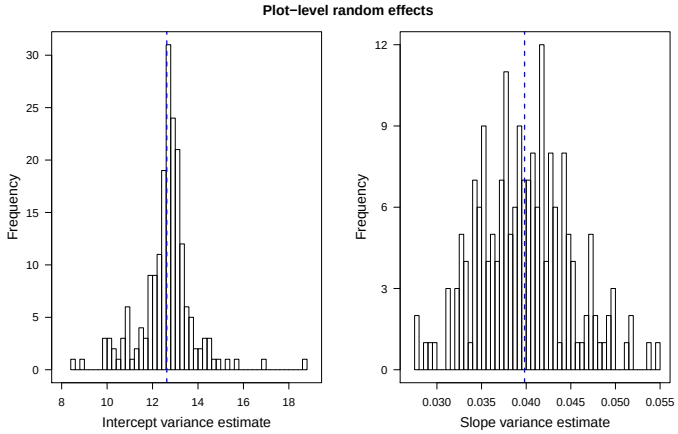


Figure 3.2: Histogram of random effects for Model 2. Broken vertical lines give bootstrap mean

Similarly, the histograms of bootstrap estimates of both tree- and plot-level random-effect for intercept and slope of Model 3 are given in Figure 3.3.



(a)



(b)

Figure 3.3: Histogram of random effects at (a) tree-level and (b) plot-level for Model 3. Broken vertical lines give bootstrap mean

Receiver operating characteristics (ROC) plot of all three models was generated in Figure 3.4 to graphically visualize and select the best classifier among the three logistic regression models in Table 3.2. The ROC method assesses model performance in threshold-independent manner and compare various models.

Area Under Curve (AUC) was highest for Model 3 (AUC = 0.946) and it was smallest for Model 1 (AUC = 0.842). It suggested that the Model 3 (3-level random effects logistic)

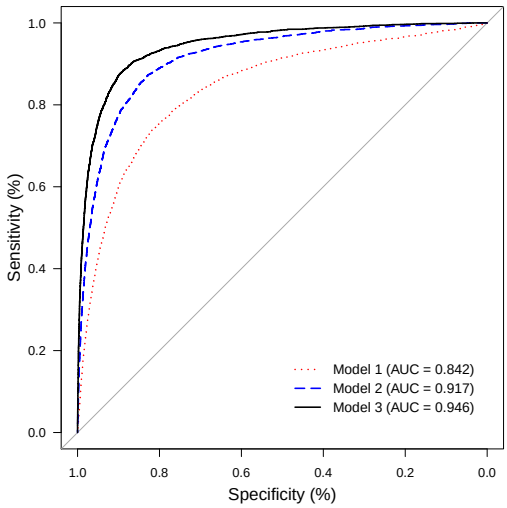


Figure 3.4: ROC curve and AUC of three logistic regression models

had the highest discrimination accuracy and Model 1 (fixed-effects logistic) had the least discrimination accuracy (Harrell, 2001).

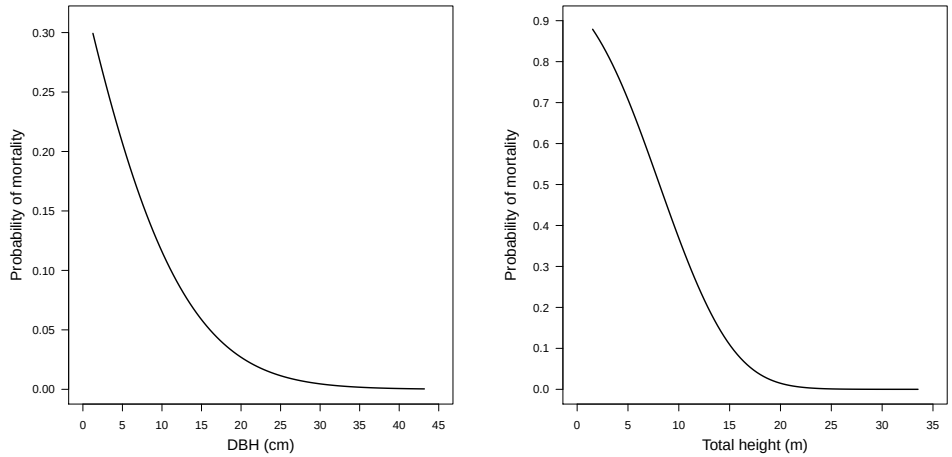
Separate models were fitted to Coastal Plain and Piedmont regions. The parameter estimates (standard errors) of fixed effects and random-effect parameters for three models in the Coastal Plains and Piedmont are given in Table 3.3. The estimates of tree-level random-effect parameters in the Piedmont were close to 0 but it does not mean the absence of variation between the trees in a plot. There was small amount of variability between the trees and the level of tree-to-tree variability may not be sufficient enough to warrant the incorporation of tree-level random effects in Model 3. The estimates of fixed-effects in Model 3 and Model 2 were similar. However, the likelihood ratio test of Model 3 for the Piedmont region ($H_0 : \sigma^2_{\upsilon_{0ij}} = \sigma^2_{\upsilon_{4ij}} = \sigma^2_{\upsilon_{0ij}\upsilon_{4ij}} = 0$) indicated that the Model 3 with both the tree- and plot-level random effects fitted the data better ($\chi^2_3 = 22.53$, $p = <0.0001$) than Model 2 with only plot-level random effects. ROC plots of three models for Coastal Plain and Piedmont are given in Figure D.1 in Appendix D. Model 3 had the highest discrimination accuracy in both the regions.

Table 3.3: Parameter estimates (SE) of fixed effects and different random-effects parameters for Coastal Plain and Piedmont

Parameters	**Coastal Plain**			**Piedmont**		
	Model 1	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
(Intercept)	−2.4870 (0.2831)*	−2.1985 (0.6005)*	−3.7944 (0.6642)*	−3.0817 (0.4112)*	−5.4531 (0.808)*	−6.7145 (0.8639)*
DBH	−0.5136 (0.0239)*	−0.2848 (0.0327)*	−0.3269 (0.0373)*	−0.1808 (0.0369)*	−0.1556 (0.0457)*	−0.1814 (0.0506)*
TH	−0.0283 (0.0265)	−0.9537 (0.0507)*	−0.8542 (0.0551)*	−0.2837 (0.0422)*	−0.7012 (0.0688)*	−0.6323 (0.0739)*
A	0.1963 (0.0170)*	−0.1307 (0.0416)*	0.0698 (0.0467)	0.1264 (0.0264)*	−0.0806 (0.0579)	0.0904 (0.0614)
BAP	0.1077 (0.0049)*	0.4640 (0.0195)*	0.4274 (0.0203)*	0.1171 (0.0082)*	0.4370 (0.0274)*	0.3983 (0.0276)*
DBH × TH	0.0123 (0.0013)*	0.0055 (0.0018)*	0.0064 (0.0021)*	−0.0020 (0.0024)	0.0003 (0.0029)	0.0011 (0.0033)
TH × A	−0.0074 (0.0011)*	0.0133 (0.0022)*	0.0046 (0.0024)	0.0015 (0.0018)	0.0049 (0.0032)	−0.0021 (0.0034)
Plot-level						
$\sigma^2_{v_{0i}}$: Int. variance		11.9011	14.1721		9.9337	10.7842
$\sigma^2_{v_{0i}v_{4i}}$: Int.-slope covar.		-0.5254	-0.6569		-0.5025	-0.5724
$\sigma^2_{v_{4i}}$: Slope variance		0.0354	0.0418		0.0330	0.0369
Tree-level						
$\sigma^2_{v_{0ij}}$: Int. variance			1.1327			0.0600
$\sigma^2_{v_{0ij}v_{4ij}}$: Int.-slope covar.			-0.0837			-0.0109
$\sigma^2_{v_{4ij}}$: Slope variance			0.0062			0.0020
AIC	10972.60	9299.93	9232.97	6128.36	5440.96	5424.43
Log Likelihood	-5479.30	-4639.96	-4603.49	-3057.18	-2710.48	-2699.22

*p-value < 0.05

Change in average marginal probability of mortality across the range of *DBH*, *total height*, *stand age* and *basal area* are given in Figure 3.5 and Figure 3.7. The probabilities of mortality in these plots were predicted by Model 3. For a particular variable of interest, a range of evenly spaced values within its range were selected and the average marginal predicted probability of mortality were obtained across the range. These probabilities were plotted against the variable of interest.



(a) (b)

Figure 3.5: Predicted probability of mortality predicted by Model 3 against (a) DBH (cm) and (b) total tree height (m)

As expected the probability of mortality declined with increasing tree DBH with rest of the variables held constant (Figure 3.5 (a)). The probability decreased with total height as well (3.5 (b)). The probability of mortality declined rapidly in juvenile tree trees (*DBH* upto about 25 cm and *total height* upto about 20 m). Mortality rate has U-shaped relationship with *DBH* i.e. mortality decreases with increasing *DBH* and it starts to increase again with further increase in tree size (Monserud and Sterba, 1999). But the selected Model 3 could not capture the expected increase in mortality rate. Lack of data on older stands has been given as possible reason for not capturing the expected U-shaped pattern in mortality in

some studies (e.g. Adame et al., 2010; Timilsina and Staudhammer, 2012). The largest *DBH* tree in the data for this study was about 43 cm and the oldest stand was 45 years.

The probability of mortality against *DBH* and *total height* were plotted for different values of stand age and stand basal area (Figure 3.6).

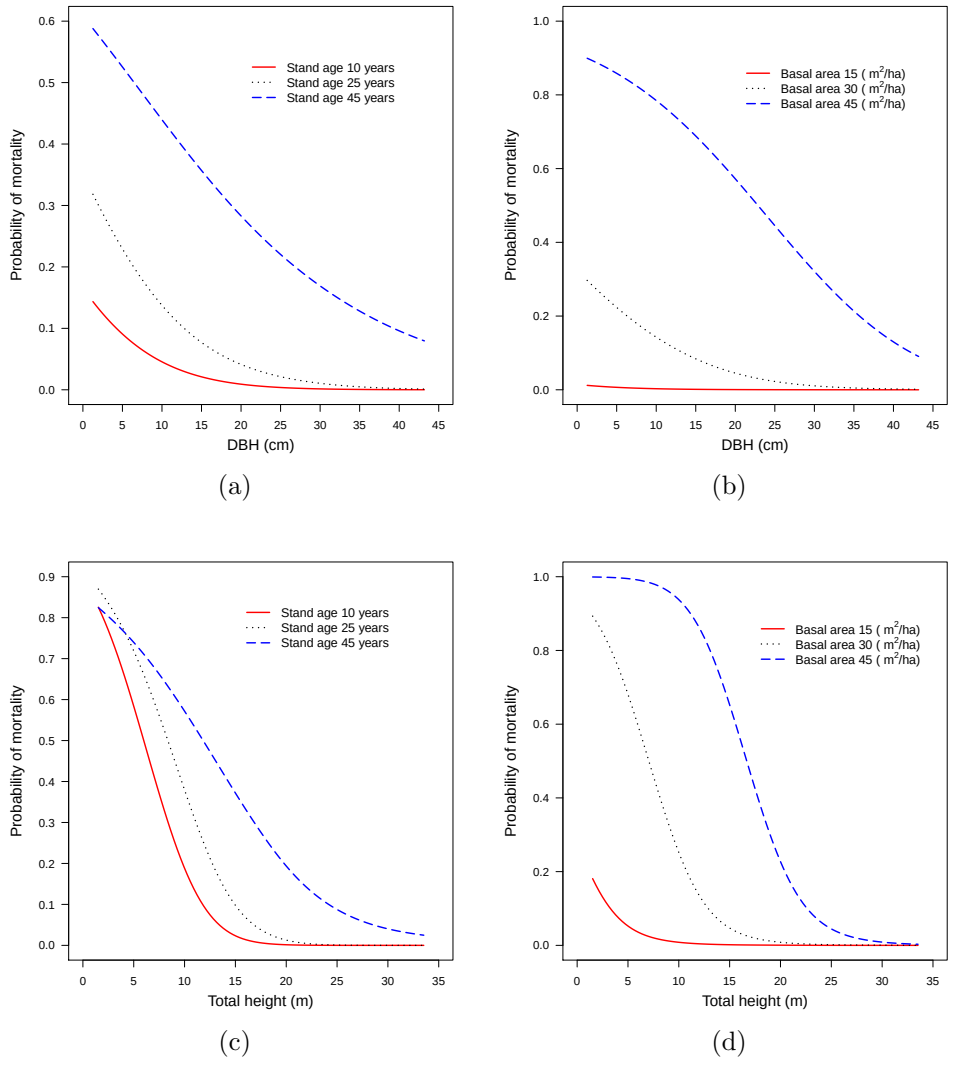
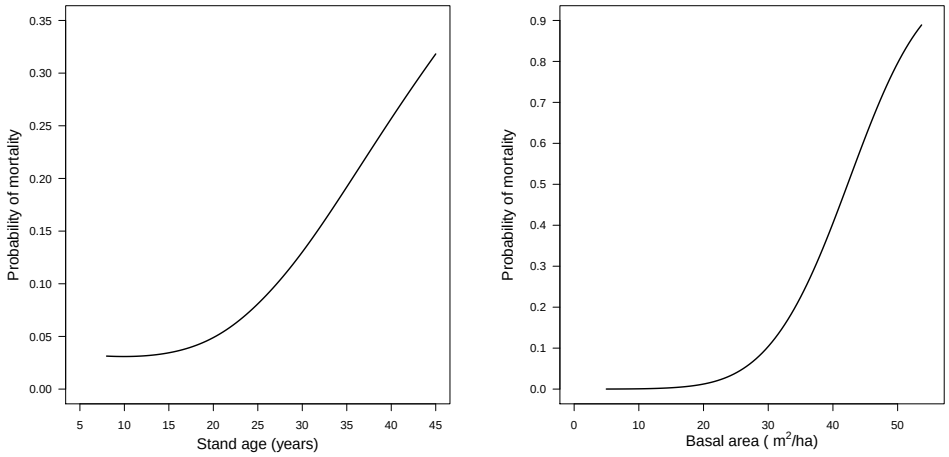


Figure 3.6: Predicted probability of mortality predicted by Model 3 against (a) DBH (cm) for different stand ages, (b) DBH (cm) for different stand basal area, (c) total tree height (m) for different stand ages and (d) total tree height (m) for different stand basal area

The probability of mortality was higher for smaller trees (low DBH) in older stand (Figure 3.6 (a)). It can be attributed to suppression by larger trees. With respect to total height, all smaller trees had higher probability of mortality regardless of stand ages. However, there was a difference in the probability of mortality with an increase in total height (Figure 3.6 (c)). The mortality probability seemed to be more sensitive to the effects of stand basal area. The trend of mortality probability being quite high for small trees was more pronounced in dense stands than in mature stands due to self-thinning. Small trees showed lower mortality probability at low basal area (Figure 3.6 (b) and Figure 3.6 (d)). Timilsina and Staudhammer (2012) reported similar results; however, Yao et al. (2001) observed that small trees showed higher mortality regardless of basal area or stand age for aspen, white spruce and lodgepole pine in Alberta boreal mixedwood forests.

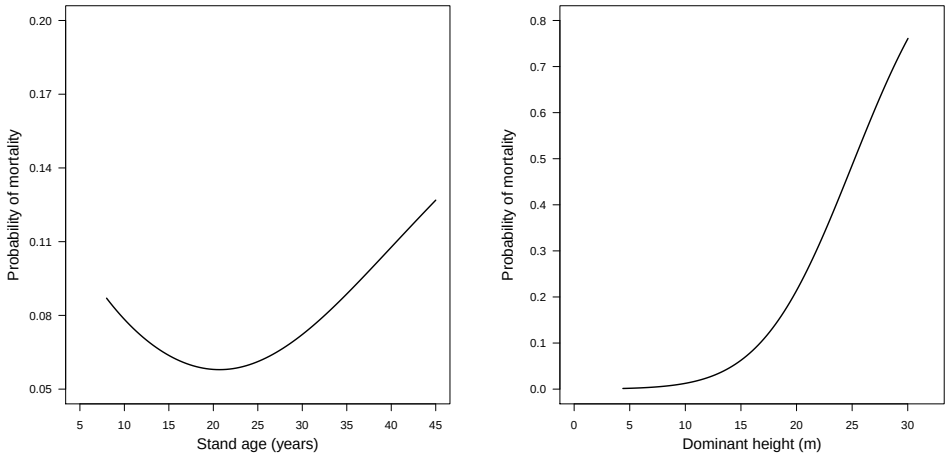


(a) (b)

Figure 3.7: Predicted probability of mortality predicted by Model 3 against (a) stand age (years) and (b) stand basal area (m^2 ha^{-1})

Curve showing the monotonic relationship between the probability of mortality and the stand age increased slowly and picked up rapidly with increasing stand age (Figure 3.7 (a)). However, when dominant height of stand was added as additional predictor in Model 3

(Equation (3.24)), the U-shaped mortality trend was observed with stand age (Figure 3.8 (a)) possibly indicating that dominant height was enough to include possible curvilinear relationship between mortality and stand age. A similar mortality trend with stand age was observed by Chen et al. (2008) and with DBH by Monserud and Sterba (1999). The parameter estimates of model with dominant height, refereed here after as Model 3a, is given in Table 3.4.



(a) (b)

Figure 3.8: Predicted probability of mortality predicted by Model 3a against (a) stand age (years) and (b) dominant height (m)

Table 3.4: Parameter estimates (SE) of model with dominant height (Model 3a)

Parameters	**Model 3a**
β_0: (Intercept)	-5.6210 (0.4457)*
β_1: PR	-0.7928 (0.2516)*
β_2: DBH	-0.2462 (0.0247)*
β_3: TH	-0.7599 (0.0285)*
β_4: A	-0.0547 (0.0277)*
β_5: BAP	0.3103 (0.0144)*
β_6: HD	0.4773 (0.0502)*
β_7: DBH $\times$ TH	0.0033 (0.0015)*
Plot-level	
$\sigma^2_{v_{0i}}$: Int. variance	9.6922
$\sigma^2_{v_{0i}v_{4i}}$: Int.-slope covar.	-0.4813
$\sigma^2_{v_{4i}}$: Slope variance	0.0301
Tree-level	
$\sigma^2_{v_{0ji}}$: Int. variance	0.8680
$\sigma^2_{v_{0ij}v_{4ij}}$: Int.-slope covar.	-0.0737
$\sigma^2_{v_{4ij}}$: Slope variance	0.0063
AIC	14609.41
Log Likelihood	-7290.70

*p-value < 0.05

Dominant height combined with stand age is an indicator of site quality. The probability of mortality increased with dominant height indicating that the higher mortality was related to better site quality. Zhao et al. (2007) found similar result for loblolly pine growing in Piedmont/Upper Coastal Plain. Results from studies of other species are consistent with this result (e.g. Eid and Tuhus, 2001; Yao et al., 2001). Empirical evidence suggests that density-dependent mortality in plantations starts earlier in better sites and increases with site productivity (Diéguez-Aranda et al., 2005).

The probability of mortality increased slowly with increasing stand basal area and the probability increased rapidly after stand basal area of about 23 m^2 ha^{-1} (Figure 3.7 (b)). Tree size and density information are incorporated in stand basal area hence, it is a good measure

of stand crowding. Trees in a stand with larger basal area are more likely to experience competition-induced mortality than trees in a stand with smaller basal area given a regular spacing pattern (Yang et al., 2003). Stand basal area is used as an indicator of stand-level competition and it has been reported to influence tree mortality (e.g. Yao et al., 2001). The results in this study are supported by other studies as well. Similar plots of average marginal predicted probability of mortality for the four variables were plotted for each physiographic region in Figure D.2 in Appendix D. In all the four variables, average marginal predicted probability of mortality was higher for the Coastal Plain than the Piedmont.

3.4.1 Validation of fitted models

Fang (2011) demonstrated that the marginal Akaike information criterion (AIC) is asymptotically equivalent to the leave-one-cluster-out cross-validation. Hence, AICs of fitted models were compared instead of performing leave-one-cluster-out cross-validation to assess model fits and select model from a family of hierarchical models in Table 3.2. AIC was defined as

$$AIC = -2 \log \text{likelihood} + 2K \tag{3.25}$$

where K is number of model parameters.

Model 3 had the smallest AIC (14678.49) or equivalently Model 3a with AIC of 14609.41. Model 2 had slightly higher AIC (14793.99) than Model 3 and Model 1 with largest AIC (17185.02).

The predictive ability of the models were compared using Nagelkerke's (1991) R_N^2 that was computed from the original data (Table 3.5).

Table 3.5: Index for quantifying predictive ability

Model	R_N^2
Model 1	0.2607
Model 2	0.3614
Model 3	0.3675
Model 3a	0.3710

The difference in index of predictive ability between Model 3 (or Model 3a) and Model 2 was small, possibly indicating slightly higher predictive strength for Model 3 (or Model 3a) over Model 2.

Receiver operating characteristics (ROC) plots of four models were generated using the validation data (Figure 3.9). There were total of 9736 observations in the validation data and 38567 observations in the model fit data. The area under the ROC curve of Model 3, Model 3a and Model 2 were similar and above 0.90 indicating their excellent and higher degree of discrimination than Model 1 with the least discrimination accuracy.

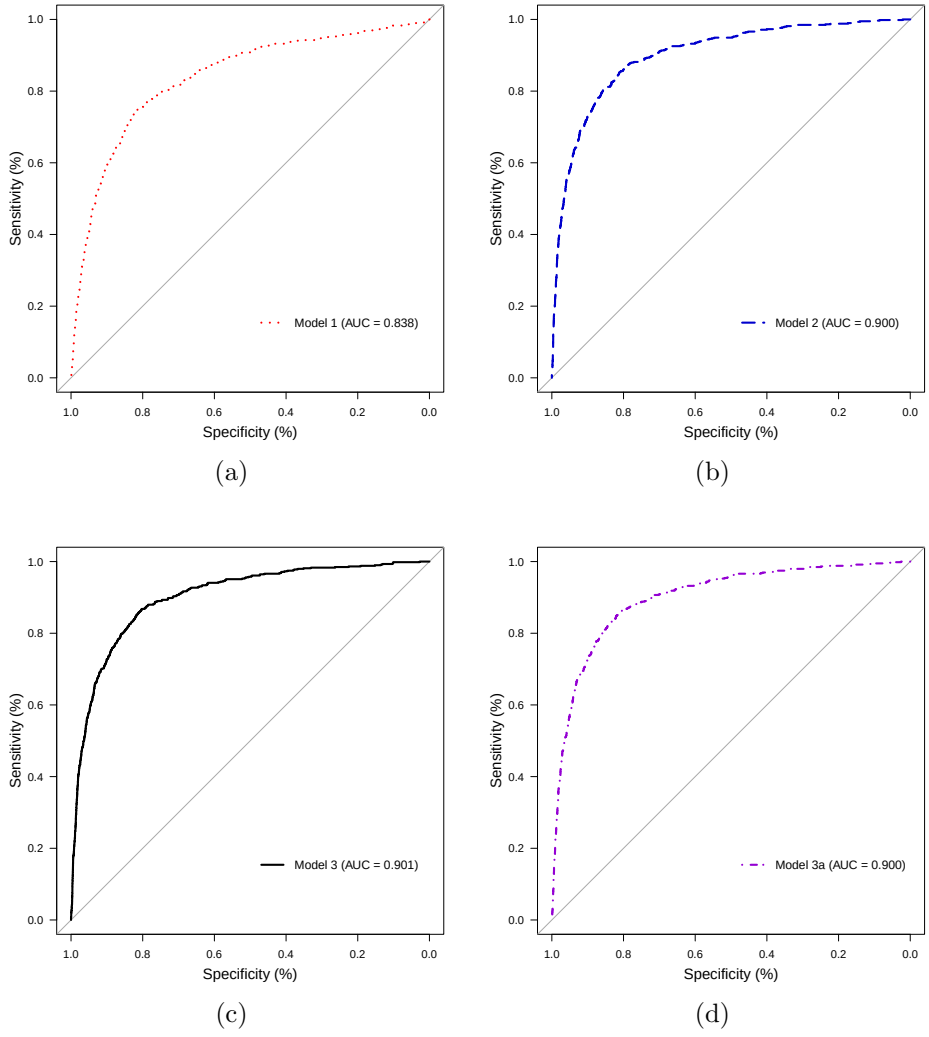


Figure 3.9: ROC curve and AUC of (a) Model 1, (b) Model 2, (c) Model 3 and (d) Model 3a

Calibration plots for the cross-validation data, which contain information on D_{xy}, c (or AUC), R^2_N, D, U and Q, are plotted in Figure 3.10. The Somers' D_{xy} was highest for Model 3 indicating it's superior predictive discrimination. The value of c (or area under the ROC) for Model 3 was 0.945, highest among all model. The generalized R^2_N of Model 3 was the highest, suggesting higher predictive strength of Model 3. Similarly, indexes D and Q were highest for Model 3. All the indexes and statistics in the plots were larger for Model 3

indicating it's better performance among them.

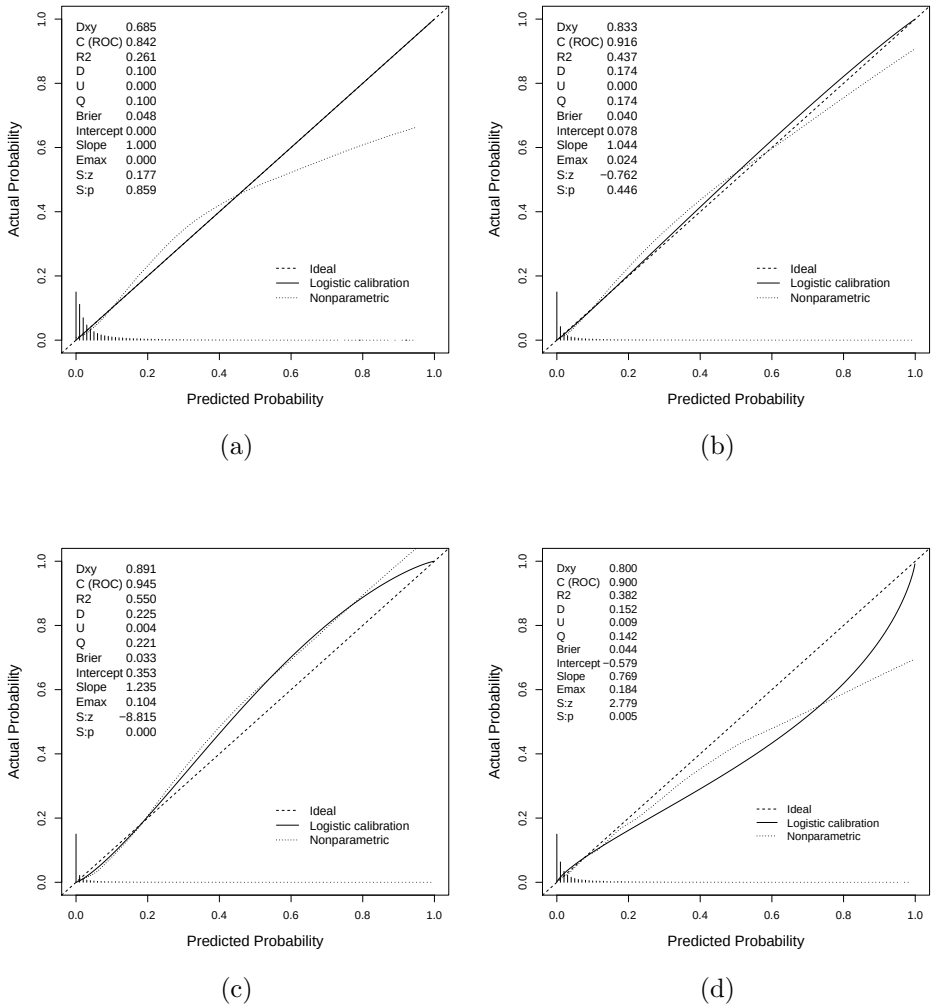


Figure 3.10: Validation of logistic models (a) Model 1, (b) Model 2, (c) Model 3 and (d) Model 3a

Prediction bias associated with three models in Table 3.2 and a model in Table 3.4 across the range of data values for *DBH*, *total height*, *stand age* and *basa*l area were obtained in Figure 3.11. Bias was in general lowest for Model 3 and Model 3a across all values of all predictor variables. Groom et al. (2012) observed that inclusion of a random intercept grouped by installation in mixed-effects individual-tree mortality models for Douglas-fir stands in the

Pacific Northwest significantly reduced model bias across all predictor variables relative to the fixed-effects model. Across a range of *DBH* classes, mean bias was higher at lower and higher classes for all models (Figure 3.11a). Across *total height* class ranges, mean bias was highest for Model 1 at lower classes (Figure 3.11b). For *stand age* classes, bias was higher for all models at higher classes (Figure 3.11c) and bias was generally higher for Model 1 at all classes of *stand basal area* (Figure 3.11d).

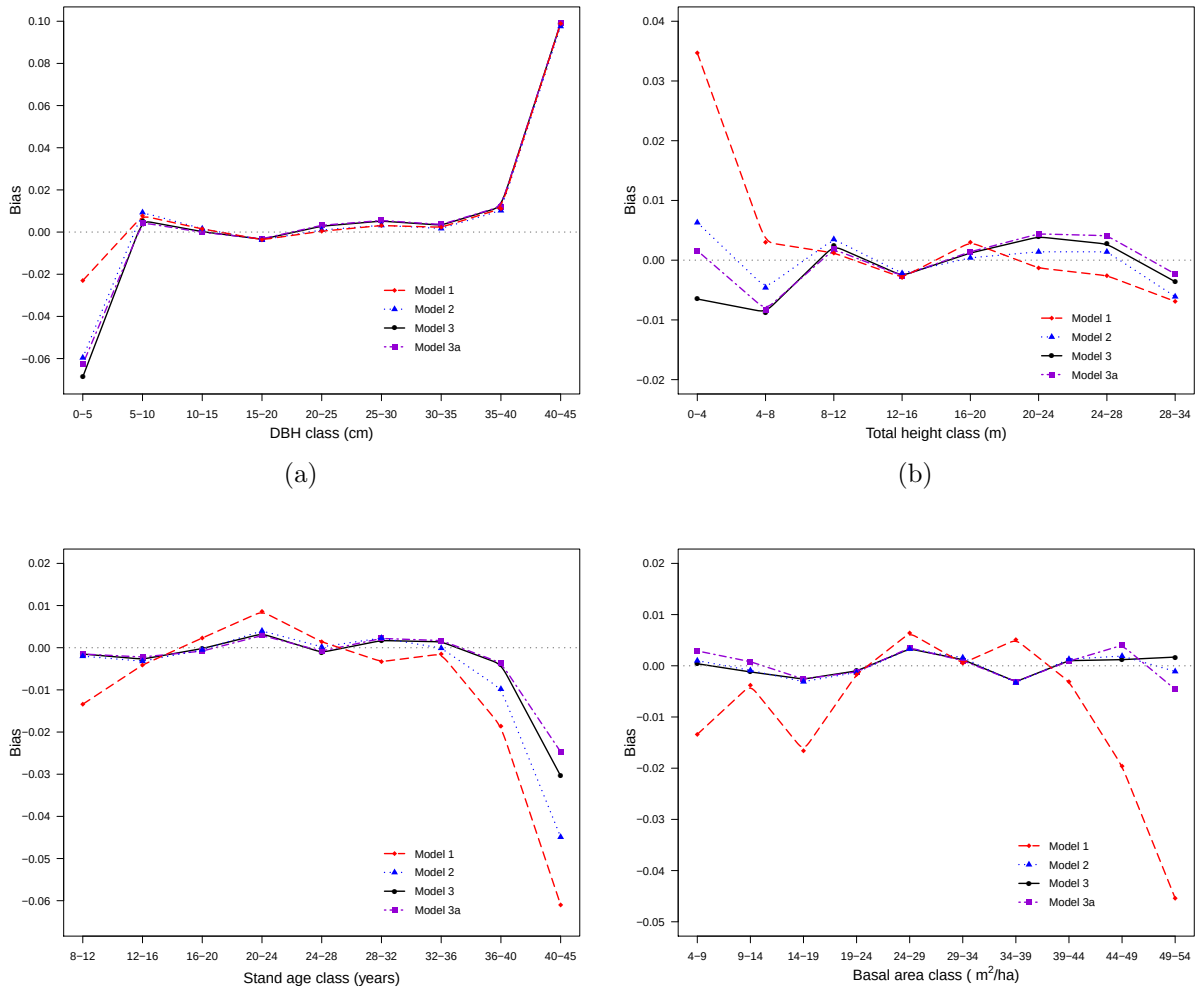


Figure 3.11: Prediction bias from four models across range of data for (a) DBH (cm), (b) total height (m), (c) stand age (years) and (d) stand basal area (m^2 ha^{-1})

3.5 Conclusion

Logistic models for prediction of mortality of individual trees was developed in this study. Multiple predictor variables related to tree size, stand density and competition were used to develop mortality function that is deterministic and empirical in nature. Logistic model with fixed effects only and with random effects added were fitted simultaneously. Among the first three models fitted, the three-level logistic regression (Model 3) had the smallest AIC and the largest log likelihood value among all the fitted models. The AUC for Model 3 was the highest with 0.946. Multilevel mixed effects models gave better predictions than the fixed effects model; however the model fits and predictions were further improved by taking into account the full hierarchical structure of the data. The probability of mortality declined with increasing tree *DBH* and *total height* and it increased with increasing *stand age* and *basal area*. Inclusion of *dominant height* in Model 3 rendered U-shaped mortality pattern with stand age and this variable was significant (Table 3.4). However, Model 3 was selected over Model 3a. First reason was *dominant height* was moderately to highly correlated with other variables in the model i.e. 0.62 with *DBH*, 0.91 with *TH*, 0.85 with *stand age* and 0.76 with *stand basal area*. The second reason was to maintain parsimony in the model.

References

Adame, P., Río, M., and Cañellas, I. (2010). Modeling individual-tree mortality in Pyrenean oak (*Quercus pyrenaica* Willd.) stands. *Annals of Forest Science*, 67(8):810–810.

Affleck, D. L. R. (2007). Mixed and modified poisson models for the analysis of stand-level mortality. *Canadian Journal of Forest Research*, 36(11):2994–3006.

Alenius, V., Hökkä, H., Salminen, H., and Jutras, S. (2003). Evaluating estimation methods for logistic regression in modelling individual-tree mortality. In Amaro, A., Reed, D., and Soares, P., editors, *Modelling Forest Systems*, pages 225–236. CABI Publishing, Cambridge, MA, USA.

Avila, O. B. and Burkhart, H. E. (1992). Modeling survival of loblolly pine trees in thinned and unthinned plantations. *Canadian Journal of Forest Research*, 22(12):1878–1882.

Buford, M. A. and Hafley, W. L. (1985). Probability distributions as models for mortality. *Forest Science*, 31(2):331–341.

Burkhart, H. E. and Tomé, M. (2012). *Modeling Forest Trees and Stands*. Springer.

Chen, H. Y., Fu, S., Monserud, R. A., and Gillies, I. C. (2008). Relative size and stand age determine *Pinus banksiana* mortality. *Forest Ecology and Management*, 255(12):3980 – 3984.

Clutter, J. L., Fortson, J. C., Pienaar, L. V., Brister, G. H., and Bailey, R. L. (1983). *Timber management: a quantitative approach*. John Wiley & Sons Inc, New York, NY.

Crecente-Campo, F., Marshall, P., and Rodríguez-Soalleiro, R. (2009). Modeling non-catastrophic individual-tree mortality for *Pinus radiata* plantations in northwestern Spain. *Forest Ecology and Management*, 257(6):1542–1550.

Diéguez-Aranda, U., Castedo-Dorado, F., Álvarez-González, J. G., and Rodŕiguez-Soalleiro, R. (2005). Modelling mortality of Scots pine (*Pinus sylvestris* L.) plantations in the northwest of Spain. *European Journal of Forest Research*, 124(2):143–153.

Eid, T. and Øyen, B. H. (2003). Models for prediction of mortality in even-aged forest. *Scandinavian Journal of Forest Research*, 18(1):64–77.

Eid, T. and Tuhus, E. (2001). Models for individual tree mortality in Norway. *Forest Ecology and Management*, 154(1-2):69–84.

Fang, Y. (2011). Asymptotic equivalence between cross-validations and Akaike Information Criteria in mixed-effects models. *Journal of Data Science*, 9(1):15–21.

Gibbons, R. D. and Hedeker, D. (1997). Random effects probit and logistic regression models for three-level data. *Biometrics*, 53(4):1527–1537.

Groom, J. D., Hann, D. W., and Temesgen, H. (2012). Evaluation of mixed-effects models for predicting Douglas-fir mortality. *Forest Ecology and Management*, 276(0):139–145.

Hamilton, D. A. (1974). Event probabilities estimated by regression. Res. Pap. INT-152, USDA For. Ser.

Harrell, Jr., F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer-Verlag, New York.

Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Statistics in Medicine*, 22(9):1433–1446.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal data analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Kiernan, D., Bevilacqua, E., Nyland, R., and Zhang, L. (2009). Modeling tree mortality in low- to medium-density uneven-aged hardwood stands under a selection system using generalized estimating equations. *Forest Science*, 55(4):343–351.

Kobe, R. K. and Coates, K. D. (1997). Models of sapling mortality as a function of growth to characterize interspecific variation in shade tolerance of eight tree species of northwestern British Columbia. *Canadian Journal of Forest Research*, 27(2):227–236.

Liu, L. C. and Hedeker, D. (2006). A mixed-effects regression model for longitudinal multivariate ordinal data. *Biometrics*, 62(1):261–268.

Ma, Z., Peng, C., Li, W., Zhu, Q., Wang, W., Song, X., and Liu, J. (2013). Modeling individual tree mortality rates using marginal and random effects regression models. *Natural Resource Modeling*, 26(2):131–153.

Monserud, R. A. and Sterba, H. (1999). Modeling individual tree mortality for Austrian forest species. *Forest Ecology and Management*, 113(2-3):109–123.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692.

Reineke, L. H. (1933). Perfecting a stand density index for even-aged stands. *Journal of Agricultural Research*, 46:627–638.

Somers, G. L., Oderwald, R. G., Harms, W. R., and Langdon, G. O. (1980). Predicting mortality with a Weibull distribution. *Forest Science*, 26(2):291–300.

Timilsina, N. and Staudhammer, C. L. (2012). Individual tree mortality model for slash pine in Florida: a mixed modeling approach. *Southern Journal of Applied Forestry*, 36(4):211–219.

Vanclay, J. K. (1995). Synthesis: growth models for tropical forests: a synthesis of models and methods. *Forest Science*, 41(1):7–42.

Weiskittel, A. R., Hann, D. W., Kershaw, J. A., and Vanclay, J. K. (2011). *Forest growth and yield modeling*. Wiley-Blackwell, Chicester, UK, 2 edition.

Woollons, R. C. (1998). Even-aged stand mortality estimation through a two-step regression process. *Forest Ecology and Management*, 105(1):189–195.

Yang, Y., Titus, S. J., and Huang, S. (2003). Modeling individual tree mortality for white spruce in Alberta. *Ecological Modelling*, 163(3):209 – 222.

Yao, X., Titus, S. J., and MacDonald, S. E. (2001). A generalized logistic model of individual tree mortality for aspen, white spruce, and lodgepole pine in Alberta mixedwood forests. *Canadian Journal of Forest Research*, 31(2):283–291.

Yoda, K., Kira, T., Ogawa, H., and Hozumi, K. (1963). Self thinning in overcrowded pure stands under cultivated and natural conditions. *Journal of Biology, Osaka City University*, 14:106–129.

Zhao, D., Borders, B., Wang, M., and Kane, M. (2007). Modeling mortality of second-rotation loblolly pine plantations in the Piedmont/Upper Coastal Plain and Lower Coastal Plain of the southern United States. *Forest Ecology and Management*, 252(1-3):132–143.

Chapter 4

Modeling Loblolly Pine (*Pinus taeda* L.) Clustered Survival Time with Time-dependent Covariates and Shared Frailties

Abstract

Tree mortality is an important component of forest growth and yield models due it significant role in controlling stand dynamics. Accurate prediction of mortality is an important part of forest growth and yield prediction systems. Repeated measure data collected from permanent sample plots established in 1980/81 across the natural range of loblolly pine in the Atlantic Coastal Plain, Gulf Coastal Plain and Piedmont were used. One of the main objectives of this study was to explain the survival of loblolly pine trees using time-varying covariates such as diameter at breast, total tree height, crown ratio, stand age, stand basal area and dominant height. Individual-tree

mortality was modeled using semiparametric proportional hazards regression. Shared frailty model, mixed model extension of Cox proportional hazards model, was used to account for unobserved heterogeneity not explained by the observed covariates in the Cox model. Fixed covariate physiographic region; tree-level covariates total tree height (m), DBH (cm) and crown ratio; and stand-level covariates stand age (years), basal area (m^2 ha^{-1}) and stand dominant height (m) were the significant covariates in the model. Gamma frailty model had smaller AIC and hence it fitted the data significantly better than lognormal frailty. Indexes of predictive ability computed from both the original data and bootstrap samples were higher for the Cox model with frailty suggesting the greater predictive strength of the shared gamma frailty model as compared to the regular Cox proportional hazards model.

4.1 Introduction

Accurate prediction of mortality is an important part of forest growth and yield prediction systems. However, mortality is one of the least understood components of the system (Hamilton, 1986) due to complex interactions between different factors such as environmental, physiological, pathological, and random events. The development of tree mortality models commonly requires data from the remeasurement of permanent sample plots (Clutter et al., 1983). Mortality being rare and stochastic, repeated measure data from permanent sample plots frequently contain a large proportion of plots with no occurrence of mortality and unequal measurement intervals.

A number of different approaches have been used in predicting tree survival (or mortality). Early stand-level mortality models provided stand estimates but advancement in computing techniques saw development of individual tree mortality models where each tree is assigned an estimate of probability of survival (Clutter et al., 1983). In 1970's and before, linear and polynomial functions were used to model mortality; for example, Staebler (1953) used linear functions to predict percent survival from age, site index, and mean diameter and Lee (1971) applied linear regression analysis to data from existing yield tables to predict mortality rates of lodgepole pine. Whole-stand mortality models have been commonly developed using derivative of the generalized Gamma distribution (Weibull or exponential) or the difference equation approach. Somers et al. (1980) used the Weibull distribution to predict mortality of young natural loblolly pine stands in South Carolina. Mortality functions derived from differential equations, expressing basic ecological principles of stand dynamics, possess desirable properties of a mortality model (Clutter et al., 1983; Diéguez-Aranda et al., 2005). Clutter and Jones (1980) developed a survival function based on a difference equation with an assumption that mortality rate is proportional to stand age and density

$$\frac{1}{N}\frac{dN}{dA} = \alpha A^{\delta} N^{\beta} \tag{4.1}$$

where N is number of trees per unit area at age A, $\frac{dN}{dA}$ is instantaneous mortality rate at age A, and α, δ, and β are parameters.

Integrating over the initial conditions that $N_2 = N_1$ when $A_2 = A_1$ gives

$$N_2 = \left[N_1^{\beta_1} + \beta_2 \left(A_2^{\beta_3} - A_1^{\beta_3} \right) \right]^{\frac{1}{\beta_1}} \tag{4.2}$$

This whole-stand model has been used later by others, often with some modification, in several growth prediction systems. Mortality functions derived from differential equations are generally nonlinear and convergence in estimation may sometimes be difficult to achieve. Data from permanent plot systems often show no mortality even over periods of several years (Woollons, 1998). Discarding data from the plots with no occurrence of mortality introduces significant bias and the mortality is overestimated. Retaining all the data in analyses results in underestimation of mortality (Woollons, 1998; Eid and Øyen, 2003). Stand-level mortality is based on an assumption of Gaussian distribution among the observations, which is rarely appropriate for repeated measure data that exhibit varying degree of dispersion and skewness (Affleck, 2007). Forest mortality analyses often focus on logistic regression modeling at the individual tree-level. Hamilton (1974) introduced the logistic function as an individual tree mortality model and since then it has been widely used for many tree species (e.g. Monserud, 1976; Buchman, 1979; Hamilton, 1986; Avila and Burkhart, 1992; Vanclay, 1995; Yao et al., 2001; Zhao et al., 2007).

Flexibility, ease of use, and straightforward interpretability might be the reasons in prevalence of the logistic function for modeling individual tree mortality (Rose et al., 2006).

However, the logistic mortality models do not directly account for changes that occur in the time-dependent covariates over time. It is commonplace in forestry to encounter tree- and stand-level covariates (e.g. diameter at breast height, crown class, density, etc.) that vary over time; using the dynamic covariate information in the mortality model seems appropriate. One method of incorporating dynamic variables would be survival analysis technique with time-dependent covariates. Mortality analytical techniques commonly used in forestry lack methodology for including time-dependent covariates, hypothesis testing, censored observations and testing the effects of covariates. Survival analysis techniques allow for censoring of observations, inclusion of time-dependent covariates, testing the assumption of a constant hazard function and dealing with non-normal distributions (Klein and Moeschberger, 2003).

Most survival analysis studies do not focus on regular (or non-catastrophic) mortality of trees and address dynamic nature of time-dependent covariates. This study modeled regular mortality of individual trees and also accounted for the time-dependent nature of individual tree- and stand-level covariates using Cox proportional hazards regression. The objective of this chapter was to develop an individual-tree mortality model for even-aged stands of loblolly pine throughout its geographic range in the United States using Cox proportional hazards model. Both plot and tree information from permanent plots were used to develop a survival model to estimate the probability of a tree surviving to a given time period. Survival time of trees in a plot may be correlated due to clustering effect. None of the past survival analysis studies of individual tree mortality has addressed this issue of correlation. This study used shared frailty model extension of the Cox proportional hazards model to study clustered survival time with time-dependent covariates.

4.2 Data

The region wide thinning study data used in Chapter 2 was used this study. The details of the plot establishment and data is described in Section 2.2 in Chapter 2. 171 plots of the 186 total plot establishment were used in the analyses.There were 105 plots in the Coastal Plain and 66 plots in the Piedmont region. The 15 plots out of total 186 establishments were dropped because some had incidence of insect attack and other were measured once only. There were 9,961 loblolly pine trees in those 171 plots.

The event of interest was the tree death and the time-to-event variable is the number of years between the tree death and the plot establishment. For trees that survived up to the eight measurements, the time variable was recorded as a censored observation. An event indicator is used to represent the status of the tree at the end of the eight measurement. Information on year of tree death was available even though the tree variables were recorded every three year. Physiographic region (Coastal Plain and Piedmont) is the time invariant variable and diameter at breast height (DBH), total tree height, crown ratio, stand age, stand basal area per ha and height of dominant stand are the time varying explanatory variables used in the analysis. The stand-level variables age, stand basal area per ha and dominant height are the three main drivers of forest stand dynamics i.e. stand age, stand density and site quality, respectively (Burkhart and Tomé, 2012). For tree-level variables, DBH, total tree height and crown ratio were used. Summary statistics for permanent plots in the Coastal Plain and Piedmont region are given in Table 4.1. Number of dead and live trees (δ), survival proportion, and mean (standard deviation), minimum and maximum of tree and stand-level variables are given.

DBH, total tree height and crown ratio development for randomly selected 200 trees in unthinned plots are shown in Figure 4.1.

Table 4.1: Summary statistics of tree and stand-level characteristics in the Coastal Plain and Piedmont

Region	No. of trees		Survived proportion	Variable	Mean (SD)	Min	Max
	Survived (δ=0)	Dead (δ=1)					
Coastal Plain	4049	1912	0.6792	DBH (cm)	17.69 (5.66)	1.27	43.18
				Total height (m)	15.38 (4.54)	1.52	33.53
				Crown ratio	0.36 (0.11)	0.02	0.94
				Stand-level			
				Age (year)	22.59 (6.98)	8.00	45.00
				Dominant height (m)	16.53 (4.46)	4.17	30.02
				Basal area (m^2 ha^{-1})	25.79 (8.60)	2.96	53.70
Piedmont	3009	1002	0.7502	DBH (cm)	17.15 (5.19)	1.52	39.88
				Total height (m)	14.96 (4.05)	1.52	28.96
				Crown ratio	0.36 (0.10)	0.02	0.85
				Stand-level			
				Age (year)	23.15 (6.77)	9.00	43.00
				Dominant height (m)	15.79 (3.69)	5.96	26.50
				Basal area (m^2 ha^{-1})	26.14 (8.58)	4.57	49.65

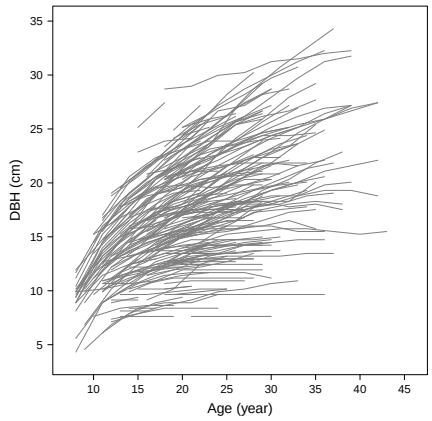
4.3 Methods

4.3.1 Survival analysis preliminaries

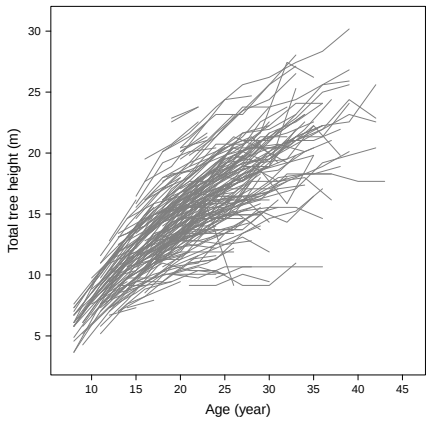
Survival analysis techniques model the time it takes for the events of interest to occur (time-to-event) and focuses on the distribution of survival times. However, several alternative time scales have been considered in past studies. Woodall et al. (2005) used ΔDBH (increase in DBH from initial forest inventory) rather than time since entry to the study as a natural time scale. The survival and hazard functions in survival analysis are used to quantify the probability distribution of time-to-event phenomenon in a population (Klein and Moeschberger, 2003). The survival function is defined by

$$S(t) = P(T > t) = \int_t^\infty f(u)\, du \tag{4.3}$$

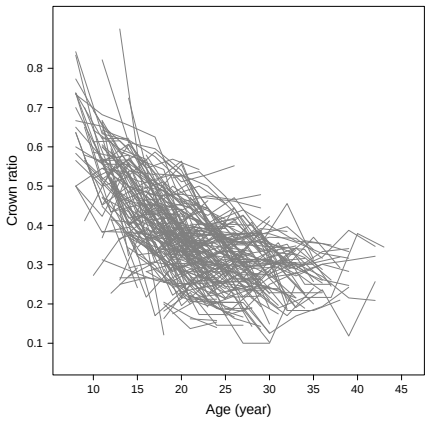
where $S(t)$ is the probability of a tree to survive to time t, f is the probability density function



(a)



(b)



(c)

Figure 4.1: Development of (a) DBH (cm), (b) total tree height (m) and (c) crown ratio for a subset of trees

of the even time t . If T is continuous random variable with the cumulative distribution function (cdf)

$$F(t) = P(T \leq t) = \int_0^t f(u)\, du \tag{4.4}$$

then $S(t)$ is a continuous, strictly decreasing function. A common parametric distribution used is the Weibull distribution due to its greater flexibility in accommodating increasing,

decreasing, and constant hazard functions (Klein and Moeschberger, 2003).

The hazard function, also called the *force of mortality* or *instantaneous event (death) rate*, is defined by

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \tag{4.5}$$

where $h(t)$ is the instantaneous mortality rate of a tree at time t, given that the tree survives until time t.

Using the law of conditional probability, Equation (4.5) becomes

$$h(t) = \frac{f(t)}{S(t)} \tag{4.6}$$

The corresponding cumulative hazard function is defined by

$$H(t) = \int_0^t h(u)\, du \tag{4.7}$$

The empirical (non-parametric) estimator, proposed by Kaplan and Meier (1958) and also called the Product-Limit estimator, provides simple estimates of the survival function when censoring is present. Let the tree deaths occur at D distinct times $t_1 < t_2 < \cdots < t_D$. This estimator is defined as following for all values of t in the data range

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1, \\ \displaystyle\prod_{t_i \leq t} \left[1 - \frac{d_i}{Y_i}\right] & \text{if } t_1 \leq t. \end{cases} \tag{4.8}$$

where d_i is the number of events (or deaths) at time t_i, Y_i is the number of trees that are at

risk at time t_i (i.e. the number of trees that are alive at t_i or experience the event at time t_i and Y_i does not include censored observations).

The quantity $\frac{d_i}{Y_i}$ provides an estimate of the conditional probability that a tree that survives to just prior to time t_i experience the event (or death) at time t_i.

The variance of this estimator is estimated by Greenwood's formula

$$\hat{V}\left[\hat{S}(t)\right] = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)} \tag{4.9}$$

The common estimator of the cumulative hazard function is the Nelson-Aalen estimator which was first suggested by Nelson (1972) and later rediscovered by Aalen (1978). This estimator is given by

$$\tilde{H}(t) = \begin{cases} 0 & \text{if } t < t_1, \\ \sum_{t_i \leq t} \frac{d_i}{Y_i} & \text{if } t_1 \leq t. \end{cases} \tag{4.10}$$

The variance of this estimator is given by

$$\tilde{V}\left[\tilde{H}(t)\right] = \sum_{t_i \leq t} \frac{d_i}{Y_i^2} \tag{4.11}$$

The specific feature that distinguishes survival analyses from other statistical analysis is its ability to handle data censoring and time-dependent covariates (Hougaard, 1999). It makes use of both censored and uncensored observations. Tree survival data from permanent plot systems are examples of Type I censoring where the event (or death) is observed only if it occurs prior to some prespecified censoring time. For right-censored data, information for tree i ($i = 1, \ldots, n$) can be conveniently represented by pairs of random variables (T_i, δ_i),

where T_i is the minimum of the event time t_i and censoring time c_i ($T_i = \min(t_i, c_i)$ and δ_i is the censoring indicator ($\delta_i = 1$ if $t_i \leq c_i$ and $\delta_i = 0$ if $t_i > c_i$)). The hazard function used in survival analysis helps to understand the mechanism of failure or death (Hougaard, 1999).

4.3.2 Cox proportional hazards model

Since Cox (1972) introduced the proportional hazards model, models based on the hazard function have dominated survival analysis due to the ease with which technical difficulties such as censoring and truncation can be handled in time-to-event data. Let $h(t)$ denote the hazard of a tree at time t with covariate vector $\boldsymbol{Z'} = (Z_1, \ldots, Z_p)$, then the proportional hazards model specifies that

$$h(t) = h_0(t) \exp(\boldsymbol{\beta^t Z}) \tag{4.12}$$

where $h_0(t)$ is a baseline hazard function and $\boldsymbol{\beta}$ is a parameter that describes the importance of the covariates.

The baseline hazard rate is treated nonparametrically. Cox (1972) suggested an estimation method that removed the effect of $h_0(t)$, letting it be completely unspecified and the analysis concentrating on the effect of the covariates. When time-dependent covariates are present, the Cox regression model can be used but it no longer satisfies the proportional hazards assumption. Therefore, extended Cox regression model is used instead which is represented by Equation (4.13)

$$h_{ij}(t) = h_0(t) \exp\left[\boldsymbol{\beta^t Z}_{ij}(t)\right] \tag{4.13}$$

where $\boldsymbol{Z}_{ij}(t)$ includes both the time-independent and time-dependent covariates.

4.3.2.1 Parameter estimation using partial likelihood

Let the data of size n consisting of the triple $(T_j, \delta_j, \boldsymbol{Z_j})$, $j = 1, \ldots, n$ where T_j is the time on study for the jth tree, δ_j is the event indicator for the jth tree ($\delta_j = 1$ if the event has occured and $\delta_j = 0$ if the lifetime is right-censored) and $\boldsymbol{Z_j} = (Z_{j1}, \ldots, Z_{jp})^t$ is the vector of covariates for the jth tree at time t.

Suppose that there are no ties between the event times. Let $t_1 < t_2 < \cdots < t_D$ denote the D distinct, ordered, event times. Let d_i be the number of deaths at t_i (here $d_i = 1$ for all i). Let $R(t_i)$ be the risk set i.e. a set of all individuals that are at risk just prior to t_i. Let $Z_{(i)k}$ be the kth covariate associated with the individual with event time t_i.

The hazard function is specified by

$$h(t \mid \boldsymbol{Z}) = h_0(t) \exp(\boldsymbol{\beta^t Z}) = h_0(t) \exp\left(\sum_{k=1}^{p} \beta_k Z_k\right) \tag{4.14}$$

The partial likelihood based on the above hazard function can be obtained as following. The conditional probability that an individual with covariates $Z_{(i)}$ dies at time t_i, given the individuals in $R(t_i)$ dies at this time, is given by

$$
\begin{aligned}
P\left[\text{individual dies at } t_i \mid \text{one death at } t_i\right] &= \frac{P\left[\text{individual dies at } t_i \mid \text{survival to } t_i\right]}{P\left[\text{one death at } t_i \mid \text{survival to } t_i\right]} \\
&= \frac{h\left[t_i \mid \boldsymbol{Z_{(i)}}\right]}{\sum_{j \in R(t_i)} h\left[t_i \mid \boldsymbol{Z_j}\right]} \\
&= \frac{h_0(t_i) \exp\left[\boldsymbol{\beta^t Z_{(i)}}\right]}{\sum_{j \in R(t_i)} h_0(t_i) \exp\left[\boldsymbol{\beta^t Z_j}\right]} \\
&= \frac{\exp\left[\boldsymbol{\beta^t Z_{(i)}}\right]}{\sum_{j \in R(t_i)} \exp\left[\boldsymbol{\beta^t Z_j}\right]}
\end{aligned}
$$

The partial likelihood is formed by multiplying these conditional probabilities over all deaths as

$$
\begin{aligned}
L(\boldsymbol{\beta}) &= \prod_{i=1}^{D} \frac{\exp\left[\boldsymbol{\beta^t Z_{(i)}}\right]}{\sum_{j \in R(t_i)} \exp\left[\boldsymbol{\beta^t Z_j}\right]} \\
&= \prod_{i=1}^{D} \frac{\exp\left(\sum_{k=1}^{p} \beta_k Z_{(i)k}\right)}{\sum_{j \in R(t_i)} \exp\left(\sum_{k=1}^{p} \beta_k Z_j k\right)}
\end{aligned} \tag{4.15}
$$

The partial likelihood in Equation (4.15) is not a full likelihood function but it has almost all the desirable properties of likelihood functions. Cox and others have shown that it can be treated as a ordinary likelihood to derive (partial) maximum likelihood estimates of $\boldsymbol{\beta}$.

The log of the partial likelihood is,

$$\begin{aligned} LL(\boldsymbol{\beta}) &= \ln\left[L(\boldsymbol{\beta})\right] \\ &= \sum_{i=1}^{D}\sum_{k=1}^{p}\beta_k Z_{(i)k} - \sum_{i=1}^{D}\ln\left[\sum_{j\in R_{(t_i)}}\exp\left(\sum_{k=1}^{p}\beta_k Z_{jk}\right)\right] \end{aligned} \tag{4.16}$$

The (partial) MLEs are found by maximizing Equation (4.16) or equivalently Equation (4.15). The likelihood function equation does not depend on the baseline hazard rate $h_0(x)$ and hence the inferences may be made on the effects of the explanatory variables without knowing $h_0(x)$.

When there are ties (two or more occurrences of the event or death) between the event times, there are several suggestions for constructing the partial likelihood. When the number of ties is not large, Breslow (1974) has derived partial likelihood function expressed as

$$L_1(\boldsymbol{\beta}) = \prod_{i=1}^{D}\frac{\exp\left[\boldsymbol{\beta^t s_i}\right]}{\left[\sum_{j\in R_i}\exp\left(\boldsymbol{\beta^t Z_j}\right)\right]^{d_i}} \tag{4.17}$$

where $\mathbf{s}_i$ is the sum of the vectors $\boldsymbol{Z}_j$ over all individuals that died at t_i i.e. $\mathbf{s}_i = \sum_{\mathbb{D}}\boldsymbol{Z}_j$, $\mathbb{D}$ is the set of all individuals that died at time t_i and $\boldsymbol{R}_i$ is the set of all individuals at risk just prior to t_i.

Efron (1977) suggested a partial likelihood function of the form

$$L_2(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\left[\boldsymbol{\beta^t s_i}\right]}{\prod_{j=1}^{d_i}\left[\sum_{k\in R_i}\exp\left(\boldsymbol{\beta^t Z_k}\right) - \frac{j-1}{d_i}\sum_{k\in \mathbb{D}i}\exp\left(\boldsymbol{\beta^t Z_k}\right)\right]} \tag{4.18}$$

Efron's and Breslow's likelihoods are quite close when the number of ties present in the data is small.

4.3.2.2 Validating the fitted Cox model

Cox proportional hazards model was validated using bootstrapping procedure. Efron and Tibshirani (1993) describe bootstrap resampling procedure for obtaining unbiased estimates of future model performance without sacrificing sample size. With the "simple bootstrap", a model is fitted repeatedly in a bootstrap sample and the model performance is evaluated on the original sample (Efron and Tibshirani, 1993). Bootstrap estimates the bias due to overfitting or the "optimism" in the final model, which is then subtracted from the index of accuracy derived from the original sample to obtain a bias-corrected or overfitting-corrected estimate of predictive accuracy of the model. The bootstrap validation of the Cox model was done following procedures described in Harrell (2001). From the original sample of size n, sample with replacement of size n is drawn. A model is obtained from the bootstrap sample and is applied to the original sample. An estimate of optimism is then calculated as a difference of the accuracy index from the bootstrap sample and the index calculated on the original sample. This process is repeated for multiple bootstrap replications (number of repetitions $B = 200$ or more) to obtain an average optimism, which when subtracted from the final model fit's apparent accuracy gives the overfitting-corrected estimate. A good measure of the fitted model's predictive discrimination is Somers' rank correlation coefficient D_{xy} or c index, which is a probability of concordance between predicted probability and response or

generalized "receiver operating characteristic" ROC area (Harrell, 2001). The concordance index c and Somers' D_{xy} rank correlation are related as

$$D_{xy} = 2(c - 0.5) \tag{4.19}$$

A model with $D_{xy} = 0$ means it makes random prediction and the model with $D_{xy} = 1$ makes the perfectly discriminating predictions. Similarly, a value of c of 0.5 means random prediction and a value of 1 means perfect prediction. *Slope* of calibration plot, a plot of observed response against predicted response, can also be used as a measure of overfitting. *Slope* is also called a *shrinkage factor*. When model with parameter estimates obtained from one dataset is applied to predict responses on another dataset, overfitting causes the slope of the clibartion plot to be less than one. With the Cox regression model, *slope shrinkage* is obtained by bootstrapping the apparent estimate of $\gamma = 1$ in the model

$$h(t) = h_0(t) \exp(\gamma \boldsymbol{\beta^t Z}) \tag{4.20}$$

The bootstrap estimate of γ also leads to indexes of unreliability (U), discrimination (D) and overall quality ($Q = D - U$). Unreliability index (U) measures how far the model maximum log likelihood is from the log likelihood evaluated at "frozen" regression coefficient ($\gamma = 1$)

$$U = \frac{LR(\hat{\gamma} \boldsymbol{\beta^t Z}) - LR(\boldsymbol{\beta^t Z})}{L^0} \tag{4.21}$$

where L^0 is the -2 log likelihood for the null model (i.e. a model with no predictive information). Discrimination index (D) is derived from the -2 log likelihood at the shrunken linear predictor, penalized for estimating one parameter (γ)

$$D = \frac{LR(\hat{\gamma}\boldsymbol{\beta^t Z}) - 1}{L^0} \tag{4.22}$$

An overall quality index (Q) that penalized discrimination for unreliability is computed as

$$Q = D - U = \frac{LR(\boldsymbol{\beta^t Z}) - 1}{L^0} \tag{4.23}$$

The closer the value of U to 0, the better it is and the higher the value of D and Q, the better.

4.3.3 Semiparametric shared frailty model

Proportional hazards models are based on the assumption of homogeneous population and independence among observations from different subjects. However, in many applications, the population under study is not homogeneous. Unobserved heterogeneity or variation not explained by observed covariates is dealt with by introducing random effects or "frailty" in the survival model. Vaupel et al. (1979) introduced the term "frailty" to explain the existence of essential differences between individuals. A frailty could be an unobservable random effect shared by subjects within a cluster. In a permanent sample plot system, trees in the same plot (cluster) may share similar non-observed environmental factors that affect their survival. The random effect modeling in survival analysis can also be applied to deal with repeated measurements on permanent sample plots. The most common model for a frailty is a shared frailty model extension of the Cox's proportional hazards regression model. It is a conditional independence model that assumes all event times in a cluster are independent given the frailty variables.

Let i ($i = 1, \ldots, n$) be number of plots (clusters) and j ($j = 1, \ldots, n_i$) be number of trees

in plot i. Let t_{ij} be the event time/censoring time for the jth tree in plot i. Let δ_{ij} be the event indicator and the number of observed events in the ith cluster is $d_i = \sum_{j=1}^{n_i} \delta_{ij}$. The hazard rate for the jth tree in the ith plot, given the shared frailty, is defined as

$$h_{ij}(t) = h_0(t) \exp\left[\boldsymbol{\beta^t Z_{ij}}(t) + w_i\right] \tag{4.24}$$

where $h_0(t)$ is a baseline function, $\boldsymbol{\beta}$ is a regression coefficient vector, $\boldsymbol{Z_{ij}}(t)$ is a vector of covariates at time t, and w_i is a random effect for the ith plot.

The hazard function in Equation (4.24) can be rewritten as

$$h_{ij}(t) = h_0(t) u_i \exp\left[\boldsymbol{\beta^t Z_{ij}}(t)\right] \tag{4.25}$$

where $u_i = \exp(w_i)$ is called the frailty for the ith cluster.

Models (4.24) and (4.25) are conditional hazard models given the u_i's. The baseline hazard function $h_0(t)$ is estimated nonparametrically. The frailties u_i are assumed to be identically and independently distributed with density function f_U. Various frailty models have been developed and suggested in the literature. Any distribution with a positive random variable can be used to model frailty. Lognormal and one-parameter gamma models for the distribution of the frailty were considered in this study.

In the lognormal frailty model, the random effects w_i in Equation (4.24) are assumed to follow normal distribution with mean zero and variance γ i.e. $w_i \sim \mathcal{N}(0, \gamma)$. The corresponding frailty has a lognormal distribution given by

$$f_U(u, \gamma) = \frac{1}{u\sqrt{2\pi\gamma}} \exp\left(-\frac{(\log u)^2}{2\gamma}\right) \tag{4.26}$$

with $\gamma > 0$. The mean and variance of the frailty are given by $E(U) = \exp(\frac{\gamma}{2})$ and $Var(U) = \exp(2\gamma) - \exp(\gamma)$.

In the gamma frailty model, the frailties u_i in Equation (4.25) are assumed to be identically and independently distributed with one-parameter gamma distribution as

$$f_U(u, \theta) = \frac{u^{\frac{1}{\theta}-1} e^{-\frac{u}{\theta}}}{\Gamma\left(\frac{1}{\theta}\right) \theta^{\frac{1}{\theta}}} \tag{4.27}$$

where Γ is gamma function.

The corresponding density for W is

$$f_W(w) = \frac{\{\exp(w)\}^{\frac{1}{\theta}} \exp\{-\exp(w)/\theta\}}{\theta^{\frac{1}{\theta}} \Gamma\left(\frac{1}{\theta}\right)} \tag{4.28}$$

The gamma frailty model is the most popular frailty model due to its mathematical tractability. Here $\mathrm{E}(U) = 1$ and $\mathrm{Var}(U) = \theta$. The individuals in a cluster i with $u_i > 1$ ($u_i < 1$) are frail (strong) i.e. at higher risk or lower risk, respectively. The variance of frailty distribution (θ) gives information on the heterogeneity among clusters. The large variance means greater heterogeneity in the individual hazards. Larger variance also indicate stronger association within clusters and the association between cluster members are measured by Kendall's τ given by

$$\tau = \frac{\theta}{(\theta + 2)} \text{ with } SE(\tau) = \frac{2SE(\theta)}{(\theta + 2)^2}$$

Based on the hazard function in Equation (4.25), the cumulative hazard function is given by

$$H_{ij}(t) = \int_0^t h_{ij}(s)ds = u_i H_{ij}^a(t) \tag{4.29}$$

where $H_{ij}^{a}(t) = \int_0^t h_0(s) \exp\left[\boldsymbol{\beta^t Z_{ij}}(s)\right] ds$. The survival function is given by

$$S_{ij}(t) = \exp\left[-H_{ij}(t)\right] \tag{4.30}$$

4.3.3.1 Parameter estimation using penalized partial likelihood

The penalized approach is based on a modification of the Cox partial likelihood to include both the regression coefficients and the frailties. The frailties are treated as additional coefficients, which are then constrained by adding a penalty terms to the log-likelihood. With $h_0(t)$, the unspecified baseline hazard function in Equation (4.25), full likelihood for the lognormal and gamma frailty model is given by Equation (4.31) and Equation (4.32), respectively

$$l_{ppl}(\boldsymbol{\beta}, \mathbf{w}, \gamma) = l_{part}(\boldsymbol{\beta}, \mathbf{w}) + l_{pen}(\gamma, \mathbf{w}) \tag{4.31}$$

$$l_{ppl}(\boldsymbol{\beta}, \mathbf{w}, \theta) = l_{part}(\boldsymbol{\beta}, \mathbf{w}) + l_{pen}(\theta, \mathbf{w}) \tag{4.32}$$

In the penalized partial likelihood method the second part of the likelihood is considered to be a penalty term and is given by

$$l_{pen}(\mathbf{w}) = -\sum_{i=1}^{n} \log f_W(w_i) \tag{4.33}$$

The penalty term for lognormal frailty is given by

$$l_{pen}(\gamma, \mathbf{w}) = \frac{1}{2} \sum_{i=1}^{n} \left(\frac{w_i^2}{\gamma} + \log(2\pi\gamma) \right) \tag{4.34}$$

and for gamma frailty, it is given by

$$l_{pen}(\theta, \mathbf{w}) = -\frac{1}{\theta}\sum_{i=1}^{n}(w_i - \exp(w_i)) \tag{4.35}$$

The penalized partial likelihood for the frailty model is given by

$$l_{ppl}(\boldsymbol{\beta}, \mathbf{w}, \theta) = \log\left[\prod_{i=1}^{n}\prod_{j=1}^{n_i}\left[\frac{u_i \exp\{\boldsymbol{\beta^t Z_{ij}}(t_{ij})\}}{\sum_{R(t_{ij})} u_q \exp\{\boldsymbol{\beta^t Z_{qs}}(t_{ij})\}}\right]^{\delta_{ij}}\right] + \sum_{i=1}^{n}\log f_W(w_i) \tag{4.36}$$

where $R(t_{ij})$ is summation over all (q, s) indices i.e. the sum over all individuals in the risk set at time t_{ij}.

The maximization of the penalized partial log-likelihood for both the frailty models consists of an inner and an outer loop. In the inner loop, the Newton-Raphson method is used to maximize $l_{ppl}(\boldsymbol{\beta}, \mathbf{w}, \gamma)$ (or $l_{ppl}(\boldsymbol{\beta}, \mathbf{w}, \theta)$), for a provisional value of γ (or fixed value of θ), to obtain estimates for $\boldsymbol{\beta}$ and $\mathbf{w}$. However, the penalized partial log-likelihood approach uses different outer loop for the lognormal and gamma frailty distribution. In the outer loop of lognormal frailty, the restricted maximum likelihood estimator for γ is obtained by using the best linear unbiased predictors (BLUPs). In the gamma frailty, since the restricted maximum likelihood estimator for θ is not available, the outer loop is based on the maximization of a profile marginal likelihood for θ. The process is iterated until convergence. The details of the algorithm can be found in Duchateau and Janssen (2008). The parameter estimates obtained from EM algorithm and partial penalized likelihood approach for the semiparametric gamma frailty model are same (Duchateau and Janssen, 2008).

4.3.3.2 Index of predictive ability

Log likelihood can be used to obtain a unitless measure of predictive ability for a Cox proportional hazards model with frailty. The first index is R index, which is the square root of the penalized fraction of log likelihood explained. Let L is -2log likelihood for the fitted model and L^0 is -2log likelihood for a null model with no predictive information. R^2 for Cox PH model is defined by

$$R^2 = \frac{(LR - 2p)}{L^0} \tag{4.37}$$

where $LR = L^0 - L$, the log likelihood ratio statistics for testing the global null hypothesis that $\hat{\boldsymbol{\beta}} = 0$. A perfectly predictive model gives R^2 near one and a model that does not discriminate between short and long survival times gives near zero. However, R^2 is too sensitive to the distribution of censoring times and a more complex measure has been suggested (Harrell, 2001).

A generalized R_N^2 index suggested by Nagelkerke (1991) that ranges from 0 to 1 is given by

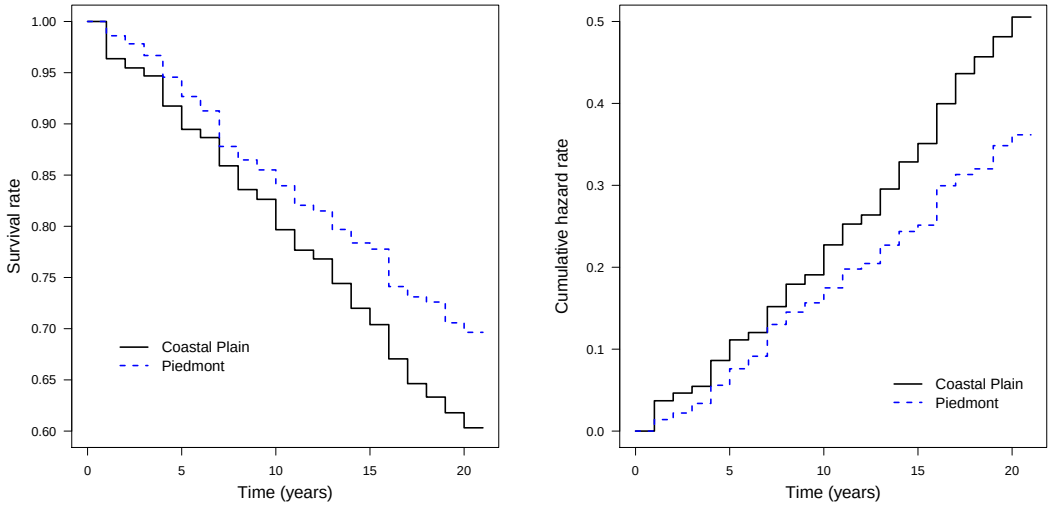
$$R_N^2 = \frac{1 - \exp\left(-\frac{LR}{n}\right)}{1 - \exp\left(-\frac{L^0}{n}\right)} \tag{4.38}$$

where n is the total number of observation.

4.4 Results and discussion

4.4.1 Cox proportional hazards regression

Plots of the estimated Kaplan-Meier survival function and the corresponding cumulative hazard function for the unthinned plots in Coastal Plain and Piedmont are given in Figure 4.2(a) and Figure 4.2(b), respectively. The survival curves in Figure 4.2(a) were obtained from the raw Kaplan-Meier survival fits.



(a) (b)

Figure 4.2: (a) *Kaplan-Meier* survival curves and (b) hazard curves for Coastal Plain and Piedmont

Trees in the Coastal Plain were at high risk of death than the trees in the Piedmont. Similarly, trees in the Coastal Plain experienced higher accumulated risk than the trees in the Piedmont. A score test of equality of hazard rates in the two physiographic regions rejected the null hypothesis of no difference ($\chi^2_1 = 59.64$ and $p < 0.0001$). When degree of freedom is 1 and the single covariate is categorical, the score test from a Cox model is identical to the log-rank test (Therneau and Grambsch, 2000).

Cox proportional hazards regression was fitted to the data; Table 4.2 shows the parameter estimates (standard error) and hazard ratio (HR) with the corresponding 95% confidence interval. The hazard ratio is an estimate of the ratio of the hazard rates in two groups and tells us how much more likely one individual is to die than another at any particular point in time.

Table 4.2: Parameter estimates and 95% confidence intervals of HR for Cox model

Variable	**Estimate (SE)**	p**-value**	**HR**	**95% CI of HR**	
				Lower	**Upper**
Physio. region	-0.3666 (0.0448)	< 0.0001	0.6931	0.6348	0.7566
Total height (m)	-0.2828 (0.0122)	< 0.0001	0.7537	0.7358	0.7720
DBH (cm)	-0.1376 (0.0085)	< 0.0001	0.8715	0.8571	0.8861
Age (year)	-0.0147 (0.0066)	0.0269	0.9854	0.9727	0.9983
Crown ratio	-5.8837 (0.2579)	< 0.0001	0.0028	0.0017	0.0046
Basal area (m^2 ha^{-1})	0.0251 (0.0040)	< 0.0001	1.0254	1.0173	1.0335
Dominant height (m)	0.3696 (0.0131)	< 0.0001	1.4472	1.4105	1.4849

From the Wald tests in Table 4.2, all the covariates were significantly associated with the survival of trees. The covariate physiographic region is an indicator variable and Coastal Plain was a baseline physiographic region. Crown ratio is assumed to add very little information after DBH and total height are considered in the model and likelihood ratio test was done to test the significance of covariate crown ratio. From the likelihood ratio statistic (χ^2_1 = 538.74 and p-value <0.0001), *crown ratio* was highly significant suggesting *crown ratio* was significantly associated with an effect on survival. Crown conditions have been found to improve the predictive ability of individual tree mortality model (Dobbertin and Brang, 2001). Magnussen et al. (2005) observed a significant influence of crown class on hazard rate of death of white spruce in the Prince George Forest Region of British Columbia, Canada.

The exp(β) term indicates the hazard ratio for one unit increase in a variable. For total height, exp($\beta_{total\ height}$) is 0.75, so that increase in total height by 1 m leads to a reduction in

risk of death by about 25 % among the surviving trees. The survival rate was higher for the Piedmont region with corresponding relative risk of 0.6931. In Figure 4.2(a), the Piedmont has higher survival rates. This difference in survival rate may be due to various climate and soil nutrient and drainage factors such as certain nutrient limitations. Much of the poorly drained soils in the southeastern Coastal Plain suffer Nitrogen and Phosphorous deficiency (Fox et al., 2007). Growth responses of loblolly pine to Phosphorous fertilization on wet Coastal Plain sites are much larger than the responses on Piedmont sites (Schultz, 1997) and nutrient and density management with southern pines minimizes competition-related mortality losses (Jokela et al., 2010). Zhao et al. (2007) stated that the observed higher mortality of loblolly pine in lower Coastal Plain in their study might be related to poorly to moderately well-drained soils inherent in that physiographic region.

The estimates of the regression coefficients of total tree height, DBH, age and crown ratio indicate that trees with larger values of those variables tend to have higher survival rate. Thus keeping the value of the other covariates fixed, increasing the total height by 1 m reduces the risk by almost 25% and so on. The estimates of regression coefficients of stand basal area and dominant height show that trees in stands with higher basal area and site index tend to have lower survival rate. Empirical evidence suggests that density-dependent mortality in plantations starts earlier in better sites and increases with site productivity (Diéguez-Aranda et al., 2005). Uzoh and Mori (2012) reported similar results in their study of managed even-aged stands of ponderosa pine in the western United States. They observed that the risk of a tree dying decreased with increasing DBH and the risk increased with increasing average height of the 5 tallest trees in the plot and site index.

Relationship between each covariate and the log hazard of death was examined using the plot of covariate against the log hazard ration in Figure 4.3. Y-axis shows $X\hat{\beta}$ and the covariates not plotted were set to reference values.

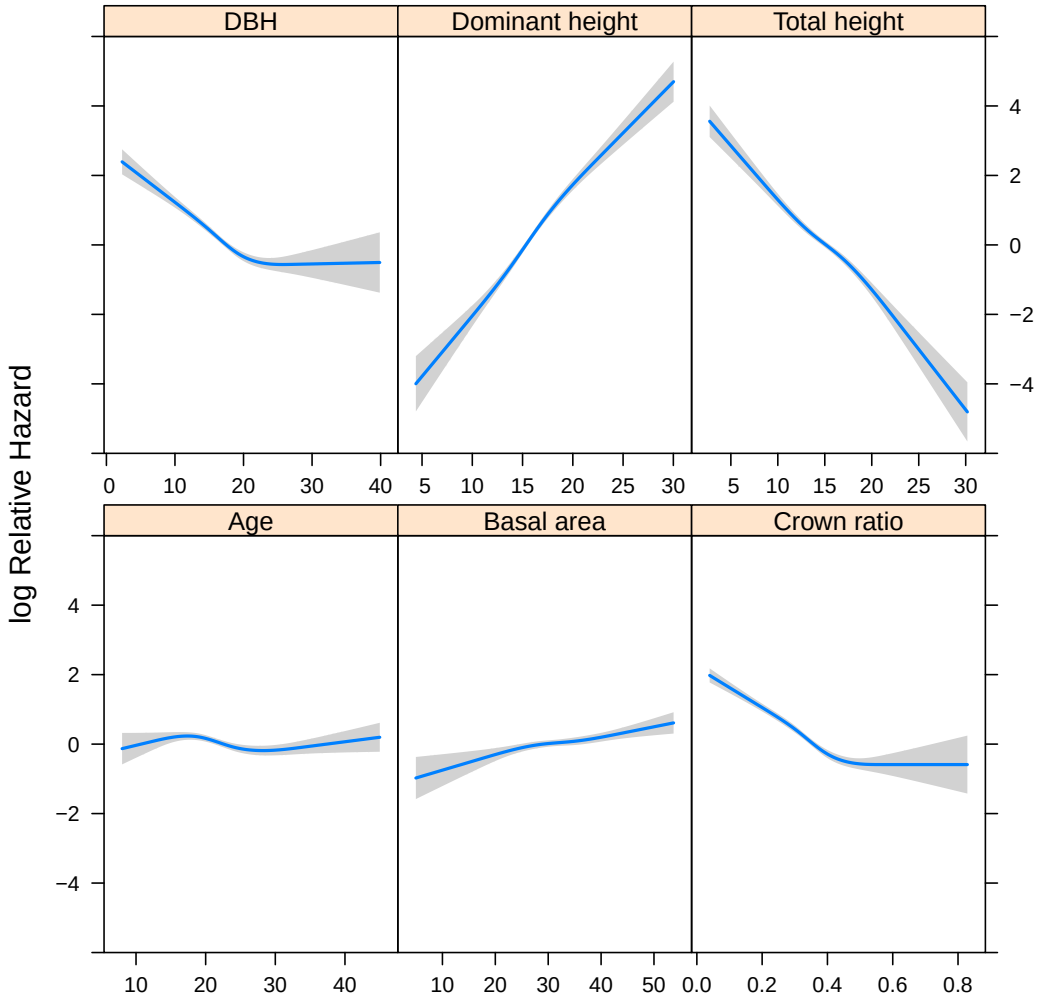


Figure 4.3: Shape of each covariate on log hazard of death with 95% confidence bands

In Figure 4.3, log hazard ratio decreases with increasing values of *total height*, *DBH* and *crown ratio*. But the log hazard ratio becomes constant after certain values for *DBH* and *crown ratio*. Log hazard ratio remains relatively constant for *age*. The log hazard ratio keeps increasing with increasing values of *dominant height* and *basal area*.

The standard errors of $\hat{\boldsymbol{\beta}}$ in Table 4.2 are model-based standard errors that were computed assuming independence among the observations. Since each tree in a plot contributed multiple measurements, this assumption obviously does not hold. Marginal model approach that takes into account the intrasubject correlation in the data was also fitted. This approach is very much similar to the generalized estimating equations (GEE) approach of Liang and Zeger (1986). The parameter estimates, model-based standard error, robust standard error, hazard ratio and their 95% confidence interval are given in Table 4.3

Table 4.3: Parameter estimates and 95% confidence intervals of HR for marginal Cox model

Variable	**Estimate (SE)**	**Robust SE**	***p*-value**	**HR**	**95% CI of HR**	
					Lower	**Upper**
Physio. region	-0.3666 (0.0448)	0.1014	0.0003	0.6931	0.5681	0.8455
Total height (m)	-0.2828 (0.0122)	0.0233	< 0.0001	0.7537	0.7200	0.7890
DBH (cm)	-0.1376 (0.0085)	0.0140	< 0.0001	0.8715	0.8478	0.8958
Age (year)	-0.0147 (0.0066)	0.0159	0.3560	0.9854	0.9552	1.0166
Crown ratio	-5.8837 (0.2579)	0.4768	< 0.0001	0.0028	0.0011	0.0071
Basal area (m^2 ha^{-1})	0.0250 (0.0040)	0.0102	0.0140	1.0254	1.0051	1.0461
Dominant height (m)	0.3696 (0.0131)	0.0317	< 0.0001	1.4472	1.3601	1.5399

The robust variance is a sandwich estimator of form ABA, where A is the ordinary variance and B is a correction term, familiar in GEE methods. The estimates of coefficients and their standard errors in Table 4.3 are same as the estimates in Table 4.2 that were obtained assuming independence among observations. However the robust standard errors in Table 4.3 are the proper standard error of the coefficients that were derived taking into account the fact that observations from a tree within a plot are not independent. The uncorrected standard errors are highly deflated. With correlated data the robust sandwich estimate of

variance is substantially larger than the model-based variance (Therneau and Grambsch, 2000). After taking into account the correlation in the data, the covariate *age* was non significant suggesting the hazard rate may not be affected by stand age.

Separate Cox regression models were fitted to Coastal Plain and Piedmont regions to investigate the effects of the covariates on hazard rate at physiographic level. Parameter estimates (standard error) and hazard ratio with the corresponding 95% confidence interval are given in Table 4.4.

Table 4.4: Parameter estimates and 95% confidence intervals of HR of Cox model for Coastal Plain and Piedmont

Region	**Variable**	**Estimate (SE)**	***p*-value**	**HR**	**95% CI of HR**	
					Lower	**Upper**
Coastal Plain	Total height (m)	-0.2691 (0.0151)	< 0.0001	0.7641	0.7418	0.7870
	DBH (cm)	-0.1489 (0.0106)	< 0.0001	0.8616	0.8440	0.8796
	Crown ratio	-6.0390 (0.3136)	< 0.0001	0.0024	0.0013	0.0044
	Basal area (m^2 ha^{-1})	0.0363 (0.0046)	< 0.0001	1.0369	1.0276	1.0463
	Dominant height (m)	0.3380 (0.0140)	< 0.0001	1.4021	1.3640	1.4412
Piedmont	Total height (m)	-0.3192 (0.0212)	< 0.0001	0.7267	0.6971	0.7575
	DBH (cm)	-0.1218 (0.0141)	< 0.0001	0.8853	0.8612	0.9102
	Age (year)	-0.0488 (0.0131)	0.0002	0.9524	0.9282	0.9772
	Crown ratio	-5.6651 (0.4287)	< 0.0001	0.0035	0.0015	0.0080
	Dominant height (m)	0.4876 (0.0270)	< 0.0001	1.6283	1.5444	1.7168

In Coastal Plain all the covariates except the stand *age* were significantly associated with an effect on survival. Likelihood ratio test was performed on the *age* covariate and the test result ($\chi^2_1 = 0.32$ and p-value=0.5708) showed that adding *age* covariate did not result in a significant improvement in model fit. Similarly, in Piedmont the stand *basal area* was not significantly associated with the survival and the likelihood ratio test also supported the result ($\chi^2_1 = 1.06$ and p-value=0.3024). Having fit Cox model to Coastal Plain and Piedmont separately, the estimated distribution of survival rates were plotted in Figure 4.4(a) and Figure 4.4(b), respectively. The survival functions were estimated at the mean values of the covariates in each region.

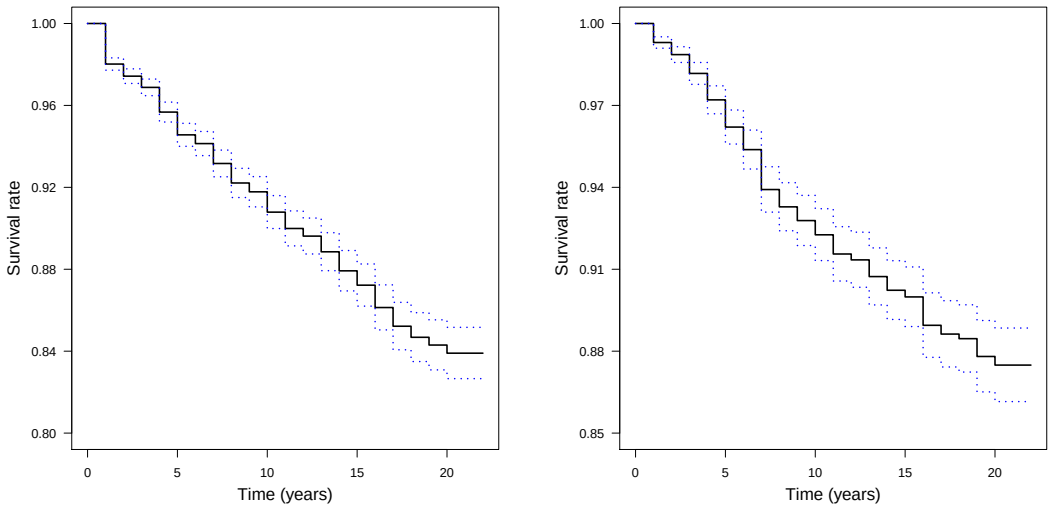


Figure 4.4: Estimated survival function for the Cox regression for (a) Coastal Plain and (b) Piedmont. The dotted lines show 95% confidence interval around the survival fucntion

4.4.1.1 Model diagnostics

Graphical diagnostics for violation of the assumption of proportional hazards were done on all covariates. For categorical covariate physiographical regions the proportional hazards assumption was assessed by plotting $\log(-\log(S(t)))$ against time separately for the Coastal Plain and Piedmont (Figure 4.5).

The two curves of the logarithm of the Nelson-Aalen estimators did not cross suggesting that proportional hazards assumption was satisfied. The graphical diagnostics for other time-dependent covariates were based on the scaled Schoenfeld residuals. The main idea behind this diagnostic is that if the proportional hazards assumption holds for a particular covariate then the Schoenfeld residuals for that covariate will not be related to survival time. The plots in Figure 4.6 do not show any evidence of deviation from the proportionality assumption for any covariate since the plotted curves are roughly constant over time for all the covariates.

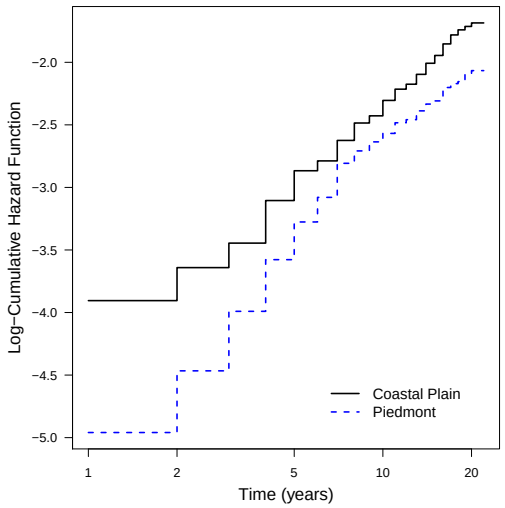


Figure 4.5: Plot of $\log(-\log(S(t)))$ against time for

4.4.1.2 Validation of Cox model

Validation of the Cox regression model in Table 4.2 was done, the bootstrap estimates with $B = 200$ resamplers are given in Table 4.5.

Table 4.5: Bootstrap estimates of some statistical indexes from validation of Cox model

Index	**Original Sample**	**Training Sample**	**Test Sample**	**Optimism**	**Corrected Index**	n
D_{xy}	0.7518	0.7525	0.7517	0.0008	0.7510	200
Slope	1.0000	1.0000	0.9935	0.0065	0.9935	200
D	0.1222	0.1226	0.1220	0.0007	0.1215	200
U	0.0000	0.0000	0.0001	−0.0001	0.0001	200
Q	0.1222	0.1227	0.1219	0.0008	0.1214	200

In Table 4.5, "training' refers to accuracy when the model was evaluated on the bootstrap sample used to fit the model, "test" refers to the accuracy when the fitted model was applied to the original sample and n is the number of bootstrap samples i.e. B. The "optimism" for all indexes was very small. Somers' D_{xy} rank correlation between predicted log hazard and observed survival time was 0.75. From D_{xy} there is no serious overfitting and the corrected

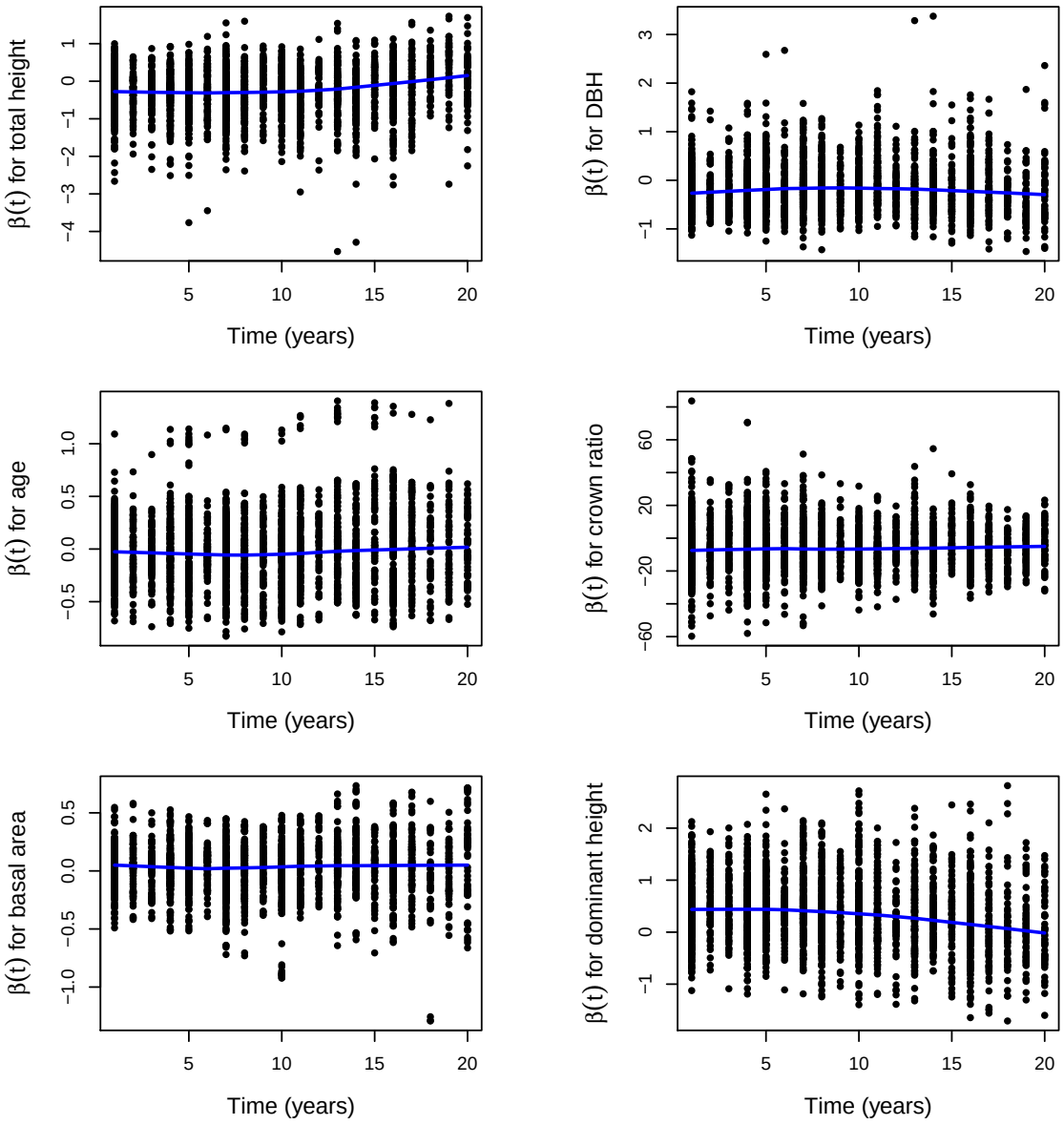


Figure 4.6: Plots of scaled Schoenfeld residuals against time for total height, DBH, age, crown ratio, basal area and dominant height

D_{xy} is the unbiased estimate of future predictive discrimination which is 0.75. The slope shrinkage factor (0.99) is very close to 1 indicating almost no overfitting. The values of other indexes D, U and Q are not troublesome either.

4.4.2 Semiparametric frailty model

The semiparametric frailty models were fitted by partial penalized likelihood approach. The lognormal and gamma distributions were chosen for the frailty. The parameter estimates (standard error), 95% confidence interval of hazard ratio and estimate of the variance of random effects (γ) and variance of frailties (θ) for both distributions are given in Table 4.6. The standard error of γ and θ is the bootstrap standard error calculated from 200 bootstrap samples.

Table 4.6: Parameter estimates and 95% confidence intervals of HR for Cox model with gamma frailty

Frailty	**Variable**	**Estimate (SE)**	***p*-value**	**HR**	**95% CI of HR**		**AIC**
					Lower	**Upper**	
Lognormal	Physio. region	-0.5819 (0.1351)	< 0.0001	0.5588	0.4288	0.7283	41517.45
	Total height (m)	-0.3459 (0.0142)	< 0.0001	0.7076	0.6881	0.7276	
	DBH (cm)	-0.1550 (0.0091)	< 0.0001	0.8564	0.8412	0.8719	
	Age (year)	0.0267 (0.0190)	0.1600	1.0271	0.9896	1.0661	
	Crown ratio	-4.7249 (0.2813)	< 0.0001	0.0089	0.0051	0.0154	
	Basal area (m^2 ha^{-1})	0.0795 (0.0076)	< 0.0001	1.0827	1.0667	1.0990	
	Dominant height (m)	0.3268 (0.0255)	< 0.0001	1.3865	1.3189	1.4576	
	γ	0.542 (0.0864)					
Gamma	Physio. region	-0.5911 (0.1360)	< 0.0001	0.5537	0.4241	0.7228	41514.98
	Total height (m)	-0.3492 (0.0143)	< 0.0001	0.7052	0.6857	0.7253	
	DBH (cm)	-0.1547 (0.0091)	< 0.0001	0.8567	0.8415	0.8721	
	Age (year)	0.0250 (0.0197)	0.2000	1.0253	0.9865	1.0656	
	Crown ratio	-4.7075 (0.2817)	< 0.0001	0.0090	0.0052	0.0157	
	Basal area (m^2 ha^{-1})	0.0759 (0.0074)	< 0.0001	1.0789	1.0633	1.0947	
	Dominant height (m)	0.3430 (0.0252)	< 0.0001	1.4092	1.3412	1.4805	
	θ	0.558 (0.0823)					
	Kendall's τ	0.218 (0.0380)					

Both random effects parameters (γ and θ) are significant. Random effects for plots explained a significant part of the tree-level variation in hazard rates. Frailty in lognormal model with parameterization based on $E(W) = 0$ is not standardized to $E(U) = 1$ but the estimates for regression coefficients are similar in both the frailty distributions. Variances of random effects parameters in two models can not be directly compared. The parameter γ in lognormal frailty model is the variance of the random effect W ($= \log(U)$) where as the parameter θ in gamma frailty model is the variance of the frailty U. Both the variance estimates are similar

($\gamma = 0.542$ and $\theta = 0.558$) indicating the robustness regarding the frailty distribution. Magnussen et al. (2005) used the gamma distribution for shared random effects of the stand (i.e. frailties) in Cox proportional hazards model for modeling relationship between spruce budworm defoliation and survival times of white spruce. The authors observed significant random stand effects that explained the tree-level variation in hazard rates.

The densities of gamma (with $E(U) = 1$ and $Var(U) = \theta = 0.558$) and lognormal (with $E(W) = 0$ and $Var(W) = \gamma = 0.542$) frailty were plotted in Figure 4.7. The gamma distribution had slightly higher density than the lognormal distribution upto the frailty value of about 0.6 and the trend was reverse after that frailty value.

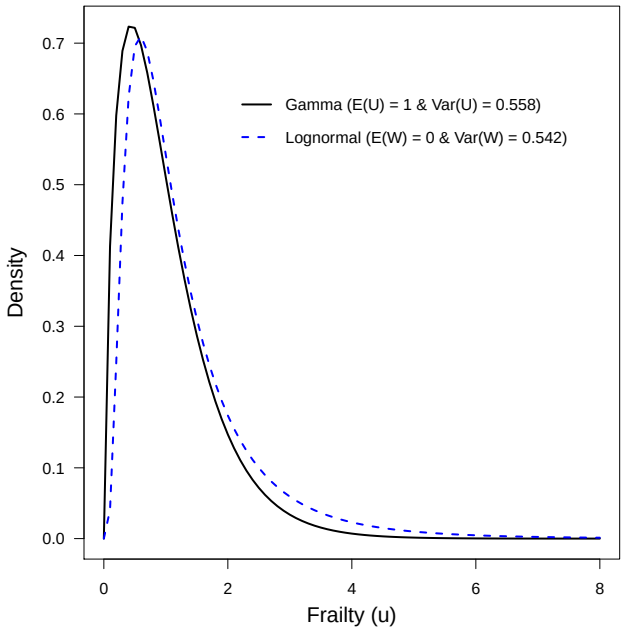


Figure 4.7: Density of gamma and lognormal frailty distribution

The gamma frailty model had smaller Akaike information criteria (AIC) and hence this distribution was selected as the distribution of frailty. Gamma frailty was chosen for the ease of results interpretation as well. As mentioned earlier, interpretation of parameter γ in lognormal frailty model is not direct but the parameter θ in gamma frailty model describes

the variance of the frailty term and it is easy to interpret. From the gamma frailty model, variance of the frailty (θ) was estimated as 0.558. The likelihood ratio test statistic for the frailty effect (i.e. H_0: $\theta = 0$ versus H_1: $\theta > 0$) was 593.21 (p-value was <0.0001) for one degree of freedom. An approximate Wald test statistic for the frailty was 990.51 (the p-value was <0.0001) on 147 degree of freedom. Both the tests indicated the significance of shared frailty effect and there appeared to be significant heterogeneity between plots in the data. The corresponding Kendall's τ (standard error) is 0.218 (0.0380) and, thus, there is on average positive correlation of about 0.22 between the tree mortality times.

Comparing the regression coefficients in Table 4.2 and Table 4.6, the estimated size of the coefficient of all the covariates except *crown ratio* and *dominant height* increased with inclusion of frailty in the model. Hence, the effect of these covariates was biased downwards when frailty effects were not taken into account. The incorporation of unobserved heterogeneity into the Cox model deattenuated the effects of the observed covariates and their standard errors. The incorporation of frailty made the effect of covariate *age* non significant (Table 4.6) suggesting that the hazard rate may not be affected by stand age. The likelihood ratio test statistic for covariate *age* was 3.22 (p-value was 0.0729). The Wald statistics for all the covariates except *age* suggested their significance on the hazard rate. The coefficients in marginal model (Table 4.3) and frailty model (Table 4.6) were not the same because the models are estimating different quantities unless the within-plot correlation is zero (Therneau and Grambsch, 2000). The marginal model estimated the population averaged relative risk and the frailty model estimated the relative risk within plots.

The confidence interval for the variance of the frailties (θ) was determined from the profile marginal likelihood (e.g. Therneau and Grambsch, 2000; Duchateau and Janssen, 2008). For the 95% confidence interval, two values of θ for which the marginal profile likelihood lies 1.92 units below the maximum profile likelihood value are taken. In Figure 4.8, the dotted

horizontal line was drawn 1.92 units below the maximum profile likelihood value and the dotted vertical lines at the intersection of the profile likelihood with the horizontal line mark the 95% confidence interval for θ which corresponds to [0.371, 0.650].

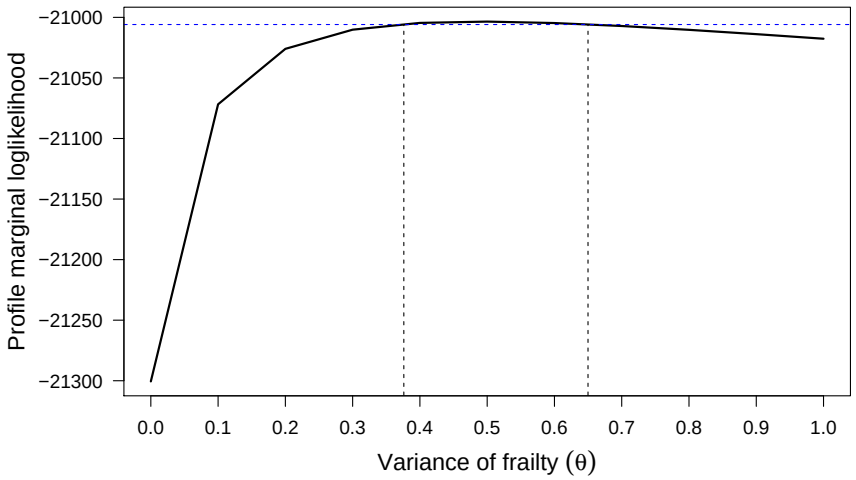


Figure 4.8: Profile marginal likelihood for θ with 95% confidence interval based on the profile marginal likelihood

Bootstrap confidence interval of θ was also calculated using 200 bootstrap samples. The 95% confidence interval was [0.414, 0.664] which is slightly narrower than the confidence interval obtained from the profile marginal likelihood above. The bootstrap distribution of θ is plotted in histogram in Figure 4.9.

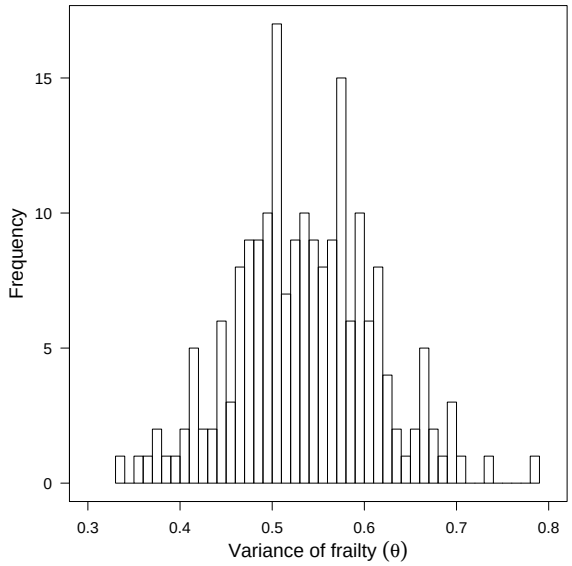


Figure 4.9: Bootstrap distribution of θ in histogram

Multiple sources of heterogeneity occur naturally in data from permanent plot systems due to the multilevel data structure inherent in the design. However, a single random effect was introduced in the model in this study when a multi-level model with random effects of plots, trees and measurement occasions would be more appropriate. The literature on estimation methods for such multi-level models has been growing recently (Magnussen et al., 2005). Permanent sample plots generate both longitudinal (repeated measurement) and survival (time to event) data. Often times such data are analyzed separately, like in this study, using well-established methods. The longitudinal variables such as DBH, crown class and stand density are correlated with tree health status and, hence, their survival endpoints. Analyzing such data separately when association exists between survival and longitudinal processes may be inappropriate and, hence, the joint modeling of longitudinal and survival data has received considerable interest lately (e.g. Henderson et al., 2000; Guo and Carlin, 2004; Tsiatis and Davidian, 2004; Li et al., 2009; Crowther et al., 2013).

4.4.2.1 Predictive ability of Cox frailty model

The predictive ability of Cox proportional hazards model with and without frailty (Table 4.2 and Table 4.6, respectively) were compared using indexes R^2 and R^2_N that are based on log likelihood of the models. R^2 and R^2_N for Cox model computed from the original data were 0.1219 and 0.1833, respectively and for Cox model with frailty were 0.1446 and 0.2149, respectively. Both the indexes of predictive ability were greater for the Cox model with frailty suggesting its greater predictive strength. R^2 and R^2_N values were also computed for both the model as a mean of estimates from 200 bootstrap samples and are given in Table 4.7. These values were greater for the Cox model with gamma frailty distribution.

Table 4.7: Bootstrap estimates of indexes for quantifying predictive ability of Cox models

Index	**Cox model**	**Cox model with frailty**
R^2	0.1228	0.1850
R^2_N	0.1450	0.2155

The predicted survival rates from the Cox model and Cox model with gamma frailty were compared for two trees in the Coastal Plain and Piedmont (Figure 4.10). Smooth curves were obtained for a young and mature loblolly pine tree in each physiographic region.

The predicted survival rates from the Cox model and Cox model with frailty were different with Cox model over predicting the survival rate for both young and mature trees in both regions. The predicted survival rate for mature tree was more closer to Kaplan-Meier curve in the Piedmont. However, Therneau and Grambsch (2000) have stated that estimation of survival probabilities in the presence of time-dependent covariates is not easy to conceptualize and they warn against using survival curves based on a time-dependent covariate with extreme caution.

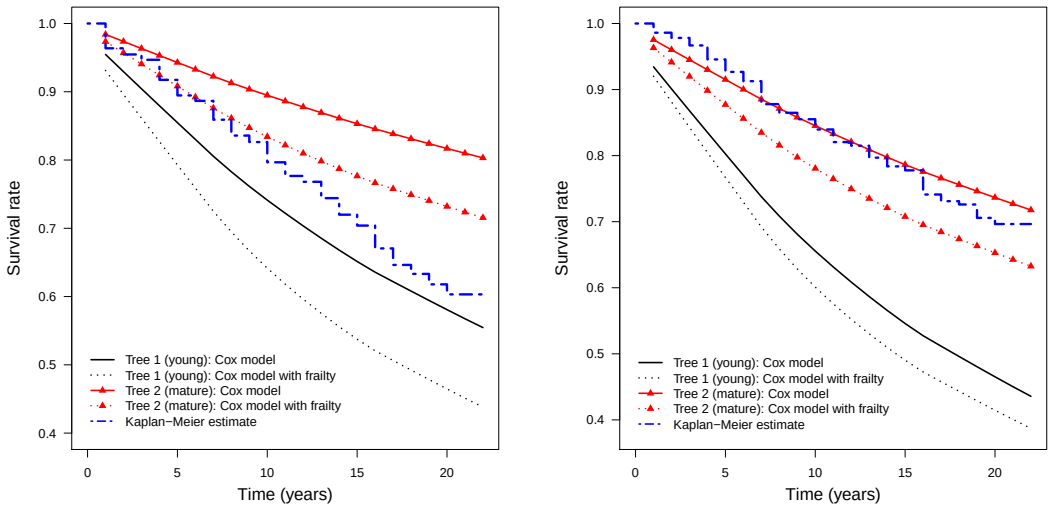


Figure 4.10: Predicted survival rates from Cox model and Cox model with frailty for young and mature loblolly pine trees in (a) Coastal Plain and (b) Piedmont.

4.5 Conclusion

Time in years since the plot establishment was used as a natural time scale for analysis of loblolly pine mortality data. The Cox proportional hazards regression was used to develop an individual tree survival model. The Kaplan-Meier survival curves, that were computed treating all trees as a cohort, suggested that the trees in the Piedmont had higher survival than the trees in the Coastal Plain. Fixed covariate physiographic region; tree-level covariates total tree height (m), DBH (cm) and crown ratio; and stand-level covariates stand age (years), basal area (m^2 ha^{-1}) and stand dominant height (m) were the significant covariates in modeling the survival of trees. Increase in total height, DBH, stand age and crown ratio lead to a reduction in relative risk of death among the surviving trees and the increase in stand age, basal area and dominant height caused the increase in the relative risk. However, the decrease or increase in the relative risk for the stand age and basal area were mild where

the hazard ratio were close to one. When the sandwich estimator, that takes into account the correlation in the data, were used for standard error, the Wald test statistic was non significant for the covariate *age* suggesting that the hazard rate may not be affected by stand age. When separate Cox models were fitted to the two physiographic regions, the stand age and stand basal were not significantly associated with an effect on survival of loblolly pine trees in the Coastal Plain and Piedmont, respectively. Gamma shared frailty model was fitted to account for the unobserved heterogeneity not explained by the covariates considered in the Cox model. The gamma frailty model had smaller AIC and hence the gamma distribution was selected as frailty distribution. Interpretation of parameter in gamma frailty model was direct and easy; this too was a reason for selecting gamma frailty model over the lognormal frailty model. The estimated size of the coefficient for most covariates increased when frailty was included in the Cox model and hence, the effect of those covariates was biased downwards when frailty effects were not taken into account. Indexes of predictive ability (R^2 and R^2_N) computed from both the original data and bootstrap samples were higher for the Cox model with frailty suggesting the greater predictive strength of the shared gamma frailty model as compared to the regular Cox model.

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Annals of statistics*, 6(4):701–726.

Affleck, D. L. R. (2007). Mixed and modified poisson models for the analysis of stand-level mortality. *Canadian Journal of Forest Research*, 36(11):2994–3006.

Avila, O. B. and Burkhart, H. E. (1992). Modeling survival of loblolly pine trees in thinned and unthinned plantations. *Canadian Journal of Forest Research*, 22(12):1878–1882.

Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics*, 30(1):89–99.

Buchman, R. G. (1979). Mortality functions. In *A generalized forest growth projection system applied to the Lake States region*, number NC-49, pages 47–55. USDA For. Ser.

Burkhart, H. E. and Tomé, M. (2012). *Modeling Forest Trees and Stands*. Springer.

Clutter, J. L., Fortson, J. C., Pienaar, L. V., Brister, G. H., and Bailey, R. L. (1983). *Timber management: a quantitative approach*. John Wiley & Sons Inc, New York, NY.

Clutter, J. L. and Jones, E. P. (1980). Prediction of growth after thinning in old field slash pine plantations. Gen. Tech. Rep SE-27, USDA For. Ser.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, 34(2):187–220.

Crowther, M. J., Abrams, K. R., and Lambert, P. C. (2013). Joint modeling of longitudinal and survival data. *Stata Journal*, 13(1):165–184.

Diéguez-Aranda, U., Castedo-Dorado, F., Álvarez-González, J. G., and Rodŕiguez-Soalleiro, R. (2005). Modelling mortality of Scots pine (*Pinus sylvestris* L.) plantations in the northwest of Spain. *European Journal of Forest Research*, 124(2):143–153.

Dobbertin, M. and Brang, P. (2001). Crown defoliation improves tree mortality models. *Forest Ecology and Management*, 141(3):271 – 284.

Duchateau, L. and Janssen, P. (2008). *The Frailty Model*. Springer, New York.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72(359):557–565.

Efron, B. and Tibshirani, R. (1993). *An introduction to the boostrap*. Chapman and Hall, New York.

Eid, T. and Øyen, B. H. (2003). Models for prediction of mortality in even-aged forest. *Scandinavian Journal of Forest Research*, 18(1):64–77.

Fox, T. R., Allen, H. L., Albaugh, T. J., Rubilar, R., and Carlson, C. (2007). Tree nutrition and forest fertilization of pine plantations in the southern United States. *Southern Journal of Applied Forestry*, 31(1):5–11.

Guo, X. and Carlin, B. P. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, 58(1):16–24.

Hamilton, D. A. (1974). Event probabilities estimated by regression. Res. Pap. INT-152, USDA For. Ser.

Hamilton, Jr., D. A. (1986). A logistic model of mortality in thinned and unthinned mixed conifer stands of Northern Idaho. *Forest Science*, 32(4):989–1000.

Harrell, Jr., F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer-Verlag, New York.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480.

Hougaard, P. (1999). Fundamentals of survival data. *Biometrics*, 55(1):13–22.

Jokela, E. J., Martin, T. A., and Vogel, J. G. (2010). Twenty-five years of intensive forest management with southern pines: Important lessons learned. *Journal of Forestry*, 108(7):338–347.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data*. Springer-Verlag, New York, NY.

Lee, Y. (1971). Predicting mortality for even-aged stands of lodgepole pine. *Forestry Chornicle*, 47(1):29–32.

Li, L., Hu, B., and Greene, T. (2009). A semiparametric joint model for longitudinal and survival data with application to hemodialysis study. *Biometrics*, 65(3):737–745.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Magnussen, S., Alfaro, R. I., and Boudewyn, P. (2005). Survival-time analysis of white spruce during spruce budworm defoliation. *Silva Fennica*, 39(2):177–189.

Monserud, R. A. (1976). Simulation of forest tree mortality. *Forest Science*, 22(4):438–444.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–965.

Rose, E. C., Hall, D. B., Shiver, B. D., Clutter, M. L., and Borders, B. (2006). A multilevel approach to individual tree survival prediction. *Forest Science*, 52(1):31–43.

Schultz, R. P. (1997). *Loblolly pine: the ecology and culture of loblolly pine (*Pinus taeda *L.)*, volume Agricultural Handbook 713. USDA Forest Service, Washington, D.C.

Somers, G. L., Oderwald, R. G., Harms, W. R., and Langdon, G. O. (1980). Predicting mortality with a Weibull distribution. *Forest Science*, 26(2):291–300.

Staebler, G. R. (1953). Mortality estimation in fully stocked stands of young-growth Douglas-fir. Res. Pap. 4, USDA For. Serv., Pacific Northwest Forest and Range Experiment Station.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834.

Uzoh, F. C. and Mori, S. R. (2012). Applying survival analysis to managed even-aged stands of ponderosa pine for assessment of tree mortality in the western United States. *Forest Ecology and Management*, 285(0):101 – 122.

Vanclay, J. K. (1995). Synthesis: growth models for tropical forests: a synthesis of models and methods. *Forest Science*, 41(1):7–42.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity on individual frailty on the dynamic of mortality. *Demography*, 16(3):439–454.

Woodall, C. W., Grambschb, P. L., and Thomas, W. (2005). Applying survival analysis to a large-scale forest inventory for assessment of tree mortality in Minnesota. *Ecological Modelling*, 189(1-2):199–208.

Woollons, R. C. (1998). Even-aged stand mortality estimation through a two-step regression process. *Forest Ecology and Management*, 105(1):189–195.

Yao, X., Titus, S. J., and MacDonald, S. E. (2001). A generalized logistic model of individual tree mortality for aspen, white spruce, and lodgepole pine in Alberta mixedwood forests. *Canadian Journal of Forest Research*, 31(2):283–291.

Zhao, D., Borders, B., Wang, M., and Kane, M. (2007). Modeling mortality of second-rotation loblolly pine plantations in the Piedmont/Upper Coastal Plain and Lower Coastal Plain of the southern United States. *Forest Ecology and Management*, 252(1-3):132–143.

Chapter 5

Summary and Recommendations

The overall goal of this research was to model loblolly pine trees mortality at both stand- and tree-level resolutions. Mortality being the most important yet least understood component of the forest growth and yield system, this research explored ways to improve the existing mortality models and attempted to borrow methods used in other areas (e.g. biomedical science) to be used in forestry. The data used in the analyses came from the remeasurement data collected from permanent sample plots established in 1980/81 across the natural range of loblolly pine in the Atlantic Coastal Plain, Gulf Coastal Plain and Piedmont. The biophysical data, used in modeling stand-level mortality, were obtained from the Oak Ridge National Laboratory, Distributed Active Archive Center.

The stand-level mortality models were constructed using two modeling approaches. In the first approach, mortality functions for directly predicting tree number reduction were developed using algebraic difference equation method called direction prediction method here. In the second approach, a two-step modeling strategy was used. In the first step, a model predicting the probability of tree death occurring over a measurement period was developed and in the second step, a function that estimates the reduction in tree number was devel-

oped. Little or no work has been done towards using climate and soil data in improving estimation of stand-level mortality. In both of the modeling approaches, the potential of using biophysical variables to refine the mortality estimates was explored. Three factors extracted from explanatory factor analysis were used as surrogates for the climate and soil information in the analyses.

In the direct prediction approach, one of the parameters of the selected model was expressed as a linear function of Heat (HI) and Drought Index (DI). In the two-step modeling approach, the biophysical variables were used in the first step where the model for predicting stand mortality probability was developed. The marginal GEE model was selected. Decision theory based method performed the best in leave-one-cluster-out cross-validation, with this model having the smallest RMSE, MAE and AIC. When the biophysical variables were incorporated in the direct prediction model, this model outperformed the decision theory based approach by exhibiting the smallest RMSE, MAE and AIC.

Tree-level mortality models were developed using multilevel mixed-effects logistic regression taking into account the full hierarchical structure of the data. The data from permanent sample plot systems are hierarchical in nature i.e. repeated measurements on each tree are nested within the tree and trees are nested within a plot. Such hierarchy, which is often ignored completely or partially in modeling tree mortality in forestry applications, was fully taken into account in this study. The likelihood ratio tests showed that three-level logistic model that accounted for full hierarchical structure of the data fitted the data significantly better that the models that either completely or partially ignored the hierarchical structure. The three-level logistic model also had the lowest AIC and the significant variables were physiographic region, DBH, total tree height, stand age and stand basal area. This model had the highest area under the ROC curve in both model fit and cross validation data. The indexes and statistics in the calibration plots for the cross-validation data were larger for the

three-level logistic model indicating its better performance and higher predictive strength among all models considered. Prediction bias across the range of different predictors was in general lower for this model.

The logistic mortality models do not directly account for changes that occur in the time-dependent covariates over time. It is common to encounter tree- and stand-level covariates that vary over time and using the covariate information that changes over time in the mortality model seems appropriate. Survival analysis techniques incorporate dynamic variables in the survival model by using time-dependent covariates. Survival analysis techniques are commonly used in biomedical areas but occasionally used in forestry applications. These methods allow for censoring of observations, inclusion of time-dependent covariates and dealing with non-normal distribution among other advantages. This study modeled regular mortality of individual trees using Cox proportional hazards regression and accounted for the time-dependent nature of individual tree- and stand-level covariates. Survival time of trees in a plot may be correlated due to clustering effect and the past survival analysis studies of individual tree mortality has never addressed this issue of correlation. This study used the shared frailty model, which is a mixed model extension of the Cox proportional hazards model, to study the clustered survival time.

Cox proportional hazards regression was fitted and the covariates physiographic region, DBH, total height, crown ratio, stand age, stand basal area and dominant height were significantly associated with the survival of trees. From the estimated Kaplan-Meier survival function, trees in the Coastal Plain were at higher risk of death than the trees in the Piedmont. The log of relative hazard decreased with DBH, total height and crown ratio and increased with dominant height, stand age and basal area. When the correlation in the data was taken into account by using robust sandwich estimator of variance, the covariate stand age was non significant suggesting the hazard rate may not be affected by stand age. The semiparametric

frailty models were fitted by partial penalized likelihood approach. The gamma frailty model had smaller AIC and hence the gamma distribution was selected as frailty distribution of the final model. Direct and easy interpretation of parameters in gamma frailty model was another reason for its selection. When the frailty effects were not taken into account, the estimated size of covariates coefficients were biased downward. The indexes of predictive ability of Cox model with frailty were greater than the Cox model without frailty suggesting greater predictive strength of the frailty model.

This study contributes to the greater body of knowledge of stand- and tree-level mortality modeling. Mortality being the most important yet least understood component of the forest growth and yield system, we are always looking for ways to improve the mortality prediction. This study opens up opportunities for further exploring the use of climate and soil information in refining mortality models. Often times during research, results obtained suggest avenues for future research. Single factor analysis of the climate and soil data was done for the study period in available data (i.e. 21 years). Factor analysis for each measurement periods could be done to examine the effects of changes of climate variables over the time. It was thought that the 21 years periods may not be sufficient to catch climate signal and it was judged adequate to use the average trend for climate variable over the time frame of observation for this analysis.

Mortality being a discreet binary event, logistic regression may be appropriate to model tree mortality. However, time to event data, when available, provides more information about tree survival. This study contributes to that aspect of forest growth and yield modeling. No research on modeling the spatial correlation among survival rates of trees at different locations has been reported. Trees in forest stands that are close in distance may be similar in terms of survival characteristics because of sharing similar environmental conditions. Future work could be done on modeling spatially correlated trees survival. Joint modeling

of longitudinal and survival data from permanent sample plots could be explored since an association exists between survival and longitudinal processes in such data.

Appendix A

Model for predicting tree number reduction for Coastal Plain and Piedmont

A.1 Mortality function for direct prediction approach obtained by fitting to data from all plots

Table A.1: Parameter estimates and fit statistics of models fitted with all plots data for direct prediction of the tree number reduction for Coastal Plain and Piedmont

Model	Equation	Parameter	Coastal Plain					Piedmont				
			Estimate (SE)	Pr > |t|	RMSE	MAE	AIC	Estimate (SE)	Pr > |t|	RMSE	MAE	AIC
M1	$N_2 = N_1\left(b_1(A_2 - A_1)\right)$	b_1	-0.0253 (0.0011)	< .0001	79.54	58.64	5311.44	-0.021 (0.0012)	< .0001	77.99	56.36	3399.00
M2	$N_2 = N_1 \exp\left(b_1\left(A_2^{b_2} - A_1^{b_2}\right)\right)$	b_1	-0.0002 (0.0001)	0.0659	72.37	52.62	5225.90	-0.0002 (0.0001)	0.1963	72.56	51.73	3357.54
		b_2	2.2939 (0.1480)	< .0001				2.2901 (0.2104)	< .0001			
M3	$N_2 = N_1 \exp\left(b_1(A_2 - A_1)\right)\left(\frac{A_2}{A_1}\right)^{b_2}$	b_1	-0.0556 (0.0033)	< .0001	72.50	52.37	5227.48	-0.0492 (0.0040)	< .0001	71.94	50.62	3352.51
		b_2	0.5977 (0.0610)	< .0001				0.5868 (0.0803)	< .0001			
M4	$N_2 = N_1 \exp\left(b_1\left(b_2^{A_2} - b_2^{A_1}\right)\right)$	b_1	-0.1522 (0.0402)	0.0002	73.07	53.68	5234.70	-0.1362 (0.0509)	0.0079	73.23	52.45	3362.99
		b_2	1.0531 (0.0061)	< .0001				1.0507 (0.0085)	< .0001			
M5	$N_2 = \left[N_1^{b_0} + b_1(A_2 - A_1)\right]^{\frac{1}{b_0}}$	b_0	0.4971 (0.1554)	0.0015	78.81	59.08	5303.95	0.2882 (0.2219)	0.1951	77.91	56.67	3399.37
		b_1	-0.4332 (0.6113)	0.4789				-0.0489 (0.1161)	0.6738			
M6	$N_2 = \left[N_1^{b_0} + b_1\left(A_2^{b_2} - A_1^{b_2}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.4141 (0.1959)	0.0351	72.08	51.62	5223.16	-0.7089 (0.2695)	0.0090	71.78	49.63	3352.17
		b_1	1.9 $\times 10^{-6}$[a]	0.4979				1.3 $\times 10^{-7}$[a]	0.6814			
		b_2	2.5315 (0.1978)	< .0001				2.7727 (0.3024)	< .0001			
M7	$N_2 = \left[N_1^{b_0} + b_1(A_2 - A_1) + b_2 \ln\left(\frac{A_2}{A_1}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.0479 (0.1582)	0.7623	72.57	52.31	5229.38	-0.3085 (0.2151)	0.1525	71.81	49.89	3352.44
		b_1	0.0019 (0.0042)	0.6500				0.0017 (0.0014)	0.2335			
		b_2	-0.0205 (0.0455)	0.6525				-0.0207 (0.0171)	0.2267			
M8	$N_2 = \left[N_1^{b_0} + b_1\left(b_2^{A_2} - b_2^{A_1}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.4081 (0.1997)	0.0416	72.80	52.76	5232.26	-0.7463 (0.2794)	0.0080	72.42	50.37	3357.41
		b_1	0.0023 (0.0027)	0.3951				0.0002 (0.0004)	0.6204			
		b_2	1.0639 (0.0085)	< .0001				1.0733 (0.0130)	< .0001			
M9	$N_2 = N_1 \exp\left(b_1 S^{b_2}(A_2 - A_1)\right)$	b_1	-0.0012 (0.0011)	0.2745	78.73	58.41	5303.05	-0.0128 (0.0198)	0.5197	78.11	56.41	3400.90
		b_2	1.0756 (0.3160)	0.0007				0.1769 (0.5518)	0.7487			

[a]SE is smaller than 1×10^{-6}

Model	Equation	Parameter	Coastal Plain					Piedmont				
			Estimate (SE)	Pr > |t|	RMSE	MAE	AIC	Estimate (SE)	Pr > |t|	RMSE	MAE	AIC
M10	$N_2 = N_1 \exp\left(b_1 \left(\frac{S}{10000}\right)^{b_2} \left(A_2^{b_3} - A_1^{b_3}\right)\right)$	b_1	-0.6533 (1.0906)	0.5494	70.63	52.10	5204.60	-0.0009 (0.0031)	0.7584	72.64	51.80	3359.24
		b_2	1.3104 (0.2632)	< .0001				0.2713 (0.4944)	0.5836			
		b_3	2.3667 (0.1469)	< .0001				2.2880 (0.2100)	< .0001			
M11	$N_2 = N_1 \exp\left(b_1 S^{b_2} (A_2 - A_1)\right) \left(\frac{A_2}{A_1}\right)^{b_3}$	b_1	-0.0103 (0.0038)	0.0069	70.83	51.47	5207.14	-0.030 (0.0185)	0.1032	71.98	50.72	3353.85
		b_2	0.6002 (0.1243)	< .0001				0.1754 (0.2158)	0.4171			
		b_3	0.6330 (0.0601)	< .0001				0.5911 (0.0806)	< .0001			
M12	$N_2 = N_1 \exp\left(b_1 S^{b_2} \left(b_3^{A_2} - b_3^{A_1}\right)\right)$	b_1	-0.0030 (0.0026)	0.2432	71.34	53.14	5213.69	-0.0781 (0.1141)	0.4939	73.34	52.53	3364.83
		b_2	1.3173 (0.2672)	< .0001				0.1985 (0.5019)	0.6928			
		b_3	1.0572 (0.0062)	< .0001				1.0506 (0.0085)	< .0001			
M13	$N_2 = \left[N_1^{b_0} + b_1 S^{b_2} (A_2 - A_1)\right]^{\frac{1}{b_0}}$	b_0	0.4057 (0.1537)	0.0086	78.22	58.91	5298.07	0.3078 (0.2350)	0.1913	78.03	56.65	3401.33
		b_1	-0.0137 (0.0259)	0.5966				-0.0845 (0.2829)	0.7654			
		b_2	0.9080 (0.3116)	0.0037				-0.1208 (0.5704)	0.8324			
M14	$N_2 = \left[N_1^{b_0} + b_1 \left(\frac{S}{10000}\right)^{b_2} \left(A_2^{b_3} - A_1^{b_3}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.9166 (0.2183)	< .0001	69.14	49.39	5186.05	-0.9183 (0.2942)	0.0020	71.50	49.36	3350.89
		b_1	0.0056 (0.0110)	0.6118				1.8×10^{-5} (0.0001)	0.7820			
		b_2	1.9465 (0.3123)	< .0001				1.0345 (0.5702)	0.0707			
		b_3	2.9983 (0.2295)	< .0001				2.8873 (0.325)	< .0001			
M15	$N_2 = \left[N_1^{b_0} + b_1 S^{b_2} (A_2 - A_1) + b_3 \ln\left(\frac{A_2}{A_1}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.2434 (0.1567)	0.1210	70.70	50.99	5206.53	Hessian is singular				
		b_1	0.00041 (0.0002)	0.1066								
		b_2	0.6361 (0.1221)	< .0001								
		b_3	-0.0286 (0.0127)	0.0248								
M16	$N_2 = \left[N_1^{b_0} + b_1 \left(\frac{S}{10000}\right)^{b_2} \left(b_3^{A_2} - b_3^{A_1}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.8649 (0.2182)	< .0001	69.98	50.50	5197.17	-0.9507 (0.3057)	0.0021	72.20	50.17	3356.61
		b_1	15.559 (28.549)	0.5860				0.0249 (0.0867)	0.7743			
		b_2	1.900 (0.3155)	< .0001				0.9782 (0.5807)	0.0931			
		b_3	1.0853 (0.0102)	< .0001				1.0782 (0.0135)	< .0001			

A.2 Mortality function for use in two-step approach obtained by fitting to data from plots with occurrence of mortality

Table A.2: Parameter estimates and fit statistics of models fitted with plots data with occurrence of mortality for prediction of the tree number reduction in two-step approach for Coastal Plain and Piedmont

Model	Equation	Parameter	Coastal Plain					Piedmont				
			Estimate (SE)	Pr > |t|	RMSE	MAE	AIC	Estimate (SE)	Pr > |t|	RMSE	MAE	AIC
M1	$N_2 = N_1 \exp\left(b_1(A_2 - A_1)\right)$	b_1	-0.0285 (0.0012)	< .0001	78.60	56.73	4629.64	-0.0251 (0.0012)	< .0001	75.35	53.01	2839.10
M2	$N_2 = N_1 \exp\left(b_1\left(A_2^{b_2} - A_1^{b_2}\right)\right)$	b_1	-0.0005 (0.0002)	0.0611	72.63	52.10	4567.50	-0.0006 (0.0005)	0.1944	72.10	50.47	2818.28
		b_2	2.0885 (0.1440)	< .0001				1.9560 (0.2061)	< .0001			
M3	$N_2 = N_1 \exp\left(b_1(A_2 - A_1)\right)\left(\frac{A_2}{A_1}\right)^{b_2}$	b_1	-0.0563 (0.0035)	< .0001	72.70	52.16	4568.19	-0.0499 (0.0051)	< .0001	71.87	50.28	2816.74
		b_2	0.5661 (0.0677)	< .0001				0.5415 (0.1069)	< .0001			
M4	$N_2 = N_1 \exp\left(b_1\left(b_2^{A_2} - b_2^{A_1}\right)\right)$	b_1	-0.2370 (0.0662)	0.0004	73.18	52.89	4573.56	-0.2840 (0.1199)	0.0186	72.40	50.77	2820.39
		b_2	1.0449 (0.0060)	< .0001				1.0376 (0.0083)	< .0001			
M5	$N_2 = \left[N_1^{b_0} + b_1(A_2 - A_1)\right]^{\frac{1}{b_0}}$	b_0	0.6557 (0.1386)	< .0001	76.75	56.91	4611.57	0.3041 (0.2008)	0.1312	75.18	53.72	2838.98
		b_1	-1.9864 (2.3612)	0.4007				-0.0689 (0.1453)	0.6358			
M6	$N_2 = \left[N_1^{b_0} + b_1\left(A_2^{b_2} - A_1^{b_2}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.0727 (0.1871)	0.6976	72.71	51.93	4569.36	-0.3926 (0.2516)	0.1200	71.89	49.36	2817.84
		b_1	1.7×10^{-5} [a]	0.2959				5.5×10^{-6} [a]	0.6193			
		b_2	2.1339 (0.1914)	< .0001				2.2268 (0.2823)	< .0001			
M7	$N_2 = \left[N_1^{b_0} + b_1(A_2 - A_1) + b_2 \ln\left(\frac{A_2}{A_1}\right)\right]^{\frac{1}{b_0}}$	b_0	0.1217 (0.1566)	0.4373	72.73	52.30	4569.59	-0.2211 (0.2209)	0.3179	71.87	49.71	2817.74
		b_1	-0.0160 (0.0378)	0.6729				0.0023 (0.0013)	0.0737			
		b_2	0.1577 (0.3689)	0.6693				-0.0266 (0.0142)	0.0618			
M8	$N_2 = \left[N_1^{b_0} + b_1\left(b_2^{A_2} - b_2^{A_1}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.0361 (0.1890)	0.8486	73.27	52.81	4575.52	-0.4101 (0.2580)	0.1133	72.18	49.56	2819.871
		b_1	0.0063 (0.0232)	0.7848				0.0034 (0.0056)	0.5416			
		b_2	1.0458 (0.0082)	< .0001				1.0498 (0.0119)	< .0001			
M9	$N_2 = N_1 \exp\left(b_1 S^{b_2}(A_2 - A_1)\right)$	b_1	-0.0019 (0.0017)	0.2418	77.82	56.53	4622.67	-0.0072 (0.0101)	0.4736	75.38	53.40	2840.29
		b_2	0.9375 (0.2952)	0.0016				0.4432 (0.4952)	0.3717			

[a] SE is smaller than 1×10^{-4}

Model	Equation	Parameter	Coastal Plain					Piedmont				
			Estimate (SE)	Pr > |t|	RMSE	MAE	AIC	Estimate (SE)	Pr > |t|	RMSE	MAE	AIC
M10	$N_2 = N_1 \exp\left(b_1 \left(\frac{S}{10000}\right)^{b_2} \left(A_2^{b_3} - A_1^{b_3}\right)\right)$	b_1	-0.8091 (1.3241)	0.5415	70.90	51.74	4549.18	-0.0136 (0.0417)	0.7447	72.09	50.51	2819.22
		b_2	1.2205 (0.2592)	< .0001				0.4756 (0.4666)	0.3090			
		b_3	2.1707 (0.1437)	< .0001				1.9528 (0.2051)	< .0001			
M11	$N_2 = N_1 \exp\left(b_1 S^{b_2} (A_2 - A_1)\right) \left(\frac{A_2}{A_1}\right)^{b_3}$	b_1	-0.0109 (0.0042)	0.0103	71.09	51.40	4551.30	-0.0233 (0.0158)	0.1400	71.82	50.31	2817.38
		b_2	0.5867 (0.1311)	< .0001				0.2726 (0.2354)	0.2478			
		b_3	0.6047 (0.0668)	< .0001				0.5469 (0.1069)	< .0001			
M12	$N_2 = N_1 \exp\left(b_1 S^{b_2} \left(b_3^{A_2} - b_3^{A_1}\right)\right)$	b_1	-0.0059 (0.0050)	0.2402	71.45	52.58	4555.44	-0.0847 (0.1178)	0.4728	72.42	50.87	2821.51
		b_2	1.2231 (0.2619)	< .0001				0.4364 (0.4695)	0.3536			
		b_3	1.0492 (0.0061)	< .0001				1.0373 (0.0083)	< .0001			
M13	$N_2 = \left[N_1^{b_0} + b_1 S^{b_2} (A_2 - A_1)\right]^{\frac{1}{b_0}}$	b_0	0.5747 (0.1399)	< .0001	76.36	56.99	4608.50	0.2730 (0.2119)	0.1989	75.31	53.80	2840.79
		b_1	-0.1442 (0.2363)	0.5421				-0.0266 (0.0814)	0.7444			
		b_2	0.6690 (0.2836)	0.0188				0.2210 (0.5083)	0.6641			
M14	$N_2 = \left[N_1^{b_0} + b_1 \left(\frac{S}{10000}\right)^{b_2} \left(A_2^{b_3} - A_1^{b_3}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.5236 (0.2113)	0.0136	70.41	50.39	4544.66	-0.5797 (0.2745)	0.0357	71.57	49.11	2816.69
		b_1	0.0279 (0.0515)	0.5873				0.0005 (0.0018)	0.7630			
		b_2	1.5934 (0.3045)	< .0001				0.9312 (0.5314)	0.0810			
		b_3	2.5556 (0.2226)	< .0001				2.3422 (0.2929)	< .0001			
M15	$N_2 = \left[N_1^{b_0} + b_1 S^{b_2} (A_2 - A_1) + b_3 \ln\left(\frac{A_2}{A_1}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.0845 (0.1577)	0.5921	71.15	51.20	4553.01	-0.3441 (0.2343)	0.1432	71.65	49.59	2817.19
		b_1	0.0005 (0.0004)	0.2147				0.0005 (0.0007)	0.4450			
		b_2	0.6004 (0.1315)	< .0001				0.3712 (0.2302)	0.1082			
		b_3	-0.0288 (0.0232)	0.2158				-0.0179 (0.0168)	0.2869			
M16	$N_2 = \left[N_1^{b_0} + b_1 \left(\frac{S}{10000}\right)^{b_2} \left(b_3^{A_2} - b_3^{A_1}\right)\right]^{\frac{1}{b_0}}$	b_0	-0.4532 (0.2107)	0.0321	71.10	51.45	4552.44	-0.5891 (0.2815)	0.0375	71.92	49.41	2819.08
		b_1	33.781 (57.952)	0.5603				0.3202 (1.0230)	0.7546			
		b_2	1.5371 (0.3065)	< .0001				0.8835 (0.5349)	0.1000			
		b_3	1.0645 (0.0098)	< .0001				1.0544 (0.0124)	< .0001			

Appendix B

Exploratory factor analysis of climate and soil data for Coastal Plain and Piedmont

B.1 Variables loaded in each factor with their corresponding loadings for Coastal Plain

Table B.1: Obliquely rotated factor loadings for Coastal Plain

Variable	Factor1	Factor2	Factor3
Summer Growing Degree Days (5^0C baseline)	**0.89110**	-0.01470	0.23318
Summer Mean Maximum Temperature (^{0}C)	**0.84332**	-0.23867	0.32934
Growing Season Growing Degree Days (5^0C baseline)	**0.84302**	0.28084	0.10701
Annual Growing Degree Days (5^0C baseline)	**0.82225**	0.29463	0.17766
Mean Annual Temperature (^{0}C)	**0.81451**	0.30943	0.17839
Annual maximum temperature (^{0}C)	**0.81051**	0.22791	0.24133
Annual minimum temperature (^{0}C)	**0.78650**	0.38535	0.10541
July Mean Maximum Temperature (^{0}C)	**0.78101**	-0.34809	0.27817
Length of Growing Season (days)	**0.74950**	0.36583	-0.02104
January Mean Maximum Temperature (^{0}C)	**0.71257**	0.36422	0.15224
Mean Growing Season Temperature (^{0}C)	**0.70543**	0.01538	0.34885
Summer Precipitation (mm)	-0.04736	**0.89423**	-0.15046
Growing Season Precipitation (mm)	0.30443	**0.82906**	0.11213
Number of Days in Growing Season with Precipitation $\geq$ 13mm	0.33730	**0.79378**	0.12620
Annual Precipitation (mm)	0.11255	**0.72751**	0.36417
Summer Dryness Index	0.28380	**-0.81443**	0.24090
Growing Season Dryness Index	0.26611	**-0.87579**	-0.07853
Percent silt	-0.14750	-0.01993	**0.89074**
Soil Available Water Storage Capacity 0 to 150 cm	-0.13659	0.12102	**0.67884**
Percent Clay	0.07555	-0.06944	**0.63646**
Percent Sand	0.26627	0.09894	**-0.90284**
Eigenvalues	11.1198	4.0579	2.0487
Variance explained by each factor	11.1198	4.0579	2.0487
Number of variables	11	6	4

B.2 Variables loaded in each factor with their corresponding loadings for Piedmont

Table B.2: Obliquely rotated factor loadings for Piedmont

Variable	Factor1	Factor2	Factor3
Annual minimum temperature(^{0}C)	**0.84678**	0.13262	-0.24941
Mean Annual Temperature(^{0}C)	**0.84278**	0.06500	-0.28271
Annual Growing Degree Days(5^{0}C baseline)	**0.84222**	0.03276	-0.27996
Growing Season Growing Degree Days(5^{0}C baseline)	**0.83066**	0.05154	-0.25572
Annual maximum temperature(^{0}C)	**0.81648**	-0.00434	-0.30980
Summer Growing Degree Days(5^{0}C baseline)	**0.78496**	-0.16718	-0.31241
Length of Growing Season (days)	**0.70227**	0.22397	-0.16429
Summer Mean Maximum Temperature(^{0}C)	**0.69188**	-0.37312	-0.31327
January Mean Maximum Temperature(^{0}C)	**0.68123**	0.27163	-0.21832
Mean Growing Season Temperature(^{0}C)	**0.64437**	-0.24270	-0.30046
July Mean Maximum Temperature(^{0}C)	**0.61031**	-0.41239	-0.26450
Growing Season Precipitation (mm)	0.22643	**0.91385**	-0.03351
Number of Days in Growing Season with Precipitation $\geq$ 13mm	0.30364	**0.84238**	-0.12428
Summer Precipitation (mm)	0.06502	**0.81352**	-0.09828
Annual Precipitation (mm)	0.11166	**0.79448**	0.03249
Summer Dryness Index	0.25123	**-0.73809**	0.02975
Growing Season Dryness Index	0.30848	**-0.81911**	-0.10190
Percent Clay	0.15162	-0.07205	**0.97441**
Soil Available Water Storage Capacity 0 to 150 cm	0.29523	-0.08418	**0.95580**
Percent Silt	0.05586	0.05619	**0.83049**
Percent Sand	0.06722	-0.01292	**-0.60820**
Eigenvalues	10.3606	4.6262	1.7454
Variance explained by each factor	11.1198	4.0579	2.0487
Number of variables	11	6	4

B.3 Kernel density plots of three factors obtained by EFA for Coastal Plain and Piedmont

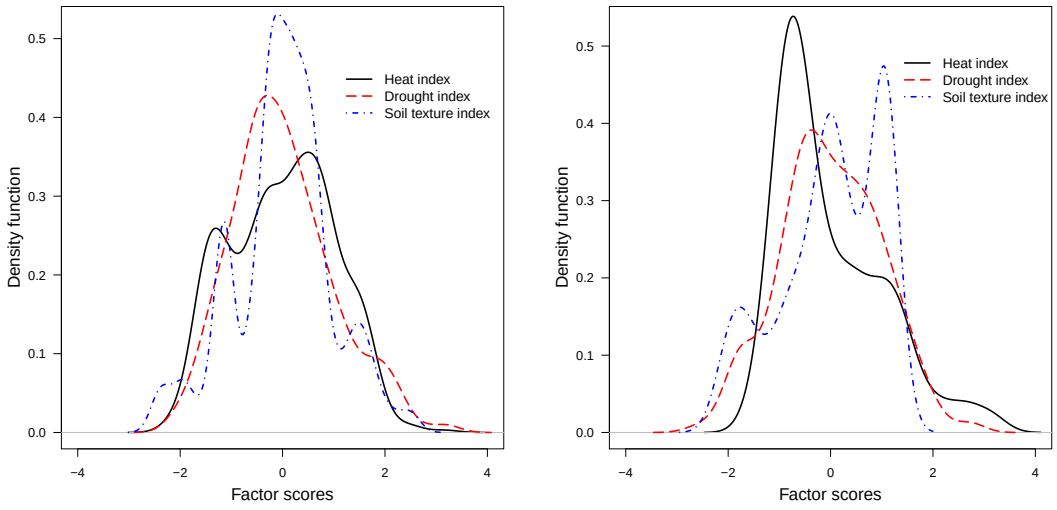


Figure B.1: Kernal density plots of three factors for (a) Coastal Plain and (b) Piedmont

Appendix C

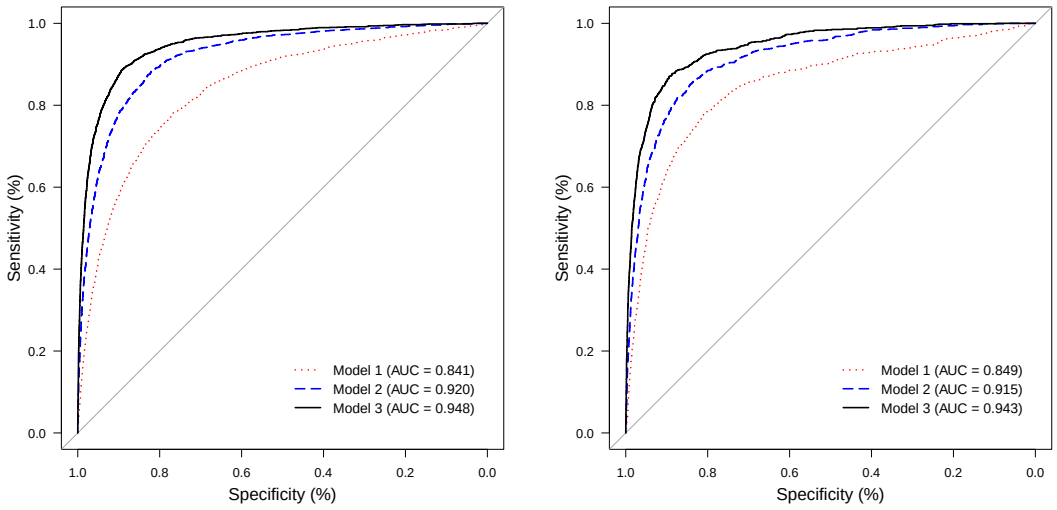
Bootstrap estimates of parameter and 95% confidence interval

Table C.1: Parameter estimates and confidence interval of fixed effects and different random-effects parameters

Parameters	**Model 2**	**Model 3**
β_0: Intercept	-4.2745 [-5.4247, -3.2317]	-4.2132 [-5.1975, -3.4096]
β_1: PR	-1.0792 [-1.5698, -0.5012]	-1.2589 [-1.8752, -0.6530]
β_2: DBH	-0.3181 [-0.3529, -0.2815]	-0.2637 [-0.3073, -0.2161]
β_3: TH	-0.6618 [-0.6987, -0.6285]	-0.7835 [-0.8465, -0.7200]
β_4: A	0.0542 [-0.0226, 0.1416]	0.0635 [0.0110, 0.1234]
β_5: BAP	0.3979 [0.3597, 0.4332]	0.4146 [0.3830, 0.4474]
β_6: DBH × TH	0.0084 [0.0064, 0.0103]	0.0040 [0.0011, 0.0068]
β_7: TH × A	-4.1 $\times 10^{-5}$ [-0.0017, 0.0019]	0.0031 [0.0001, 0.0056]
Plot-level		
$\sigma^2_{\upsilon_0}$: Int. variance	10.6911 [8.4748, 12.8150]	12.6265 [10.0375, 14.6702]
$\sigma^2_{\upsilon_0\upsilon_5}$: Int.-slope covar.	-0.5223 [-0.6685, -0.3944]	-0.6171 [-0.7471, -0.4786]
$\sigma^2_{\upsilon_5}$: Slope variance	0.0340 [0.0260, 0.0439]	0.0398 [0.0308, 0.0501]
Tree-level		
$\sigma^2_{\upsilon_0}$: Int. variance		0.3465 [0.1010, 0.7074]
$\sigma^2_{\upsilon_0\upsilon_5}$: Int.-slope covar.		-0.0332 [-0.0568, -0.0146]
$\sigma^2_{\upsilon_5}$: Slope variance		0.0032 [0.0022, 0.0047]

Appendix D

D.1 ROC curves for two physiographic regions



(a) (b)

Figure D.1: ROC curve and AUC of three models for (a) Coastal Plain and (b) Piedmont

D.2 Predicted probability of mortality for two physiographic regions

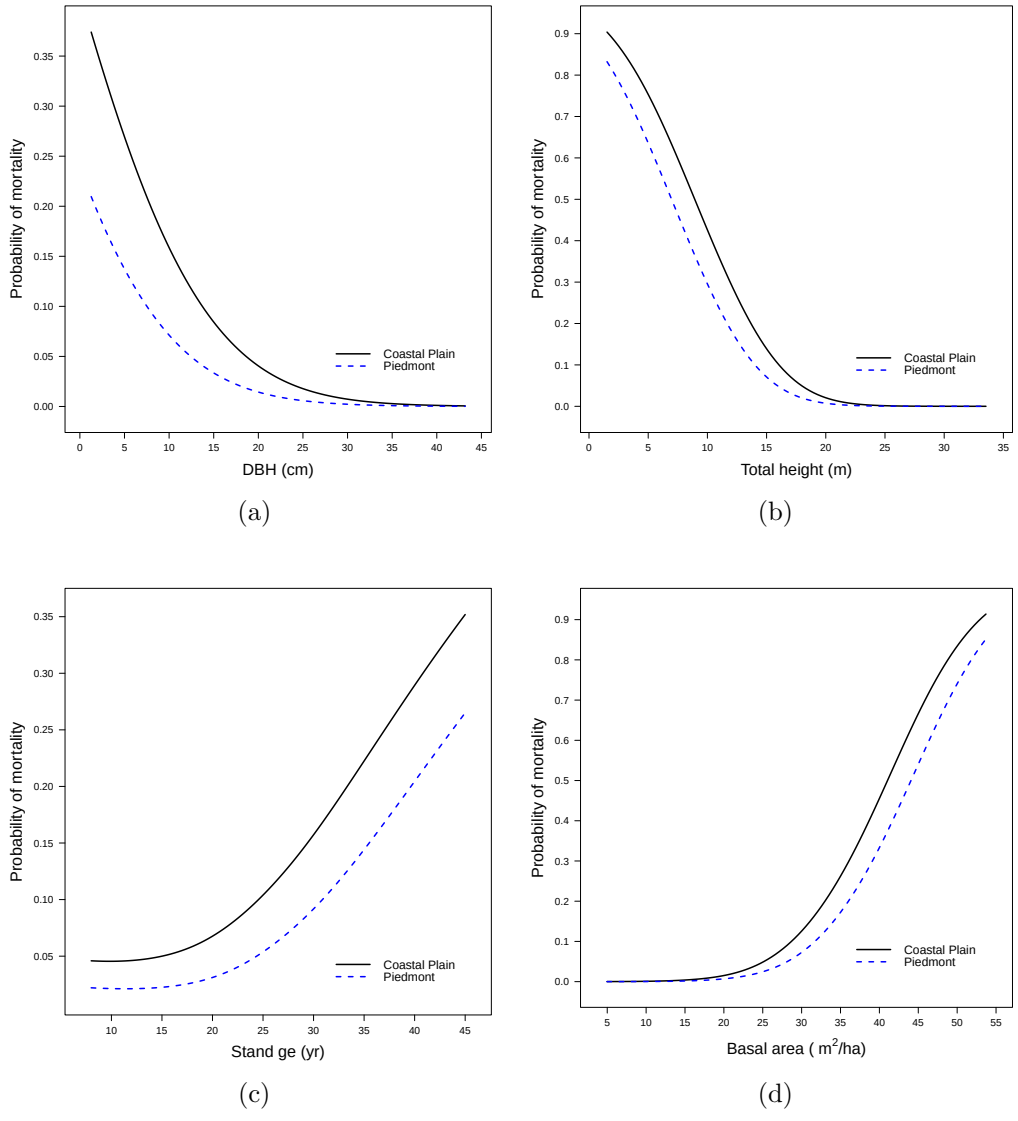


Figure D.2: Predicted probability of mortality predicted by Model 3 against (a) DBH (cm), (b) total tree height (m), (c) stand age (years) and (d) stand basal area (m^2 ha^{-1}) for Coastal Plain and Pidemont