

2008

Towards greater accuracy in individual-tree mortality regression

Clara Antón Fernández
Michigan Technological University

Follow this and additional works at: <https://digitalcommons.mtu.edu/etds>



Part of the [Forest Sciences Commons](#)

Copyright 2008 Clara Antón Fernández

Recommended Citation

Antón Fernández, Clara, "Towards greater accuracy in individual-tree mortality regression", Dissertation, Michigan Technological University, 2008.
<https://digitalcommons.mtu.edu/etds/124>

Follow this and additional works at: <https://digitalcommons.mtu.edu/etds>



Part of the [Forest Sciences Commons](#)

**TOWARDS GREATER ACCURACY IN INDIVIDUAL-TREE
MORTALITY REGRESSION**

By

CLARA ANTÓN FERNÁNDEZ

A DISSERTATION

Submitted in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

(Forest Science)

MICHIGAN TECHNOLOGICAL UNIVERSITY

2008

Copyright © Clara Antón Fernández 2008

This dissertation, "Towards greater accuracy in individual-tree mortality regression " is hereby approved in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY in the field of Forest Science.

School of Forest Resources and Environmental Science

Signatures:

Dissertation Advisor _____

Dr. Robert E. Froese

Committee _____

Dr. David D. Reed

Dr. James B. Pickens

Dr. Qiuying Sha

Dean _____

Dr. Margaret R. Gale

Date _____

ABSTRACT

Background mortality is an essential component of any forest growth and yield model. Forecasts of mortality contribute largely to the variability and accuracy of model predictions at the tree, stand and forest level. In the present study, I implement and evaluate state-of-the-art techniques to increase the accuracy of individual tree mortality models, similar to those used in many of the current variants of the Forest Vegetation Simulator, using data from North Idaho and Montana. The first technique addresses methods to correct for bias induced by measurement error typically present in competition variables. The second implements survival regression and evaluates its performance against the traditional logistic regression approach.

I selected the regression calibration (RC) algorithm as a good candidate for addressing the measurement error problem. Two logistic regression models for each species were fitted, one ignoring the measurement error, which is the “naïve” approach, and the other applying RC. The models fitted with RC outperformed the naïve models in terms of discrimination when the competition variable was found to be statistically significant. The effect of RC was more obvious where measurement error variance was large and for more shade-intolerant species. The process of model fitting and variable selection revealed that past emphasis on DBH as a predictor variable for mortality, while producing models with strong metrics of fit, may make models less generalizable.

The evaluation of the error variance estimator developed by Stage and Wykoff (1998), and core to the implementation of RC, in different spatial patterns and diameter distributions, revealed that the Stage and Wykoff estimate notably overestimated the true variance in all simulated stands, but those that are clustered. Results show a systematic bias even when all the assumptions made by the authors are guaranteed. I argue that this is the result of the Poisson-based estimate ignoring the overlapping area of potential plots around a tree. Effects, especially in the application phase, of the variance estimate justify suggested future efforts of improving the accuracy of the variance estimate.

The second technique implemented and evaluated is a survival regression model that accounts for the time dependent nature of variables, such as diameter and competition variables, and the interval-censored nature of data collected from remeasured plots. The performance of the model is compared with the traditional logistic regression model as a tool to predict individual tree mortality. Validation of both approaches shows that the survival regression approach discriminates better between dead and alive trees for all species.

In conclusion, I showed that the proposed techniques do increase the accuracy of individual tree mortality models, and are a promising first step towards the next generation of background mortality models. I have also identified the next steps to undertake in order to advance mortality models further.

ACKNOWLEDGMENTS

As in any long journey, during the course of this dissertation I have met many people on my way. Some of them have been essential towards its completion, and I would like to take this opportunity to thank them.

I forever will be in debt to my advisor, Dr. Robert Froese, the compass and the map on my journey. I have benefited from his patience, enthusiasm, care, and knowledge; I am eternally grateful to him for being such a great mentor. I would also like to thank the members of my committee: Dr. Dave Reed, Dr. Jim Pickens, Dr. Qiuying Sha, and Dr. Ann Maclean, for their valuable comments, reviews, and guidance.

For his forbearance during the everlasting writing months and for his unconditional love and support I owe Mitko more than I would ever be able to express.

I would also like to gratefully acknowledge the colleges and friends who have made the journey much more enjoyable. My colleagues at the lab have made these three years very pleasurable through their company and discussions. I particularly would like to thank Lacey, Jessie, Bharat, and Chris Miller. I much enjoyed every coffee break.

I am indebted to Archana for her truthful friendship, for introducing me into the Indian cuisine, and for sustaining me mentally and physically during the last months of my studies. I would also like to thank my good friends Nikki and Jessie, for their always-positive attitude, and for the great times we shared around a bottle of wine.

There have been many people who have made me feel welcome in this country, but I specially would like to thank the warmth and care that Erin and Robert Froese offered me from the beginning, the invaluable help Bharat and Sarita provided me when I first arrived to Houghton, and the hospitality of the Hispanic community that made me feel a little closer to home.

I would not be here today without the mental, emotional and financial support of my parents, Ricardo and Laly, who have paved the way towards this PhD. I would like to express my most sincere gratitude for their unconditional support.

TABLE OF CONTENTS

ABSTRACT	iii
ACKNOWLEDGMENTS	v
TABLE OF CONTENTS	vii
LIST OF TABLES	x
LIST OF FIGURES	xi
INTRODUCTION TO THE DISSERTATION	1
1 Forest models	1
2 The mortality component	4
3 Modelling mortality	5
4 Improving individual tree mortality models	6
5 References	9
CHAPTER 1. A MEASUREMENT ERROR MODEL OF INDIVIDUAL-TREE MORTALITY	12
1 Introduction	12
1.1 Modelling Tree Mortality	15
1.2 Measurement Error in Simulation Systems	16
1.3 Objectives	18
2 Materials and methods	19
2.1 Regression Calibration	19
2.2 Calibration and Validation Data	21
2.3 Model Development and Evaluation	22
2.4 Model Validation	23
2.5 Benchmark Comparisons	25
3 Results	26
3.1 Regression Calibration and Data Screening	27
3.2 Model Calibration	31
3.3 Model Validation	33
4 Discussion	37
4.1 Variable selection	37
4.2 Measurement error and regression calibration	39
4.3 Benchmark comparisons	40
4.4 Assumptions, improvements and future work	41
5 Conclusions	43
6 References	44
CHAPTER 2. SURVIVAL REGRESSION AS AN ALTERNATIVE FOR MODELLING INDIVIDUAL TREE MORTALITY	49

1	Introduction	49
1.1	Survival analysis techniques	50
1.2	Survival analysis in forestry	53
1.3	Logistic regression versus survival analysis	55
2	Objectives	58
3	Methodology	59
3.1	Approach	59
3.2	Sparling methodology	60
3.3	Software	62
4	Data	63
5	Results	66
5.1	Fitting phase	66
5.2	Validation phase	68
6	Discussion	70
6.1	Data and model form	70
6.2	Fitting phase	71
6.3	Validation phase	72
7	Conclusions	75
8	References	76
CHAPTER 3. EVALUATION OF THE SENSITIVITY OF A POISSON BASED LOCAL COMPETITION VARIANCE ESTIMATE TO SPATIAL PATTERNS		79
1	Introduction	79
1.1	The variance of local competition variables	81
1.2	Sensitivity of the model to PBA variance	83
1.3	Objectives and Approach	84
2	Creating spatially-explicit simulated stands	86
2.1	Spatial point process	86
2.2	Assigning diameters	88
2.3	Randomization	91
2.4	Prototypical stands	91
3	Methods	94
3.1	Sampling simulation	94
3.2	Sensitivity analysis using field data	95
4	Results	97
4.1	Comparisons of variance estimators	97
4.2	Shrinkage factor	99
4.3	Fitting phase	104
4.4	Application phase	106
5	Discussion	109
5.1	Shrinkage factor	112
5.2	Fitting phase	113
5.3	Application phase	114

6	Conclusions	116
7	References	117
SUMMARY AND CONCLUSIONS		120
1	References	122

LIST OF TABLES

Table 1-1. Correlation among some candidate variables for the calibration dataset. All correlations are statistically significant ($p < 0.001$). _____	31
Table 1-2. Summary of parameter estimates, bootstrap standard errors and p-value for the naive and RC models fitted using the calibration dataset. Log odds ratio increment is approximately the interquartile of the distribution of the variable for the three species in the calibration dataset, where for CR, PBAL is 10.2 m^2/ha , and for PBAL, CR is 40 %. _____	32
Table 1-3. Goodness of fit measurements for the models fit to the calibration data _____	33
Table 1-4. Measures of overall explanatory power (R^2_N), reliability and discrimination for the four models applied to the validation dataset. Temesgen and Mitchell (2005) did not fit a model to <i>A. grandis</i> . _____	35
Table 2-1. Summary of the three datasets: total number of trees, observed number of dead trees, percentage of dead trees, and mean observation period in years. _____	64
Table 2-2. Summary of the models for the logistic regression and survival regression approaches. _____	66
Table 2-3. Measures of predictive power for the calibration data. _____	67
Table 2-4. Likelihood ratio test where H_0 : no change in mortality rate with time. _____	68
Table 2-5. Measures of overall explanatory power (R^2_N), reliability and discrimination for the four models applied to the validation dataset. _____	68
Table 3-1. R index of clustering. _____	93
Table 3-2. Summary of the results for REAL, S&W, and RAW per stand type based on average values of the 10 realizations. Standard deviations are the mean over all plots and stands of the standard deviations of the PBA variance per plot. _____	98
Table 3-3. Root mean square error (RMSE) for the variance of PBA variance. _____	104
Table 3-4. Average variances of competition variables of the two datasets used in Chapter 1 calculated following Stage and Wykoff (1998). _____	105
Table 3-5. Parameters and standard errors of the coefficients (in parenthesis) of the <i>P. contorta</i> model fitted using different multiplier for the variance estimate. All coefficients had a $p < .001$ at $\alpha = 0.05$. The standard errors were calculated by 500 bootstrap. _____	105
Table 3-6. Characteristics of the selected validation stands. _____	107

LIST OF FIGURES

- Figure 1-1. An illustration of the effects of measurement error on bias, loss of power, and obscuring of features in the data. Hypothetical increment data presented on the left are augmented with simulated measurement error in the predictor, on the right. After Carroll et al. (2006) created by Robert Froese. _____ 14
- Figure 1-1. A hierarchical distribution of sample trees by species, status, and data set. The area of each rectangle is proportional to the number of individual trees in that category. _____ 26
- Figure 1-2. Boxplots of the four candidate variables for the calibration (top) and validation (bottom) datasets, by species. _____ 27
- Figure 1-3. RC adjustments relative to the observed measurement of error-prone predictors. Solid lines are smoothed estimates of the median, and dashed lines are smoothed estimates of the lower and upper quartiles, respectively. 28
- Figure 1-4. Kernel density estimates of observed mortality as a function of each of the candidate predictor variables for the calibration and validation data sets. ___ 30
- Figure 1-5. Distribution of bootstrap PBAL estimates of model coefficients for the naïve and RC model, for the three species studied. _____ 33
- Figure 1-6. The distribution of observed and predicted mortality with respect to the observed PBAL, for the three species examined. _____ 36
- Figure 2-1. A graphical representation of the size and attributes of the calibration dataset. The area of each bar is proportional to the number of trees in the category. _____ 64
- Figure 2-2. Distribution of the variables for the three datasets. For the survival dataset each measurement of each tree used to fit the model has been considered a separate measurement. _____ 65
- Figure 2-3. Sensitivity plots. Predictions for 10 years. _____ 67
- Figure 2-4. Kernel density estimates of observed and predicted mortality for the survival and logistic models in the validation dataset. _____ 69
- Figure 2-5. An illustration for the simple linear regression case of higher explanatory power with lower slope. The linear regression model in the left panel (bold line) has a slope of 0.81 and R^2 of 0.91, while the model on the right panel (bold line) has a slope of 0.69 and R^2 of 0.98. _____ 74
- Figure 3-1. One of the 10 realizations of each typified stand. Circles represent tree diameter. The diameters are not shown in real scale but the scale is similar for all plots. _____ 92
- Figure 3-2. Distribution of differences between average stand REAL and RAW variance estimates. _____ 98
- Figure 3-3. Effects of the shrinkage factor (SF) in the difference between the variance estimate and REAL for *young* stands. _____ 100

Figure 3-4. Effects of the shrinkage factor (SF) in the difference between the variance estimate and REAL for <i>Poisson</i> stands. _____	101
Figure 3-5. Effects of the shrinkage factor (SF) in the difference between the variance estimate and REAL for <i>old</i> stands. _____	102
Figure 3-6. Effects of the shrinkage factor (SF) in the difference between the variance estimate and REAL for <i>plantation</i> stands. _____	103
Figure 3-6. 10-year sensitivity plot for the <i>P. contorta</i> model. _____	106
Figure 3-8. Probability of mortality in a 5-year period with different PBAL variance estimates for 4 different stands in the validation dataset. The mean for all the trees in the stand is shown in bold. _____	108

INTRODUCTION TO THE DISSERTATION

*El hombre de estos campos que incendia los pinares
su despojo aguarda como botín de guerra
antaño hubo raído los negros encinares
talado los robustos robledos de la sierra.
Hoy ve sus pobres hijos huyendo de sus lares
la tempestad llevarse los limos de la tierra
por los sagrados ríos hacia los anchos mares
y en los páramos malditos, trabaja, sufre y yerra.
Antonio Machado (1875-1939)*

*The man of this countryside who burns the pinewoods
and awaits the remains like battle-won booty,
in the past has razed the black ilex woods
and cut down the robust oak stands of the sierra.
Today he sees his poor children fleeing their homes,
the storm carrying away the topsoil
down the sacred rivers to the wide seas;
and on accursed barren plains he labors, suffers and roams.*

Most of us, city dwellers, have started to forget how much we depend on forests. We rely on forests for clean water and air, wood supply, grazing and fodder, landscape, habitat for wild game and fish, to stabilize soil, and for recreational outdoor opportunities. The unprecedented rapid increase in human population of the last century due to medical advances and the vast increase in agriculture productivity, multiplies our demands on the forests. At the same time, this unprecedented growth in human population is increasing the competition for the use of forested land. As a result, large amounts of woodlands are being degraded, urbanized or converted to agricultural land. The remaining forests have to meet the demands of the population. In this scenario, the multi-objective optimization of forest resources is a key factor to meet the demands of the population and preserve the potential of forested lands.

Optimization of forest resources requires, in most cases, forest management, which includes human interventions. To lead the forest towards a multi-objective goal is not an easy task. Many factors affect the productivity of the forest in different areas, and many of these factors interact with each other. Moreover, unlike agriculture, forestry has long rotation periods, which make predicting interventions effects even harder. Hence, the need of tools and models that assist forest managers in their decisions.

1 Forest models

There are six different basic types of forest vegetation simulation models: forest growth and yield models, ecological gap models, ecological compartment models (resources

fluxes), process/mechanistic models, vegetation distribution models, and hybrid models (Monserud, 2003).

Forest growth and yield models form the oldest and broadest class. Forest yield models are usually statistical models that have traditionally been used to estimate diameter at breast height (DBH), tree height, and total volume. The most important advantage of this type of model is their ability to predict tree and stand characteristics in detail. However, forest yield models are not linked to underlying causes of productivity: the carbon and nutrient cycles, or the moisture regime and climate (Monserud, 2003). Consequently, these statistical models are unable to simulate either the impacts of future climate change on forest stands or the forest growth dynamics of some regions where historical data are not available (Peng, 2000; Schwalm and Ek, 2001). This is considered their fundamental disadvantage.

Ecological gap models are built by and for ecologists (Monserud, 2003). These models study successional population dynamics by estimating the establishment, diameter growth, and mortality of each tree on a small plot, usually equivalent to a gap or opening formed by a dominant tree's death. Gap models prediction timeframe is usually 100-1000 years, which makes validation a difficult task.

Ecological compartment (resources fluxes) models main purpose is to predict fluxes from one ecosystem compartment to another (Godfrey, 1986). These models rarely link directly to forest management (Monserud, 2003).

Process or mechanistic models are built mainly as a tool for scientific explanation rather than prediction. They attempt to model the growth processes based on the underlying causes of productivity. They usually require large amounts of detailed environmental data and complex calculations, and involve physiological observations difficult to measure, which limits their ability to scale up to the stand or the region (Hinckley *et al.*, 1996 in Monserud, 2003) and makes validation difficult. Another drawback of this type of models is that the knowledge of some important physiological processes is very limited, and some of the processes are so poorly understood that an accurate physiological model is not yet possible (Hinckley *et al.*, 1996).

Vegetation distribution models predict the occurrence of a specific vegetation zone or biome for each point on the landscape. Since they predict potential vegetation, they have been widely used to address questions of vegetation distribution under climate change. However, due to the broad scales they manage it is difficult to incorporate any management effects. The model “biome” (Prentice *et al.*, 1992) is one of the most well known vegetation distribution models.

Hybrid models are an attempt to merge the best of process based models and the empirical-based forest yield models. Although this type of model is promising, further development is needed before they can be considered a realistic option (Monserud, 2003).

Among these types of models, forest yield models are the most widely used in forest management. Forest yield models not only provide an efficient way to forecast resources but, more importantly, have the ability to make detailed predictions of tree and stand dynamics. This detail on stand structure, species composition, and silvicultural treatment response make this type of model well suited to analyzing alternative methods for compatible forest management.

Within forest yield models, the Forest Vegetation Simulator (FVS) is not only one of the most widely used, but also covers an important area in North America. FVS is an individual-tree, nonspatial, stand growth model built around a set of empirically derived equations of diameter growth, height growth, crown ratio, regeneration and mortality (Dixon, 2003).

Forest growth and yield models typically consist of several components that work together to simulate forest growth and management actions. The three main components are: establishment model, diameter/basal area increment, and mortality model.

Mortality plays a critical role in forest dynamics. The death of a tree directly reduces density measures, and affects the social position of the remaining trees, which in turn defines diameter growth, potential regeneration and probability of mortality of the residual trees. Although mortality plays a defining role in stand structure and dynamics,

it still remains one of the least understood components of the growth and yield models (Hamilton, 1986; Guan and Gertner, 1991b; Lutz and Halpern, 2006).

2 The mortality component

Mortality might be classified according to the intensity of the event. For example, mortality is often divided into “background” mortality and “catastrophic” mortality (Lugo and Scatena, 1996).

Background mortality affects individual trees and it is usually associated with senescence (aging), competition, and succession (Lugo and Scatena, 1996). In background mortality, tree death often represents a point in a continuum. When a tree cannot mobilize or acquire enough resources to support its respiratory requirements or fight the attack of opportunistic agents, the tree eventually dies. The immediate cause of death might be a root disease, an insect attack, or other agent, but those agents are just the last, but secondary, factor in a series of debilitating factors; e.g., drought, nutrients deficiency (Franklin *et al.*, 1987).

A tree allocates its resources depending on innumerable factors such as temperature, water, light and nutrient availability, reproductive stage, presence of herbivores, or competition. Biomass allocation will differ depending on the type of competition (Bazzaz, 1996). If, for example, a tree has been overtopped by other trees it will allocate more energy and biomass to try to grow faster and it will decrease its allocation to root growth and production of chemical defenses, which will make it more prone to pathogen attacks. If competition takes place below ground, as it might happen in dry stands, the tree will allocate more resources on increasing its root system. In any of the above cases the tree that cannot meet its respiratory requirements finally dies of starvation or an attack of one or various opportunistic agents.

Catastrophic mortality affects large-scale events and occurs when a forest is impacted by an external force such as fire, hurricanes, or epidemic attacks of certain insects. Catastrophic mortality is usually unpredictable.

Forest models commonly simulate catastrophic and background mortality separately. This dissertation focuses on background mortality, specifically in aspatial individual tree mortality models. Such models characterize the mortality of individual trees, in contrast to the ones that model mortality at the stand level, or ignore the absolute or relative position of individual trees.

Several functions and methods have been used to model background mortality, including the Weibull function (e.g. Somers *et al.*, 1980), logistic function (e.g. Hamilton and Edwards, 1976), negative exponential (e.g. Moser, 1972), gamma function (Kobe and Coates, 1997), Richard's function (e.g. Buford and Hafley, 1985), the exponential (e.g. Kobe and Coates, 1997), and more recently, techniques such as recursive partitioning, also known as classification and regression trees (e.g. Fan *et al.*, 2006), artificial neural networks (e.g. Guan and Gertner, 1991a), and survival analysis (e.g. Woodall *et al.*, 2005a). From the above alternatives, the logistic function is the most common (Monserud, 1976; Vanclay, 1995; Monserud and Sterba, 1999). Furthermore, Guan and Gertner (1991b) suggested that the best function to model individual tree mortality modelling might be the logistic function.

3 Modelling mortality

Mortality is difficult to model or predict. It is rare and highly stochastic, and yet an essential component of any stand growth model and any timber management system (Lee, 1971; Monserud, 1976). For example, Gertner (1989) found in a study of the precision of predictions made with the STEMS model (Belcher and Brand, 1982) that most of the variability of the predictions of either the number of trees per hectare or the basal area per hectare was associated with the mortality component of the model he studied. Furthermore, the proportion of the prediction variability due to mortality increased as the projection period increased. Guan and Gertner (1991b) considered the above situation common in stand simulation modelling.

The difficulties of acquiring adequate data to model mortality have forced modelers to use non-optimal solutions. In North America, the most widely used forest growth simulator framework, the Forest Vegetation Simulator (FVS) (Dixon, 2003), predicts

mortality in many of its variants using a Stand Density Index (SDI) based mortality model (Dixon, 1986; Johnson and Dixon, 1986). The SDI mortality model simulates some background mortality at the tree level, using a logistic regression function of tree DBH. Other background mortality due to competition is simulated by constraining stand-level stocking using size-density relationships. Stand-level mortality estimates are dispersed among trees following logic from the TWIGS model (Buchman, 1983; Buchman and Lentz, 1983; Buchman *et al.*, 1983; Teck and Hilt, 1990) or on shade tolerance of individual species, social position, and crown ratio. Related equations have not been empirically fitted; this approach essentially constrains individual-tree simulations using generally accepted stand-level emergent properties.

Other FVS variants, such as the Northern Idaho variant (Bush and Brand, 1995), simulate mortality using the Prognosis type mortality model (Stage, 1973). Dixon (2003) notes that this approach is used in variants only where there were enough inventory data suitable for developing the equations. This approach simulates background mortality using a logistic regression model, with a set of predictor variables including diameter at breast height and estimated diameter increment.

4 Improving individual tree mortality models

This dissertation aims to advance individual tree mortality models. I focus on two main aspects: correcting for measurement error in competition variables, and studying new forms for the mortality model that allow more flexibility exploring and modelling time dependent effects.

Correcting for measurement error in mortality models

Measurement error affects all FVS component models that include competition variables estimated by sample survey. Essentially, because some predictor variables in component models are estimated using samples, they are estimated with error. This measurement error can cause bias in the predictions of the models and can cause loss of power in detecting relationships among variables while fitting the model (Stage and Wykoff, 1998; Carroll *et al.*, 2006). Most FVS components include competition variables,

and are, therefore, subject to measurement error source bias. Recently, methods have been proposed to address the problem but have been confined to the linear statistical model (e.g., the Prognosis large-tree diameter equation; Stage and Wykoff 1998). Since The Prognosis type mortality model of FVS uses a logistic model, as do many other mortality models, the methodology applied by Stage and Wykoff (1998) is not applicable. The objective of this first chapter is to build a methodology to correct for measurement error in individual tree mortality models.

Logistic versus survival analysis for mortality models

Individual tree mortality has been typically modelled using the conventional logistic model, which does not directly account for changes in the covariate values over time. Yet, many of these covariates do vary over time (e.g. diameter, crown ratio, density, and climatic variables). The most natural alternative to handle such covariates would be to use survival analysis techniques with time-dependent covariates. There are already some examples of the application of survival analysis techniques in forestry (e.g. Volney, 1998; Rose Jr *et al.*, 2004; Woodall *et al.*, 2005a; Woodall *et al.*, 2005b), and some of them even include time dependent covariates (e.g. Rose *et al.*, 2006). However, none of them aims to model or study background mortality in individual tree mortality models, and addresses time dependent covariates. Moreover, to my knowledge, survival analysis has never been compared to logistic regression in the context of individual tree mortality models.

The objective of my second chapter is to use survival analysis techniques to fit a model that accounts for the time dependent nature of some of the variables and the interval-censored nature of data collected from remeasurements, and to compare this model with the traditional logistic regression model as a tool to predict individual tree mortality.

On the variance of local competition variables

The core of the measurement error correction for both diameter increment/basal area and mortality models is the variance of the error. Stage and Wykoff (1998) derived equations to estimate these variances. To derive the estimate of the variance error they

had to make some assumptions. For example, they assumed a random spatial distribution (Poisson distribution) of the trees and their diameters.

The objective in this chapter is to study the behavior of the variance of local competition variables in different spatial patterns and diameter distributions and to compare it with the behavior of Stage and Wykoff's (1998) approximation to the variance. It is also the objective of this chapter to study the sensitivity of the mortality models to the error variance.

Summary

The objectives of these three studies together aim to advance the accuracy of individual tree mortality models. When considered in the larger context of forest growth and yield models, such as FVS, improving the accuracy of the mortality component should result in more realistic forecasts of stand structures and dynamics. This is because the number, size and social position of dead trees directly influence the variables that define stand structure and dynamics, and thus feed back into other model components, such as those that simulate dimensional change and regeneration.

5 References

- Belcher, D.M., Brand, G.J., 1982. A description of STEMS-- the stand and tree evaluation and modeling system. Gen. Tech. Rep. NC-79. U.S. Department of Agriculture, Forest Service, North Central Research Station, St. Paul, MN.
- Bazzaz, F.A., 1996. *Plants in Changing Environments*. Cambridge University Press.
- Buchman, R.G., 1983. Survival predictions for major Lake States tree species. North Central Forest Experiment Station, USDA For. Serv. iii + 7.
- Buchman, R.G., Lentz, E.L., 1983. More Lake States tree survival predictions. Research Note NC-312. North Central Forest Experiment Station, USDA For. Serv. 6.
- Buchman, R.G., Pederson, S.P., Walters, N.R., 1983. A tree survival model with application to species of the Great Lakes region. *Canadian Journal of Forest Research* 13, 601-608.
- Buford, M.A., Hafley, W.L., 1985. Modeling the probability of individual tree mortality. *Forest Science* 31, 331-341.
- Bush, R., Brand, G.J., 1995. Lake States TWIGS geographic variant of the Forest Vegetation Simulator. Accessed Sep 17, 2008 from <http://ftp.fs.fed.us/pub/fmsc/ftp/fvs/docs/overviews/lsvr.txt>.
- Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M., 2006. *Measurement error in nonlinear models*. Chapman and Hall/CRC Press, Boca Raton, FL, 488 p.
- Dixon, G.E., 1986. Prognosis mortality modeling. Internal Rep. USDA, For. Serv., Forest Management Service Center, Fort Collins, CO. 10.
- Dixon, G.E., 2003. *Essential FVS: A User's Guide to the Forest Vegetation Simulator*. Internal Rep. USDA For. Serv., Fort Collins, CO. 193.
- Fan, Z., Kabrick, J.M., Shifley, S.R., 2006. Classification and regression tree based survival analysis in oak-dominated forests of Missouri's Ozark highlands. *Canadian Journal of Forest Research* 36, 1740-1748.
- Franklin, J.F., Shugart, H.H., Harmon, M.E., 1987. Tree Death as an Ecological Process. *Bioscience* 37, 550-556.
- Gertner, G., 1989. The need to improve models for individual tree mortality. In: Rink, G., Budelsky, C.A. (Eds.), *Seventh Central Hardwood Forest Conference*. USDA For. Serv., Southern Illinois University at Carbondale, March 5-8, 1989, pp. 59-61.
- Godfrey, K., 1983. *Compartmental models and their applications*. Academic Press, New York, NY, 291 p.
- Guan, B.T., Gertner, G., 1991a. Modeling Red Pine Tree Survival with an Artificial Neural Network. *Forest Science* 37, 1429-1440.

-
- Guan, B.T., Gertner, G., 1991b. Using a parallel distributed processing system to model individual tree mortality. *Forest Science* 37, 871-885.
- Hamilton, D.A., Jr., 1986. A logistic model of mortality in thinned and unthinned mixed conifer stands of northern Idaho. *Forest Science* 32, 989-1000.
- Hamilton, D.A., Jr., Edwards, B.M., 1976. Modeling the probability of individual tree mortality. Res. Pap. INT-185. USDA For. Serv. 22.
- Hinckley, T.M., Sprugel, D.G., Batista, J.L.F., Brooks, J.R., Brubaker, L.B., Compton, J., Erickson, H.E., Little, R.L., Maguire, D.A., McCarter, J.B., McKay, S.J., Pass, D., Peterson, D.W., Reed, J.P., Tacey, W., Wilkinson, L.E., Whytemare, A., 1996. Use of the JABOWA family of individual-tree based models for exploration of forest responses to global climate change. National Council of the Paper Industry for Air and Steam Improvement, Inc. II1-II54.
- Johnson, R.A., Dixon, G.E., 1986. Adapting models: problems and benefits from incorporating stand density index into the Prognosis Model. Internal Rep. USDA For. Serv., Fort Collins, CO. 17.
- Kobe, R.K., 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, 227-236.
- Lee, Y.J., 1971. Predicting mortality for even-aged stands of lodgepole pine. *The Forestry Chronicle* 47, 29-32.
- Lugo, A.E., Scatena, F.N., 1996. Background and Catastrophic Tree Mortality in Tropical Moist, Wet, and Rain Forests. *Biotropica* 28, 585-599.
- Lutz, J.A., Halpern, C.B., 2006. Tree mortality during early forest development: a long-term study of rates, causes, and consequences. *Ecological Monographs* 76, 257-275.
- Monserud, R.A., 1976. Simulation of forest tree mortality. *Forest Science* 22, 438-444.
- Monserud, R.A., 2003. Evaluating forest models in a sustainable forest management context. *Forest Biometry, Modelling and Information Sciences* 1, 35-47.
- Monserud, R.A., Sterba, H., 1999. Modeling individual tree mortality for Austrian forest species. *Forest Ecology and Management* 113, 109-123.
- Moser, J.W., 1972. Dynamics of an uneven-aged forest stand. *Forest Science* 18, 184-191.
- Peng, C., 2000. Understanding the role of forest simulation models in sustainable forest management. *Environmental Impact Assessment Review* 20, 481-501.
- Prentice, I.C., Cramer, W., Harrison, S.P., Leemans, R., Monserud, R.A., Solomon, A.M., 1992. A Global Biome Model Based on Plant Physiology and Dominance, Soil Properties and Climate. *Journal of Biogeography* 19, 117-134.
- Rose, E.C., Hall, D.B., Shiver, B.D., Clutter, M.L., Borders, B., 2006. A multilevel approach to individual tree survival prediction. *Forest Science* 52, 31-43.

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

Schwalm, C., Ek, A.R., 2001. Climate change and site: relevant mechanisms and modeling techniques. *Forest Ecology and Management* 150, 241-257.

Somers, G.L., Oderwald, R.G., Harms, W.R., Langdon, O.G., 1980. Predicting mortality with a Weibull distribution. *Forest Science* 26, 291-300.

Stage, A.R., 1973. Prognosis model for stand development. Res. Pap. RP-INT-137. USDA For. Serv. 32.

Stage, A.R., Wykoff, W.R., 1998. Adapting distance-independent forest growth models to represent spatial variability: effects of sampling design on model coefficients. *Forest Science* 44, 224-238.

Teck, R.M., Hilt, D.E., 1990. Individual-tree probability of survival model for the Northeastern United States. Research Paper Northeastern Forest Experiment Station. NE-642. USDA For. Serv. 10.

Vanclay, J.K., 1995. Growth models for tropical forests: a synthesis of models and methods. *Forest Science* 41, 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, 1784-1793.

Woodall, C.W., Grambsch, P.L., Thomas, W., 2005a. Applying survival analysis to a large-scale forest inventory for assessment of tree mortality in Minnesota. *Ecological Modelling* 189, 199-208.

Woodall, C.W., Grambsch, P.L., Thomas, W., Moser, W.K., 2005b. Survival analysis for a large scale forest health issue: Missouri oak decline. *Environmental Monitoring and Assessment* 108, 295-307.

CHAPTER 1. A MEASUREMENT ERROR MODEL OF INDIVIDUAL-TREE MORTALITY

1 Introduction

Trees compete with neighboring vegetation for scarce resources such as light, water and nutrients. The availability of these resources needed by a tree to exist in a given site is called growing space (Smith, 1986). Competition for growing space between individual trees plays an important role in forest stand dynamics. Specifically, competition leads to a reduction in growth, reproduction and/or chances of survival of at least one of the competing individuals (Begon *et al.*, 1996). Hence, including competition variables in models of tree diameter growth and mortality has been both useful and important.

Competition is commonly conceptualized at two levels, the tree and the stand level. **Stand competition variables** represent average competition in the stand, although they may also characterize competition from areas not represented at the sample location. This is especially true in water limited situations (Bormann, 1957; Stage, 2003). **Local competition variables**, on the other hand, are estimates of the competition in the vicinity of the tree. They add resolution to the competition, particularly in light limited ecosystems (Stage, 2003), and are usually estimated from sample plots.

Traditionally, characterizing local and stand competition has been attempted through density measurements derived from sampling plots; e.g. point basal area and trees per acre. The sampling procedure used to estimate competition generates error, commonly known as **measurement error**, which is error due to inaccurate measurements, error-in-variables, and, in the forestry literature, sampling error, probably because this type of error is generated in the sampling process (e.g. Stage and Wykoff, 1998). The estimates obtained through sampling may be unbiased with respect to the true value of the competition variable, but the variance of the estimates change with the spatial structure of the trees, tree diameter distribution, and plot size. The spatial structure of the trees

affect the variance of the competition variables through the number of trees present in each plot; for example, a clustered stand, where trees tend to be clumped, will have higher error variance than one with a very regular distribution of the trees, i.e. a plantation, since the density of trees varies highly from the dense clumps to the bare, or almost bare, space between clumps. When competition variables take into account the density and tree sizes, i.e. stand basal area (SBA), the variance is also affected by the diameter distribution. For instance, old stands will have higher variances than young stands with very similar small diameter trees. The plot size also influences the variance of competition variables. All else equal, smaller plots will tend to have larger variances than large plots, since the larger the plot, the larger the area is being considered, and the closer to the mean the measurement will be. Plot sizes do not change usually during the simulation of the stands, but plot size of the simulated stands may differ from the plots used to fit the data. Plot sizes may also differ among stands in the fitting dataset.

When variables with measurement error are used to model and predict diameter growth or mortality, and the measurement error is not accounted for, the effects on the models can be nontrivial. Summarizing the effects, Carroll et al. (2006) describe what they call the “triple whammy of measurement error”: (1) it causes bias in parameter estimates; (2) it leads to a sometimes profound loss of power for detecting interesting or important effects; and, (3) it masks features of the data, making graphical analysis difficult. All three are straightforward to illustrate by example. In the left panel of Figure 1-1, a unimodal function typical of those used to characterize diameter increment as a function of size is presented (e.g., Wykoff 1990). In the right panel, the same data, with the predictor variable perturbed to simulate measurement error, is shown. In this example, both graphically and statistically, most of the relationship is masked by the error.

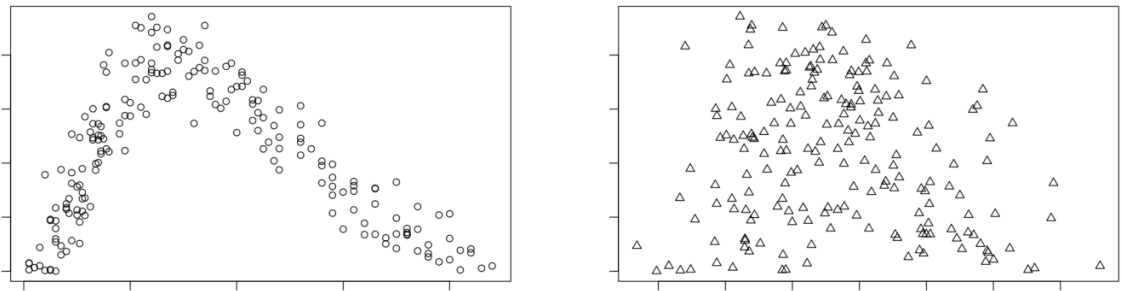


Figure 1-1. An illustration of the effects of measurement error on bias, loss of power, and obscuring of features in the data. Hypothetical increment data presented on the left are augmented with simulated measurement error in the predictor, on the right. After Carroll et al. (2006) created by Robert Froese.

Treatment of measurement error has been the focus of extensive literature in medicine (e.g., Rosner *et al.*, 1989; Prentice, 1996; Zeger *et al.*, 2000), but few are the examples that can be found in forest modelling. Curtis *et al.* (1974) were among the first to discuss the applicability of measurement error models in a forestry setting. Although they did not directly address the problem, they noted that error in the independent variable affects regression coefficients and is a source of the differences between site index estimation curves and height growth curves. Another example of measurement error in site index equations is given by Smith and Watts (1987), who found when developing a site index model for black spruce (*Picea mariana* Mill.) that accounting for measurement error in height increased prediction accuracy. Goelz and Burk (1992; 1996) and Wang *et al.* (2004) also explored the effects and treatment of measurement error in nonlinear or simultaneous site index systems. Tang and Wang (2002) presented a two-stage measurement error model for simultaneous equations. Hann and Zumrawi (1991), Kangas (1997), and Hynynen and Ojansuu (2003) examined the impact of error associated with variables influenced by sampling design, namely competition variables, on the accuracy of diameter and basal area increment model predictions, while Stage and Wykoff (1998) and Kangas (1998) also examined the impact on coefficients estimated in model development.

Notably, the effect of measurement error in individual-tree mortality models has received no obvious attention, despite these models sharing the same issues with competition

variables and being important components of forest simulation systems. For example, Buchman et al. (1983), Hamilton (1986), Teck and Hilt (1990), Monserud and Sterba (1999), Yang et al. (2003), and Temesgen and Mitchell (2005) all developed mortality models that rely upon competition variables subject to measurement error. In this chapter, I evaluate a comprehensive approach to accounting for measurement error in mortality model development and application.

1.1 Modelling Tree Mortality

Mortality is a discrete event; a tree may be either alive or dead. Although there are a few methodologies to model this type of behavior, logistic regression is the most common in individual-based forest models (Hasenauer *et al.*, 2001; Rose Jr, 2002). The logistic model has the form:

$$P_m = p(y = 1|x) = \frac{1}{1 + e^{-\beta x}} = \frac{e^{\beta x}}{1 + e^{\beta x}}$$

where P_m , is probability of mortality, and βx is a linear combination of the regression coefficients and the predictors. The logistic model implicitly assumes that either all remeasurement intervals are similar, or the differences in remeasurement intervals do not affect the outcome (Callas *et al.*, 1998). Because it is common in permanent plot data to have variable remeasurement periods and the probability of mortality increases with time, I adopted the modification suggested by Monserud (1976) to account for uneven remeasurement periods:

$$P_m = 1 - \left[1 - \frac{1}{1 + e^{-\beta x}}\right]^t$$

where t is the length of the observation period in years. This modification of the logistic model treats the survival probability as a compound interest phenomenon (Monserud, 1976).

1.2 Measurement Error in Simulation Systems

Measurement error associated with competition variables affect most simulation systems where competition variables estimated through sampling are included. These simulation systems involve many of the basal area/diameter increment and mortality models.

Stage and Wykoff (1998) proposed a solution for the basal area increment problem based upon a method of moments estimator presented by Fuller (1987). Their approach, which they call “Structural Based Prediction” (SBP), has two phases: estimation and application. In the estimation phase, the error in predictor variables is estimated and explicitly accounted for in parameter estimation, closely following Fuller’s theory. In application and to generate predictions, the coefficients of the regression equation are dynamically modified, case-by-case, using current estimates of error in the simulated system. A constraint is that certain intermediate matrix computations must be stored as they are necessary to compute the modification. Stage and Wykoff (1998) demonstrated SBP using the Prognosis basal area increment equation (Wykoff, 1990), though the method is applicable in general to linear models.

Measurement error methods require having some knowledge about the error variance. The four most common ways to gain this knowledge are (Carroll *et al.*, 2006):

- external validation data, where a parallel study measures the “gold standard”, the true value, and the error-prone estimate;
- instrumental data, where another variable, the instrumental variable, is observable in addition to the unbiased measurement;
- internal replicates, where two or more estimates for the same individual are available in the study;
- and external estimation of the variance, where the variance is estimated from other studies or derived from previous knowledge.

Since it is not common in forest inventories to measure the “gold standard”, or to have replicates of the measurements, the variance of the competition variables is most easily estimated externally.

External estimation procedures without replicated measurements or spatial data require simplifying assumptions (Smith and Watts, 1987; Stage and Wykoff, 1998; Lappi, 2005). To estimate error in stand basal area (SBA), Stage and Wykoff (1998) used the usual sample-based estimator, considering each single sampling point in the stand an estimate of SBA. For basal area in the vicinity of the tree being simulated (Point Basal Area or PBA), they derived an estimator assuming a spatially random (Poisson) tree distribution, with a spatially variable parameter, and tree diameters independent of their location and of the location of the rest of the trees.

Fuller’s (1987) method-of-moments estimator and Stage and Wykoff’s (1998) SBP approach are restricted to linear models and, therefore, they are not applicable to the intrinsically nonlinear logistic case. For nonlinear models, the recent literature describes five main approaches of accounting for measurement error in modelling. They include: regression calibration (Carroll and Stefanski, 1990; Pierce and Kellerer, 2004), simulation extrapolation (Cook and Stefanski, 1994; Stefanski and Cook, 1995), score function methods (Stefanski and Carroll, 1987; Stefanski, 1989; Nakamura, 1990), likelihood and quasilielihood (Carroll *et al.*, 1984; Wang *et al.*, 1996), and Bayesian methods (Gustafson, 2004). The first two approaches are widely applicable, general methodologies, although they result in estimators that are only approximately consistent. The last four approaches share the advantage of resulting in estimators that are fully consistent more generally than the first two approaches.

Both regression calibration (RC) and simulation extrapolation (SIMEX) are straightforward and potentially applicable in fitting any regression model. The main difference between the two methods is that SIMEX relies on computer simulation to estimate the parameters, while on the RC approach the variables measured with error are replaced with conditional expected values, and then used as if the error did not exist. These differences result in two important advantages of the RC approach: first, RC is computationally less intensive than SIMEX; second, because in RC the data are

corrected before fitting the model, contrary to SIMEX, correcting for measurement error in the application phase is straightforward. Notably, when RC is applied to linear regression and the error variance is estimated externally (i.e., not using replicated measurements), RC reproduces Fuller's (1987) method-of-moments estimator (Carroll *et al.*, 2006). Instead of using unaltered data and adjusting the model, as in Stage and Wykoff's approach, with RC the data are adjusted before developing or applying an unaltered model. Additionally, no data from the fitting dataset have to be stored to correct for measurement errors in a new application.

1.3 Objectives

The overall goal was to investigate the effect of applying measurement error methodologies in individual-tree mortality modelling in forestry, using Prognosis as a model system. The overall objective is developed in three specific objectives:

- (1) implement a measurement error methodology for nonlinear models in individual tree mortality models,
- (2) assess the effects of the proposed methodology on the model by comparing a model fitted ignoring measurement error and a model fitted with the competition variables corrected for measurement error, and
- (3) evaluate the differences in performance of the two models when applied to an independent validation dataset.

2 Materials and methods

The overall approach was to fit the same candidate mortality model, with and without applying the RC method, and compare the results. Two data sets were obtained from the Northern Rocky Mountain region for the analysis; one was used for calibration and the other for validation. The models, fitted with the calibration dataset, were used to predict mortality in the validation dataset, so the performance of both approaches could be compared.

2.1 Regression Calibration

The **classical additive error** model is usually defined as $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where \mathbf{X} is the unobservable, true value of the error-prone predictor and \mathbf{U} is the observation error. Then, the essence of RC is estimating \mathbf{X} , using the correspondent surrogate \mathbf{W} , along with \mathbf{Z} , which are the error-free predictor variables. In application for individual tree mortality models, this means that measured PBA, basal area of larger trees in the vicinity of the trees (PBAL) and/or SBA and any other assumed error-free variables (e.g., diameter at breast height, crown ratio) are used to predict the correspondent real value of PBA, PBAL and/or SBA.

A best linear approximation for \mathbf{X} for the classical additive error model was independently derived by both Carroll and Stefanski (1990) and Gleser (1990), and it is defined as follows when the error variance (Σ_{uu}) is estimated externally:

$$E(\mathbf{X}_i | \mathbf{Z}_i, \mathbf{W}_i) \approx \hat{\mu}_w + (\hat{\Sigma}_{xx}, \hat{\Sigma}_{xz}) \begin{bmatrix} \hat{\Sigma}_{xx} + \hat{\Sigma}_{uu} & \hat{\Sigma}_{xz} \\ \hat{\Sigma}_{xz}^t & \hat{\Sigma}_{zz} \end{bmatrix}^{-1} \begin{pmatrix} \mathbf{W}_i - \hat{\mu}_w \\ \mathbf{Z}_i - \hat{\mu}_z \end{pmatrix}$$

$$\text{Where } \hat{\Sigma}_{uu} = \frac{\sum_{i=1}^n (\hat{\Sigma}_{uu})_i}{n}$$

$$\hat{\Sigma}_{zz} = \frac{\sum_{i=1}^n (\mathbf{Z}_i - \hat{\mu}_z)(\mathbf{Z}_i - \hat{\mu}_z)^t}{n - 1}$$

$$\hat{\Sigma}_{xz} = \frac{\sum_{i=1}^n (\mathbf{W}_i - \hat{\mu}_w)(\mathbf{Z}_i - \hat{\mu}_z)^t}{n - 1}$$

$$\hat{\Sigma}_{xx} = \frac{\sum_{i=1}^n (\mathbf{W}_i - \hat{\mu}_w)(\mathbf{W}_i - \hat{\mu}_w)^t}{n - 1} - \hat{\Sigma}_{uu}$$

$$\hat{\mu}_w = \hat{\mu}_x = \bar{\mathbf{W}}.$$

$$\hat{\mu}_z = \bar{\mathbf{Z}}.$$

Σ in general denotes a covariance matrix for two random variables indicated by subscript, and $i=1,2,\dots,n$ represents the set of all individual observations. Note that a linear relationship between \mathbf{X} and (\mathbf{W}, \mathbf{Z}) is assumed.

The error variance (Σ_{uu}) was estimated adopting the same procedures as developed by Stage and Wykoff (1998). The authors explain these procedures in detail, which I do not reproduce here. Note that the estimates assume that: (1) tree locations follow a random (Poisson) pattern, and thus, independence of the location of each tree from the location of the rest of the trees; (2) the diameter of each tree is independent of the location of the rest of the trees; (3) the parameter of the Poisson forest varies spatially; and (4) the tree counts at a single sampling point are each estimates of the parameters of the local population.

The literature provides little guidance on the selection of \mathbf{Z} , other than to include all available variables (e.g., Hunter *et al.*, 1996; Carroll *et al.*, 2006; Kim *et al.*, 2006). Therefore, all of the candidate predictors in the RC correction were used; these variables were the two covariates measured with error, PBAL and SBA, and the two error-free covariates, DBH and crown ratio (CR). In a few cases, the adjusted values of PBAL and SBA were negative, which I replaced with 0 before model fitting.

In summary, the error variance was estimated following Stage and Wykoff (1998). Then, the RC algorithm was implemented separately for each species using PBAL, SBA, DBH and CR as predictors, obtaining the estimates of the measurement-error-free variables.

2.2 Calibration and Validation Data

Two data sets, obtained from the same geographic region as those used in prior development of the Prognosis model, were used for model development and testing. The calibration set was the larger of the two, and came from the USDA Forest Service, Rocky Mountain Region “Permanent Plot” program. This set includes an arbitrary sample of treated, regenerating stands in northern Idaho and western Montana, though in each stand control (untreated) conditions are also monitored.

The calibration data set is made up of clusters of 3 fixed-area plots installed in arbitrarily selected stands, which usually had been selected for pre-commercial thinning. No less than one control and three treated clusters were randomly located in a given stand. Plots were of 8.0 m radius and spaced 1.5 chains (1 chain = 66 ft. or 20.12 m) apart on a line. On each plot, species, DBH and crown length were recorded for all trees larger than a breakpoint DBH that varied by stand (usually 7.62 or 12.7 cm). At establishment, 10-year inside bark radial increment was measured for a sub-sample of one tree, by species, in each 5.08 cm diameter class, beginning at 7.62 cm. Sample data were obtained from stands located on the Idaho Panhandle (Kanisku, Coeur d’Alene and St. Joe), Flathead, Kootenai and Lolo US National Forests. Two re-measurement periods of approximately 10 years were available for many stands.

The validation data are from the USDA Forest Inventory and Analysis (FIA) program in the same geographic region. Relatively few of the FIA plots established in Idaho and Montana have been re-measured since the last periodic inventory, conducted in the 1980s in Idaho and 1990s in Montana. These “old” design plots used a cluster of 5, 7 or 10 variable-radius point samples at each field location and a constant basal area factor of $9.183 \text{ m}^2 \cdot \text{ha}^{-1}$. In 1999, FIA began a transition to a new inventory program, using a “new” design based on a cluster of four 7.3 m radius fixed-area subplots (Bechtold and Patterson, 2005). Some new design plots were established prior to 1999 under a Forest Health Monitoring (FHM) program. Under both designs, trees of all species greater than 12.7 cm DBH were measured for DBH, height and crown length, among other variables, and ten-year inside bark radial increment was measured using increment cores for a sub-sample that included two trees, by species, for each 5.1 cm diameter class. When

and where the new design plots are established at an old FIA field location, trees that were previously measured on the old plot are identified and measured one last time. Because only 10% of the new design plots are measured each year, and not all new design plots will be established at an old field location, only a small subset of FIA data are presently available with two field measurements.

Because the two datasets vary substantially in species representation, diameter and stand characteristics, the analysis was restricted to a subset of the available data. The three species best represented in both data sets were selected for the study: Douglas-fir (*Pseudotsuga menziesii* [Mirb.] Franco), grand fir (*Abies grandis* [Dougl. ex D. Don] Lindl.), and lodgepole pine (*Pinus contorta* Dougl. ex Loud.). The two sets also represent different management histories; the calibration set is the larger, principal set, but emphasizes managed conditions and stands sometimes in early stages of development. To make the validation set more comparable, I reserved only trees and stands where both SBA and DBH were less than the 99th percentile of the same variables in the calibration set.

2.3 Model Development and Evaluation

My approach emphasized discerning the effect of RC and not developing an operationally useful mortality model. To develop a credible foundation, I screened a simple set of potential predictor variables, and their usual transformations, commonly included in operational models (e.g., Hamilton, 1986; Monserud and Sterba, 1999; Temesgen and Mitchell, 2005). These included DBH, CR, $CR^{0.5}$, PBAL, SBA, and $SBA^{0.5}$. I generated kernel density estimates of mortality as a function of each predictor variable, by species, and used these as a guide to develop two logistic models for each species. The first used the measured values of the competition variables, ignoring measurement error; I call this the naïve approach. The second model was developed after correcting the competition variables for the measurement error with regression calibration, which I call the RC approach. All models were fit using maximum likelihood in the statistical computing software R (R Development Core Team 2008). Standard errors for model coefficients were estimated using bootstrapping. Log odds ratio for

each variable using the interquartile range found in the calibration data were computed to facilitate the interpretation of the effect of coefficients in the models.

Traditionally in the forestry literature the goodness-of-fit criteria has included the χ^2 test, and more recently the Hosmer-Lemeshow test (Hosmer and Lemeshow, 1980). Recent studies, however, do not recommend these tests due to their instability (Hosmer *et al.*, 1997; Alenius *et al.*, 2003). The quality of fit of the two models was instead evaluated using the generalized R^2_N index of Nagelkerke (1991) and Cragg and Uhler (1970) to characterize the general predictive ability. Further characterization of quality of fit was achieved by calculating the probability of concordance, **c** or **c-index**, (Harrell *et al.*, 1982; Harrell *et al.*, 1984), which is a measure of the model's predictive discrimination. The c-index is computed by taking all possible pairs of trees, such that one is alive and the other dead; then, the index is the proportion of such pairs with the dead tree having a higher predicted probability of mortality than the live one. A widely used measure of diagnostic discrimination, the area under a receiver operating characteristic curve (AUC) is identical to the c-index (Hanley and McNeil, 1982). The c-index takes values from 0.5 to 1, where a value of 0.5 indicates random predictions, and a value of 1 indicates perfect separation.

2.4 Model Validation

To evaluate the differences in performance of the two versions of the models I applied naïve and RC models to an independent dataset. I used the first measurement from this validation dataset to predict mortality for each species, and I compared the results to the observed mortality in the second measurement to analyze the predictive accuracy of both approaches.

Predictive accuracy of a logistic model has two components: reliability and discrimination (Harrell *et al.*, 1984). **Reliability** measures the extent to which the predictive probabilities agree with or differs from the observed frequencies. **Discrimination** addresses how well the model can discriminate between dead and alive trees. I used the slope and intercept estimates from Cox's measures of calibration and refinement (Cox, 1958; Harrell and Lee, 1991; Miller *et al.*, 1991) as indices of reliability

and discrimination, respectively. I augmented these with the additional metrics U, the unreliability index (Harrell and Lee, 1991), D, the discrimination index (Harrell and Lee, 1991), and the c-index used to evaluate goodness-of-fit in the calibration data.

Cox's measures of calibration and refinement (Cox, 1958; Harrell and Lee, 1991; Miller *et al.*, 1991) characterize the relationship between observed and predicted outcomes. This is based on fitting a separate binary logistic model to the validation sample, where the response is the observed outcome and the predictor is the probability forecast for that case from the model being validated. The new calibration equation is defined as:

$$Prob(Y_i|P_i) = \frac{1}{1 + e^{-(a+bL_i)}}$$

where P_i is the predicted probability from the logistic model being validated, and

$$L_i = \text{logit}(P_i) = \log\left(\frac{P_i}{1 - P_i}\right)$$

Ideally, for the model being validated, the calibration equation would have an intercept of zero and a slope of one (Miller *et al.*, 1991). The intercept, a , denotes the overall calibration of the model if the slope, b , is one; otherwise, a denotes the calibration at $p=0.5$. In any case, $a>0$ indicates that the predictive probability is in general too low, and $a<0$ indicates that the predictive probability is too high.

The **unreliability index U** can be calculated as (Harrell and Lee, 1991):

$$U = \frac{L(0,1) - L(a,b) - 2}{n}$$

Where $L(a,b)$ is the minimum likelihood for all a,b and $L(0,1)$ is the value of the likelihood at $a=0$, $b=1$. A perfectly reliable model would have an unreliability index $U=0$.

The **discrimination index D** is calculated as (Harrell and Lee, 1991):

$$D = \frac{L(a, 0) - L(a, b) - 1}{n}$$

Where $L(a,0)$ is minimum likelihood for all a when $b=0$. D has expected value 0 if there is no discrimination.

2.5 Benchmark Comparisons

To place the naïve and RC models in context, I compared these models to two extant mortality equations in the region. The first was the current Prognosis mortality model, developed by Hamilton (1986), and the second a set of mortality equations developed by Temesgen and Mitchell (2005) for a Prognosis variant for southeastern British Columbia, an area immediately adjacent to my study area.

Hamilton's model (1986) uses the 10-year diameter increment from the antecedent period to the one being projected as an explanatory variable. Because this variable was not available in the validation data, it was imputed using the Prognosis basal area increment equation, re-calibrated using FIA periodic inventory data from the 1980s and 1990s (Froese, 2007; Froese and Robinson, 2007). While this introduces some error, my objective was only to provide a general context for the RC results, not to rigorously evaluate alternative models. The results for Hamilton's model should be considered accordingly.

3 Results

The two datasets selected for this study differ markedly in size, mortality rate, and tree and stand characteristics. The calibration dataset consist of 25,344 trees with an average mortality proportion of 5.0%, whereas the validation dataset includes 4,007 trees with a mean mortality proportion of 15.3% (Figure 1-1). *P. contorta*, the predominant species in the calibration set, has the largest differences in mortality rates between datasets. *A. grandis* represents, on the other hand, the smallest proportion of trees in both sets, while *P. menziesii* dominates the validation dataset. The datasets also differ in tree and stand characteristics (Figure 1-2). The largest differences in the distribution of the candidate variables between the calibration and validation datasets are found in *P. contorta*, which has also the narrowest distribution of DBH and PBAL in the former dataset.

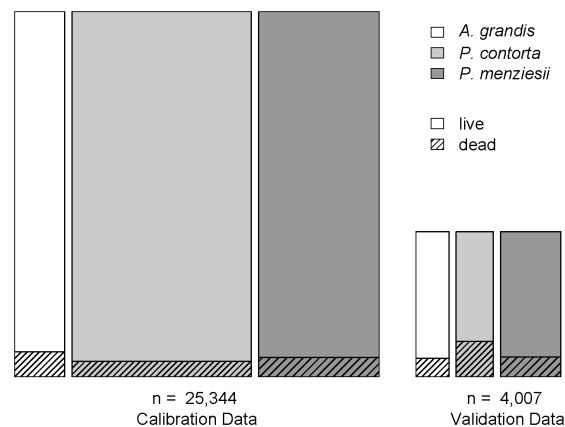


Figure 1-1. A hierarchical distribution of sample trees by species, status, and data set. The area of each rectangle is proportional to the number of individual trees in that category.

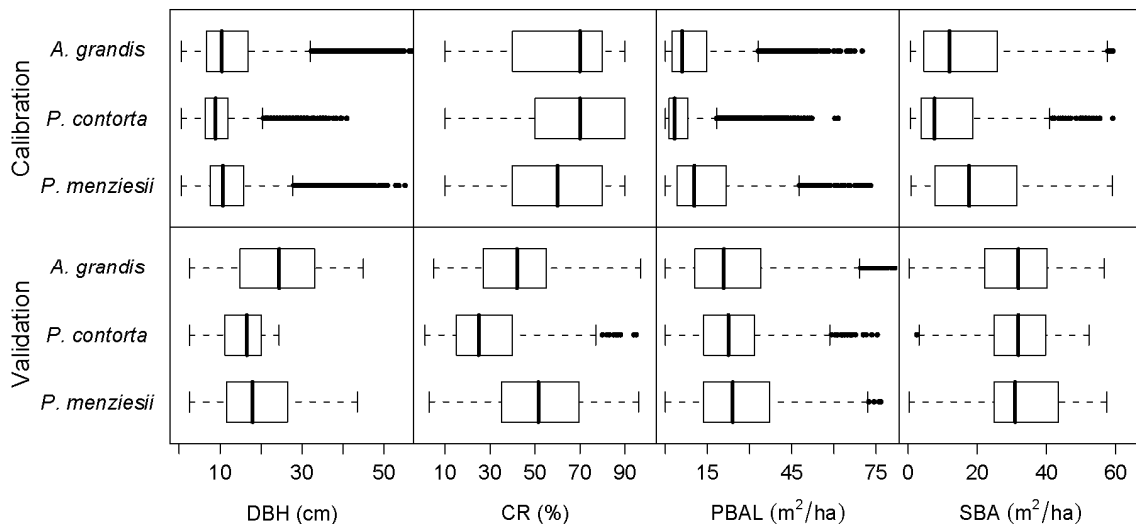


Figure 1-2. Boxplots of the four candidate variables for the calibration (top) and validation (bottom) datasets, by species.

3.1 Regression Calibration and Data Screening

The effect of RC on the error-prone variables is, in general, of shrinkage towards the mean (Figure 1-3). In other words, for large values of the observed variable, \mathbf{W} , the predicted values of the corresponding unobserved error-free variable, \mathbf{X} , tend to be smaller than \mathbf{W} , and vice versa. The corrections were similar among the three species, but vary between variables and datasets. The shrinkage towards the mean effect was more noticeable for PBAL than for SBA, and it was larger for the validation set. Notably, the estimated error variance was much larger for PBAL than for SBA in both sets, and much larger for the validation than for the calibration set for both variables.

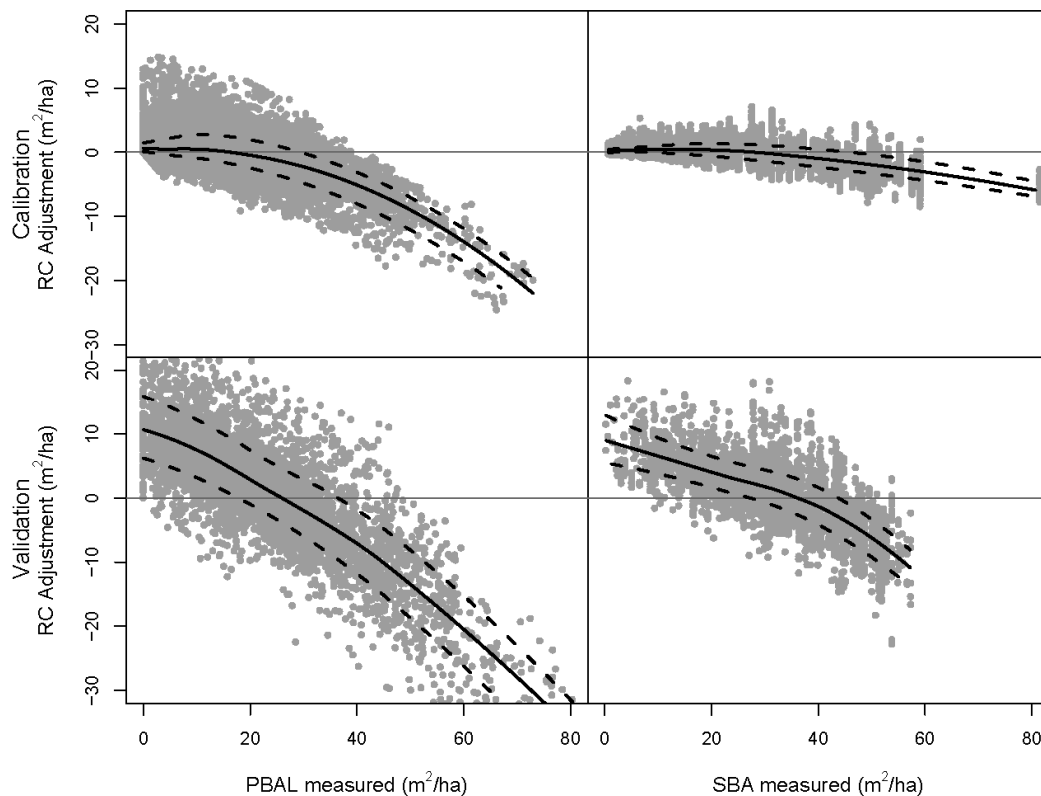


Figure 1-3. RC adjustments relative to the observed measurement of error-prone predictors. Solid lines are smoothed estimates of the median, and dashed lines are smoothed estimates of the lower and upper quartiles, respectively.

Variable screening revealed the trends of mortality along the candidate variables (Figure 1-4). The direction and slope were roughly congruent in both datasets for all variables, although the differences in average mortality rate result in parallel lines at different mortality rates, especially visible in SBA for *A. grandis* and *P. contorta*. The only, but eminent, difference between calibration and validation mortality trends was found in DBH, where the validation dataset shows a clear tendency to mortality reduction with increasing DBH, while the calibration dataset shows, if any, a trend of higher mortality with increasing DBH. In general, mortality decreases rather sharply for CR values between 0 and 40% approximately, after which the descent becomes more moderate. *A. grandis* shows the most abrupt fall, while *P. menziesii* shows the weakest, though all the species and datasets show a markedly descent in mortality with CR. Competition

variables exhibit an increase in mortality with increasing competition, with *P. contorta* being the species with the most pronounced slope.

The effect of RC in respect with mortality trends differed among species, sets and predictor variables (Figure 1-5c,d). Generally, the effect of RC, if any, was to steepen the trend. Similarly to Figure 1-3, when there was an RC effect it was always greater for the validation set than it was for the calibration set, and larger for PBAL than for SBA. For example, RC had no obvious effect on the trend in mortality across SBA for *A. grandis*, which was quite flat, in either set. In contrast, RC had a noticeable effect on the trend across PBAL and SBA for *P. contorta* that was particularly large for the validation set.

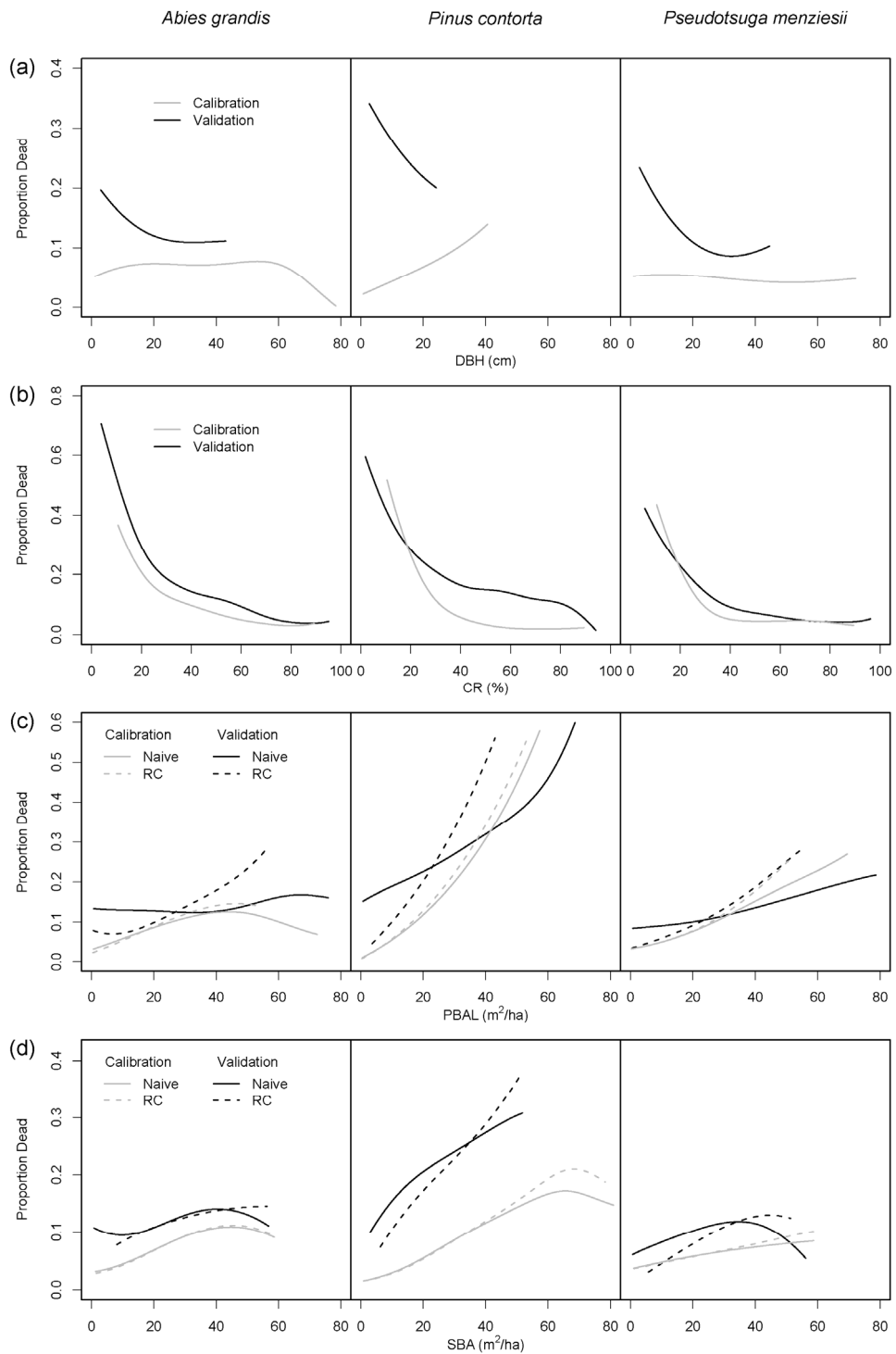


Figure 1-4. Kernel density estimates of observed mortality as a function of each of the candidate predictor variables for the calibration and validation data sets.

3.2 Model Calibration

Initial tests suggested that all variables excluding DBH were promising predictors and followed the expected biological trend. Consistently with results from screening, diameter variables were statistically significant for all species, but as already suggested by the screening the coefficients indicated increasing mortality with increasing diameter, contrary to most mortality models in use (e.g., Hamilton, 1990; Monserud and Sterba, 1999; Temesgen and Mitchell, 2005). Additionally, model performance in the validation was downgraded when diameter variables were included in the model. Therefore DBH variables were dropped as candidate predictors. Test fits also showed that it was easy to construct complex, statistically significant models that performed poorly on the validation data. I suspected this was due to a mixture of overfitting and parameter instability due to collinearity among candidate predictors. Correlations among independent variables were high, and slightly increased when RC was applied (Table 1-1).

Table 1-1. Correlation among some candidate variables for the calibration dataset. All correlations are statistically significant ($p < 0.001$).

	DBH	PBAL	PBAL-RC	SBA ^{0.5}	(SBA-RC) ^{0.5}	CR ^{0.5}
DBH	1					
PBAL	0.22	1				
PBAL-RC	0.25	0.97	1			
SBA ^{0.5}	0.53	0.74	0.82	1		
(SBA-RC) ^{0.5}	0.55	0.77	0.84	1	1	
CR ^{0.5}	-0.44	-0.69	-0.77	-0.74	-0.77	1

Including both competition variables lead to either instability in the coefficients (*P. menziesii*), either competition variable becoming not significant ($p > 0.05$) in naïve and RC models (*P. contorta*), or was significant in naïve models but not in RC (*A. grandis*). Therefore, the combination of CR^{0.5} and competition variables that yielded the most parsimonious model was selected, namely PBAL. The final models were therefore fitted for all the species with PBAL and CR^{0.5}. The coefficient estimates for both variables were statistically different from zero ($\alpha = 0.05$) for all three species and for both naïve and RC fits, with the exception of PBAL in the model for *A. grandis*. For all three species the predicted mortality was more sensitive to CR^{0.5} than it was to PBAL; estimated log

odds ratio for PBAL ranged from -0.11 – 0.55, while log odds ratios for CR^{0.5} were in the range of -1.50 – -3.18. For consistency, I retained PBAL in the model for *A. grandis* though PBAL was not a significant predictor under either the naïve or RC approaches.

Table 1-2. Summary of parameter estimates, bootstrap standard errors and p-value for the naïve and RC models fitted using the calibration dataset. Log odds ratio increment is approximately the interquartile of the distribution of the variable for the three species in the calibration dataset, where for CR, PBAL is 10.2 m²/ha, and for PBAL, CR is 40 %.

Species	Model	Predictor	Coefficient	Std. Err.	p-value	Log odds ratio
<i>Abies grandis</i>	Naïve	Intercept	-1.4076	0.3798	<0.001	
		PBAL	-0.0095	0.0053	0.087	-0.10
		CR ^{0.5}	-1.5808	0.1525	<0.001	-3.16
	RC	Intercept	-1.3711	0.4619	0.002	
		PBAL	-0.0104	0.0073	0.158	-0.11
		CR ^{0.5}	-1.5892	0.1682	<0.001	-3.18
<i>Pinus contorta</i>	Naïve	Intercept	-3.5446	0.3249	<0.001	
		PBAL	0.0461	0.0045	<0.001	0.47
		CR ^{0.5}	-1.0763	0.1265	<0.001	-2.15
	RC	Intercept	-3.7717	0.3565	<0.001	
		PBAL	0.0534	0.0055	<0.001	0.55
		CR ^{0.5}	-1.0074	0.1372	<0.001	-2.01
<i>Pseudotsuga menziesii</i>	Naïve	Intercept	-3.7795	0.3368	<0.001	
		PBAL	0.0189	0.0047	<0.001	0.19
		CR ^{0.5}	-0.7485	0.1297	<0.001	-1.50
	RC	Intercept	-3.4640	0.3937	<0.001	
		PBAL	0.0136	0.0068	0.043	0.14
		CR ^{0.5}	-0.8531	0.1442	<0.001	-1.71

The effect of RC was small, but noticeable, on the model fit to the calibration data (Table 1-2). RC affected all coefficients, including the intercept and the coefficient for CR^{0.5}, a predictor assumed to be error free and not adjusted by the RC procedure. For *P. contorta*, RC increased the sensitivity of predicted probability to PBAL and decreased it to CR^{0.5}; for *P. menziesii*, RC produced the opposite result. RC also invariably increased the estimated standard error of each of the coefficients (Figure 1-5), though not sufficiently to change any conclusions about significance ($\alpha = 0.05$). The effect of RC on goodness of fit metrics was negative, but very small (Table 1-3).

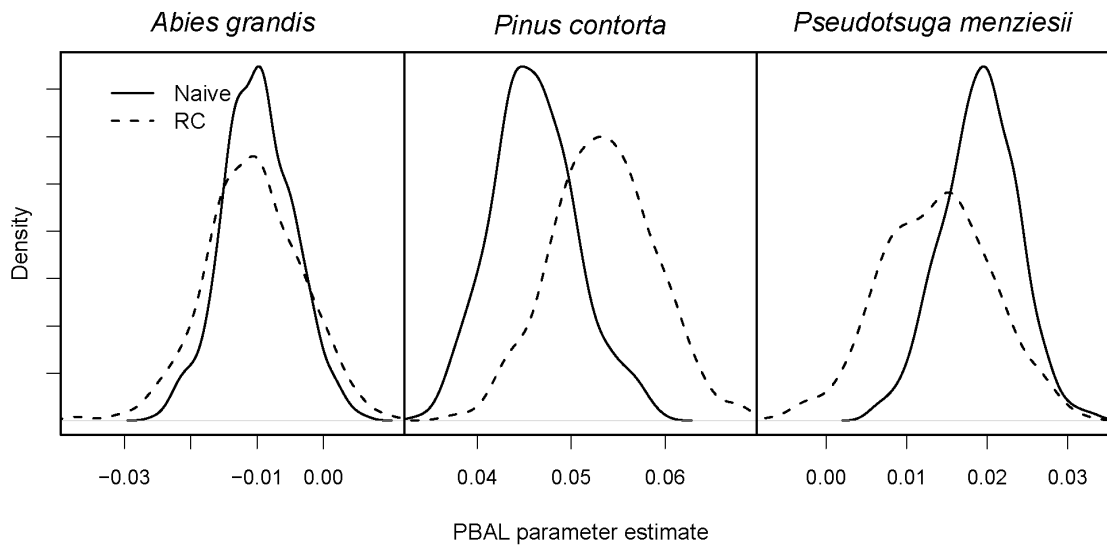


Figure 1-5. Distribution of bootstrap PBAL estimates of model coefficients for the naïve and RC model, for the three species studied.

Table 1-3. Goodness of fit measurements for the models fit to the calibration data

Species	Model	c-index	R^2_N	Residual Deviance
<i>Abies grandis</i>	Naïve	0.698	0.094	1,683.7
	RC	0.697	0.093	1,684.7
<i>Pinus contorta</i>	Naïve	0.759	0.179	3,843.6
	RC	0.759	0.178	3,849.0
<i>Pseudotsuga menziesii</i>	Naïve	0.642	0.054	3,491.3
	RC	0.640	0.050	3,503.4

3.3 Model Validation

The effects of RC were much more noticeable in the model validation than they were in the model calibration, particularly for *P. contorta* (Table 1-4). For this species, RC improved the model performance as measured by R^2_N and all five of the metrics of reliability and discrimination. The RC model was also superior to Hamilton’s model as judged by R^2_N , D and the c-index. Temesgen’s model for *P. contorta* performed worst among alternatives, including the naïve model. Results for *P. menziesii* were less consistent; RC improved the model for *P. menziesii*, by margins similar to the

improvements for *P. contorta*, as judged by R^2_N and the three metrics of discrimination. However, the two metrics of reliability indicated worse performance under RC than observed for the naïve model. In contrast to the results for *P. contorta*, the RC model for *P. menziesii* was inferior to Hamilton's model based on all metrics, except the slope of the calibration function, but still superior to Temesgen's model in all metrics but U. For *A. grandis*, RC resulted in a uniformly slight degradation in model performance, though both the naïve and RC models were superior to Hamilton's model based on all but one metric.

Table 1-4. Measures of overall explanatory power (R^2_N), reliability and discrimination for the four models applied to the validation dataset. Temesgen and Mitchell (2005) did not fit a model to *A. grandis*.

Species	Model	R^2_N	Reliability		Discrimination		
			Intercept	U	Slope	D	c-index
<i>Abies grandis</i>	Naïve	0.218	1.140	0.107	0.976	0.124	0.786
	RC	0.210	1.138	0.112	0.967	0.119	0.780
	Hamilton	0.144	-0.013	0.140	0.554	0.080	0.730
<i>Pinus contorta</i>	Naïve	0.120	-0.153	0.112	0.508	0.083	0.675
	RC	0.122	0.005	0.089	0.595	0.085	0.684
	Hamilton	0.085	-0.136	0.022	0.671	0.058	0.650
	T&M ¹	0.050	-0.844	0.278	0.200	0.033	0.648
<i>Pseudotsuga menziesii</i>	Naïve	0.097	0.018	0.011	0.846	0.049	0.693
	RC	0.112	0.410	0.013	0.998	0.057	0.703
	Hamilton	0.132	-0.258	0.003	0.806	0.067	0.722
	T&M ¹	0.086	-0.396	0.007	0.724	0.043	0.670

¹ Temesgen and Mitchell

Kernel density plots of observed mortality against CR and PBAL in the validation set showed that the naïve, RC and Hamilton's model all captured the mortality trends to various degrees, but generally forecast mortality rates lower than those observed in the data (Figure 1-6). The effect of RC was small with respect to CR for all three species, and small with respect to PBAL for *A. grandis*. In contrast, RC had a large effect for *P. contorta* and *P. menziesii*, tempering the predicted mortality particularly beyond about 30 m² ha⁻¹ and bringing the trend in line with that observed in the validation set. Only the RC and Hamilton's model approximate the observed trend in all cases, with the RC model superior to Hamilton's model in many cases.

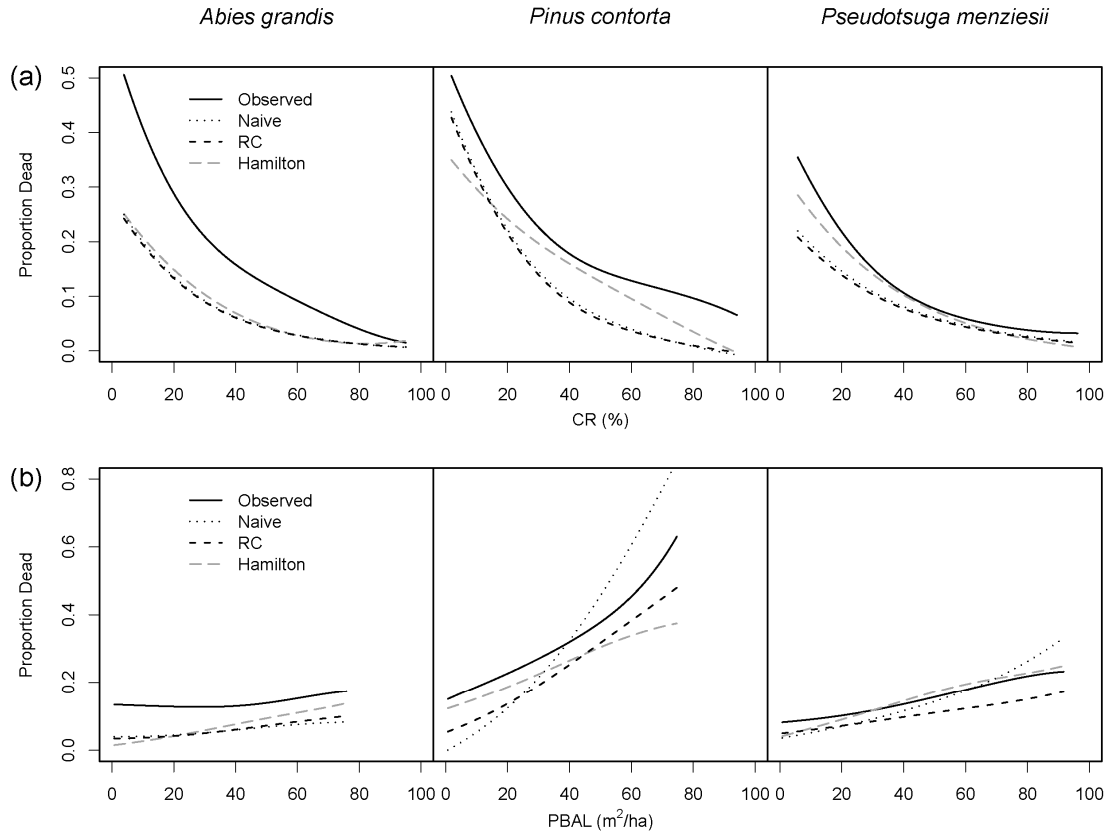


Figure 1-6. The distribution of observed and predicted mortality with respect to the observed PBAL, for the three species examined.

4 Discussion

In this chapter the regression calibration approach was selected to correct for measurement error in individual tree mortality models. This methodology was demonstrated using two datasets from the northern Idaho and western Montana. The largest of these datasets was used for calibrating alternative models that either ignore or address measurement error in competition variables, and the smallest of the datasets was used for the validation of the models. The sets differ among other characteristics in plot size, management, diameter distribution, species composition, average mortality rate and stocking. These broad differences have proved to be an asset in illustrating the importance of correcting for measurement error in model calibration and application. Initially, I expected that the fundamental differences between the two data sets would be an unavoidable limitation, as the calibration set clearly emphasizes young, managed stands quite in contrast in structure and mortality rate from the FIA data, which as random samples are representative of the average forest condition in the region (Figure 1-1, Figure 1-1). However, this disparity in the datasets turned out to be an asset for two reasons. First, the relationship between mortality rate and DBH or SBA is clearly different in the two data sets, and this has implications for variable selection in mortality modelling, in general. More importantly, the intrinsic difference in variability of tree and stand conditions between the two sets meant that the RC effect was actually much more obvious in application of the fitted model. In other words, had I investigated RC based only on the model calibration I might have been inclined to conclude that the utility of the method was limited.

4.1 Variable selection

One of the variables that most individual tree mortality models have in common is DBH. For example, Buchman et al., (1983), Hamilton (1986), Monserud and Sterba (1999), Eid and Tuhus (2001), Yang et al. (2003), Moore et al. (2004), Zhao et al. (2004) and Temesgen and Mitchell (2005) all included DBH or one or more transformations of DBH in their models. According to Hawkes (2000) the use of diameter as a covariate in mortality models is mainly because statistical analysis shows a relationship between

size and mortality in forest yield tables, representing competitive self-thinning. The calibration dataset involves almost exclusively pre-commercially thinned stands; in these stands, self-thinning has been interrupted through artificial reduction in density (Smith *et al.*, 1996), breaking the empirical link between DBH and mortality rate, at least until stand density recovers. This would explain the lack of trend between mortality and DBH in the calibration set. However, Monserud and Sterba (1999) argued that DBH is a measure of the ability of a tree to compete for scarce resources. Monserud and Sterba did not cite any work to support this argument, and a literature review has not revealed any empirical study that supports this theory. I hypothesized that the clear relationship between DBH and mortality presented by Monserud and Sterba (1999) and apparent also in the validation set (Figure 1-4) is mostly associated with relationships between stand density and size, and not with the increased ability of larger trees to compete for scarce resources that, although probable, is a much weaker relation than the former. My theory would only explain the lack of a defined trend between mortality and DBH but not the decreasing trend visible at least in two species. Preliminary work showed that the negative trend is a typical example of confounding variables, where CR is correlated with DBH and mortality, and DBH seems to be negatively correlated with mortality until CR is included in the model. Exploratory tests showed negative DBH coefficients, decreasing mortality with increasing diameter, in both *P. contorta* and *P. menziesii* models, when DBH was the only variable included in the model. When models were fitted with both CR and DBH, the DBH coefficients were statistically not significant (*P. menziesii*) or significant ($p < 0.05$), but had small log odds ratio (*P. contorta*).

These hypotheses are speculative, and should be investigated further. However, if supported, they suggest that mortality models intended for application in managed stands should eschew DBH, despite empirical evidence of utility, and rely instead upon more direct proxies indicative of vigor and competition, such as CR and PBAL. Results presented in this chapter suggest that this approach can perform as well or better than models that rely on DBH as a predictor variable. The RC model was superior to Hamilton's model for *A. grandis* and superior to Temesgen's and arguably similar in performance to Hamilton's model for *P. contorta* and *P. menziesii* (Table 1-3).

Variable selection has proven challenging due mainly to overfitting, collinearity among covariates, and inadequacy of selection criteria for logistic regression. Test fits showed that it was easy to construct complex, statistically significant models with superior model selection criteria, e.g. residual deviance, that performed poorly on the validation data. Astrud et al. (2008) found similar results while studying the appropriate level of complexity for a simulation model using an annual radial increment model. Astrud et al. (2008) noted that more complex models with higher traditional model selection criteria do not result in the best predictive model and suggest that the most traditional or common ways to select variables or models might not be the most appropriate since they tend to produce overfitting. Overfitting can be partially overcome by using resampling (Harrell, 2001); e.g. bootstrap; but it does not guarantee a parsimonious model and is no substitute for a validation with an independent dataset. Thus, it is fundamental to assess the predictive ability of the models with independent data since, as I have shown in this study (Table 1-3 and Table 1-4), better values of model selection criteria such as R^2_N or the c-index, do not necessarily mean better performance as predictive models.

4.2 Measurement error and regression calibration

At least two parts of Carroll et al.'s (2006) "triple whammy" were clearly evident in my analysis. The third part, loss of power for detecting important effects, was not apparent; all predictor variables I examined that were statistically significant ($\alpha=0.05$) in the RC approach were already significant in the naïve approach. I did, however, find clear evidence that measurement error masked features in the data (Figure 1-4) and introduced bias in parameter estimates (Figure 1-6, Table 1-3). Differences between the calibration and validation sets have likely enhanced these conclusions. For example, the effect of RC was relatively small on features in the calibration data, such as the trend between mortality and PBAL (Figure 1-5c), likely because these stands had been made deliberately homogeneous through pre-commercial thinning. In contrast, the more structurally complex validation data had larger measurement error, larger adjustment through RC (Figure 1-3), and a more obvious RC effect on the observed trend between mortality and PBAL (Figure 1-5c). Notably, after RC was applied, the observed trend

with PBAL in the validation data was driven towards that observed in the calibration data.

More obvious was the effect that measurement error had on apparent bias in the parameter estimates. It is important to note that the treatment of measurement error is important only where error-prone variables have shown utility as predictor variables. For example, for *A. grandis* PBAL was not a useful predictor of mortality even under the RC approach; any trend between PBAL and mortality seemed already accounted for by the much stronger relationship between mortality and $CR^{0.5}$. For the other two species, however, the RC approach led to obvious improvements in nearly all measures of predictive ability on the validation data (Table 1-3). The improvement in both metrics of discrimination is especially desirable, showing that the RC models are better able to capture the functional relationship between the competition variables and tree mortality. This implies greater accuracy in future simulations of plot-level dynamics, as well as population behaviour, leading to more realistic simulation of stand structures as they evolve with time. Evidence of the RC effect was also likely enhanced by differences between the calibration and validation data. Had the two data sets been similar, the effect of measurement error would have been similar and the bias in parameter estimates relative to the error free values would have been difficult to detect. In other words, RC has the effect of standardizing data for differences in measurement error, leading to more consistent simulations among stand structures, whether these differences are due to different geography, site quality, management, or even within-stand development through time.

4.3 Benchmark comparisons

The overall objective in this chapter was to investigate the effect of applying measurement error methodologies in individual-tree mortality modelling in forestry. Thus, the focus in the validation phase was not on the quality of the predictions, *per se*, but on the differences in performance between the two approaches. Considering the differences between the two datasets, and especially taking into account the peculiarities of the calibration set, it was surprising to see how Hamilton (1986) model never outperformed either the naïve or the RC model in all measures.

The RC approach benefits are most consistent for *P. contorta*, where each measure of overall explanatory power, reliability and discrimination improved with respect to the naïve approach. Improvements are also clear for *P. menziesii*, although not as consistent as in *P. contorta*. Shade-tolerance is associated with the statistical significance of the competition variable (Table 1-1), and it seems to be related also with the magnitude of the improvements brought by the RC approach. The most shade-intolerant of the three species, *P. contorta*, benefits the most from the RC approach, followed by *P. menziesii* in benefits and in shade-tolerance. For *A. grandis*, the most tolerant species, the improvements of RC are absent since the one variable with measurement error, PBAL, is not even statistically significant.

4.4 Assumptions, improvements and future work

One of the keys to adequately applying RC is the correct estimation of Σ_{uu} . I have adopted the methodology described by Stage and Wykoff (1998). However, Stage and Wykoff already mentioned that a refinement of the variance model is needed. Moreover, Lappi (2005) critiques many aspects of Stage and Wykoff's methodology. Stage and Wykoff assumed the classical measurement error (see section 2.1 in Materials and methods), where the error term is distributed independently of the true value, versus Berkson measurement error, where the error term is distributed independently of the observed value. Lappi suggests that since the local density around a tree affects tree growth, Stage and Wykoff's (1998) assumption that the measurement error is uncorrelated with the true value is not applicable for local density, and therefore the assumption of classical measurement error is incorrect. Lappi also notes that the assumption of Poisson locations of trees is not likely valid. However, Lappi does not suggest any solution to these problems and, even though the estimations of Σ_{uu} proposed by Stage and Wykoff might not be perfect, the results seem coherent. Furthermore, since there are no obvious better alternatives at this time I consider Stage and Wykoff's approach the best option.

A closer examination of how variables are selected and how the RC correction is structured is also needed. Little guidance is available in the literature, other than to include all available covariates (e.g., Hunter *et al.*, 1996; Carroll *et al.*, 2006; Kim *et al.*,

2006) and that the relationship between \mathbf{X} and (\mathbf{W}, \mathbf{Z}) is assumed to be linear (Carroll *et al.*, 2006). I included DBH and CR in the RC correction for PBAL and SBA, though the former are tree-level variables used to correct the latter, which are not. It may be that transformations of \mathbf{Z} are required or desirable to meet the linearity assumption. Also, I applied RC by species to the entire data set; in conceivable applications, this would be repeated each model projection cycle, allowing the RC correction to change as stand structure and estimated measurement error variances change (Stage and Wykoff, 1998). However, since measurement error should also vary stand by stand, it may be that a more sophisticated algorithm would allow for a more accurate specification of Σ_{uu} and a more effective RC adjustment in both the calibration and application phases.

5 Conclusions

Despite limited data, correction for measurement error using Regression Calibration improved the predictive ability of individual tree mortality models, particularly in terms of discrimination, and changed how mortality responds to competition variables. These effects were larger when the models included variables with high error variance and when the species were shade intolerant. Available calibration data from intensively managed stands also revealed that past emphasis on DBH as a predictor variable, while producing models with strong metrics of fit, may make models less generalizable. An errors-in-variables solution via RC should be compared to Stage and Wykoff's (1998) structural based prediction, particularly in the context of the diameter increment model. Because RC corrections are before models are fit or applied, it may offer a unified approach. The practical and theoretical tradeoffs between methods need to be examined and quantified.

6 References

- Alenius, V., Hökkä, H., Salminen, H., Jutras, S., 2003. Evaluating estimation methods for logistic regression in modelling individual-tree mortality. In: Amaro, A., Reed, D., Soares, P. (Eds.), *Modelling Forest Systems* CABI Publishing, Oxford.
- Astrud, R., Coates, K.D., Hall, E., 2008. Finding the appropriate level of complexity for a simulation model: An example with a forest growth model. *Forest Ecology and Management* doi: 10.1016/j.foreco.2008.07.016.
- Bechtold, W.A., Patterson, P.L., 2005. The enhanced Forest Inventory and Analysis program—national sampling design and estimation procedures. Gen. Tech. Rep. GTR-SRS-80. USDA For. Serv., Asheville, NC.
- Begon, M., Harper, J.L., Townsend, C.R., L, J., 1996. *Ecology: Individuals, Populations and Communities*. Blackwell Publishing.
- Bormann, F.H., 1957. Moisture Transfer between Plants through Intertwined Root Systems. *Plant Physiology* 32, 48.
- Buchman, R.G., Pederson, S.P., Walters, N.R., 1983. A tree survival model with application to species of the Great Lakes region. *Canadian Journal of Forest Research* 13, 601-608.
- Callas, P.W., Pastides, H., Hosmer, D.W., 1998. Empirical comparisons of proportional hazards, poisson, and logistic regression modeling of occupational cohort data. *American Journal of Industrial Medicine* 33, 33-47.
- Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M., 2006. *Measurement error in nonlinear models*. Chapman and Hall/CRC Press, Boca Raton, FL.
- Carroll, R.J., Spiegelman, C., Lan, K.K., Bailey, K.T., Abbot, R.D., 1984. On errors-in-variables for binary regression models. *Biometrika* 71, 19-26.
- Carroll, R.J., Stefanski, L.A., 1990. Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of American Statistical Association* 85, 652-663.
- Cook, J.R., Stefanski, L.A., 1994. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association* 89, 1314-1328.
- Cox, D.R., 1958. Two further applications of a model for binary regression. *Biometrika* 45, 562-565.
- Cragg, J.G., Uhler, R., 1970. The demand for automobiles. *Canadian Journal of Economics* 3, 386-406.
- Curtis, R.O., DeMars, D.J., Herman, F., R., 1974. Which dependent variable in site index-height-age regressions? *Forest Science* 20, 74-87.

-
- Eid, T., Tuhus, E., 2001. Models for individual tree mortality in Norway. *Forest Ecology and Management* 154, 69-84.
- Froese, R.E., 2007. Unpublished data, available from the author. Unpublished manuscript, available from the authors,
- Froese, R.E., Robinson, A.P., 2007. A validation and evaluation of the Prognosis individual-tree basal area increment model. *Canadian Journal of Forest Research* 37, 1438-1449.
- Fuller, W.A., 1987. *Measurement error models*. Wiley, New York, NY.
- Gleser, L.J., 1990. Improvements of the naive approach to estimation in non-linear errors-in-variables regression models. In: Brown, P.J., Fuller, W.A. (Eds.), *Statistical Analysis of Measurement Error Models and Application*. American Mathematics Society, Providence.
- Goelz, J.C.G., Burk, T.E., 1992. Development of a well-behaved site index equation: jack pine in north central Ontario. *Can. J. Forest Res* 22, 776-784.
- Goelz, J.C.G., Burk, T.E., 1996. Measurement error causes bias in site index equations. *Can. J. Forest Res* 26, 1585-1593.
- Gustafson, P., 2004. *Measurement error and misclassification in statistics and epidemiology*. CRC/Chapman & Hall.
- Hamilton, D.A., Jr., 1986. A logistic model of mortality in thinned and unthinned mixed conifer stands of northern Idaho. *Forest Science* 32, 989-1000.
- Hamilton, D.A., Jr., 1990. Extending the range of applicability of an individual tree mortality model. *Canadian Journal of Forest Research* 20, 1212-1218.
- Hanley, J., McNeil, B., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29-36.
- Hann, D.W., Zumrawi, A.A., 1991. Growth model predictions as affected by alternative sampling-unit designs. *Forest Science* 37, 1641-1655.
- Harrell, F.E., 2001. *Regression modelling strategies with applications to linear models, logistic regression, and survival analysis*. Springer, New York, NY.
- Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A., 1982. Evaluating the yield of medical tests. *Journal of American Medical Association* 247, 2543-2546.
- Harrell, F.E., Lee, K.L., 1991. Using logistic model calibration to assess the quality of probability of predictions. Unpublished manuscript, available from the authors, f.harrell@vanderbilt.edu.
- Harrell, F.E., Lee, K.L., Califf, R.M., Pryor, D.B., Rosati, R.A., 1984. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine* 3, 143-152.
- Hasenauer, H., Merkl, D., Weingartner, M., 2001. Estimating tree mortality of Norway spruce stands with neural networks. *Advances in Environmental Research* 5, 405-414.

-
- Hawkes, C., 2000. Woody plant mortality algorithms: description, problems and progress. *Ecological Modelling* 126, 225-248.
- Hosmer, D.W., Hosmer, T., Le Cessie, S., Lemeshow, S., 1997. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine* 16, 965-980.
- Hosmer, D.W., Lemeshow, S., 1980. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics-Theory and Methods* 9, 1043-1069.
- Hunter, D.J., Spiegelman, D., Adami, H.O., Beeson, L., van den Brandt, P.A., Folsom, A.R., Fraser, G.E., Goldbohm, R.A., Graham, S., Howe, G.R., 1996. Cohort Studies of Fat Intake and the Risk of Breast Cancer--A Pooled Analysis. *New England Journal of Medicine* 334, 356.
- Hynynen, J., Ojansuu, R., 2003. Impact of plot size on individual-tree competition measures for growth and yield simulators. *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere* 33, 455-465.
- Kangas, A.S., 1997. On the prediction bias and variance in long-term growth projections. *Forest Ecology and Management* 96, 207-216.
- Kangas, A.S., 1998. Effect of errors-in-variables on coefficients of a growth model and on prediction of growth. *Forest Ecology and Management* 102, 203-212.
- Kim, E.H.J., Willett, W.C., Colditz, G.A., Hankinson, S.E., Stampfer, M.J., Hunter, D.J., Rosner, B., Holmes, M.D., 2006. Dietary Fat and Risk of Postmenopausal Breast Cancer in a 20-year Follow-up. *American Journal of epidemiology* 164, 990.
- Lappi, J., 2005. Plot size related measurement error bias in tree growth models. *Canadian Journal of Forest Research* 35, 1031-1040.
- Miller, M.E., Hui, S.L., Tierney, W.M., 1991. Validation techniques for logistic regression models. *Statistics in Medicine* 10, 1213-1226.
- Monserud, R.A., 1976. Simulation of forest tree mortality. *Forest Science* 22, 438-444.
- Monserud, R.A., Sterba, H., 1999. Modeling individual tree mortality for Austrian forest species. *Forest Ecology and Management* 113, 109-123.
- Moore, J.A., Hamilton, D.A., Jr., Xiao, Y., Byrne, J., 2004. Bedrock type significantly affects individual tree mortality for various conifers in the inland Northwest, U.S.A. *Canadian Journal of Forest Research* 34, 31-42.
- Nagelkerke, N.J.D., 1991. A note on a general definition of the coefficient of determination. *Biometrika* 78, 691-692.
- Nakamura, T., 1990. Corrected score functions for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika* 77, 127-137.
- Pierce, D.A., Kellerer, A.M., 2004. Adjusting for covariate errors with nonparametric assessment of the true covariate distribution. *Biometrika* 91, 863-876.

Prentice, R.L., 1996. Measurement Error and Results From Analytic Epidemiology: Dietary Fat and Breast Cancer. *Journal of the National Cancer Institute* 88, 1738-1747.

R Development Core Team, 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rose Jr, C.E., 2002. Modeling and allocating forestry survival: a loblolly pine case study. Univ. Georgia. Dissertation. 176 pp.

Rosner, B., Willett, W.C., Spiegelman, D., 1989. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine* 8, 1051-1069.

Smith, D.M., 1986. *The practice of silviculture*. Wiley, New York.

Smith, D.M., Larson, B.C., Kelty, M.J., Ashton, P.M.S., 1996. *The practice of silviculture: applied forest ecology*, 9th edition. Wiley, New York, NY.

Smith, V.G., Watts, M., 1987. An assessment of the structural method of deriving a black spruce site equation. *Canadian Journal of Forest Research* 17, 1181-1189

Stage, A.R., 2003. How forest models are connected to reality: evaluation criteria for their use in decision support. *Canadian Journal of Forest Research* 33, 410-421.

Stage, A.R., Wykoff, W.R., 1998. Adapting distance-independent forest growth models to represent spatial variability: effects of sampling design on model coefficients. *Forest Science* 44, 224-238.

Stefanski, L.A., 1989. Unbiased estimation of a nonlinear function of a normal mean with application to measurement error model. *Communications in Statistics, Series A* 18, 4335-4358.

Stefanski, L.A., Carroll, R.J., 1987. Conditional scores and optimal scores in generalized linear measurement error models. *Biometrika* 74, 703-716.

Stefanski, L.A., Cook, J.R., 1995. Simulation-extrapolation: The measurement error jackknife. *Journal of the American Statistical Association* 90, 1247-1256.

Tang, S.Z., Wang, Y.H., 2002. A parameter estimation program for the error-in-variable model. *Ecological Modelling* 156, 225-236.

Teck, R.M., Hilt, D.E., 1990. Individual-tree probability of survival model for the Northeastern United States. Research Paper Northeastern Forest Experiment Station. NE-642. USDA For. Serv. 10.

Temesgen, H., Mitchell, S.J., 2005. An individual tree mortality model for South-eastern British Columbia. *Western Journal of Applied Forestry* 20, 101-109.

Wang, N., Carroll, R.J., Liang, K.Y., 1996. Quasilikelihood estimation in measurement error models with correlated replicates. *Biometrics* 52, 401-411.

Wang, Y.H., Huang, S.M., Yang, R.C., Tang, S.Z., 2004. Error-in-variable method to estimate parameters for reciprocal base-age invariant site index models. *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere* 34, 1929-1937.

Wykoff, W.R., 1990. A basal increment model for individual conifers in the Northern Rocky mountains. *Forest Science* 36, 1077-1104.

Yang, Y., Titus, S.J., Huang, S., 2003. Modeling individual tree mortality for white spruce in Alberta. *Ecological Modelling* 163, 209-222.

Zeger, S.L., Thomas, D., Dominici, F., Samet, J.M., Schwartz, J., Dockery, D., Cohen, A., 2000. Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environ Health Perspect* 108, 419–426.

Zhao, D.H., Borders, B., Wilson, M., 2004. Individual-tree diameter growth and mortality models for bottomland mixed-species hardwood stands in the lower Mississippi alluvial valley. *Forest Ecology and Management* 199, 307-322.

CHAPTER 2. SURVIVAL REGRESSION AS AN ALTERNATIVE FOR MODELLING INDIVIDUAL TREE MORTALITY

1 Introduction

Individual tree mortality is most commonly modeled with logistic regression, but other methods have been historically used in forestry to model this binary outcome. These include the Weibull function (e.g. Somers *et al.*, 1980), negative exponential (e.g. Moser, 1972), gamma function, Richard's function (e.g. Buford and Hafley, 1985), the exponential (e.g. Kobe and Coates, 1997), and more recently, techniques such as recursive partitioning, also known as classification and regression trees (e.g. Fan *et al.*, 2006), neural networks (e.g. Guan and Gertner, 1991), and survival analysis (e.g. Woodall *et al.*, 2005a). The prevalence of logistic regression in models of individual tree mortality, among other options to model dichotomous outcomes (Hasenauer *et al.*, 2001; Rose Jr, 2002), seems to be related to its straightforward parameter interpretation (Rose *et al.*, 2006).

Event occurrence and dichotomous outcomes are also the focus of other fields, such as epidemiology and biology. In those fields, survival analysis techniques are much more commonly used than other alternatives. Two reasons that might have contributed to the lack of interest in survival analysis in forestry include: (1) the usual lack of data on tree age, that has led some authors to conclude that survival analysis techniques cannot be used (Flewelling and Monserud, 2002); and (2) the intrinsic peculiarities of most of the available data on mortality. These peculiarities include the time-dependent nature of the key variables traditionally used to predict background mortality, and the interval censored character of most mortality data.

Data in permanent plots are usually collected by revisiting the site after a period of time, such as 5 or 10 years. This yields interval censored mortality data. That is, the exact date of a tree death is not known; the only information available is that a tree died within

certain interval or that it survived at least until the end of the interval. The second characteristic that distinguishes tree mortality from usual epidemiology cases is that some of the most important forestry covariates, such as diameter, crown ratio and competition, change with time; i.e., are time-dependent variables. These characteristics, especially when they are concurrent, are uncommon in other fields, and have delayed the application of survival analysis techniques in tree mortality modelling.

1.1 Survival analysis techniques

In survival analysis, the technique focuses on modelling time until an event of interest. This approach can be used to answer questions such as:

- Which proportion of trees will survive through time t ?
- How a change in competition increases or decreases the odds of survival?
- Does thinning have an effect on mortality?
- How long does the effect of thinning on mortality last?
- Is mortality rate constant through time?

Following, I present a survey of basic ideas behind survival analysis, drawing from Harrell (2001) and Collett (1994).

One of the key characteristics of survival analysis techniques is that they can handle censored data. Censoring occurs when the exact time of the event is not known. **Left censoring** occurs when the only information available about the time of the event is that it occurred before a certain point in time; e.g. the tree was already dead in the first measurement. **Right censoring** occurs when the follow-up of the event has finished before the event has occurred; e.g. the last time the plot was revisited the tree was still alive, so there is no available information about when the tree died, it is only known that the tree was still alive at the last measurement. **Interval censoring** occurs when there is no exact information of the time of the event but it is known that it has happened within a certain time interval; e.g. the tree died between measurements. In forestry, right and interval censoring are very common. Inventories are usually performed in a periodic

manner; therefore, they yield interval censored data when trees died between measurements and are right censored data otherwise.

In survival analysis, survival is usually defined by any of three functions: the survival function, the hazard function, or the cumulative hazard function. These functions all describe, in different ways, the same distribution. The **survival function**, $S(t)$, is a decreasing function that defines the probability that a tree dies after a certain time t , and it is, therefore, bounded between 0 and 1. It can also be interpreted as the probability that the tree survives until certain time t . $F(t)$, the **cumulative distribution function**, is thus the probability that the individual dies before a certain time t .

$$S(t) = P(T > t) = 1 - P(T \leq t)$$

$$F(t) = P(T \leq t) = 1 - S(t)$$

The **hazard function**, $\lambda(t)$, is the event rate at time t conditional on survival until time t or later. In individual tree mortality terms, the hazard function $\lambda(t)$ gives the probability of a tree dying during a very small time interval, assuming that the tree has survived up to the beginning of the interval. The hazard function is especially useful in comparing the risk of death among different classes or groups. For example, one might examine whether risk varied between two different values of local competition or crown ratio.

$$\lambda(t) = \lim_{u \rightarrow 0} \frac{P(t < T \leq t + u | T > t)}{u} = \frac{F(t)dt}{S(t)} = \frac{f(t)}{S(t)}$$

The **cumulative hazard function**, $\Lambda(t)$, is the cumulative risk up to time t . This function is non-decreasing and can take any value between zero and infinity.

$$\Lambda(t) = \int_0^t \lambda(v)dv = -\log S(t)$$

$$S(t) = e^{-\Lambda(t)}$$

The simplest technique for survival analysis is the **homogeneous failure time distribution**, where there are no covariates to differentiate between subjects, but a survival distribution function is assumed. Nonparametric estimation of the survival distribution with no covariates is also possible. The **Kaplan-Meier estimator** (Kaplan and Meier, 1958) and the Altschuler-Nelson estimator (Altshuler, 1970; Aalen, 1976; Fleming and Harrington, 1984; Nelson, 2000) are the two most common estimators in this category.

Survival analysis techniques that allow for covariates are called **survival regression models**. The two most widespread survival regression models are the proportional hazards model and the accelerated failure time model. In a **proportional hazards (PH) model**, the effect of a covariate is multiplicative with respect to the hazard. The hazard function does not need to be specified if using the **Cox's semiparametric proportional hazard model** (Cox, 1972). **Accelerated failure time (AFT) models** differ from the proportional hazard models in that the predictors act multiplicatively with respect to survival time. An advantage of AFT models over PH models is that the regression parameters in an AFT models are robust towards ignored covariates, which is not the case for PH models (Lambert *et al.*, 2004). The exponential and the Weibull distributions are the only distributions that can describe both a proportional hazards model or an accelerated failure time model.

PH models can be expressed in the form:

$$h_x(t) = g(x)h_0(t)$$

where x is a vector of explanatory variables, $x = \{x_1, x_2, \dots\}$, $g(x)$ is a function of x independent of the time variable t , and $h_0(t)$ is the baseline hazard function, the hazard function for an individual for whom the values of all explanatory variables that make up the vector x are zero.

Among PH models, Cox's semiparametric model is one of the most widely used multivariate survival methods. The most important advantage of this approach is that, without knowing the baseline hazard, the coefficients for each covariate can still be calculated. Therefore, there is no assumption about the shape of the hazard function.

The Cox model is more efficient than parametric models when the assumptions about the parametric model are not true. However, in the presence of heavily interval censored data, as it is the case in most tree mortality datasets, parametric models are robust and generally more informative than their non-parametric counterparts (Lindsey, 1998).

The Weibull model is the most widely used parametric survival model. It can be used in AFT or PH approaches. If the survival time has a Weibull distribution, which has scale parameter λ and shape parameter γ , then the hazard function is such that:

$$h_0(t) = \lambda\gamma t^{\gamma-1}$$

The hazard function for the i^{th} individual has a Weibull distribution with scale parameter $\lambda e^{\beta x_i}$ and shape parameter γ :

$$h_{x,i}(t) = e^{\beta x_i} \lambda \gamma t^{\gamma-1}$$

AFT models can be expressed in the form:

$$h_x(t) = g(x)h_0(g(x)t)$$

and the AFT Weibull model has the form:

$$h_{x,i}(t) = e^{\beta x_i} \lambda \gamma (e^{\beta x_i} t)^{\gamma-1} = (e^{\beta x_i})^\gamma \lambda \gamma t^{\gamma-1}$$

so that the survival time of this individual has a Weibull function with parameters $\lambda e^{\beta x_i \gamma}$ and shape parameter γ .

1.2 Survival analysis in forestry

Examples of survival analysis techniques used to study individual tree mortality in diverse contexts can be found in the literature; however, most of them are focused on forest health related issues. For example, Preisler and Slaughter (1997) used the log-normal distribution to analyze the simultaneous effects of tree characteristics and tree locations on the survival of individual trees in stands affected by annosum root disease. Volney (1998) studied tree mortality following a jack pine budworm outbreak using a

nonparametric proportional hazards model and Cox's proportional hazard model. He and Alfaro (2000) used an accelerated failure time model to analyze resistance of white spruce to the attack of white pine weevil. More recently, survival analysis techniques were used to develop an individual tree mortality model to help detect, monitor and mitigate large-scale forest health issues (Woodall *et al.*, 2005b). This last model has the peculiarity that the driving variable is DBH increment instead of the usual variable, time.

There are also examples of forestry applications using survival analysis that are not focused on forest health related issues. Burgman *et al.* (1994) used a Cox model to develop background mortality models for mountain ash (*Eucalyptus regnans*) and alpine ash (*Eucalyptus delegatensis*) in Australia. The time scale used was the difference between the age at study entry and the age at death or present age. Rose *et al.* (2004) developed a method for deriving whole-stand survival models. Rose *et al.* (2006) used survival analysis techniques to model and predict individual tree survival using permanent plots in loblolly pine plantations. In this study, measurements were taken annually and the driving variable was age of the stand. Additionally, the Rose *et al.* (2006) model considered silvicultural treatment and tree and stand characteristics effects on survival. Fan *et al.* (2006) analyzed tree survival in oak forests using a combination of classification and regression tree and survival analysis, using the nonparametric Kaplan-Meier estimator.

All of the above models, except for the Woodall *et al.* (2005a; 2005b) and Burgman *et al.* (1994) models, use age or time from a certain event, such an insect outbreak, as the driving variable. Unfortunately, when background mortality is modelled in uneven aged stands neither of these direct measures of time can be used. A more unconventional approach was suggested by Woodall *et al.* (2005a), who used DBH as the driving variable instead. Woodall *et al.* noted the high potential that survival analysis offers and that survival analysis techniques have not been used extensively in forestry due to the dearth of detailed time and age data for large-scale inventories. However, traditional logistic regression models assume constant hazard, that is, if we assume that all important covariates are included in the model, mortality rate is constant in time. If one makes a similar assumption for survival regression models, an initial point of time can be defined arbitrarily, similarly to logistic regression. For example, we can define as time

zero the first time the plot was measured, i.e. the first measurement, and consider time as the driving variable, similarly to Burgman et al. (1994).

Most mortality models and studies using survival analysis techniques do not focus on background individual tree mortality, and the ones that do are not applicable to uneven age stands or do not address the interval censored characteristics of the dataset and the time-dependent nature of most of the variables. In fact, from the aforementioned models only Rose et al. (2006) addressed interval censoring and the time-dependent nature of the variables. Most surprising is that, although the advantages of survival analysis techniques to analyze large-scale inventory tree mortality are apparent (Woodall *et al.*, 2005a; Woodall *et al.*, 2005b), there is no comparison in the literature between survival regression models and the traditional logistic regression approach.

1.3 Logistic regression versus survival analysis

Most fundamental covariates in individual tree mortality models are time-dependent (e.g. DBH, PBA, PBAL, SBA, CR). However, conventional logistic regression does not allow the inclusion of time-dependent covariates; instead, the values of the covariates at the beginning of the interval are used and assumed to be constant through the period. Leffondré et al. (2003) evaluated logistic regression in matched-control data with time-dependent covariates and observed that logistic regression estimates were less accurate in some scenarios, especially when the covariates were correlated. In their study the logistic regression estimates were always very close to the true parameters for the fixed covariates, but time-dependent covariate parameters were over-estimated in some cases and under-estimated in others. Therefore, one would expect that taking into account time-dependent information could help to more accurately assess the impact of the covariates in tree mortality.

Since logistic regression cannot handle time-dependent variables, dealing with more than 2 measurements per tree poses some difficulties. In logistic regression the problem can be addressed in more than one way, none of them ideal. First, one can just ignore all measurements but two consecutive measurements. In this way we are ignoring expensive available mortality data and the always small number of dead trees would be

smaller than it could potentially be, wasting existing information. Conversely, the estimates would be unbiased. Second, we can consider only the first and last measurements, ignoring the change in time-dependent covariates. In this case since we are using values from the first measurement and ignoring any change in between, our conclusions might be biased. Third, we could consider the first two consecutive measurements for live trees, and for dead trees at any measurement we consider the last measurement the tree was seen alive and the measurement it was seen for the first time dead. In this case, the bias from using not updated covariates will be smaller but we might be overestimating the mortality rate because we are ignoring that some dead trees were observed for more than one period. For example, if it is known that the tree survived 15 years after the first measurement and it died sometime between the 15th and 20th year after the first observation, and we only take into account the last 5 years (15th – 20th years), we are overestimating the mortality rate. Fourth, we consider two non-consecutive measurements, similarly to the treatment of the data for chapter 1. With this approach we would be discarding some of the costly mortality data, but less than with the first option, and we would be ignoring changes in the time-dependent variables. These problems would be solved if we can use an approach that accounts for time-dependent and interval-censored data.

There are no practical and comparable evaluations in the literature, to my knowledge, between logistic regression and AFT survival regression models. However, there are a few papers on comparability of logistic regression and PH models. The main conclusion that all of these papers reach is that, when the follow-up period is short and the event is rare, proportional hazard models, mainly Cox models, and logistic regression yield similar solutions (Green and Symons, 1983; Annesi *et al.*, 1989; Callas *et al.*, 1998).

Survival analysis techniques not only have the potential to handle time-dependent and interval censored data, but they allow testing the assumption of a constant hazard function and modelling changing hazard functions (Collett, 2003). In a similar way that spatial analysis of prediction residuals is a versatile tool to study the effects of missing or incorrectly modeled variables in the prediction of diameter or mortality (Froese and Robinson, 2007), survival analysis techniques can be a strategic tool to study defining

variables that change with time. For example, effects of droughts, winds, thinning, unusual high/low temperature periods, or el niño/la niña effect.

The potential benefits from the use of survival analysis techniques to model individual tree mortality models seem substantial, but it first has to be proven that the survival analysis approach is able to perform at least at the same level as the traditional logistic regression approach.

2 Objectives

There is no comparison in the literature that I am aware of between survival regression models and the traditional logistic regression approach for modelling individual tree mortality. Given that survival analysis has theoretical benefits in this context and has shown promise in some studies, the goal of this chapter was to undertake such a comparison. Comprising this goal, my objectives were to:

- 1) assess the differences between the logistic regression and the survival regression approach by comparing the fit of both models,
- 2) test the hypothesis of constant hazard in the fitting dataset,
- 3) compare the predictive ability of both approaches when applied to an independent validation dataset.

3 Methodology

3.1 Approach

The general strategy to compare the traditional logistic approach to the survival regression approach is to compare the naïve mortality models fitted in chapter 1 with new survival regression models fitted using the same original dataset and variables. Because of the difference in the methods the dataset used to fit the logistic regression is a subset of the set used to fit the survival regression. The difference between the two datasets is that in the survival regression set all available data for each species will be used; that is, any tree with at least 2 measurements will be part of the calibration dataset.

Various semiparametric and parametric models for interval-censored data allow for time-dependent covariates. However, none of these have yet been implemented in R (R Development Core Team, 2008). I have selected a method based on Odell et al. (1992) that allows for fixed and time-dependent covariates, left, right, interval and non-censored events. This method is flexible and possesses a direct computational solution, and is described in Sparling et al. (2006).

Data from the first measurement included dead trees. In theory, these dead trees could be included in the model as left censored data, since the Sparling et al. (2006) methodology allows for this type of censoring. However, when the model includes time-dependent variables, as it is this case here, it requires data for an initial time when it is known that the individual was event free, e.g. the tree was alive. Given that such information was unavailable, dead trees at the first measurement are excluded from the analysis. In the case of the survival analysis model all trees were assumed to enter the study at time zero, even though the actual calendar year might vary by stand.

As a measure of overall predictive ability I chose the generalized R^2_N (Cragg and Uhler, 1970; Nagelkerke, 1991), and as a measure of predictive discrimination I chose the c-

index. In order to be able to compare fairly both approaches, I calculated these indices with respect to their ability to predict mortality, and not survival.

The models' ability to predict mortality was tested on an independent validation dataset (described in chapter 1). The performance of the models was assessed with measures of discrimination and unreliability described in chapter 1.

3.2 Sparling methodology

Sparling et al. (2006) provide the likelihood function, hazard and cumulative hazard function, and a section on model diagnostics that include the deviance statistic. As a brief description of the model I present the likelihood function used to fit the models, the hazard function, and the cumulative hazard function, used to calculate the predictions.

The likelihood function is defined as:

$$L = \prod_{i=1}^n \left\{ f_i(t_i | z_i, Y_{i[(t_i)]})^{\delta_{E_i}} \times F_i(t_i | z_i, Y_{i[(t_i)]})^{\delta_{L_i}} \times (1 - F_i(t_i | z_i, Y_{i[(t_i)]}))^{\delta_{R_i}} \times \left[F_i(t_{R_i} | z_i, Y_{i[(t_{R_i})]}) - F_i(t_{L_i} | z_i, Y_{i[(t_{L_i})]}) \right]^{\delta_{I_i}} \right\}$$

Where

$\delta_{R_i} = 1$ if right censored at time $t_i < x_i$, 0 otherwise

$\delta_{L_i} = 1$ if left censored at time $t_i > x_i$, 0 otherwise

$\delta_{I_i} = 1$ if interval censored with $t_{L_i} < x_i \leq t_{R_i}$, 0 otherwise

$\delta_{E_i} = 1$ if has an event observed exactly at time $t_i = x_i$, 0 otherwise

z_i is a vector of p fixed covariates for the i th subject

$Y_{i[t]}$ denotes the sequence of time-dependent covariate values up to time t for the i th subject.

The general form of the hazard function is:

$$\lambda(\tau_{ij}|z_i, y_{ij}) = \frac{\alpha \beta_{ij} \tau_{ij}^{\alpha-1}}{(1 + \beta_{ij} \tau_{ij})^k}$$

where $\beta_{ij} = e^{\theta + \gamma' z_i + \eta' y_{ij}}$ $\beta_{ij} = e^{\theta + \gamma' z_i + \eta' y_{ij}}$, and $\alpha > 0$ and k are general hazard function parameters. In particular, $k = 0$ yields a Weibull hazard for event times, $k = 1$ yields a log-logistic hazard, $\alpha = 1$ and $k = 0$ yields a negative binomial distribution. The hazard function for $k = 0$ is constant at $\alpha = 1$, increasing for $\alpha > 1$, and decreasing $0 < \alpha < 1$.

The survival function is calculated as

$$S(t) = e^{-\Lambda(t)}, \text{ where}$$

$$\Lambda(t) = \sum_{j=0}^{k_i-1} I_{\{\tau_{i(j+1)} \leq t\}} [\mu_{ij}[\tau_{i(j+1)}] - \mu_{ij}[\tau_{ij}]]$$

and at time u $\mu_{ij}(u) = \frac{[1 + \beta_{ij} u^\alpha]^{1-\kappa}}{1-\kappa}$ and for $\kappa \neq 1$, and for $\kappa = 1$ $\mu_{ij}(u)_{\kappa=1} = \ln [1 + \beta_{ij} u^\alpha]$

The optimal value of κ can be estimated from the data, but if the event rate is low there is inadequate information about the true shape of the hazard function to allow precise estimation of κ . Recall that $\kappa = 0$ yields a Weibull hazard, and that for $\kappa = 0$ and $\alpha = 1$ the hazard function is constant. Since we are trying to model background mortality we are assuming constant hazard, therefore I selected the Weibull constant hazard form ($\kappa = 0$, $\alpha = 1$), with hazard function:

$$\lambda(\tau_{ij}|z_i, y_{ij}) = \frac{1 \beta_{ij} \tau_{ij}^{1-1}}{(1 + \beta_{ij} \tau_{ij})^0} = \beta_{ij} = e^{\theta + \gamma' z_i + \eta' y_{ij}}$$

For the statistical significance of the estimated parameters of the covariates the Wald test was used. Usually, the likelihood ratio test is preferred to the Wald test because the latter is sensitive to problems in the estimated variance-covariance matrix in the full model (Harrell, 2001). However, because the survival regression models fitting process consume a non trivial amount of time and the likelihood ratio test would require fitting

each model at least four times (one for each variable, one for the intercept and one the full model), the Wald test was considered a sufficient approximation to the true significance test. The Wald test was calculated following Sparling et al. (2006) and the original code in SAS written by the authors. The Wald test to test $\beta = \beta_0$ is defined as:

$$W = \frac{(\beta - \beta_0)^2}{var(\beta)}$$

Where β is the estimated parameter and β_0 is the hypothesized value.

3.3 Software

Data analysis, simulations and graphs for this chapter were done in R environment (R Development Core Team, 2008). Data manipulation was carried out in MySQL (2008) on the Microsoft WindowsXP platform.

Since no package was available in the R environment to fit the proposed model, I followed the Sparling et al. (2006) methodology and the original SAS code the authors made publically available, to code the algorithms in R and to fit the models. One of the challenges of writing the code to optimize non-standard functions is to test that the code is error free. Unfortunately the data used by Sparling et al. (2006) were not publicly available, so an alternative to test the accuracy of the code was needed. The approach I used was to contrast the gradient function at several different points with that computed numerically by the *grad* function in *numDeriv* library version 2006.4.1 (Gilbert, 2008). The likelihood was compared to the likelihood obtain by fitting a similar equation, but ignoring the time-dependent nature of the variables, with the *survreg* function in *survival* library version 2.34-1 (Therneau and Lumley, 2008).

4 Data

The calibration dataset used to fit the models in chapter 1 had a maximum of 4 measurements per tree. For the logistic regression, as previously discussed in chapter 1, only measurements 1-3 or 2-4 for each plot were selected. The objective for choosing these remeasurements in the logistic regression was to select similar intervals long enough to include enough mortality data but short enough that the variables wouldn't change significantly. For the survival regression calibration dataset there are no constraints on the length or number of periods considered. Therefore, all available trees with at least 2 measurements (i.e. 1 remeasurement period) were included in the survival calibration dataset.

The differences in the ability of both approaches to handle time-dependent variables resulted, therefore, in differences between calibration datasets. These differences are obvious (Table 2-) in both the total number of trees and the total number of dead trees. The number of trees available for each species in the survival calibration dataset is more than 50% larger than the available trees for the correspondent species in the logistic calibration dataset. In contrast, the proportion of dead trees in the logistic calibration dataset for the three species is always larger than the correspondent survival dataset. Differences in the average period of observation for trees of each species do not vary substantially, although it tends to be higher in the survival dataset. Differences in mortality percentage between both datasets are only negligible for *P. contorta*.

Similarly to chapter 1, only trees and stands where both SBA and DBH were less than the 99th percentile of the same variables in the calibration set were reserved. The final validation dataset differed from the one used in chapter 1 in that the 99th percentile was calculated from the survival calibration set instead of the logistic calibration set.

Table 2-1. Summary of the three datasets: total number of trees, observed number of dead trees, percentage of dead trees, and mean observation period in years.

		<i>A. grandis</i>	<i>P. contorta</i>	<i>P. menziesii</i>
Calibration	Total	3,641	12,971	8,732
	Dead	250	552	470
	Logistic Dead (%)	6.9	4.3	5.4
	Mean obs. Period	11.1	10.9	10.8
Calibration	Total	7,114	20,322	13,617
	Survival Dead	309	830	638
	Survival Dead (%)	4.3	4.1	4.7
	Mean obs. period	11.1	11.6	11.4
Validation	Total	1,062	1,290	1,914
	Dead	135	317	213
	Dead (%)	12.7	24.6	11.1
	Mean obs. period	6.4	6.4	9.1

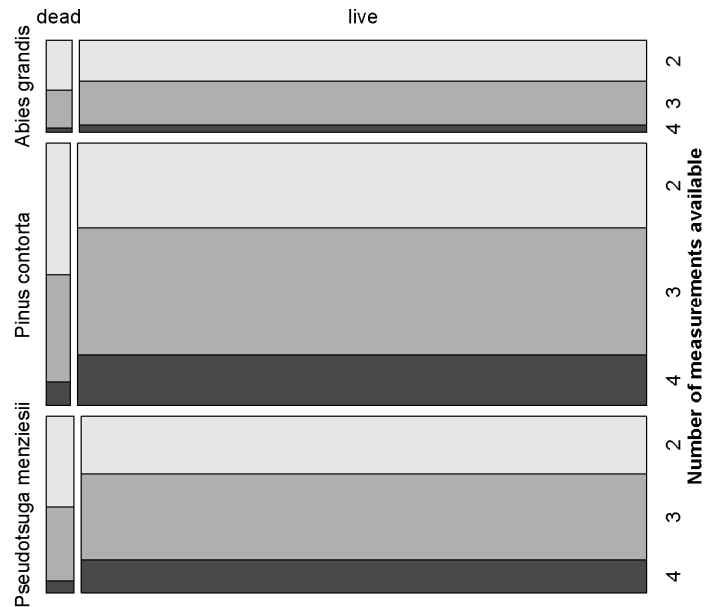


Figure 2-1. A graphical representation of the size and attributes of the calibration dataset. The area of each bar is proportional to the number of trees in the category.

The proportion of trees with more than 3 remeasurements is low (Figure 2-1), especially in the case of *A. grandis*.

The distribution of the main variables for the three datasets is presented in Figure 2-1 and Figure 2-2. The dataset used for validation differs substantially from the datasets used for calibration in the distribution of all the variables. The divergence between both

calibration datasets is in general larger for both competition variables than for DBH and CR.

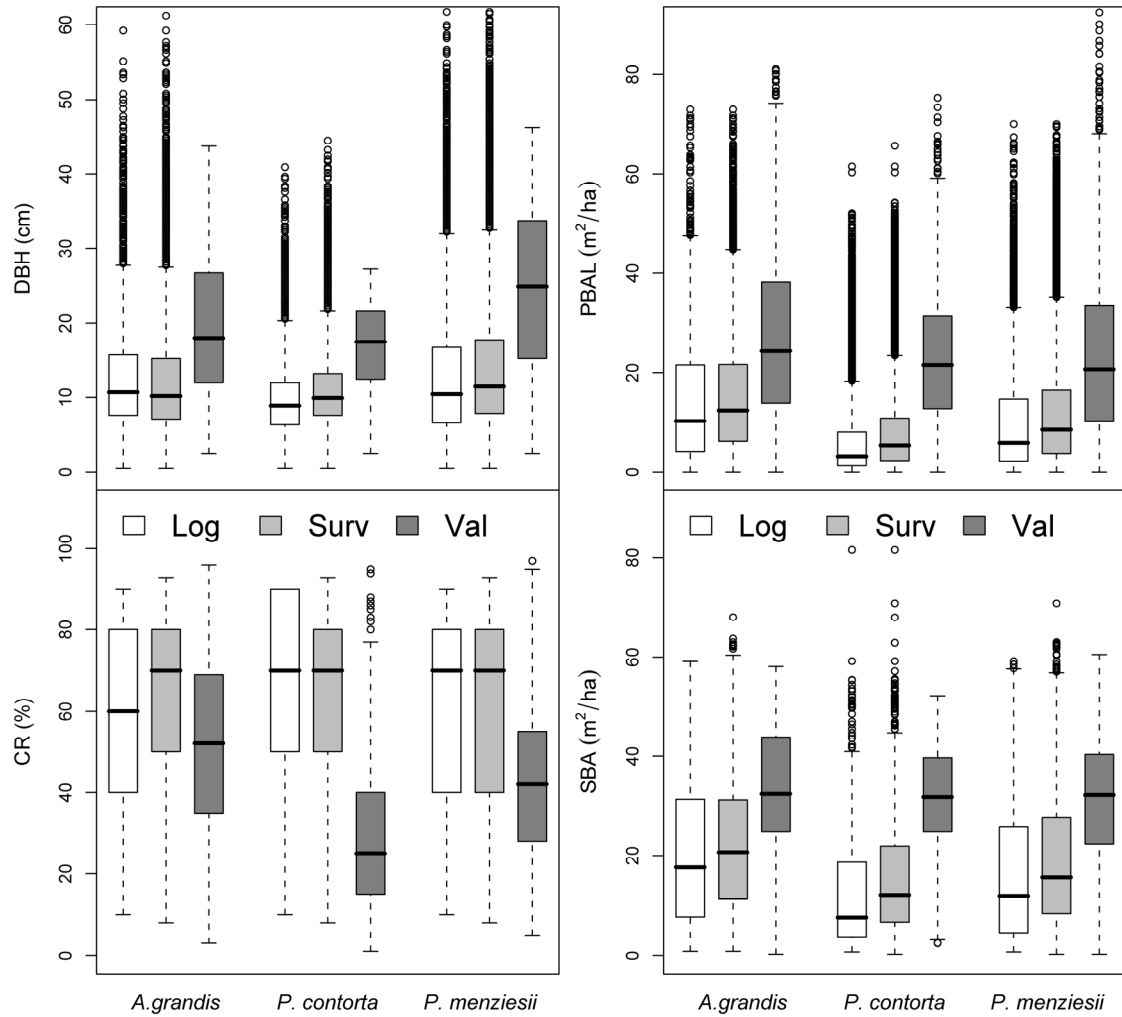


Figure 2-2. Distribution of the variables for the three datasets. For the survival dataset each measurement of each tree used to fit the model has been considered a separate measurement.

5 Results

5.1 Fitting phase

Direct comparison of estimated parameters (Table 2-) for the survival regression and logistic regression approaches is not possible since they are not parameterized using similar equations. Nevertheless, it is possible to compare both approaches by studying how they differ in their reactions to changes in the covariates (Figure 2-3). The largest differences between survival and logistic models appear for the case where CR and PBAL are small. In general, the logistic models predict lower probability of mortality for trees with small CR than do survival models, but predict higher mortality for suppressed trees with CR 40% or larger. Logistic regression models appear to be more sensitive to PBAL than survival models for *P. contorta* and *P. menziesii* and less sensitive for *A. grandis*.

	LOGISTIC			SURVIVAL		
	Parameter	Std. Err.	p-value ¹	Parameter	Std. Err.	p-value ²
	<i>Abies grandis</i>			<i>Abies grandis</i>		
Intercept	-1.4076	0.3798	0.001	-1.0137	0.3336	0.002
PBAL	-0.0095	0.0053	0.087	-0.0181	0.0052	0.001
CR ^{1/2}	-1.5808	0.1525	<0.001	-1.7381	0.1222	<0.001
	<i>Pinus contorta</i>			<i>Pinus contorta</i>		
Intercept	-3.5446	0.3249	<0.001	-2.7577	0.2332	<0.001
PBAL	0.0461	0.0045	<0.001	0.0347	0.0035	<0.001
CR ^{1/2}	-1.0763	0.1265	<0.001	-1.4074	0.0835	<0.001
	<i>Pseudotsuga menziesii</i>			<i>Pseudotsuga menziesii</i>		
Intercept	-3.7796	0.3368	<0.001	-2.9964	0.2620	<0.001
PBAL	0.0189	0.0047	0.001	0.0130	0.0036	<0.001
CR ^{1/2}	-0.7485	0.1297	<0.001	-1.1151	0.0937	<0.001

¹ Likelihood ratio test; ² Wald test

Table 2-2. Summary of the models for the logistic regression and survival regression approaches.

The statistical significance of the estimated parameters is similar in both approaches. There is only one exception; PBAL for *A. grandis* is not significant ($\alpha=0.05$) for the logistic model but is significant for the survival model. Notice that the p-values correspond to different significance tests.

Another way to evaluate the candidate models is with respect to measures of fitness of the models to the calibration datasets. While differences in deviance are not directly comparable due the differences in the datasets used to fit the models, the c-index and the R^2_N can give an idea of the explanatory ability of the models. The R^2_N is always larger in the logistic models, while the relative difference in c-index varies across species.

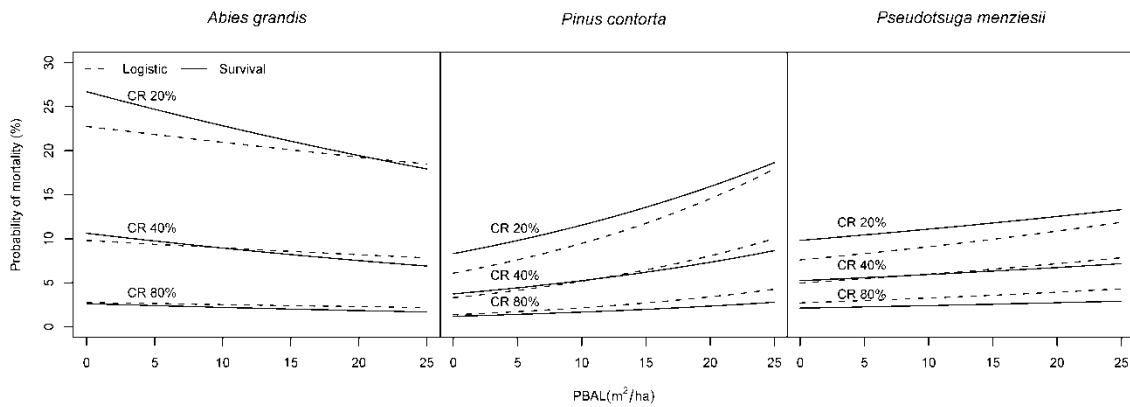


Figure 2-3. Sensitivity plots. Predictions for 10 years.

Species		R^2_N	c-index
<i>Abies grandis</i>	Logistic	0.698	0.094
	Survival	0.675	0.095
<i>Pinus contorta</i>	Logistic	0.759	0.179
	Survival	0.712	0.163
<i>Pseudotsuga menziesii</i>	Logistic	0.642	0.054
	Survival	0.597	0.076

Table 2-3. Measures of predictive power for the calibration data.

All survival regression models have been fitted assuming constant hazard function ($\alpha = 1$); that is, it is explicitly assumed that the hazard does not change with time. Actually, this assumption can be tested and results from this test are shown in Table 2-. Using a 95% significance level, the null hypothesis could not be rejected for *A. grandis* and *P. menziesii*, but the test for *P. contorta* rejected the null hypothesis, suggesting increasing hazard with time.

Species	α	p-value
<i>Abies grandis</i>	1.08	0.329
<i>Pinus contorta</i>	1.10	0.039
<i>Pseudotsuga menziesii</i>	0.93	0.160

Table 2-4. Likelihood ratio test where H_0 : no change in mortality rate with time.

5.2 Validation phase

The ultimate goal of individual tree mortality equations in the context of growth and yield models is to accurately predict mortality in a variety of stands. Hence, the most adequate approach to validate a model is to test its performance in an independent dataset. When survival regression models were tested in the independent validation dataset described in chapter 1 the results, when compared to those of logistic regression models, vary by species, but, overall, survival models were able to better discriminate alive and dead trees, and explained more variability (Table 2-). The differences were always larger for *P. contorta* and *P. mensiezii* than for *A. grandis*. Measures of reliability were better for logistic regression models than for survival in *A. grandis* and *P. mensiezii*, but worse for *P. contorta*.

Species	Model	R^2_N	Reliability		Discrimination		
			Intercept	U	Slope	D-index	c-index
<i>Abies grandis</i>	Logistic	0.220	1.165	0.106	0.983	0.124	0.788
	Survival	0.225	1.246	0.131	0.958	0.127	0.793
<i>Pinus contorta</i>	Logistic	0.111	-0.150	0.131	0.488	0.077	0.668
	Survival	0.121	-0.125	0.127	0.501	0.084	0.677
<i>Pseudotsuga menziesii</i>	Logistic	0.094	0.011	0.014	0.832	0.048	0.689
	Survival	0.109	0.044	0.017	0.827	0.056	0.701

Table 2-5. Measures of overall explanatory power (R^2_N), reliability and discrimination for the four models applied to the validation dataset.

When predictions for logistic and regression models are compared along variables differences between both models are small for CR, but meaningful for PBAL. The largest differences are found in *P. contorta* and *P. menziesii* (Figure 2-4). Overall, the kernel

density estimates of the survival models more closely follow the observed mortality than the logistic regression models (Figure 2-4b).

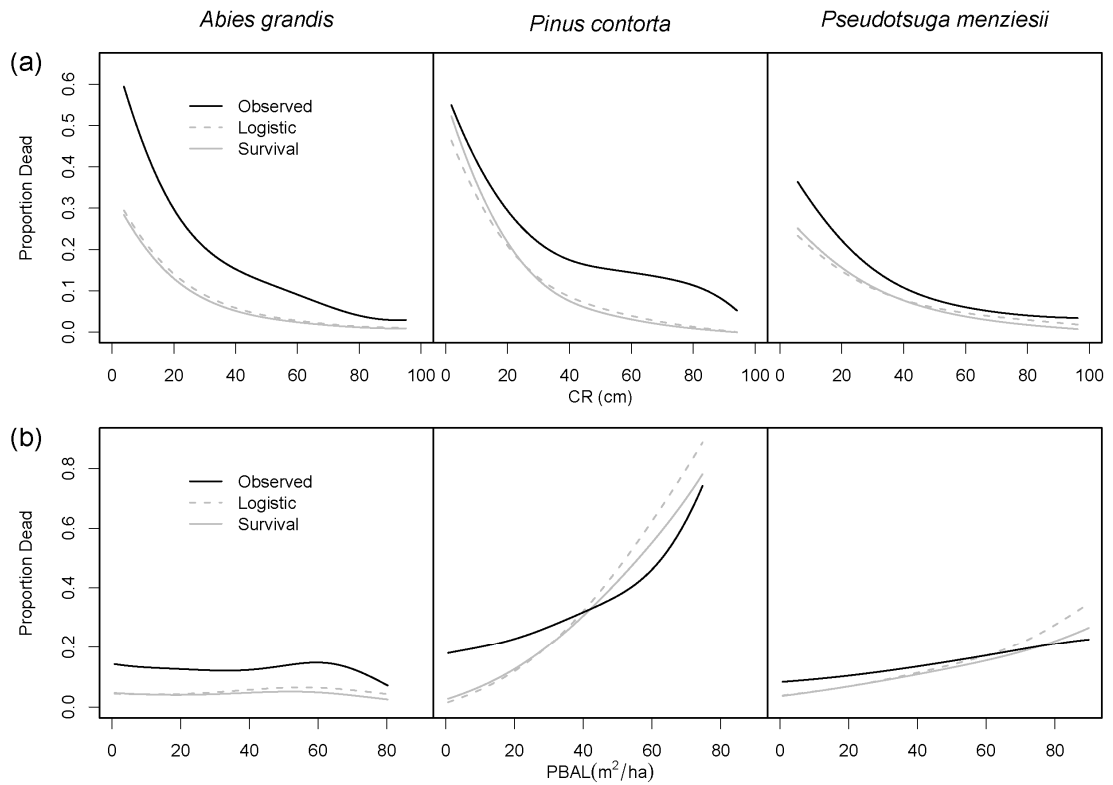


Figure 2-4. Kernel density estimates of observed and predicted mortality for the survival and logistic models in the validation dataset.

Suppressed trees with PBAL over 40 have the largest differences in predicted mortality shown between survival and logistic models. For these trees the logistic model generally predicts higher mortality than the survival model.

6 Discussion

6.1 *Data and model form*

One of the multiple challenges that individual tree mortality modelers face is the scarcity of quality data, if not data in general. Survival regression, an alternative to the traditional approach using logistic regression, maximizes the use of the information in the available datasets because there is no restriction on the length or number of observation periods it can handle, and because it updates the values of the covariates every time new information is available. The original logistic model implicitly assumes that the remeasurement periods are similar or that the length of the period does not affect the outcome (Callas *et al.*, 1998), which is not usually the case when tree mortality is being modelled. Ingram and Kleinman (1989) cautioned against using logistic regression when remeasurement periods vary because it can seriously bias estimates, and preliminary results of this study that included consecutive and alternate remeasurement periods resulted in inferior model fit and prediction accuracy (results not shown), supporting Ingram and Kleinman's (1989) opinion.

Logistic regression models that use remeasurement periods larger than 1 year to model annual mortality assume constant probability of mortality within the period studied. Cao (2000) argued that this assumption is not correct because the probability varies annually with changes in stand and tree attributes, and proposed an iterative method to calculate annual diameter growth and survival that accounts for yearly changes in stand and tree characteristics. In his study, he showed that the iterative method yields better mean relative and absolute deviation and better log-likelihood values than the traditional method, although he did not validate his results using an independent dataset. Nevertheless, his results, corroborated by Flewelling and Monserud (2002), and theoretical arguments seem robust enough to deserve further analysis in the survival regression context. Cao's survival logistic equations only considered age, basal area and diameter as predictors, and these latter two predictors were themselves based on a diameter increment equation calibrated using the diameter at the beginning and at the end of the period considered. It is unclear if the author did or did not consider the same

tree in each interval as a separate observation. In the survival regression context Cao's approach would cause incongruities since the diameter predicted at time $t+a$ with an equation based on the values of the variable at time t will not match the observed value of the variable at time $t+a$. This does not necessarily poses a problem towards the quality of the fitted model, and the potential improvement in the model might very well overcome any initial drawback. A possible alternative approach for model fitting is to interpolate the values of the variable between the two measurements. Fortunately the selected survival regression approach for this chapter allows for updates in time-dependent variables without updates in the status of the tree.

The Weibull form for the survival regression selected for this chapter has several advantages and one main possible drawback. The Weibull survival equation is one of the most widely used parametric functions in survival analysis, it can describe constant and monotone hazard functions, and it can be interpreted as a PH or AFT model. I consider all of these characteristics an advantage. However, the Weibull survival regression cannot model non-monotone, e.g. bathtub shape, hazard functions. This type of hazard functions would not likely appear when modelling non-background mortality, because the effect of a fire, heavy winds, etc, is likely to result in high initial hazard that decreases with time. However, when exploring the possible effects of time-dependent variables in background mortality it might be useful to have more flexibility in the shape of the hazard function. For example, if one is studying the effect of a severe 5-year drought, it might be interesting to study the background mortality before and after the drought at the same time, and examine the recovery curve in mortality after the drought. The methodology described and the code written for this chapter are fairly flexible and would allow modelling of non-monotone shapes of the hazard function without further development efforts.

6.2 Fitting phase

Differences in the significance of the PBAL parameters for *A. grandis* can have two origins. First, the p-value for the logistic model is not very large, and the increase in power due to the major increase in total number of observed trees and dead trees could cause the p-value to get close to the p-value for the survival model. The second possible

cause of differences is the differences between the log-likelihood ratio test and the Wald test already pointed out in the methodology section. As noted by Harrell (2001) the Wald test is sensitive to problems in the estimated variance-covariance matrix in the full model, and to the way the parameter appears in the model.

An interesting possibility of survival analysis is to test the hypothesis of constant hazard functions. When I tested this hypothesis for the survival models (Table 2-) two out of the three species studied could be assumed to have constant hazard. *P. contorta*, in contrast, was found to have increasing mortality with time at $\alpha = 0.05$. In the context of background mortality models, and taking into account that the year of establishment of the plot (year zero) is potentially different for each stand, this might indicate the need to study the possible leverage of one or more stands with 4 remeasurements. If I was studying the effect of a treatment in stands where all data have the same starting year (e.g. 1989), this would indicate that the treatment had a statistically significant negative result on the survival probability of *P. contorta* trees.

6.3 Validation phase

Overall, the survival regression approach results in a consistent improvement of the discrimination and overall explanatory power of the models that is encouraging. Notably, survival regression models better allocate tree mortality, assigning more probability of mortality to trees that actually die than the logistic regression models. Correctly allocating mortality among trees is important because the size and social position of the dead trees influence the variables that define stand structure and dynamics. The death of a tree directly reduces density measures, and affects the social position of the remaining trees, which in turn defines diameter growth, potential regeneration and probability of mortality of the residual trees. Thus, the correct allocation of tree mortality of the survival regression models results in models that produce more realistic stand structure and dynamics.

Reliability measures improved for *P. contorta*, but worsen for *A. grandis* and *P. menziesii*. These results are in concordance with differences observed in mortality rates in the calibration datasets. Species with the highest differences in reliability also had the

highest difference in mortality ratio between the two datasets. For *P. contorta*, the only species with similar mortality rate in both datasets, reliability improves using the survival model, suggesting that with similar mortality rates in the other two species reliability might as well improve in survival models. This high sensitivity of empirical mortality models to the characteristics of the calibration dataset resulting in temporal and spatial inflexibility is one of the arguments made against empirical mortality models (Hawkes, 2000). Some authors have suggested a shift in emphasis toward modelling for exploration and explanation, rather than to predict with high levels of accuracy as a tool to overcome the problems in mortality modelling related with the poor understanding of tree mortality processes and the chronic shortage of data (Hawkes, 2000). Survival regression would be a valuable tool on both fronts since, as presented in the introduction and methodology sections of this chapter, it uses more efficiently the data available (Ingram and Kleinman, 1989) and is also a catalyst for the study of environmental time-dependent variables.

At first, it might be disconcerting to see that the slopes decrease but the D-index and c-index increase. This actually has an easy explanation. Recall from chapter 1 that the D-index is a likelihood ratio test to test the hypothesis that the coefficient is statistically different from zero; that is, is an indicator of the explanatory power of the model. Even with a higher slope, the explanatory power of one model can be lower than the explanatory power of the other model if the variability of the predictions around the observed values is high. This concept is easier to describe in the simple linear case ($y = a+bx$). Figure 2-5 reproduces the data and the fitted line for two simple linear models. On the left panel the slope is higher, but the variability around the predicted values is also higher than in the right panel, yielding higher slope but lower R^2 . In the left panel, x values correspond better on average to the actual y variables, but for a similar predicted x value the corresponding y can vary markedly. The ability of the model to explain variability in the dependent variable is better in the model represented on the right panel, a higher x corresponds more often with higher y , although values of x do not correspond necessary with similar magnitude of y .

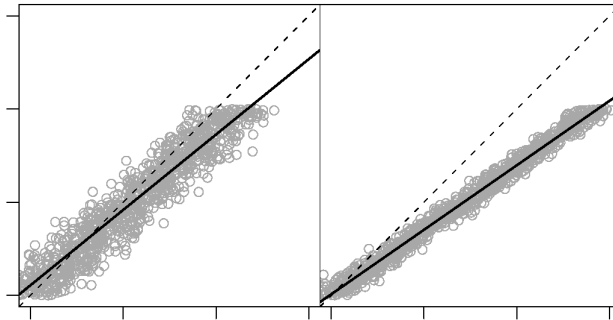


Figure 2-5. An illustration for the simple linear regression case of higher explanatory power with lower slope. The linear regression model in the left panel (bold line) has a slope of 0.81 and R^2 of 0.91, while the model on the right panel (bold line) has a slope of 0.69 and R^2 of 0.98.

One of the main objectives of this chapter was to compare the performance of survival analysis against the traditional logistic regression approach to model background mortality. In this chapter it has been shown that survival analysis models have the capacity to outperform the traditional logistic regression approach, leaving the door open to the study of the influence of other time-dependent variables or events in non catastrophic mortality. The methodology presented is able to model a relatively wide variety of hazard functions, including Weibull and log-logistic hazard, and a negative binomial distribution of event times. This flexibility is a strength of the proposed methodology in terms of exploring changes in background mortality related with time-dependent variables. Further research on this study area should include the implementation of a similar approach to Cao's (2000) iterative method adjusted to survival regression.

7 Conclusions

The potential advantages of survival regression models to model individual tree mortality have been reported in several studies (e.g. Burgman *et al.*, 1994; Rose Jr, 2002; Rose Jr *et al.*, 2004; Woodall *et al.*, 2005a; Woodall *et al.*, 2005b; Rose *et al.*, 2006).

However, to my knowledge there has not been a survival regression individual tree background mortality model that addresses the time-dependent nature of most of the covariates and the interval-censored character of tree death events, and that is applicable to uneven age mixed-species stands. In this chapter a survival regression background mortality model for individual trees that allows for time-dependent variables and interval censored events has been fitted to a relatively broad geographic area dataset and it has been compared to the traditional logistic regression approach. The empirical comparison of both approaches has proved that the survival regression model is superior, in terms of overall explanatory power, and discrimination, to the logistic regression approach to model individual tree background mortality.

8 References

Sun Microsystems, 2008. MySQL 5.0.

Aalen, O.O., 1976. Nonparametric inference in connection with multiple decrement models. *Scandinavian Journal of Statistics* 3, 15-27.

Altshuler, B., 1970. Theory for the measurement of competing risks in animal experiments. *Mathematical Biosciences* 6, 11-11.

Annesi, I., Moreau, T., Lellouch, J., 1989. Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies. *Statistics in Medicine* 8, 1515-1521.

Buford, M.A., Hafley, W.L., 1985. Modeling the probability of individual tree mortality. *Forest Science* 31, 331-341.

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

Callas, P.W., Pastides, H., Hosmer, D.W., 1998. Empirical comparisons of proportional hazards, poisson, and logistic regression modeling of occupational cohort data. *American Journal of Industrial Medicine* 33, 33-47.

Cao, Q.V., 2000. Prediction of annual diameter growth and survival for individual trees from periodic measurements. *Forest Science* 46, 127.

Collett, D., 1994. *Modeling Survival Data in Medical Research*. Chapman and Hall/CRC Press, London.

Collett, D., 2003. *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC Press, London.

Cox, D.R., 1972. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B* 34, 187-220.

Cragg, J.G., Uhler, R., 1970. The demand for automobiles. *Canadian Journal of Economics* 3, 386-406.

Fan, Z., Kabrick, J.M., Shifley, S.R., 2006. Classification and regression tree based survival analysis in oak-dominated forests of Missouri's Ozark highlands. *Canadian Journal of Forest Research* 36, 1740-1748.

Fleming, T.R., Harrington, D.P., 1984. Nonparametric estimation of the survival distribution in censored data. *Communications in Statistics-Theory and Methods* 13, 2469-2486.

Flewelling, J.W., Monserud, R.A., 2002. Comparing methods for modelling tree mortality. USDA For. Serv. Fort Collins, USA: Rocky Mountain Research Station., Fort Collins, Colorado, USA, 12-14 February 2002, pp. 168-177.

-
- Froese, R.E., Robinson, A.P., 2007. A validation and evaluation of the Prognosis individual-tree basal area increment model. *Canadian Journal of Forest Research* 37, 1438-1449.
- Gilbert, P., 2008. numDeriv, version: 2006.4-1. Available at: <http://cran.r-project.org/>.
- Green, M.S., Symons, M.J., 1983. A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. *The Journal of Chronic Diseases* 36, 715-723.
- Guan, B.T., Gertner, G., 1991. Modeling Red Pine Tree Survival with an Artificial Neural Network. *Forest Science* 37, 1429-1440.
- Harrell, F.E., 2001. Regression modelling strategies with applications to linear models, logistic regression, and survival analysis. Springer, New York, NY.
- Hasenauer, H., Merkl, D., Weingartner, M., 2001. Estimating tree mortality of Norway spruce stands with neural networks. *Advances in Environmental Research* 5, 405-414.
- Hawkes, C., 2000. Woody plant mortality algorithms: description, problems and progress. *Ecological Modelling* 126, 225-248.
- He, F., Alfaro, R.I., 2000. White pine weevil attack on white spruce: a survival time analysis. *Ecological Applications* 10, 225-232.
- Ingram, D.D., Kleinman, J.C., 1989. Empirical comparisons of proportional hazards and logistic regression models. *Statistics in Medicine* 8, 525-538.
- Kaplan, E.L., Meier, P., 1958. Nonparametric estimation from incomplete observations. *Journal of American Statistical Association* 53, 457-481.
- Kobe, R.K., 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, 227-236.
- Lambert, P., Collett, D., Kimber, A., Johnson, R., 2004. Parametric accelerated failure time models with random effects and an application to kidney transplant survival. *Statistics in Medicine* 23, 3177-3192.
- Leffondré, K., Abrahamowicz, M., 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, 3781-3794.
- Lindsey, 1998. A Study of Interval Censoring in Parametric Regression Models. *Lifetime Data Analysis* 4, 329-354.
- Moser, J.W., 1972. Dynamics of an uneven-aged forest stand. *Forest Science* 18, 184-191.
- Nagelkerke, N.J.D., 1991. A note on a general definition of the coefficient of determination. *Biometrika* 78, 691-692.
- Nelson, W., 2000. Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics* 42, 12-25.

Odell, P.M., Anderson, K.M., D'Agostino, R.B., 1992. Maximum Likelihood Estimation for Interval-Censored Data Using a Weibull- Based Accelerated Failure Time Model. *Biometrics* 48, 951-959.

Preisler, H.K., Slaughter, G.W., 1997. A stochastic model for tree survival in stands affected by annosum root disease. *Forest Science* 43, 78-86.

R Development Core Team, 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

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

Rose Jr, C.E., 2002. Modeling and allocating forestry survival: a loblolly pine case study. Univ. Georgia. Dissertation. 176 pp.

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

Somers, G.L., Oderwald, R.G., Harms, W.R., Langdon, O.G., 1980. Predicting mortality with a Weibull distribution. *Forest Science* 26, 291-300.

Sparling, Y.H., Younes, N., Lachin, J.M., Bautista, O.M., 2006. Parametric survival models for interval-censored data with time-dependent covariates. *Biostatistics* 7, 599-614.

Therneau, T., Lumley, T., 2008. survival: Survival analysis, including penalised likelihood, version: 2.34-1. Available at: <http://cran.r-project.org/>.

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

Woodall, C.W., Grambsch, P.L., Thomas, W., 2005a. Applying survival analysis to a large-scale forest inventory for assessment of tree mortality in Minnesota. *Ecological Modelling* 189, 199-208.

Woodall, C.W., Grambsch, P.L., Thomas, W., Moser, W.K., 2005b. Survival analysis for a large scale forest health issue: Missouri oak decline. *Environmental Monitoring and Assessment* 108, 295-307.

CHAPTER 3. EVALUATION OF THE SENSITIVITY OF A POISSON BASED LOCAL COMPETITION VARIANCE ESTIMATE TO SPATIAL PATTERNS

1 Introduction

Competition among neighboring trees plays an important role in forest stand dynamics. Because competition affects each of the basic components of most tree growth models (Begon *et al.*, 1996), it is broadly included in forest growth and yield simulation systems (Hann and Zumrawi, 1991). Competition variables, such as stand or point basal area, are, like most measured variables, subject to measurement error. The main source of error for competition variables arises from the variability inherent to point estimation (Stage and Wykoff, 1998) that originates when competition variables are estimated through sampling.

The potential effects of measurement error when ignored in regression model development are well known, and include bias in the coefficients of the model parameter estimates and bias in model predictions (Canavan and Hann, 2004). The recognition of these critical latent risks has motivated sustained attention in forestry literature for the last five decades (e.g. Jaakkola, 1967; Hann and Zumrawi, 1991; Stage and Wykoff, 1998; Lappi, 2005). Most of these studies examined the effects of alternative sampling-unit designs on the sensitivity of the models to the competition variables (Jaakkola, 1967; Stage and Wykoff, 1998; Hynynen and Ojansuu, 2003; Lappi, 2005), and on growth-model predictions (Hann and Zumrawi, 1991; Stage and Wykoff, 1998; Ledermann and Eckmüllner, 2004); but few of them propose a method that addresses totally or partially the negative effects of measurement error. Ledermann and Eckmüllner (2004) proposed the record splitting method to adjust the resolution of the basal-area-in-larger trees competition variable during simulation. The record splitting method does not, however, address other problems caused by measurement error in model fitting, such as bias in the coefficients or the lack of validity of the test for the statistical significance

of the coefficients of the model. Stage and Wykoff (1998), in contrast, proposed a solution to address the adverse effects of measurement error in both the fitting and the prediction phases of the models, which they called “Structural Based Prediction” (SBP). The methodology they proposed was based on Fuller’s method of moments regression estimator (Fuller, 1987), which assumes, as well as many other approaches to correct for measurement error, some knowledge of the variance of the measurement error (Fuller, 1987; Carroll *et al.*, 2006). The correct estimation of the variance is one of the most critical steps in all of the most widely used approaches for the correction of measurement error (Fuller, 1987; Carroll *et al.*, 2006).

Estimates of the variance can be obtained from internal replicates, that is, when several measurements of the competition variable are available one can estimate the variance with the usual sample-based estimator. Standard forest inventories do not include replicates of local competition variables, and doing so would multiply the cost of the inventories; however, when more than one plot is measured in each stand it is possible to estimate the variance of the stand competition from those replicates. Other options when internal replicates are not available include estimating the variance from external studies or deriving it from the data available in the inventory. Because the variance changes with the spatial structure, diameter distribution and plot size (Stage and Wykoff, 1998) the former option would require herculean work and data that are usually not available on inventories. The latter involves assumptions about the spatial pattern of tree locations, and about the diameter correlation among neighboring trees. This last approach was selected by Stage and Wykoff (1998) to implement the only methodology developed until now that corrects for measurement error in competition variables in both the calibration and application phases of diameter increment / basal area models. Even though the variance equations developed by the mentioned authors are the only ones available at the moment, and albeit the variance estimates are a key component of the approach, there has been no study of the performance of the equations with respect to the estimation of the real variance in simulated or real stands.

1.1 The variance of local competition variables

As mentioned earlier, Stage and Wykoff's approach requires certain assumptions to derive the variance. One of the main assumptions is that tree locations and tree diameters follow a random (Poisson) pattern. Thus, it assumes independence of the location of each tree from the location of the rest of the trees, independence of the diameter of each tree from the diameters and locations of the rest of the trees, and hence, that the tree counts in each diameter class is itself a Poisson-distributed variable. These assumptions are not met in practice (Stage and Wykoff, 1998; Lappi, 2005), but so far there is no better feasible approximation to the variance of measurement error in local competition variables.

With the above assumptions and an asymptotic approximation to the variance of the percentile component by Wilks (1962), Stage and Wykoff derived an estimate of the variance of two local competition variables, point basal area (PBA) and point basal area of larger trees (PBAL) (see Stage and Wykoff (1998) for details).

1.1.1 Sampling local competition variance

The sampling variance of PBA in the neighborhood of the t th tree is estimated according to Stage and Wykoff (1998) as:

$$var_t(PBA) = \frac{SBA}{PBA} c^2 \left[\left(\sum_{j=1}^r \frac{M_j DBH_j^4}{a_j} \right) - \frac{DBH_t^4}{a_t} \right]$$

Where DBH is diameter at breast height, $c = \pi/[4(10,000 \text{ cm}^2/\text{m}^2)]$ for diameters in cm and area in hectares, a is the area of the plot, M is the number of trees per area, and r is the number of tree records left in the inventory sample point at the current time step.

The reasoning behind the last term of the equation is that when we sample the local competition around a tree, every sample includes that particular tree, and thus its distribution is not a true Poisson distribution but it is truncated since all potential plot locations including a certain tree always include at least that tree.

Froese (2003) argued that Stage and Wykoff erroneously omitted the M_t term on the last part of the equation, and thus:

$$var_t(PBA) = \frac{SBA}{PBA} c^2 \left[\left(\sum_{j=1}^r \frac{M_j DBH_j^4}{a_j} \right) - \frac{M_t DBH_t^4}{a_t} \right]$$

The point basal area of larger trees (PBAL) is defined as $(1-p_t) \cdot PBA$, where p_t is the percentile position of the tree based on the total stand inventory. The variance of PBAL is derived from an estimate of the variance of 1 minus percentile (vPCT) by Wilks (1962). In this chapter, unless otherwise stated, when I refer to Stage and Wykoff's equation I will be referring to the above corrected version.

An estimate of the PBAL variance for tree t is:

$$var_t(PBAL) = PBA^2 \cdot vPCT_t + (1 - p_t^2) var_t(PBA) - vPCT_t \cdot var_t(PBA)$$

where $T_t = \frac{p_t(1-p_t)}{r}$, and r is the number of trees in the plot.

1.1.2 The shrinkage factor

Stage and Wykoff claimed that samples with higher than average counts per point were more frequently overestimating the local mean and variance, and that samples with lower than average counts per point were more frequently underestimating the true mean and variance. To overcome this problem, they added a shrinkage factor (SF) which is a multiplier to the PBA variance. The suggested SF was:

$$SF = \frac{PV SBA + (1 - PV)PBA}{PBA}$$

Where PV is a constant taking a value between 0 and 1. Stage and Wykoff found a value for PV empirically, based on the criterion of minimizing the error of the estimate of the resulting model. The selected PV was 1, so $SF = SBA/PBA$. The main effect of the shrinkage factor is to shrink the estimated variance from the single sample towards a central value for the stand (Stage and Wykoff, 1998).

The SF is applied to all stands regardless of their spatial structure, diameter distribution or variance. Stage and Wykoff did not investigate whether this shrinkage factor was necessary for all types of stands and conditions, or if a correction for certain spatial structures was needed. Therefore, many questions remain unanswered. For example, since the SF was empirically derived one might query if the SF should be recalculated for each dataset to minimize the error of the estimate of the resulting model. In fact, since the under or overestimation of the local variance varies with the spatial pattern it might be necessary to adjust the SF stand by stand. For example, the variability of the differences between the estimates and the true local variances will be higher in a clustered stand than in a regular stand, so the SF adjustment needed to correct the under and overestimation in these two stands might be different.

1.2 Sensitivity of the model to PBA variance

Stage and Wykoff (1998) performed a simulation to test the sensitivity of the projections using Structural Based Prediction to the Poisson variance model. They introduced a variance multiplier (VM) in the model that scaled the variance of local competition variables. Four different VM values were tested: 0.0, 0.5, 1.0, 2.0, and they found that 140-year volume projections with non-zero variance multipliers (VM = 0.5, 1.0, 2.0) varied less than 6% among themselves, but the projection assuming no error-variance (VM) was only 80% of that obtained with the rest of the values. They also noted that increasing the VM forced PBA and PBAL coefficients towards zero, and seemed to decrease growth predictions at low stand densities and increase largely growth predictions at high stand densities. While this sensitivity analysis provided some useful information, it was rather arbitrary. An analysis where the variance was derived from a typical range of actual or simulated actual stand structures could provide a more realistic approximation of the sensitivity in genuine applications.

In regression calibration (RC) as well as in the fitting phase of Stage and Wykoff's algorithm, the average variance of all trees is used to correct the adverse effects of measurement error (see Chapter 1). Hence, overestimation of the variance in some trees or stands will compensate, at least partially, for the underestimation in some other

trees and stands. The bias, if any, of the estimates will be reduced, but the overall variance might still be biased, resulting in bias in model coefficients.

When we use the measurement error-corrected models to predict future mortality or basal area growth in a typical application context, the measurement error corrections are made tree-by-tree (SBP) or stand-by-stand (RC). This implies that the reduction in the bias of the fitting phase due to averaging disappears or is severely reduced during the prediction phase, and thus, the effects of the deviations of the variance estimates from the true variances might have a large effect on predictions. Thus, sensitivity of the models in the fitting and prediction phase must be studied separately.

The variance of the measurement error is a defining component of both SBP and RC methodologies. The equations used to estimate it assume several hypotheses that are known to be false. Previous analyses on the sensitivity of the models to these assumptions have been useful first approximations, but many questions remain unanswered, including the bias of the variance estimates in different spatial patterns, the impact of the SF in stands with different spatial patterns, or if the effect of the SF is affected by the diameter distribution.

1.3 Objectives and Approach

The overall goal in this chapter is to study the behavior of the variance of local competition variables in different spatial patterns and diameter distributions, and to study the sensitivity of the mortality models to deviations from the estimated variance of local competition variables.

This overall goal is divided into four specific objectives:

- examine and quantify the deviations of the variance estimate from the real variance for different spatial patterns and diameter scenarios,
- analyze the effect of the shrinkage factor used by Stage and Wykoff (1998) in the estimate of the variance for different spatial patterns,

-
- study the sensitivity of the coefficients in the mortality model to measurement error variance, and
 - examine the sensitivity of model predictions to measurement error variance.

My approach to these objectives has two parts. First, I create simulated, spatially-explicit stands, for which the real measurement error variance can be quantified, and to which different variance approximations can be compared. Second, I use sample plot data from the calibration and validation data sets used in Chapter 1 to test the sensitivity of variance multipliers and assumptions about the shrinkage factor.

2 Creating spatially-explicit simulated stands

The variety of spatial patterns and diameter distributions combinations found in North American forests is so large that it would be impossible to test all of them, at least in the scope of this chapter. Therefore, I selected four combinations of spatial patterns and diameter distributions that represent a variety of situations of natural stands with rather distinct patterns. To ensure and control the assumptions necessary to calculate the variance, the stands were simulated using spatial point processes (Stoyan and Penttinen, 2000). I selected four prototypical stands covering a wide range of spatial patterns, including a highly clustered stand, a Poisson forest, and stands defined by inhibition processes. A thorough introduction to point processes can be found in Diggle (2003), and for a more mathematical presentation see Stoyan et al. (1995).

The prototypical stands should be realistic not only in terms of their spatial pattern, but also in terms of their diameter distributions and correlation in diameter among neighboring trees. The first step to succeed in this goal is to recreate suitable spatial patterns, the second step is to assign diameters according to the objective stands.

2.1 *Spatial point process*

The first prototypical stand is intended to represent a young, naturally-regenerated and even-aged stand, so I will be referring to it as *young*. The *young* stand represents the early stages of development, with a clear clustering due to environmental heterogeneity, seed dispersion or competition with other species. The mean and variance of tree diameter should be relatively small. *Young* stands were simulated using the spatially homogeneous Matèrn cluster process (Matèrn, 1960, 1986). To simulate this process, three parameters must be specified; I adopted values of the intensity of the Poisson process of cluster centers of 0.01, radius of the clusters 3.0 m, and mean number of trees per cluster of 10.

The second prototypical stand recreates a Poisson forest, so it will be called *Poisson*. As a stand matures, the spatial distribution has a tendency towards regularity (Reed and Burkhart, 1985; Stoyan and Penttinen, 2000). So the *Poisson* forest represents an

intermediate state between a young stand and a mature stand. The spatial point process used to generate this stand is a homogeneous Poisson process. The variance in the tree diameters is larger than in the *young* stand, and the number of trees is smaller. The Poisson process requires only one parameter be specified; the value I used to simulate this forest is $\lambda=0.08$.

The third prototypical stand aims to recreate an “old growth” stand, and therefore I name it *old*. This stand is simulated by assuming trees are distributed according to a softcore inhibition process (Ripley and Kelly, 1977). The SBA, mean diameter and diameter variance are the largest of all stands. These stands have been generated using the Metropolis-Hastings algorithm (Metropolis *et al.*, 1953; Hastings, 1970; Chib and Greenberg, 1995), a Markov Chain algorithm, to generate simulated realizations from the pairwise soft core interaction spatial point process (Ripley and Kelly, 1977). The soft core process has three parameters: β , σ , and κ (0,1). This is an inhibitive process with the interpoint interaction decreasing smoothly with distance. The parameter β indicates the base intensity, the σ and κ parameters controls the strength of the interaction, with larger values of σ corresponding to stronger interaction, and larger values of κ corresponding to weaker interactions. The soft core point process has probability density function:

$$f(x) = \alpha \beta^n e^{-\sum_{i < j} \left(\frac{\sigma}{\|x_i - x_j\|} \right)^{\kappa/2}}$$

Where X represent the points (trees) of the pattern, n is the number of points in the pattern, α is the normalizing constant, and $\|x_i - x_j\|$ indicates the distance between x_i and x_j . The parameters I selected to simulate the *old* stand were $\beta=0.09$, $\sigma=1$, $\kappa=0.5$.

The fourth prototypical stand, *plantation*, represents a very uniform plantation which has undergone light thinning. The spatial pattern of trees is defined by a hardcore process (Matérn, 1960, 1986), which means that there is no tree within the inhibition distance of each tree, in this case 1.5 meters. The initial intensity of the Poisson process of proposed points is 0.1. The thinning process was simulated after the diameter

assignment; strips of 4 meters in width were placed each 25 meters and any tree falling within these strips was removed from the stand.

2.2 Assigning diameters

Once the tree spatial patterns have been simulated the next task is to assign realistic diameters to trees. The literature on forest spatial pattern simulation focuses mainly on reconstructing the spatial patterns of stands from tree lists. For example, Hanus et al. (1998) proposed a nonsimple sequential inhibition process to generate coordinates for a tree list based on the inhibition distance derived from the maximum crown width of an open-grown tree of the same species and DBH as the subject tree. Hann and Zumrawi (1991) following Newnham and Maloley (1970) started by simulating the random spatial patterns of trees, which they used to calculate the free growing space around each point, then sort the tree list and the areas and assign the trees to a location according to the hierarchy of both ordered lists. Pretzsch (1997) developed a stand structure generator designed for the modelling and reproduction of spatial stand structures from tree lists and qualitative data on the spatial pattern. The spatial pattern was the result of a combination of a non-homogeneous Poisson process and a hard-core process. There is also a fair amount of literature on this topic in the German language. Hui et al. (2003) proposed a method based solely on information about the angle index¹ (Hui and Gadow, 2002) and stand density to reproduce the spatial distribution of trees. Lewandowski and von Gadow (1997), also in German, presented a method based on the spatial distribution of neighbor-attributes for reproducing the structure of an heterogeneous mixed forest.

Methodologies to assign tree characteristics to a spatial point pattern without a tree list include Reed and Burkhart (1985), who proposed to generate the diameters as a random variable from the correlation of diameters of neighboring trees based on the desired overall population characteristics and the characteristics of the neighboring trees. Pukkala (1989) presented two methods to predict DBH based upon tree locations. The first method predicts DBH directly from the position of neighbors in the vicinity of the

¹ The uniform angle index is derived from the angles subtended by adjacent neighbors at the reference trees.

tree. The rationale of this first method is based on dependencies of tree DBH on tree density at different distances from the object tree. For example, if the number of trees per hectare increases towards the tree, it is probably a small individual. If, on the contrary, the stocking is very low around a tree but it increases with distance, it is probably a dominant tree. In the second method the tree size was assigned by estimating the local diameter distribution in a subarea around the object tree and sampling that distribution to obtain the diameters. More recently, Pommerening and Stoyan (2006) simulated Poisson, Matérn hard-core and a Matérn cluster process to investigate the need of edge correction in estimating indices of spatial forest structure. For the assignment of diameters they followed Nagel and Biging's (1995) methodology (in German), which transforms uniform random numbers to Weibull-distributed random numbers. The methodologies described by Pukkala (1989) and Nagel and Biging's (1995) do not require an existing tree list; they generate the tree diameters directly from the spatial position of neighbors in the vicinity of the object tree, general stand parameters, and an assumed diameter distribution.

Since the objective of this work is not to reconstruct the spatial pattern of certain tree lists but to compare the variance of local competition variables in a variety of realistic spatial patterns and diameter distributions, the methods that do not require a previous tree list, Pukkala (1989) and Nagel and Biging (1995), are adequate. I selected the first method described by Pukkala (1989) because of the difficulty in reading Nagel and Biging's (1995), written in German, and because Pukkala's first method is simpler than the second and does not suppose any major disadvantage. Pukkala's first method, which only requires information about the position of the trees and the stand density, only explained about 40% of the variability in tree diameters of the mapped stands studied by Pukkala (1989). The author found that the greatest share of variation was in the stochastic residual component, which was found to be close to normal. Therefore, Pukkala included a random variable normally distributed with zero mean and approximate standard deviation 0.7. As a result of having 70% of the variability produced by a normally distributed variable, the diameter distributions of the stands were highly normal, and the variability of diameters in *plantation* stands were higher than one would expect in such situations. Therefore, the standard deviation of the normal

random variable for the *plantation* stands was reduced to 0.1. The diameter distributions of *young* stands calculated according to Pukkala methods were within the expected range for this type of stands and within similar simulations (Pommerening and Stoyan, 2006). However, the *Poisson* and *old* stands' diameter distributions had an unusually normal shape. For these two stand types the diameters were reassigned, so their final diameter distribution was Weibull truncated.

One of the most common functions used to describe diameter distributions in forestry is the Weibull distribution, although it is not appropriate for stands with small (e.g. <7 cm) mean diameter (Merganič and Sterba, 2006). The Weibull distribution has three parameters: α , b and c . α is a location parameter, b is a scale parameter and c is shape parameter. The parameter b can be interpreted as the 63rd percentile of the diameter distribution (Bailey and Dell, 1973).

The probability density function (pdf) of the Weibull distribution is defined as:

$$f(x) = \frac{c}{b} \left(\frac{x - \alpha}{b} \right)^{c-1} e^{-\left(\frac{x - \alpha}{b} \right)^c}$$

I note that, in forestry, inventory data are usually truncated; that is, only trees above a certain diameter (T) are measured. Fortunately, equations for the truncated Weibull function are given by (Zutter *et al.*, 1986) and reproduced in Merganič and Sterba (2006):

$$f_T(x) = \frac{c}{b} \left(\frac{x}{b} \right)^{c-1} e^{-\left(\frac{T^c - x^c}{b^c} \right)}$$

$$F_T(x) = 1 - e^{-\left(\frac{T}{b} \right)^c} e^{-\left(\frac{x}{b} \right)^c}$$

$$x_T = b \left[\left(\frac{T}{b} \right)^c - \ln(1 - F_T(x)) \right]^{\frac{1}{c}}$$

where x indicates DBH, $f(x)$ the probability density function, and $F(x)$ the cumulative distribution function.

In summary, *Poisson* and *old* stands were first assigned diameters according to Pukkala (1989) methodology, including the random variable. Second, new diameters were assigned from a Weibull truncated distribution to each tree conditioned on the original diameter. In this second phase diameters were assigned according to the hierarchy of the diameters calculated by Pukkala's methodology, so dominant trees were still dominant and suppressed trees were still suppressed. For example, if the diameter of a tree was the 15th smallest tree in the diameter $F_T(x) = \frac{15}{n} 0.95$ where n is the total number of trees in the stand, and 0.95 was added to avoid extreme large trees. For the *Poisson* stands I used $b=15$, $c=0.87$, and $T=5$, for *old* stands $b=25$, $c=0.87$, and $T=5$.

2.3 Randomization

In my simulated stands, differences among observed and estimated variances could be due to the spatial pattern, to the diameter correlations or to diameter variance. In order to analyze the origin of the potential differences between estimated and true variances, one or two randomizations were performed in each prototypical stand. The first randomization consisted on permuting DBH (rdbh) to ensure that there was no DBH correlation among neighbor trees (García, 2007). The second randomization was performed over the coordinates (rxy), so all the stands with this second randomization had Poisson spatial patterns.

2.4 Prototypical stands

The results of the spatial point processes and diameter assignments are four prototypical stands (Figure 3-1). Each stand type was replicated 10 times using exactly the same parameters. However, because point processes are stochastic, each realization differs from the rest of the realizations of the same stand type in tree location, total number of trees and other tree and stand characteristics. All created stands had a square shape and an area of 1 ha.

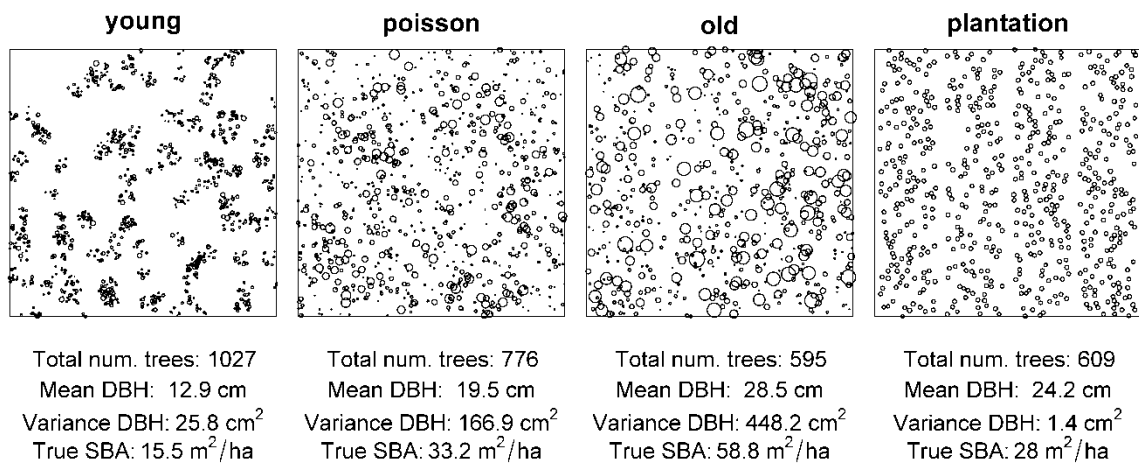


Figure 3-1. One of the 10 realizations of each typified stand. Circles represent tree diameter. The diameters are not shown in real scale but the scale is similar for all plots.

For *young* stands the variance among DBHs of the same stand is relatively small. The mean of the average DBH of the 10 *young* stands simulated is only 12.5 cm, with a standard deviation of the average DBH of 2.3 cm. Mean and standard deviation of the average DBH per stand are larger for *Poisson* stands, 19.03 cm and 5.14 cm respectively, and even larger for *old* stands, 28.3 cm and 8.9 cm respectively, while *plantation* stands have the lowest standard deviation, 0.6 cm, and one of the largest means, 24.3 cm.

All variations of the stands, non-randomized and randomized (DBH and coordinates) had in each realization the same stand characteristics, that is, same SBA, same mean and DBH variance and total number of trees.

I used the **Clark and Evans aggregation index** (Clark and Evans, 1954), R , as a crude metric of clustering in the typified stands. R is a dimensionless number, defined as the ratio of the observed mean nearest neighbor distance in the pattern to that expected in a Poisson point process of the same intensity. A value $R > 1$ suggest ordering, that is, inhibition process, and $R < 1$ suggests clustering. To correct for bias in the R index due to edge effect, the Donnelly edge correction (Donnelly, 1978) was applied. The average R index for each stand type is as follows:

Table 3-1. R index of clustering.

	non-random	rdbh	rxv
<i>young</i>	0.51	0.51	0.99
<i>Poisson</i>	1.00	1.00	-
<i>old</i>	1.14	1.14	1.00
<i>plantation</i>	1.18	1.18	1.00

In summary, a total of 40 realizations were created, that is, 10 realizations of each of the 4 prototypical stands. On each of these 40 stands two or one (*Poisson* stands) randomizations were performed, so a total of 110 stands were simulated.

3 Methods

The overall approach has two main parts. In the first one, plot sampling is simulated on the 110 spatially-explicit stands. For each tree included in each sampling plot the real measurement error variance is quantified and compared to different variance approximations. In the second part, sample plot data from the calibration and validation data sets used in Chapter 1 are used to test the sensitivity of variance multipliers and assumptions about the shrinkage factor.

For the first part of the approach three variances will be calculated for each tree included in every plot: the PBA variance following Stage and Wykoff's methodology with the shrinkage factor (hereinafter "S&W") and without the shrinkage factor (hereinafter "RAW"), and the true PBA variance.

The actual PBA variance will be calculated using bootstrapping, by placing 50 random plots around each tree and calculating the PBA for each of these plots. I assume that the variance of these 50 replicates is a good approximation to the real variance of the PBA around the subject tree; therefore, hereinafter I refer to this variance as "REAL". The plots will have the same size and shape as the sample plots.

A similar approach to Stage and Wykoff (1998) was used to test the sensitivity of model coefficients and predicted mortality to changes in the estimated variance. The values of the variances used to evaluate the sensitivity of the model will depend upon the results of the analysis of the first part of the approach on *young* stands.

Data analysis, simulations and graphs for this chapter were done in R environment (R Development Core Team, 2008) using the 'spatstat' package (Baddeley and Turner, 2005).

3.1 Sampling simulation

Six sampling plots were placed randomly in each stand. The sampling plots were selected in such way that their areas did not intersect with each other and that their centers were at least 12 meters from the edge. Around each tree falling inside any of

these six plots 50 new randomly placed plots were located. The variance of the PBA of these resampling plots was used to calculate REAL for each tree. These plots were of similar size and shape as the original sampling plots and were located in such a way that all 50 resampling plots included the tree of interest. For each of the trees falling in every sampling plot three variances were calculated: the estimated real variance (REAL), and the Stage and Wykoff (1998) approximation to PBA variance with (S&W) and without (Hann and Zumrawi, 1991) shrinkage factor.

In each stand, plot locations were recalculated, so, for example, the location of the six plots in the first stand is different for the non-random stand, the one with random DBH and the one with random location of trees.

Root mean square error (RMSE) was used to assess the difference in the precision of the variance estimates (S&W and RAW) and REAL. For example, in the case of Stage and Wykoff, the RMSE was calculated as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (REAL_i - S\&W_i)^2}{n}}$$

3.2 Sensitivity analysis using field data

The second part of my approach focuses on testing the sensitivity of model coefficients and predicted mortality to changes in the estimated variance. For analyzing the sensitivity of model coefficients to variations in the variance estimate all stands in the fitting dataset of Chapter 1 were assumed to be young stands. Several variance multipliers were selected in accordance to the results obtained during the simulation for *young* stands, and applied to the estimated variance. Coefficients and statistical significance of the coefficients were compared for the species most sensitive to competition from the three studied in Chapter 1, that is, *P. contorta*.

The evaluation of the sensitivity of the model to Stage and Wykoff variance was evaluated similarly to Stage and Wykoff's (1998) approach, but the multipliers were chosen according to the results of the first part of this chapter, that is, their magnitude will correspond approximately to the magnitudes of the differences found between

variance estimates and true variances in Chapter 1. For each multiplier a new model will be fitted using the same calibration dataset as Chapter 1 but applying the multiplier to the variance.

Four stands from the Chapter 1 validation dataset were selected arbitrarily, two of them with *P. contorta* and two of them with *P. menziesii* as the main species. The stands were selected such that per species one stand had significantly lower PBAL variance than the other. For each of these four stands mortality is estimated repetitively, each time using a different value from a range of different PBAL variance multipliers. In this phase, because the variance of SBA will be the same for each tree in each stand SBA will not be included in RC, as this makes some of the matrices not invertible, and therefore the algorithm produces an error. In conceivable applications this problem should be unlikely to be encountered, and could be avoided by using the values of all the variances, except for Σ_{uu} , and the means from the fitting RC model, in a similar approach to Stage and Wykoff (1998). Since SBA is not included in any of the models and the measurement error associated with SBA is notably smaller than that associated with PBAL, the selected approach, ignoring SBA, should not cause major deviations.

4 Results

4.1 Comparisons of variance estimators

The RAW estimates of the variance are, on average, larger than the S&W estimates for each stand, with minor exceptions (Table 3-2. Summary of the results for REAL, S&W, and RAW per stand type based on average values of the 10 realizations. Standard deviations are the mean over all plots and stands of the standard deviations of the PBA variance per plot. Table 3-2 and Figure 3-2). When differences between average stand REAL and RAW variance estimates are calculated, the maximum values range from over 1,000 for *old* stands, to as low as 14 for randomized *young* stands. In contrast, maximum differences between average stand REAL and S&W estimates range, in absolute values, from 12 in the *young* stand to over 500 in the *old* stand. Differences in individual trees are much larger than differences per stand (Figure 3-2), with values over 3,000 in *old* stands, and over 1,000 in *Poisson* stands. Individual differences are significantly reduced when the SF is used (S&W), but mainly in trees where the variance estimate is underestimating REAL.

Both approaches overestimate REAL in all stands but the clustered stands ($R < 1$, see Table 3-1). It is precisely in clustered stands, *young* stands with non-randomized coordinates, where the largest differences proportional to REAL can be found. In these stands REAL is more than 3 times the estimated S&W or RAW. In absolute value, however, the largest differences are found in *old* stands for both S&W and RAW (530 and 1,069 respectively). The bias of the PBA variance estimate tends to diminish when the location of the trees is randomized, except in the case of *plantation* stands. Contrary, randomization of DBH only decreases bias for *young* stands.

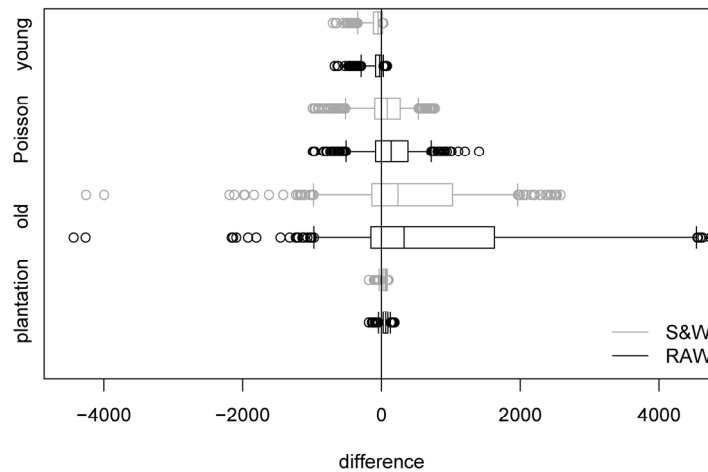


Figure 3-2. Distribution of differences between average stand REAL and RAW variance estimates.

Table 3-2. Summary of the results for REAL, S&W, and RAW per stand type based on average values of the 10 realizations. Standard deviations are the mean over all plots and stands of the standard deviations of the PBA variance per plot.

Stand type		PBA variance mean and SD (in parenthesis)			Difference with respect to REAL ¹		REAL / variance estimate	
		REAL	S&W	RAW	S&W	RAW	S&W	RAW
<i>young</i>	modeled	118 (35)	36 (4)	57 (4)	-82 (-0.69)	-61 (-0.52)	3.28	2.07
	rdbh	80 (25)	28 (4)	48 (5)	-52 (-0.65)	-32 (-0.4)	2.86	1.67
	rxy	17 (7)	29 (4)	31 (4)	12 (0.71)	14 (0.82)	0.59	0.55
<i>Poisson</i>	modeled	303 (149)	373 (85)	445 (89)	70 (0.23)	142 (0.47)	0.81	0.68
	rdbh	194 (76)	320 (74)	383 (79)	126 (0.65)	189 (0.97)	0.61	0.51
<i>old</i>	modeled	786 (349)	1,199 (310)	1,706 (359)	413 (0.53)	920 (1.17)	0.66	0.46
	rdbh	868 (348)	1,398 (400)	1,937 (471)	530 (0.61)	1,069 (1.23)	0.62	0.45
	o.rxy	762 (302)	983 (257)	1,347 (301)	221 (0.29)	585 (0.77)	0.78	0.57
<i>plantation</i>	modeled	58 (22)	101 (3)	114 (3)	43 (0.74)	56 (0.97)	0.57	0.51
	rdbh	54 (21)	103 (4)	114 (3)	49 (0.91)	60 (1.11)	0.52	0.47
	rxy	66 (23)	100 (4)	119 (3)	34 (0.52)	53 (0.8)	0.66	0.55

¹Percentage relative to REAL in parenthesis

4.2 Shrinkage factor

The effect of the SF in the estimates of the REAL variance depends on the spatial distribution of the trees and the diameters. Overestimation of the variance in plots with higher PBA than average and underestimation of the variance in plots with lower PBA than average is clear in *Poisson*, *old*, and *plantation* stands (Figures 3-4, 3-5, and 3-6). Non-random *young* stands, on the contrary, have a tendency to overestimate REAL in plots with lower PBA than average; this trend is exacerbated when the SF is used (Figure 3-3). In all the rest of the stands the shrinkage factor reduces the bias of the estimates with respect to PBA/SBA. The correction is, however, insufficient; that is, even after applying the SF, the overestimation of the variance in plots with higher PBA than average and underestimation of the variance in plots with higher PBA than average is still present.

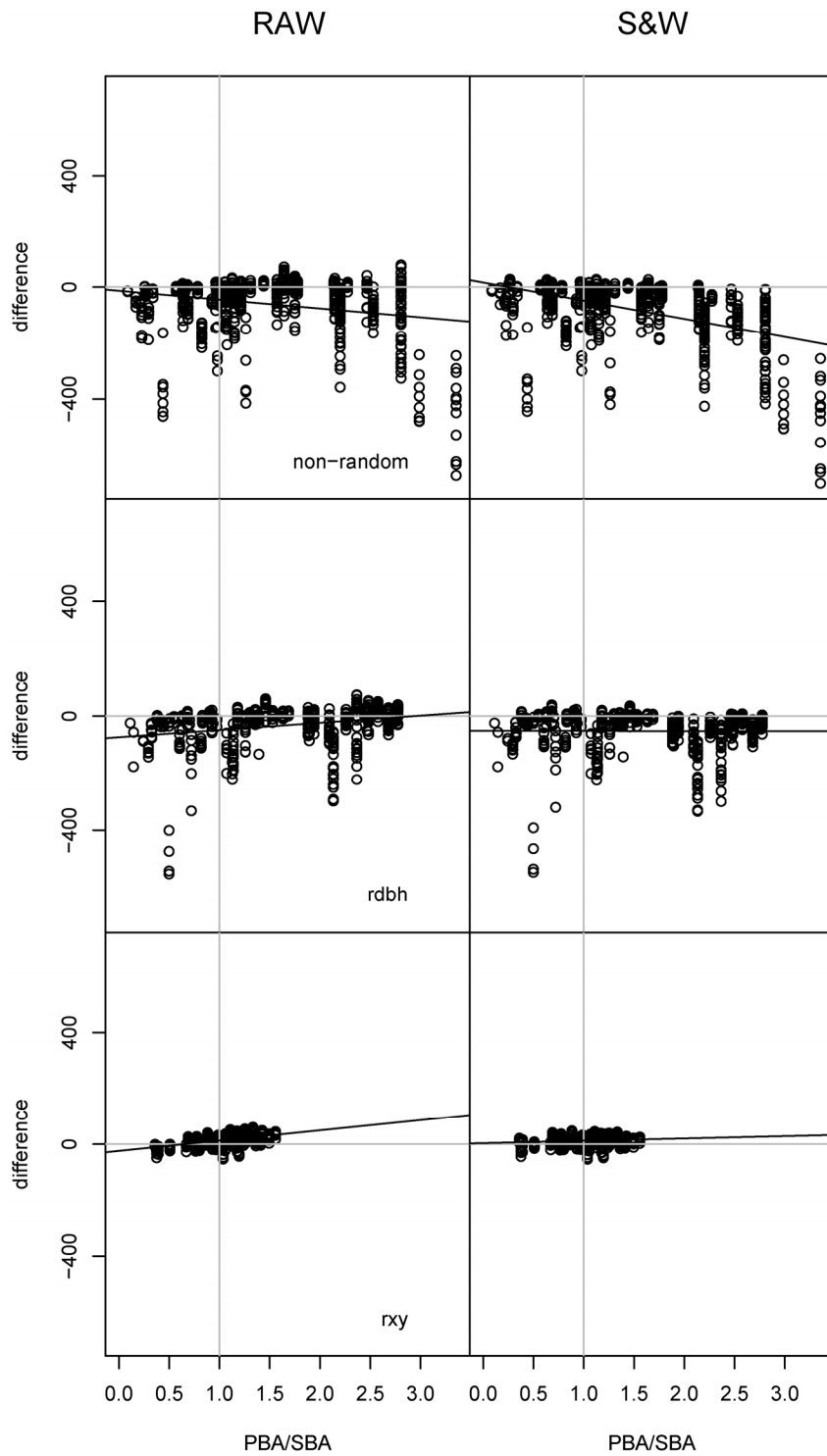


Figure 3-3. Effects of the shrinkage factor (SF) in the difference between the variance estimate and REAL for *young stands*.

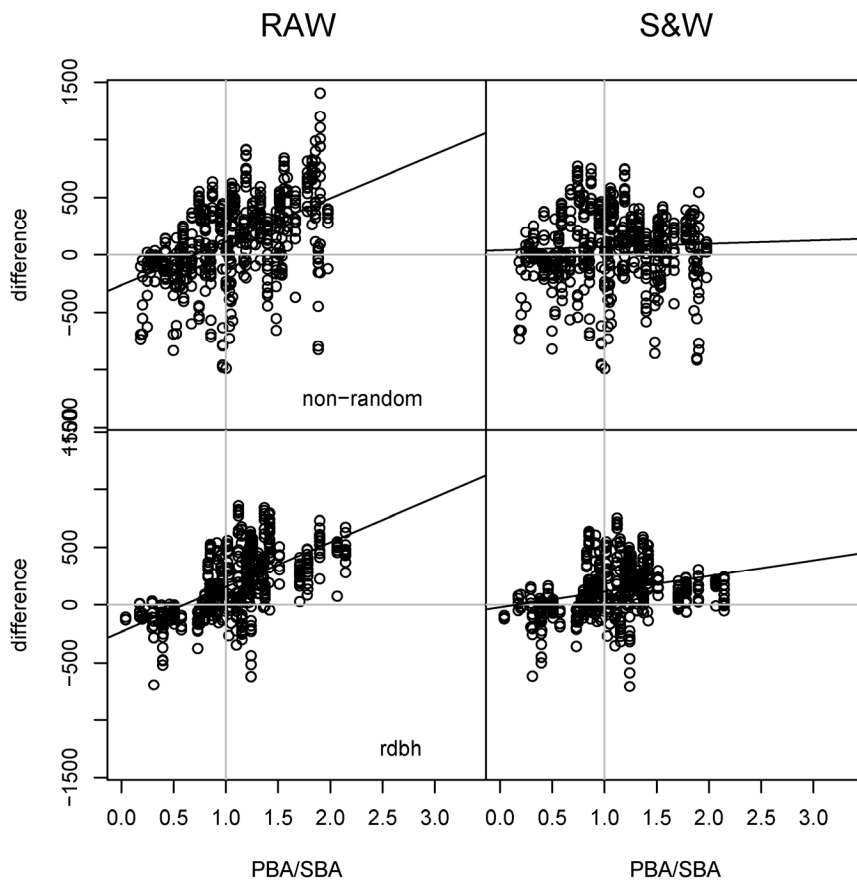


Figure 3-4. Effects of the shrinkage factor (SF) in the difference between the variance estimate and REAL for *Poisson* stands.

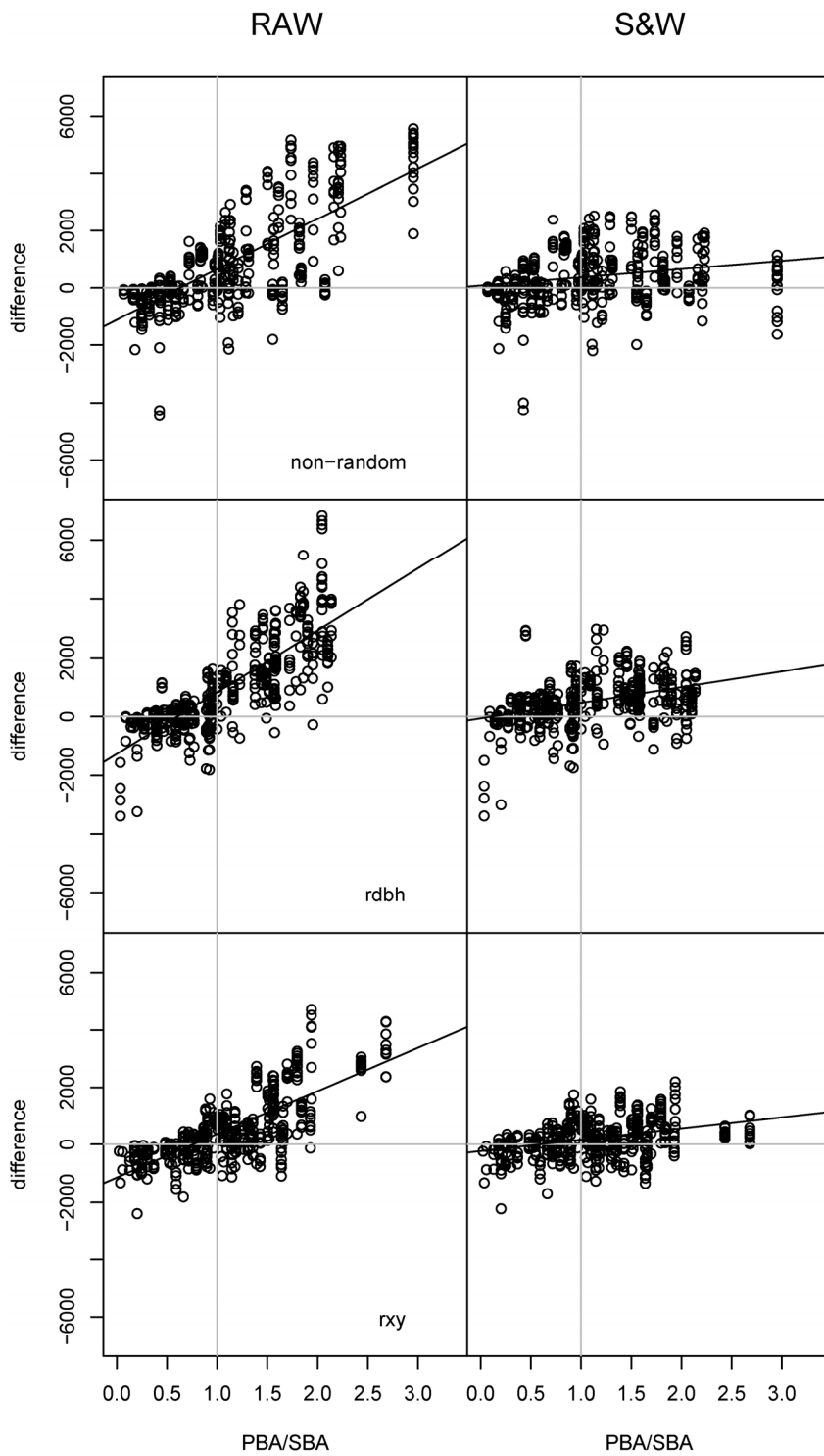


Figure 3-5. Effects of the shrinkage factor (SF) in the difference between the variance estimate and REAL for *old* stands.

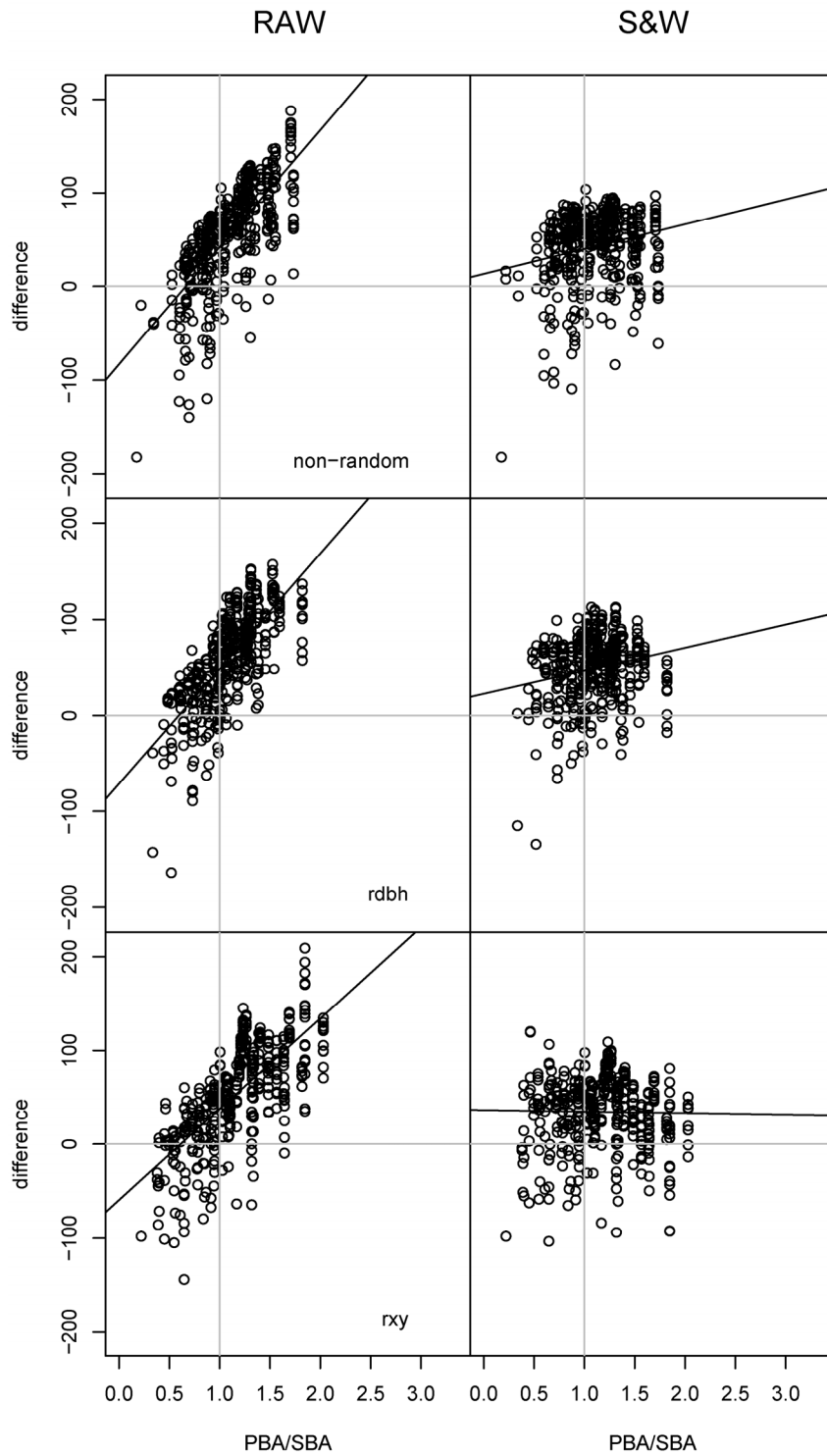


Figure 3-6. Effects of the shrinkage factor (SF) in the difference between the variance estimate and REAL for *plantation* stands.

Table 3-3. Root mean square error (RMSE) for the variance of PBA variance.

	stand type	RMSE		RMSE difference (RAW-S&W)
		S&W	RAW	
<i>young</i>	non-random DBH	138.63	122.42	-16.21
	random DBH	91.75	81.73	-10.02
	random DBH and coordinates	19.34	22.8	3.46
<i>Poisson</i>	non-random DBH	317.26	388.53	71.27
	random DBH	250.29	334.56	84.27
<i>old</i>	non-random DBH	1,036.72	1,898.19	861.47
	random DBH	988.35	1,885.72	897.37
	random DBH and coordinates	680.88	1,300.38	619.5
<i>plantation</i>	non-random DBH	57.96	78.21	20.25
	random DBH	60.18	77.69	17.51
	random DBH and coordinates	52.67	77.42	24.75

The S&W approach is, on average, more precise than the RAW approach (Table 3-). The SF (S&W) reduces RMSE considerably in *old* stands, where the variances are higher. The RAW approach, on the contrary, outperforms the S&W approach in terms of precision in *young* stands, where the average REAL variance is the smallest, but only when the spatial pattern is clustered.

4.3 Fitting phase

In relative terms the larger differences between calculated variances and estimated variances occur in the *young* clustered stands. For the non-random DBH case, REAL is more than 3 times larger than that estimated by S&W. The Chapter 1 fitting dataset contains mainly young stands, with low PBA and PBAL estimated variance (Table 3-4). Certainly, not all the stands in the fitting dataset of Chapter 1 are as clustered as the simulation presented for this chapter, so, for purposes of testing the sensitivity to the Poisson assumptions for this type of stands, I will use a multiplier for the PBAL variance with values ranging from 0.25 to 3 (0.25, 0.5, 1, 2, 3).

Table 3-4. Average variances of competition variables of the two datasets used in Chapter 1 calculated following Stage and Wykoff (1998).

	<i>PBAL variance</i>		<i>PBA variance</i>	
	Calibration	Validation	Calibration	Validation
<i>A. grandis</i>	29	136	46	186
<i>P. contorta</i>	6	116	16	173
<i>P. menziesii</i>	20	137	37	199

Table 3-5. Parameters and standard errors of the coefficients (in parenthesis) of the *P. contorta* model fitted using different multiplier for the variance estimate. All coefficients had a $p < .001$ at $\alpha = 0.05$. The standard errors were calculated by 500 bootstrap.

	Variance Multiplier					
	3	2	1	0.5	0.25	naive
Intercept	-4.013 (0.405)	-3.923 (0.358)	-3.77 (0.351)	-3.691 (0.346)	-3.651 (0.333)	-3.545 (0.308)
PBALRC	0.064 (0.007)	0.059 (0.006)	0.053 (0.005)	0.051 (0.005)	0.050 (0.005)	0.046 (0.004)
CR ^{0.5}	-0.943 (0.153)	-0.964 (0.136)	-1.007 (0.134)	-1.032 (0.133)	-1.044 (0.131)	-1.076 (0.121)
Residual deviance	3,867	3,856	3,849	3,847	3,846	3,844

Increasing the variance of local competition variables has three main effects for the *P. contorta* model (Table 3-5). The first is an increase in the magnitude of the competition variable and the intercept coefficient, and a decrease in the magnitude of the CR^{0.5} coefficient. The second is an increase in the standard errors of the variables with increasing variance, and the third is an increase in the residual deviance.

Ten-year mortality predictions (Figure 3-8) show larger differences in mortality predictions between the naïve and the 0.25 multiplier lines than between the 0.25 and 0.5 multipliers. Increasing the variance estimate increases the sensitivity of the model to local competition (PBAL); with increasing variance the model predicts higher mortality for high PBAL and lower for low PBAL. Differences are larger for low CR and tend to disappear rapidly with increasing CR.

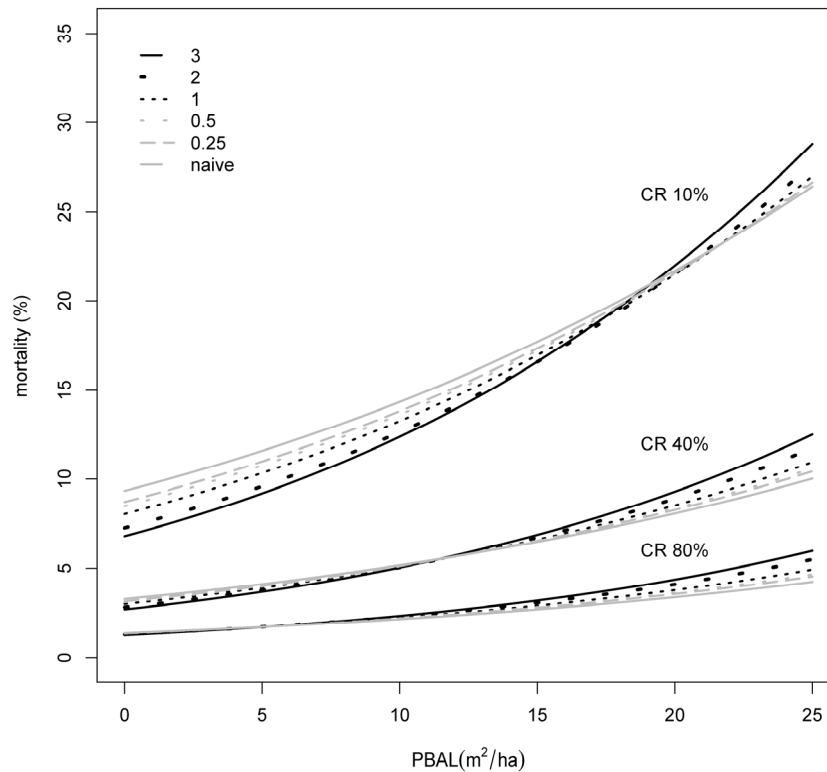


Figure 3-6. 10-year sensitivity plot for the *P. contorta* model.

4.4 Application phase

Four stands from the validation dataset of Chapter 1 were selected according to the criteria detailed in the Methods section (Table 3-6). The first two stands were used to predict mortality in *P. contorta* and the last two in *P. menziesii*. The first and second stands have similar mean DBH but different DBH variance and mean VPBAL. The third stand had higher mean DBH than the fourth and much higher DBH variance.

Table 3-6. Characteristics of the selected validation stands.

	mean DBH(cm)	variance DBH(cm)	SBA(m ² /ha)	mean PBAL variance
stand1	24.84	91.02	37.20	114.69
stand2	23.78	61.42	32.02	61.60
stand3	47.23	558.41	47.02	201.12
stand4	16.01	70.23	22.90	28.62

Predictions for individual trees offer a wide variety of reactions to the increase in variance (Figure 3-8). *P. menziesii* predictions appear indifferent to any change in the variance of the local competition variable (PBAL). On the contrary, the effect of increasing the PBAL variance estimate in *P. contorta* is generally to increase the differences between trees with higher probability of mortality and trees with lower probability. In other words, trees with relatively high predicted probability of mortality in the null model were more sensitive to the multiplier than trees with relatively low predicted probabilities. The trend is relatively consistent across trees. Therefore, on average, predicted mortality for *P. contorta* increases with increasing variance, but predicted mortality for *P. menziesii* remains practically constant.

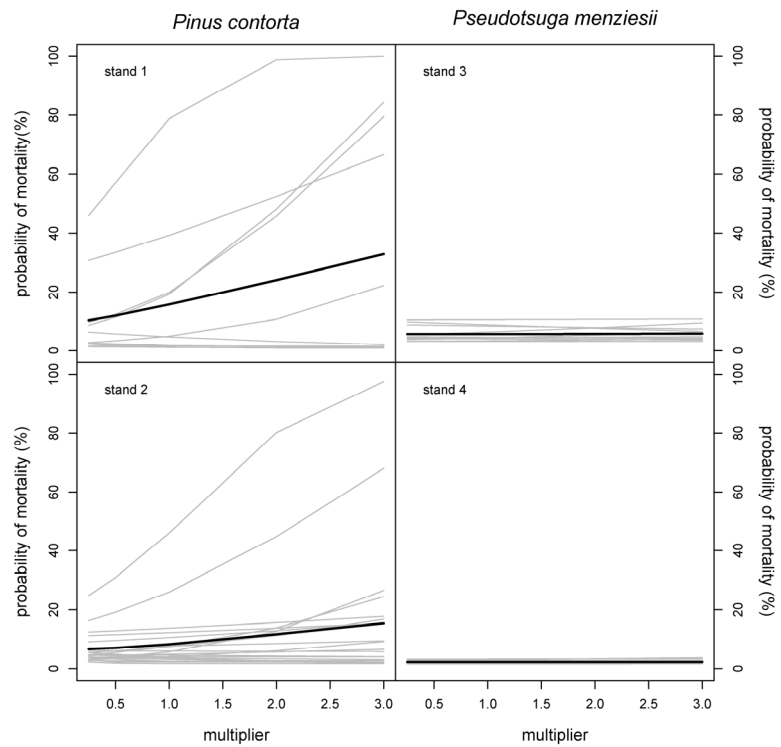


Figure 3-8. Probability of mortality in a 5-year period with different PBAL variance estimates for 4 different stands in the validation dataset. The mean for all the trees in the stand is shown in bold.

The effects of increasing variance on *P. contorta* are not similar in all trees. Some trees increase their probability of mortality sharply while others are unaffected or experience a decrease in predicted probability with increasing error variance. This asymmetry in the effects of changes in error variance alters the hierarchy of probability of mortality, so some trees that at variance multiplier 0.25 have lower probability of mortality than another tree, at variance multiplier 1 have higher probability than the other tree.

5 Discussion

The main objective in this chapter has been studying and quantifying the bias of the local competition variance, as recommended by Stage and Wykoff (1998), in different spatial patterns and diameter distributions. The analysis has revealed a serious overestimation of the true variance in most cases, including Poisson forests. I argue that this bias results from inadequacy of the simple Poisson assumption in the S&W estimator, which ignores that all possible plot samples including the tree of interest have overlapping areas. The estimated variance is also biased with respect to the spatial pattern, underestimating the variance in highly clustered stands, and considerably overestimating it in regular stands. This suggests, therefore, that accounting for the spatial patterns would improve the variance estimates. Consequences of these departures from the real variance in terms of model application are far from negligible, and justify the effort of correcting the variance to account for overlapping and for differences in spatial patterns.

Stage and Wykoff's approximations overestimate the variance of local competition variables in all stands but those with a clustered spatial structure. If we take into account that the variance estimates are derived directly from the distribution of a Poisson forest it is surprising to see that the estimates for *Poisson* stands, with or without randomized DBH, are considerably larger than the calculated variance. In the case of *Poisson* stands with randomized DBH, a Poisson forest is ensured by the simulation, and independence of DBH from the location of the rest of the trees is guaranteed by the randomization of the diameters. Even in this case, when all the assumptions are guaranteed to be met, the overestimation is substantial. I argue that the reason behind the overestimation of the variance is that the possible sampling points around a tree have overlapping areas, and this is incongruent with some of the properties of the Poisson distribution. One of these defining properties is the *memorylessness* property (Cox *et al.*, 1980). It is derived from this property that the number of events (trees) in one time interval or area is independent from the number of events (trees) in an interval or area that is disjoint from the first interval or area. However, if two areas are not

disjoint, but are actually overlapping, the number of trees in one area is not independent of the other area. Since all possible samples including a tree must have an overlapping area that includes the tree of interest, all possible sampling plots are not disjoint, and thus, not independent. The expected value is still λ , so the estimates are unbiased, but the true variance is smaller than the estimated variance. If, for example, the sampling in a Poisson forest with parameter λ is such that the expected value of the overlapping area of the plots of size A and radius r is a , the expected number of trees per plot would still be $\lambda \cdot A$ but the variance would be $\lambda \cdot (A - a)$, because the overlapping area doesn't contribute to the variance, it remains constant. Since any possible plot must include the tree of interest a is the expected overlapping area of two plots with centers within a distance r of the tree of interest.

Stage and Wykoff noted that the variance model they proposed was in apparent contradiction with the fact that the Poisson distribution admits empty plots (count zero). Since the tree of interest is always included in any sample plot that measures the local competition around that tree, the authors concluded that empty plots are actually not possible. They addressed this problem by subtracting the portion of the variance contributed by the subject tree. The problem is actually not related with a truncated Poisson but with the overlapping area of potential plots. Notably, there is no need to correct for the truncated Poisson since the tree of interest will always be included in the overlapping area, and because the only variance considered is outside the overlapping area, a true Poisson distribution can be assumed.

For all calculations of this chapter the corrected version of Stage and Wykoff's equation (Froese, 2003) has been used. The original estimate suggested by Stage and Wykoff (1998) reduced the variance under the truncated Poisson by subtracting the portion of the variance contributed by the subject tree. This corrected partially the overestimation product of overlapping. The corrected version used in this dissertation (Froese, 2003) does not include any variance related to the tree of interest, and therefore the estimated variance is smaller than the original, and the residual overestimation was smaller. If instead of the corrected version, estimates would have been calculated with the original variance, the overestimation would have been aggravated. This does not mean, though,

that any of the two suggested options to estimate a truncated Poisson variance is, in fact, correct. The variance of a Poisson truncated distribution, described elsewhere (Xie and Aickin, 1997), gives estimates that are between Froese (2003) and Stage and Wykoff (1998).

Stage and Wykoff's variance estimate assumes independence of tree diameters from the location of the remaining trees. Stage and Wykoff argued that attributes of the trees used to estimate the measurement error variances already include some of the effects of stand dynamics processes that produce the spatial autocorrelation among neighbor tree attributes. That is, if large trees tend to be close to smaller trees, the diameters of trees considered to calculate the variance will show this tendency, so if there is a large tree in the plot the rest of the trees will tend to be smaller, and in this way the diameter correlations are included in the variance estimate.

If we assume that my assignment of diameter is realistic, and Stage and Wykoff's reasoning holds, no significant differences should be found between non-random and random DBH. This is because, according to Stage and Wykoff's theory, the autocorrelation of diameter would be at least partially incorporated into the estimate. Table 3-2 shows that the precision of the estimates of non-random and random DBH are relatively similar (difference less than 5% of larger value) in *old*, and *plantation* stands. *Young* stands have higher discrepancies between non-random and random DBH, with differences between 14% and 34% of the larger value. In general the estimates are better for random DBH stands, with the only exception of *plantation* stands for SW. However, Table 3-2 shows that in all but *young* stands randomization of DBH produces larger differences between average REAL and S&W, indicating that the PBA variance is on average better estimated in non-random DBH stands than in rdbh stands, but individual differences between REAL and the estimates are larger in rdbh. This implies that a correction wouldn't be needed in the fitting phase, since SBP and RC uses average variance to correct for measurement error in the fitting phase, but, because SBP uses the individual tree measurement error variance in the application phase, better estimates might be obtained if DBH correlation is taken into account. One option to roughly estimate the DBH correlation at the plot level in a stand is to compare the variance of DBH within the plot and between plots. If the variance of DBH within plots is

smaller than the variance of DBH in the stand, then the correlation of DBH at the distances considered by the sampling plot can be considered positive, and negative otherwise (García, 2007). How to incorporate this effect in the equations needs further study.

As expected, the best estimates of the variance are for the *Poisson* forest, and the estimate worsens with departures of the R index from 1. The only stands where REAL and RAW underestimate the variance are *young* clustered stands. When the locations of the trees in *young* stands are randomized the underestimation disappears and RMSE improves (see Table 3-2 and Table 3-3). In general, the bias estimation of the variance tends to diminish when the location of the trees is randomized. Thus, considering spatial pattern is necessary to improve the estimate of the variance of local competition. The problem is then how to estimate the spatial pattern from unmapped stands. One option, as suggested by Diggle (1979) is to use plot to stand variance ratios as indicators of spatial structure. In this case, one would estimate the variance of the number of trees in the plot by assuming a Poisson forest, and estimate the variance of the number of trees between plots by simply calculating the variance of the number of trees per plot in the stand. However, if only one plot per stand is available this approach is not feasible. Other options include to estimate the degree of clustering from remotely sensed data (Utterera *et al.*, 1998; Morsdorf *et al.*, 2004).

5.1 Shrinkage factor

Stage and Wykoff (1998) reasoned that samples with higher than average trees per point will more frequently overestimate the true local mean and variance, and samples with lower than average trees per point will more frequently underestimate it. The stands simulated for this study show similar tendencies in all cases but the clustered stands.

Clustering seems to have an inverted trend, with samples with higher than average trees per point will more frequently underestimate the true local competition variance. This tendency disappears when clustered stands have random DBH, supporting the need to account for diameter correlation in the variance estimate.

The shrinkage factor improves the variance estimate (Table 3-1 and 3-2) for all stand types except the cluster stands. This exception seems to be explained by the decreasing effect of the SF on the magnitude of the variance estimates. Because cluster stands are the only stands where the real PBA is underestimated, when SF is applied the decreasing effect results in an even worse underestimation of the variance.

As the variance estimate stands the SF has an overall positive effect on the variance estimates. The SF reduces drastically the tendency to overestimate the true local variance in samples with higher than average trees per point, and vice versa, but in most cases even after applying the SF the tendency is still evident. The residual trend varies with spatial pattern and diameter autocorrelation, which indicate that once the variance estimates are adjusted for these two factors, as suggested earlier in the discussion, a new study of the effect and necessity of the SF should be performed.

5.2 Fitting phase

The sensitivity analysis of the coefficients to bias in variance estimates has exposed the potential consequences of under and overestimation of the variance, albeit with limitations imposed by the low variance of the fitting dataset. For instance, Stage and Wykoff looked at the differences in the coefficients with different variance multipliers for the basal area increment model, and found that with higher variance multipliers coefficients involving local competition were forced towards zero. Contrary to what Stage and Wykoff noted, this study shows that an increase in the variance translates in an increase in the competition coefficients, enhancing the sensitivity of the model to the competition variable for the *P. contorta* model (see Figure 3-2). The effects of potential bias in the variance estimate seem to be, therefore, of an unforeseeable sign. Effects of the coefficients for different species might differ from those found for *P. contorta*, and although the effects are not radical, the impact might be more sizeable if the dataset used had larger variances. When fitting operative models the dataset would likely include a higher variety of stand structures and ages, with higher variances than those of the fitting dataset used in this chapter, thus a more accurate variance estimate might be very beneficial.

Under or overestimation of the variance has an apparent effect on the estimated value of model coefficients, but not so much on their statistical significance. Similar to results presented by Kangas (1998), all coefficients that were clearly significant remained significant. Kangas (1998) also observed that those coefficients which were not so clearly significant were significant in some realizations but not in others. Because in this study the two variables included in the *P. contorta* model were highly significant ($p < 0.001$) it was not possible to study the effect of variance fluctuations with respect to coefficient significance.

Realistically, since we know that in most cases Stage and Wykoff's equation overestimates the variance, and since in the fitting phase the variance used is the average over all trees and stands, we can presume that unless the fitting dataset has an unusual proportion of clustered stands, the estimates of the coefficients calculated with the real variance should be fairly close to those estimated using Stage and Wykoff's equation. Differences between a multiplier of 1 and 0.5 are not very large, so coefficients shouldn't divert much from the ones obtained in Chapter 1 with a correct PBA variance estimate. However, this lack of sensitivity of the model coefficients to changes in the variance might be influenced by the overall small variance of the local competition variables in the stands of the calibration dataset. A more representative calibration dataset will have larger variances and results on the sensitivity of the model to changes in the variance might differ from the ones obtained here.

5.3 Application phase

Overall 5-year predicted mortality in stand 1 more than doubles between variance multipliers of 0.25 and 3. Taking into account that the most common individual tree simulation engines do not predict mortality in a deterministic way, but use the probability of mortality to reduce proportionally the number of trees represented by the sample tree in the stand (Dixon, 2003), the effect of under or overestimating the variance are not negligible during application. The possible effects of miscalculation of the variance of local competition variables on mortality predictions might be very large and will be likely reflected in the calibration indexes of any validation. These results differ from the small

differences found by Stage and Wykoff in 140 year old simulation for the basal area increment model.

The minor effect of the changes in the variance for the *P. menziesii* model reveals that the effect depends greatly on the sensitivity of the model to the variable. *P. contorta* is more shade intolerant, and therefore more sensitive to PBAL than *P. menziesii*. The effects on the predicted probability of mortality of PBAL variance deviations are much larger in the first than in the second species.

Another interesting point to consider is the asymmetrical effect of changes in the error variance in different trees. This unequal behavior to increases or decreases in the variance has an effect on the hierarchy of probability of mortality of the stands. Since some of the most commonly used measures of a model's predictive discrimination are based on comparing the probability of mortality of pairs of trees such that one is alive and the other dead, e.g. c-index (Harrell *et al.*, 1982; Harrell *et al.*, 1984), the asymmetrical effect will also likely affect discrimination indexes on the validation results.

In the case of SBP it has to be noted that in the application phase the correction is made tree by tree and the variance estimate used to correct measurement error is the one calculated for each individual tree. If differences in average estimated variance and REAL can be over 400, the differences between individual tree estimations for the variance and REAL can reach over 2,000 (see Figure 3-2) even after the SF. These high individual differences might produce an undesired effect on predictions. An interesting question to be addressed in the future is if these rather large differences between real and estimated variances in individual trees have such a large effect on basal area increment models as it has in mortality.

6 Conclusions

The study of the behavior of the local competition variance suggested by Stage and Wykoff (1998) has uncovered a major omission in the variance estimate. Ignoring that all the possible plot samples including a tree of interest have overlapping areas causes overestimation of the variance in most spatial patterns. Addressing this bias should be a priority in the improvement of the estimates.

This study has revealed that underestimation of the variance in highly clustered stands can be large, and that, in general, accounting for the spatial patterns and diameter correlation would improve the variance estimates. It has also been shown that the effect of the SF differs depending on the spatial structure of the stands. It is therefore concluded that taking into account the spatial structure of the trees is necessary to improve the accuracy of the variance. Approaches to estimate the spatial patterns and diameter correlations from unmapped stands have been suggested.

Although coefficients don't seem to be very sensitive, at least for this calibration dataset, to small departures from Stage and Wykoff approach, in application differences are much larger and suggest that implementing the above improvements would certainly enhance the already proven abilities of measurement error techniques to increase the predictive ability of the models.

7 References

- Baddeley, A., Turner, R., 2005. Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software* 12, 1-42.
- Bailey, R.L., Dell, T.R., 1973. Quantifying diameter distributions with the Weibull function. *Forest Science* 19, 97-104.
- Begon, M., Harper, J.L., Townsend, C.R., L, J., 1996. *Ecology: Individuals, Populations and Communities*. Blackwell Publishing.
- Canavan, S.J., Hann, D.W., 2004. The Two-Stage Method for Measurement Error Characterization. *Forest Science* 50, 743.
- Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M., 2006. *Measurement error in nonlinear models*. Chapman and Hall/CRC Press, Boca Raton, FL.
- Chib, S., Greenberg, E., 1995. Understanding the Metropolis-Hastings algorithm. *The American Statistician* 49, 327-335.
- Clark, P.J., Evans, F.C., 1954. Distance to nearest neighbour as a measure of spatial relationships in populations. *Ecology* 35, 445-453.
- Cox, D.R., Isham, V., Isham, V., 1980. *Point Processes*.
- Diggle, P.J., 1979. On parameter estimation and goodness-of-fit testing for spatial point patterns. *Biometrics* 35, 87-101.
- Diggle, P.J., 2003. *Statistical analysis of spatial point patterns*. Academic press, New York.
- Dixon, G.E., 2003. *Essential FVS: A User's Guide to the Forest Vegetation Simulator*. Internal Rep. USDA For. Serv., Fort Collins, CO. 193.
- Donnelly, K., 1978. Simulations to determine the variance and edge-effect of total nearest neighbor distance. *Simulation Methods in Archeology*. Cambridge University Press, London, 91-95.
- Froese, R.E., 2003. *Re-engineering the prognosis basal area increment for the Inland Empire*. University of Idaho, Moscow, Idaho.
- Fuller, W.A., 1987. *Measurement error models*. Wiley, New York, NY.
- García, O., 2007. Scale and spatial structure effects on tree size distributions: implications for growth and yield modelling. *Canadian Journal of Forest Research* 36, 2983-2993.
- Hann, D.W., Zumrawi, A.A., 1991. Growth model predictions as affected by alternative sampling-unit designs. *Forest Science* 37, 1641-1655.

-
- Hanus, M.L., Hann, D.W., Marshall, D.D., 1998. Reconstructing the Spatial Pattern of Trees from Routine Stand Examination Measurements. *Forest Science* 44, 125-133.
- Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A., 1982. Evaluating the yield of medical tests. *Journal of the American Medical Association* 247, 2543-2546.
- Harrell, F.E., Lee, K.L., Califf, R.M., Pryor, D.B., Rosati, R.A., 1984. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine* 3, 143-152.
- Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97-109.
- Hui, G.Y., Albert, M., Chen, B.W., 2003. Reproduction der Baumverteilung im Bestand unter Verwendung des Strukturparameters Winkelmaß β [Reproduction of the spatial distribution of trees using the structural parameter uniform angle index (Winkelmaß β)]. *Allgemeine Forst-und Jagdzeitung* 174, 109–116.
- Hui, G.Y., Gadow, K., 2002. Das Winkelmaß Herleitung des optimalen Standardwinkels. *Allgemeine Forst-und Jagdzeitung* 173, 173-177.
- Hynynen, J., Ojansuu, R., 2003. Impact of plot size on individual-tree competition measures for growth and yield simulators. *Canadian Journal of Forest Research* 33, 455-465.
- Jaakkola, S., 1967. On the use of variable size plots for increment research. In: DVFFA (Ed.), *Proceedings of XIV IUFRO-Kongress, Band VI, München*, pp. 371-378.
- Kangas, A.S., 1998. Effect of errors-in-variables on coefficients of a growth model and on prediction of growth. *Forest Ecology and Management* 102, 203-212.
- Lappi, J., 2005. Plot size related measurement error bias in tree growth models. *Canadian Journal of Forest Research* 35, 1031-1040.
- Ledermann, T., Eckmüller, O., 2004. A method to attain uniform resolution of the competition variable Basal-Area-in-Larger Trees (BAL) during forest growth projections of small plots. *Ecological Modelling* 171, 195-206.
- Lewandowski, A., von Gadow, K., 1997. Ein heuristischer Ansatz zur Reproduktion von Waldbeständen. *Allgemeine Forst-und Jagdzeitung* 169, 170-174.
- Matérn, B., 1960. Spatial variation. *Meddelanden fraan statens skogsforskningsinstitut* 49, 1-144.
- Matérn, B., 1986. *Spatial Variation*. Springer-Verlag.
- Merganič, J., Sterba, H., 2006. Characterisation of diameter distribution using the Weibull function: method of moments. *European Journal of Forest Research* 125, 427-439.
- Metropolis, N., Rosenbluth, A.E., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Perspective on "Equation of state calculations by fast computing machines". *Journal of Chemical Physics* 21, 1087-1092.

-
- Morsdorf, F., Meier, E., Kötz, B., Itten, K.I., Dobbertin, M., Allgöwer, B., 2004. LIDAR-based geometric reconstruction of boreal type forest stands at single tree level for forest and wildland fire management. *Remote Sensing of Environment* 92, 353-362.
- Nagel, J., Biging, G.S., 1995. Schätzung der Parameter der Weibullfunktion zur Generierung von Durchmesserverteilungen. *Allgemeine Forst-und Jagdzeitung* 199, 185-189.
- Newnham, R.M., Maloley, G.T., 1970. The generation of hypothetical forest stands for use in simulation studies. Inf. Rep. FMR-X-26. Can. For. Serv. For. Manage. Inst., Ottawa, ON. 41.
- Pommerening, A., Stoyan, D., 2006. Edge-correction needs in estimating indices of spatial forest structure. *Canadian Journal of Forest Research* 36, 1723-1739.
- Pretzsch, H., 1997. Analysis and modeling of spatial stand structures. Methodological considerations based on mixed beech-larch stands in Lower Saxony. *Forest Ecology and Management* 97, 237-253.
- Pukkala, T., 1989. Prediction of tree diameter and height in a Scots pine stand as a function of the spatial pattern of trees. *Silva Fennica* 23, 83-99.
- R Development Core Team, 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Reed, D.D., Burkhart, H.E., 1985. Spatial autocorrelation of individual tree characteristics in loblolly pine stands. *Forest Science* 31, 575-594.
- Ripley, B.D., Kelly, F.P., 1977. Markov point processes. *Journal of the London Mathematical Society* 15, 188-192.
- Stage, A.R., Wykoff, W.R., 1998. Adapting distance-independent forest growth models to represent spatial variability: effects of sampling design on model coefficients. *Forest Science* 44, 224-238.
- Stoyan, D., Kendall, W.S., Mecke, J., 1995. *Stochastic geometry and its applications*. Wiley, Chichester.
- Stoyan, D., Penttinen, A., 2000. Recent applications of point process methods in forestry statistics. *Statistical Science* 15, 61-78.
- Uuttera, J., Haara, A., Tokola, T., Maltamo, M., 1998. Determination of the spatial distribution of trees from digital aerial photographs. *Forest Ecology and Management* 110, 275-282.
- Wilks, S.S., 1962. *Mathematical Statistics*. New York.
- Xie, T., Aickin, M., 1997. A truncated Poisson regression model with applications to occurrence of adenomatous polyps. *Statistics in Medicine* 16, 1845-1857.
- Zutter, B.R., Oderwald, R.G., Murphy, P.A., Farrar, R., 1986. Characterizing diameter distributions with modified data types and forms of the Weibull distribution. *Forest Science* 32, 37-48.

SUMMARY AND CONCLUSIONS

This dissertation aimed to advance individual tree mortality models, focusing on two main aspects: correcting for measurement error in competition variables in mortality models, and evaluating survival analysis techniques as an alternative to modelling background individual tree mortality. In this work I have implemented state of the art techniques that have proven to enhance the performance of background mortality models.

In the first chapter the implemented algorithm to correct for measurement error in competition variables successfully increased the discrimination ability of the models where the competition variable studied was statistically significant. These improvements were larger when the species were shade intolerant. The fitting and variable selection process revealed that past emphasis on DBH as a predictor variable for mortality, while producing models with strong metrics of fit, may make models less generalizable.

Results from the second chapter suggest that not only is survival regression adequate to model background mortality, but it is superior in terms of discrimination to the traditional logistic regression. The potential of survival regression to model and test changes in mortality through time and the superiority of this approach to the logistic regression open the door to the evaluation and modelling of tree mortality responses to environmental and human disturbances through time, e.g. thinning, droughts, allowing more flexibility exploring and modelling time dependent effects. Another advantage of survival regression is the ability to optimize the use of the information available in the calibration datasets.

The core of the most commonly used measurement error correction algorithms is the error variance. In the third chapter of this dissertation I evaluated the sensitivity of the estimate proposed by Stage and Wykoff (1998) to derive the local competition error variance from a single sample plot. I uncovered a major omission in the variance estimate that results in a noticeable overestimation of the true variance in most stand types. The study also revealed that variations in the estimate has a noticeable effect on

the predicted mortality, especially for those more shade intolerant species, which justify the need to correct the source of the general overestimation and adjust the variance estimate to account for variability in spatial patterns.

One of the elements that highlights the results of this dissertation is the fortunate availability of two independent datasets, which has allowed not only to fit parsimonious models which performs similarly to models with three times more variables, but has also given credibility to the claims of improvement made in this work (Hawkes, 2000).

Mortality is still a rare and highly unpredictable event, but the improvements, especially in terms of discrimination, presented in this dissertation if implemented in growth and yield models will result in more realistic stand structures and more sensible management decisions, which can potentially turn in more efficient use of forest resources.

1 References

Hawkes, C., 2000. Woody plant mortality algorithms: description, problems and progress. *Ecological Modelling* 126, 225-248.

Stage, A.R., Wykoff, W.R., 1998. Adapting distance-independent forest growth models to represent spatial variability: effects of sampling design on model coefficients. *Forest Science* 44, 224-238.