EMPIRICAL AND THEORETICAL ANALYSES OF THE APPLICABILITY OF PROJECTION MODELS AND MIXED MODELS TO FOREST INVENTORY UPDATES

by

CHENGCAI NI

(Under the Direction of Chris J. Cieszewski)

ABSTRACT

Projecting forest inventory plays an essential role in forest management. In this study I focused on techniques for projecting forest inventories, such as projection models, stand table projection techniques, and southern annual forest inventory system (SAFIS) sample plot updates.

I developed a new stand table projection model whose model form was derived based on the same assumptions as the Pienaar & Harrison equation (1988). The new stand table model is a two random effects model. It significantly outperformed the Pienaar & Harrison stand table projection model using data for Consortium for Accelerate Pine Plantation Studies (CAPPS). The new stand table modeling technique is an integration of a new expectation function, maximum likelihood estimation, and Empirical Best Linear Unbiased Predictor (EBLUP).

I proposed the quantile regression estimator for parameters of percentile growth models. According to extensive simulation analyses, the new estimator favorably compared with ordinary least squares in terms of the first order and second order statistics, especially when error terms are heteroscedastic. Simulation results indicated that the gain from quantile regression was approximately proportional to heteroscedasticity.

In forest biometrics, it is often the situation that only one observation is available for predictions due to investment and biological limitations. Accordingly, mixed models are not necessarily superior to projection models. However, mixed models are appropriate for updating SAFIS sample plots since multiple observations will become available as the inventory cycle repeats. EBLUP can provide the best prediction in comparison with any other methodologies available through a weighted scheme that uses information on individual observations and the population mean.

INDEX WORDS: Projection Model, Mixed Effect Model, CAPPS, Projecting Forest Inventory, Stand Table Projection, Quantile Regression, Percentile-Based Stand Table Projection, FIA, SAFIS, SAFIS Plot Updating, EBLUP, Disaggregation Model, Model-Based Imputation.

EMPIRICAL AND THEORETICAL ANALYSES OF THE APPLICABILITY OF PROJECTION MODELS AND MIXED MODELS TO FOREST INVENTORY UPDATES

by

CHENGCAI NI

B.S. Beihua University, China, 1988

M. S. Beijing Forestry University, China, 1991

M. S. The University of Georgia, 2004

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2005

EMPIRICAL AND THEORETICAL ANALYSES OF THE APPLICABILITY OF PROJECTION MODELS AND MIXED MODELS TO FOREST INVENTORY UPDATES

by

CHENGCAI NI

Major Professor: Chris J. Cieszewski

Committee: Barry B. Shiver
Bruce E. Borders
Daniel Hall
Richard F. Daniels

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2005

DEDICATION

This dissertation is dedicated to my family for all of their love and support through the years.

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to my advisor Chris Cieszewski for his guidance and support over the past four plus years. I would like to extend my gratitude to members of my dissertation committee, Barry Shiver, Bruce Borders, Dan Hall, and Richard Daniels for serving on my advisory committee. Their advice on my project enabled the completion of my dissertation. The fulfillment of this degree would not have been possible without the help of my dissertation committee.

Many thanks go to Ingvar Elle, research coordinator of the WSFR Fiber Supply Assessment Unit, for his help over the past four years. I am indebted to Ali Conner and Richard Harper, Southern Research Station and USFS, for preparing and providing FIA data for Georgia.

Special thanks to D.B. Warnell School of Forest Resources and the University of Georgia Graduate School for their financial support.

My gratitude is far beyond what I can express with words.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

This study investigates various aspects of modeling forest dynamics. One of two basic objectives of modeling is to understand and test hypotheses on the mechanism that generates data. Another objective is to forecast on the basis of the understood mechanism. In forestry, desired model properties, requirements for accounting for variations among individuals, and limited observations are often conflicting.

1.1. Growth and Yield Model Classification

According to the level of detail in the stand description, forest growth and yield models are classified into three categories: whole stand models; diameter class models; and individual models. Whole stand models describe a hypothesized functional relation between the underlying stand attribute and other attributes that are used as explanatory variables, e.g., stand age, site index, trees per acre, stand basal area, and quadratic mean diameter. Diameter class models represent a refinement of whole stand models to provide more detailed information on how the underlying stand attributes are distributed across size classes.

Diameter distributions can be estimated from mathematical models, including probability density functions (e.g., Normal, Beta, Weibull, Lognormal). When a predefined pdf is used, the idea is to estimate the parameters of the pdf such that the observed diameter distribution is adequately fitted. The parameters of the pdf are estimated as a function of stand level characteristics (e.g., stand basal area, site index, stand age, quadratic mean diameter, and diameter class percentiles) through a variety of methodologies. Once a diameter distribution model is identified, it can be linked with a growth function to predict future yields by diameter classes.

Models using the individual tree as the basic unit are referred to as individual models. Individual models represent a further level of refinement over whole-stand or diameter distribution models in that they simulate the growth of individuals rather than whole stands or diameter classes. Although these models can vary greatly in how they generate a list of individuals, they generally contain three primary components: diameter growth, height growth, and crown growth. When a mortality function is incorporated with these growth functions, future yields and stand structure can be simulated. These components are typically functions of stand age, site index, and stand density (TPA or stand basal area). Examples among many others are FVS (Wykoff et al., 1982), and PTAEDA (Daniels and Burkhart, 1975).

1.2. Approaches to Accounting for Variations Among Individuals

Clutter (1963) constructed compatible growth and yield equations for loblolly pine. Two remarkable contributions to forest biometrics were made by his celebrated research work, which are the idea of compatibility and the approach to accounting for variations among individuals. Clutter identified the necessity for compatibility between growth and yield functions, arguing future yield estimates obtained from the summation of the growth function over the projection interval should equal the future yield estimate obtained from the yield function for the same interval. Clutter derived a basal area equation (equation 1.1), which ensures compatibility between growth and yield functions, since it describes the basal area increment (in logarithm scale, ln BA_2-ln BA_1) during the projection interval from stand age A_1 to A_2, and more importantly, it can be viewed as a yield equation taking variations among stands into account by using ln BA_2 to identify the stand-specific coefficient responsible for variations from stand to stand.

$$\ln BA_2 = \left(\frac{A_1}{A_2}\right)\ln BA_1 + \beta_1\left(1-\frac{A_1}{A_2}\right) + \beta_2\left(1-\frac{A_1}{A_2}\right)S \tag{1.1}$$

$$\ln BA = \beta_1 + \beta_2 S + \beta_3\left(\frac{1}{A}\right) \tag{1.2}$$

where

BA=basal area, A=stand age, S=site index, and β=parameter

Suppose one has the hypothesized basal area yield equation 1.2 and that parameter β_3 is the stand-specific parameter. The equation, $\beta_3 = (\ln BA - \beta_1 + \beta_2 S)A_1$, can be used to calculate β_3. Plugging it into equation 1.2 gives equation 1.1. Clutter (1963) used a different approach to derive equation 1.1, i.e., differentiating equation 1.2 with respect to A, getting β_3 replaced using equation 1.2, and integrating the resulting differential equation between ln BA_2 and ln BA_1, with constraint such that ln BA_2=ln BA_1, when A_1=A_2. It is easy to see that ln BA_1 has two functions: serving to increment calculation; and providing specific information on the course of the underlying individual. For simplicity I call equations such as 1.1 as projection equations and models such as $\ln BA_{ij_2} = \left(\frac{A_{ij_1}}{A_{ij_2}}\right)\ln BA_{ij_1} + \beta_1\left(1 - \frac{A_{ij_1}}{A_{ij_2}}\right) + \beta_2\left(1 - \frac{A_{ij_1}}{A_{ij_2}}\right)S_i + e_{ij_2}$ as projection models.

In comparison with corresponding yield models (e.g., $\ln BA = \beta_1 + \beta_2 S + \beta_3\left(\frac{1}{A}\right) + e$), projection models can substantially decrease RMSE, the criterion that is widely used to jointly measure both estimator variance and bias. The research works of Lynch and Murphy (1995) and Lynch et al. (1999) exemplified the decrease in RMSE.

Estimated regression function 1.3 had a fit index of 0.95 and MSE of 20.02 ft^2 while equation 1.4 had a fit index of 0.98 and MSE of 8.53. The only difference between estimated regression functions 1.3 and 1.4 is whether or not β_1 is taken as a stand-specific coefficient and replaced with its estimator (equation 1.5).

$$\left(\hat{H}_i - h\right) = \hat{\beta}_1 \left(H_D - h\right)^{\hat{\beta}_2} Exp\left(\hat{\beta}_3 {D_i}^{\hat{\beta}_4}\right) \tag{1.3}$$

$$\left(\hat{H}_{i2} - h\right) = \left(H_{i1} - h\right)\left(\frac{H_{D2} - h}{H_{D1} - h}\right)^{\hat{\beta}_2} Exp\left(\hat{\beta}_3\left({D_{2i}}^{\hat{\beta}_4} - {D_{1i}}^{\hat{\beta}_4}\right)\right) \tag{1.4}$$

$$\frac{\left(H_{i1} - h\right)}{\left(H_{D1} - h\right)^{\hat{\beta}_2} Exp\left(\hat{\beta}_3 {D_{i1}}^{\hat{\beta}_4}\right)} = \hat{\beta}_1 \tag{1.5}$$

where

H=individual height, *h*=1.3m, *HD*=dominant height, and *D*=individual diameter at breast height (dbh).

Projection models have been widely applied in forest biometrics since Clutter (1963) first proposed them, including site index (Bailey & Clutter, 1974; McDill & Amateis, 1992; Cieszewski & Bailey, 1999; Cieszewski, 2001, 2002), basal area (e.g., Sullivan & Clutter, 1972; Pienaar & Harrison, 1988), volume (Clutter1963), stand table projection (Clutter & Jones, 1980; Pienaar & Harrison, 1988), and survival models (Clutter & Jones, 1980; Pienaar & Shiver, 1981).

The projection equations, derived through either differential-integral (Clutter 1963), or algebraic difference approach (ADA, Bailey & Clutter, 1974), or generalized algebraic difference approach (GADA, Cieszewski & Bailey, 1999), have the following desirable properties: 1.) If yield is projected from A_1 to A_2, and then from A_2 to A_3, the result should be identical with a projection from A_1 to A_3; 2.) The projected value at A_2 should be invariant to the choice of A_1; 3.)

The response remains the same when projection interval is zero. It is noteworthy that there are many equations that take the form of projection equations but lack these desirable properties.

To a certain degree the fact that projection models are able to account for variations among individuals is the reason that projection models were widely applied. Equation 1.3 is a representation of the course taken by the population average of the underlying attribute (individual height), paying no attention to any specific individual, whereas equation 1.4 takes into consideration to a certain degree the particular course of the individual height with the requirement that a prior observation provide information on the course of the subject individual. A substantial decrease in MSE was observed (from 20.02 ft^2 to 8.53 ft^2, see Lynch et al., 1999 for details).

Growth and yield models are usually constructed with repeated measurement data that are collected repeatedly on sampled subjects. Variations among observations mainly are from two sources: variations among subjects and variations within subjects.

Suppose that a function $f(x,\beta)$ may be specified to model the functional relation between the response and covariate *x*, where β is parameter vector. It is reasonable to assume that the function form of $f(x,\beta)$ is common to all subjects in the sampled population, but the parameter β vary across subjects. Accordingly, the mean response function for *i*th subject can be written as $E(y_{ij}|\beta_i)= f(x_{ij},\beta_i)$ so that *j*th response of *i*th subject corresponding to covariate x_{ij} is

$y_{ij} = f(x_{ij}, \beta_i) + e_{ij}$, where e_{ij} is the error term associated with y_{ij}, generally assumed it follows normal but not necessarily iid (identically, independently distributed). The model describes the systematic variation $E(y_{ij}|\beta_i) = f(x_{ij}, \beta_i)$ and e_{ij} random variation associated with measurements on the *i*th subject. It is easy to see that variation among subjects is accounted for through β_i. Parameters may vary due to variation unexplained by $f(x, \beta)$, for example, due to unobservable or excluded explanatory variable.

Projection models in forest biometrics can be viewed as mixed models that account for the variations among individuals through one parameter. If a projection model has the desired properties discussed by Clutter (1963), it can be broken down into two components: a mixed model, $y_{ij} = f(x_{ij}, \beta_i) + e_{ij}$, where only one element of β_i varies across subjects; and a predictor for β_i on the basis of one prior observation.

Generally, one parameter is not sufficient to account for variations among subjects (individuals). Accordingly, various approaches have been proposed, among which the commonly used is reparameterization that is to be discussed later. Since estimating β_i requires prior observations, the number of prior observations, one or more, and statistical test such as likelihood ratio test jointly determine how many parameters should be random. Often, the former is a dominant factor. Bredenkamp and Gregoire (1998) modeled diameter growth using a function that was developed by Schnute (1981). Schnute showed that many existing functions are

special cases of his function (Schnute, 1981; Zeide, 1997). Despite the derivations presented by Schnute (1981), I can show that the Schnute function is no more than a four-parameter Chapman & Richards model with two expected-value parameters (Ratkowsky, 1990).

$$y = \beta_1 \left(1 - \beta_2 e^{-\beta_3 A}\right)^{\left(\frac{1}{\beta_4}\right)} \tag{1.6}$$

$$y = \left({y_1}^{\beta_4} + \left({y_2}^{\beta_4} - {y_1}^{\beta_4} \right) \left(\frac{1 - e^{-\beta_3 (A - A_1)}}{1 - e^{-\beta_3 (A_2 - A_1)}} \right) \right)^{\frac{1}{\beta_4}} \tag{1.7}$$

$$\left(\frac{y}{\beta_1}\right)^{\beta_4} - 1 = -\beta_2 e^{-\beta_3 A} \tag{1.8}$$

$$\left(\frac{y_1}{\beta_1}\right)^{\beta_4} - 1 = -\beta_2 e^{-\beta_3 A_1} \tag{1.9}$$

$$\left(\frac{y_2}{\beta_1}\right)^{\beta_4} - 1 = -\beta_2 e^{-\beta_3 A_2} \tag{1.10}$$

$$1 - \frac{\left(\frac{y}{\beta_1}\right)^{\beta_4} - 1}{\left(\frac{y_1}{\beta_1}\right)^{\beta_4} - 1} = 1 - e^{-\beta_3 (A - A_1)} \tag{1.11}$$

$$1 - \frac{\left(\frac{y_2}{\beta_1}\right)^{\beta_4} - 1}{\left(\frac{y_1}{\beta_1}\right)^{\beta_4} - 1} = 1 - e^{-\beta_3 (A_2 - A_1)} \tag{1.12}$$

$$\frac{\left(\frac{y_1}{\beta_1}\right)^{\beta_4} - \left(\frac{y}{\beta_1}\right)^{\beta_4}}{\left(\frac{y_1}{\beta_1}\right)^{\beta_4} - \left(\frac{y_2}{\beta_1}\right)^{\beta_4}} = \left(\frac{1 - e^{-\beta_3 (A - A_1)}}{1 - e^{-\beta_3 (A_2 - A_1)}} \right) \tag{1.13}$$

The four-parameter Chapman & Richards function can be written as function 1.6 and the Schnute function is function 1.7. Solving function 1.6 for $\left(\frac{y}{\beta_1}\right)^{\beta_4}$ at A_1, A_2 and A yields function 1.8, 1.9, and 1.10, respectively. Dividing function 1.8 by 1.9 and 1.10 by 1.9 gives function 1.11 and 1.12. Further, dividing function 1.11 by 1.12 yields function 1.13. Solving function 1.13 for y leads to function 1.7, the Schnute function.

Function 1.7 can be interpreted in two different ways. One interpretation is that it is a new population model, having nothing to do with accounting for variations among the individual but having two new parameters (y_1 and y_2) with different parameter biological interpretations (Ratkowsky, 1990). In contrast, another interpretation is that it is an individual function, getting variations among individuals accounted for through two prior observations (β_1, β_2 are the individual-specific parameters in equation 1.6). Different interpretations lead to different applications (see Fang & Bailey, 2001; Bailey & Clutter, 1974). The second interpretation is most widely applied since it provides great flexibility. The Schnute function demonstrates a way to account for variations among individuals, i.e., p prior observations replace p individual-specific parameters.

One approach using one observation to account for multiple varying parameters is the Generalized Algebraic Difference Approach (GADA) first proposed by Cieszewski and Bailey

(1999). It has been successfully used to derive site index models (Cieszewski & Bella, 1989; Cieszewski, 2001, 2002; Rivas et al., 2004, among many others). The GADA makes it feasible to construct a site index model that is able to yield polymorphic curves with variable asymptotes.

If an additional assumption on β_i in $y_{ij} = f(x_{ij}, \beta_i) + e_{ij}$ that $\beta_i \sim N(0, D)$ is made, where D is the covariance matrix of β_i, $y_{ij} = f(x_{ij}, \beta_i) + e_{ij}$ becomes a mixed effect model. The mixed model approach provides a flexible and powerful tool for the analysis of longitudinal data. As Pinheiro and Bates (2000) stated, the increasing popularity of the mixed model approach is due to the flexibility it offers in modeling the within-subject correlation that is often present in longitudinal data, by the handling of balanced or unbalanced data in a unified framework, and by the availability of reliable and efficient software such as Splus LME, NLME, and SAS PROC NLMIXED, and PROC MIXED for model fitting.

If projection models can be viewed as simplified mixed models, Clutter (1963) is the first implicit application of mixed models in forestry. Biging (1985) used the random parameter approach to estimate the fixed parameters of site index curves. Lappi and Bailey (1989) applied a nonlinear mixed model to predict dominant height at both plot and individual tree levels. In the Lappi and Bailey model, the random effect entered into dominant height model linearly, so their model is a nonlinear marginal model (Demidenko, 2004). Other examples are Hall and Bailey (2001), Hall and Clutter (2004), Fang and Bailey (2001), Fang et al. (2001), and Calegario et al.

(2004). In these studies, EBLUP (empirical best linear unbiased predictor) was employed to predict dominant height, volume and basal area, on the basis of multiple prior observations.

One advantage of EBLUP is that it can be used to estimate as many random effects as desired with any number of prior observations, only if the model is differentiable with respect to these random parameters. In contrast, projection models restrict the number of individual specific parameters. Generally speaking, algebraic difference approach (ADA) is able to specify one individual specific parameter to account for the variations, whereas generalized algebraic difference approach (GADA, Cieszewski & Bailey, 1999) is able to use two in practice. The ADA approach is applicable to any base equation. The GADA can specify two parameters varying among individuals and is algebraically appropriate for fractional growth functions. The ADA and GADA do not require specifying a distributional function of individual-specific parameters since the distribution is irrelevant to estimation of the individual-specific parameters.

The more observations of an individual over time are available, the better the growth trajectory can be identified. One example using two prior observations in forestry is Bredenkamp and Gregoire (1988), modeling diameter with the Schnute function (Schnute 1981). Projection models are able to use two or more observations to increase the accuracy of predictions substantially, but not as efficient as EBLUP.

The requirement for multiple observations is not always satisfied in the applications of mixed models to forestry due to economical as well as biological limitations. In fact, often only one prior observation is available for predictions in forestry. In order to account for variations among parameters using one observation, the standard approach is to use the observation to account for one varying parameter, and to model others with simple functions so that there is negligible or no variation among individuals. The Clutter (1963) basal area equation can be thought as a perfect example to illustrate this approach. Equation 1.14 is the Schumacher model (Schumacher 1939), which, since it is sigmoid and can be easily linearized so that more stand level attributes can be incorporated into it, might be the most widely used model in forestry. Equation 1.15 is the basal area model in Clutter (1963), where A is stand age, BA is basal area, BA_{20} is basal age at 20, and SI is site index. Comparing equation 1.14 and 1.15 reveals that simple linear functions $\beta_0 = b_0 + b_1 S$ and $a_1 = b_2 + b_3 BA_{20} + b_4 S$ were specified to parameter β_0 and β_1, respectively, to explain variations among individuals. There is no doubt that replacing parameters with functions is always challenged by the data and unknown mechanisms of attribute interaction. Accordingly, one prior observation was used to explain β_1 variations among stands and equation 1.1 followed, whereas $\beta_0 = b_0 + b_1 S$ is still responsible for accounting for β_0 variations among stands.

$$\ln y = \beta_0 + \beta_1 \left(\frac{1}{A} \right) \tag{1.14}$$

$$\ln BA = b_0 + b_1 S + b_2\left(\frac{1}{A}\right) + b_3\left(\frac{BA_{20}}{A}\right) + b_4\left(\frac{S}{A}\right) \tag{1.15}$$

$$\ln BA = \ln BA_1 + \left(\ln BA_2 - \ln BA_1\right)\left(\frac{A^{-1} - A_1^{-1}}{A_2^{-1} - A_1^{-1}}\right) \tag{1.16}$$

If two observations are expected to be available, on the assumption that the Schumacher function is appropriate for basal area, equation 1.16 follows. It does not even require parameter estimation since two observations have already exhausted all parameters. Although equation 1.16 is always ready for applications and may be useful in the case no model is available, it does not find a comfortable niche in practical management use due to two drawbacks. First, it completely depends on two observations of a specific individual, failing to use information from similar stands to improve predictions. Second, the course of stand basal area might be distorted seriously by error terms contained in the two observations.

1.3. Model-Based FIA Sample Plot Updates

The USDA forest service has developed an annual inventory system featuring a hexagonal grid of Forest Inventory and Analysis (FIA) sample plots to be measured in 5-year inventory cycles, with 20% of the plots to be measured each year (Lessard 2001). Because inventories are conducted over 5 year cycles, data from the plots not measured in the current year will be 1 to 4 years old. Some methods have been proposed to eliminate this lag by estimating current conditions for FIA plots. One approach to calculating annual FIA estimates is to update to

the current year data for plots measured in previous years and to base estimates on the data for all plots. Imputation and Model-based updating techniques have been developed for annual forest inventories (Lessard, 2001; McRoberts, 1999, 2001).

Modeling FIA sample plots is different from common forest modeling and prediction in many ways.

First, since sample plots are systematically distributed throughout a state, each representing about 6,000 acres forestland, among FIA plots much more variation occurs in comparison with research plots. Consequently, it is reasonable to believe that variations among FIA sample plots are much larger than that among research sample plots. One solution to this problem is to use more explanatory variables, both quantitative and qualitative, to account for this sort of variation. Unfortunately, it does not seem feasible due to the unavailability of such additional information and because the primary purpose of forest inventory is to estimate population parameters, not to obtain data for modeling. In addition, how to employ additional covariates to account for parameter variations is very challenging. Another practical approach in forestry to explain variations among individual is the algebraic difference approach, which can be viewed as a simple mixed effect model. In an FIA context, this approach is not appropriate since one random coefficient is not sufficient to account for all variations among sample plots.

Secondly, the primary objective of modeling FIA sample plots is to update plots not measured in a current year for annual forest statistics so that the prediction period will be 1 to 4 years. However, long-term prediction is also desired since SAFIS (Southern Annual Forest Inventory System) is intended to improve estimation of both the current resource inventory and changes in resources (Roesch et al., 1999).

Last but most important, model systems constructed with FIA sample plot data are to be used for updating the very sample plots that models are based on. As the USDA annual inventory proceeds in a five-year cycle, more and more measurements of each sample plot will accumulate to provide informative data for prediction of individual plots. In this sense, it is important to make the best use of previous measurements to improve sample plot prediction. Lappi and Bailey (1988) stated that no previous measurement except the most recent one could be used for prediction with a projection model. Actually, there are two approaches for which a projection model can make use of more than one previous measurement. One simple approach to using all prior observations is to project first from every single observation to the desired projection age, then simply average these predicted values to reduce the biases caused by random components within individuals. The disadvantage of this approach is that it fails to specify the model correctly since only one parameter can be assumed to be varying across individuals. Another approach, as in Ratkowsky (1990) is to specify more than one specific parameter to

accommodate more previous measurements. The number of parameters in the base model limits the number of previous measurements that can be used by this approach. More importantly, projection models fail to provide algorithms as efficient as does EBLUP.

1.4. Using EBLUP To Update FIA Sample Plots

Mixed models provide a flexible and powerful tool for the analysis of longitudinal data. Mixed models allow one to account for multiple sources of heterogeneity and correlation in data through the inclusion of random effects in the model (Hall & Clutter, 2004). Data on FIA sample plots is typically collected repeatedly through time (five-year cycle), so correlation and heterogeneity are often present in such longitudinal data.

Applications of the mixed models in forestry mainly focused on predictions. Some examples are Lappi and Bailey (1988), Hall and Bailey (2001), Hall and Clutter (2004), Fang and Bailey (2001), and Fang et al. (2001). In these articles, the EBLUP was used to predict dominant height, volume and basal area, making use of multiple previous observations.

Mixed model prediction is performed through using the EBLUP (empirical best linear unbiased predictor) to predict random effects for a given subject individual. As BLUP (best linear unbiased predictor) implies, it is unbiased and has minimum variance among all linear functions of prior observations that are unbiased (Davidian & Giltinan, 1995, p.78). EBLUP has

links to several other areas of statistical methodology, including empirical Bayes methods, Kalman filtering, selection index, and Kriging (Hall & Clutter, 2004; Robinson, 1991).

I follow Pinheiro and Bates (2000) notations to present mixed model and EBLUP. The *j*th observation on *i*th plot is modeled as model (1.17)

$$y_{ij} = f\left(\phi_i, x_{ij}\right) + e_{ij} \tag{1.17}$$

where $\phi_i = A_i\beta + B_i b_i$, i=1,…, m; j=1,…,n_i, $b_i \sim N(0, D)$; m is the number of sample plots; n_i is the number of observations on ith plot; ϕ_i is plot specific parameter vector; the matrices A_i and B_i depend on plot (also possibly depend on the values of some covariates at the jth observation); β is a p dimensional vector of fixed effects. Random effect b_i is r dimensional vector of random effects; $e_{ij} \sim N\left(0, R_{ij}\right)$ and it does not have to be iid.

In general, the problem of predicting a random variable can be shown to be that of estimating its conditional mean, given the available data. Suppose a vector of n_i observations on plot i, say $Y_i = \begin{pmatrix} y_{i,1} \\ \vdots \\ y_{i,n_k} \end{pmatrix}$, corresponding to $X_i = \begin{pmatrix} x_{i,1} \\ \vdots \\ x_{i,n_k} \end{pmatrix}$, A_i, and B_i (X_i could be a matrix), the best predictor of b_i is the conditional mean of it, given a vector of the response Y_i from ith individual, as equation 1.1 shows:

$$\hat{b}_i = DZ_i{}^T\left(Z_i DZ_i{}^T + R_i\right)^{-1}\left(Y_i - f\left(A_i\hat{\beta}, X_i\right)\right) \tag{1.18}$$

where b_i is r by 1; $Z_i = \begin{pmatrix} Z_{i,1} \\ \vdots \\ Z_{i,n_i} \end{pmatrix}$ is n_i by r; $Z_{i,j} = \frac{\partial f\left(A_i\beta + B_i b_i, x_{ij}\right)}{\partial b_i^T}$; $\left(Y_i - f\left(A_i\hat{\beta}, X_i\right)\right)$ is n_i by 1; $R_i = \begin{pmatrix} R_{i,1} \\ \vdots \\ R_{i,n_i} \end{pmatrix}$ is n_i by n_i; D is r by r; Z_i is n_i by r.

Equation 1.18 is called the best linear unbiased predictor (BLUP). When the unknown covariance parameters are replaced by their REML or ML estimates, the resulting predictor is referred as to EBLUP. Small area estimation involves using direct survey information from areas of individual interest together with information on similar or related areas. It has been found that more precise estimates can be made if information on the other areas had not been ignored (Robinson, 1991). Like small area estimation, the EBLUP can borrow some strength from similar or related individuals to improve estimates of separate individual profiles. It is interesting to note that the predicted individual response can also be expressed as a weighted average of the estimated population-averaged mean and the individual observed response profile (see Fitzmaurice, 2004). In summary, in a SAFIS context mixed models in conjunction with EBLUP have the following advantages not obtained by other approaches.

1). Heterogeneity and correlation can be readily handled by the mixed model approach if they are present.

2). The EBLUP predicts observations of the response variable such that both multiple

measurements of the subject plot and information on similar or related plots can be jointly used. In addition the mixed model can use correlation information to improve predictions (Hall & Clutter 2004). A weighting scheme is automatically determined by the EBLUP according to variations within the plot and among plots (See chapter 4 for details).

3) Parameter estimation and random effect prediction are performed simultaneously (Pinheiro & Bates, 2000, chap. 7; Hall & Clutter, 2004) In this sense, modeling FIA sample plots may be thought of as a dynamic process, and mixed model estimation might be used as an algorithm for FIA plot updating.

4) The EBLUP can provide a means to make long-term prediction and a basis for decision-making. Lessard et al. (2001) employed individual diameter growth models to update FIA sample plots. It is unlikely that their model is reliable for long-term prediction because only one observation is utilized for prediction, though it is appropriate for updating (1 to 4 years). In addition, their model depends on reference DBH and cannot grasp the real growth profile.

1.5. Study Objectives

According to availability of prior observations for prediction, I classify growth and yield model applications into two categories: one observation and multiple observations. The latter is exemplified in SAFIS and some intensive forest management. Summarily, the objectives of this study are to

- Investigate the applicability of mixed models in the case where only one prior observation is available for predictions.
- Develop new models for stand table projection and new estimation methods to provide detailed information for management.
- Present a model system to update SAFIS sample plots.
- Illustrate the advantages of EBLUP in SAFIS plot updates for long-term trend analysis through theoretical analysis.

CHAPTER 2

A NEW STAND TABLE PROJECTION MODEL

2.1. Introduction

According to the level of detail in the stand description, growth and yield models can be classified into three categories: stand level models, size class models, and individual models. Stand level models describe the relation between one stand level attribute and others. In most situations, stand level models are likely to be the most appropriate for management. Approaches that use individual trees as basic units to predict growth and yield are referred to as individual models. Individual models use both stand level attributes and detailed information specifically related to each individual, including the coordinates, diameter, height, and crown dimensions. Individual models do not find a comfortable niche in practical management due to the very detailed inventory information they required, though these models can be useful as research tools to study spatial relationships or to provide insights into stand dynamics that could improve stand level models.

Diameter distribution models represent a refinement of whole stand models because they break stand level attributes into diameter classes. Goelz (2001) stated that all growth and

yield models are diameter distribution models that differ only in regard to which diameter distribution is employed and how the distribution is projected in the future. According to Goelz (2001), diameter distribution models (Bailey & Dell, 1973; and references cited therein, Hafley & Schreuder, 1977; Border et al., 1987; Borders & Patterson, 1990; Liu et al., 2002), disaggregation models (Ritchie & Hann, 1997a, 1997b), and stand table projection models (Clutter & Jones 1980; Pienaar & Harrison 1988; Nepal & Somers, 1992) fall into this category. Two fundamental ways to project the diameter distribution are either from stand-level attributes to individual trees or in the reverse order. The diameter distribution or individual tree can be either disaggregated from the projected stand attributes (Bailey et al. 1981; Borders 1989) or projected directly and aggregated with the stand attributes (Daniels & Burkhart 1988). One common characteristic of the category is that both stand level attributes and information specifically related to each individual diameter class or tree are included as explanatory variables to account for two variation sources: stand variation and individual variation nested in a stand.

Growth and yield disaggregation is generally based on either additive or proportional allocation. Most disaggregation functions are proportion allocation functions, which allocate growth or yield in proportion to estimated contribution to the total. Proportional allocations have been applied both to growth functions (Zhang et al., 1993) and to the allocation to yield in stand table projection (Clutter & Jones, 1980; Pienaar & Harrison, 1988; Nepal & Somers, 1992;

McTague & Stansfield, 1994, 1995). Additive disaggregation functions are mainly used to allocate growth among sublevel individuals (Dhote, 1994). One potential drawback of the additive function is that negative growth predictions may result (Ritchie & Hann, 1997b).

Stand characteristics explicitly used in functions cited by Ritchie and Hann (1997b) include mean basal area or quadratic mean diameter (Clutter & Allison, 1974; Pienaar & Harrison, 1988; Harrison & Daniels, 1988; McTague & Stansfield, 1994), basal area (Campbell et al., 1979; Clutter & Jones, 1980; McTague & Stansfield, 1994, 1995; Moore et al., 1994), volume (Dahms, 1983; Zhang et al., 1993), site index and dominant height (Harrison & Daniels, 1988), and stand density indicators such as tree per acre (TPA) and stand density index (SDI) (see Ritchie & Hann, 1997b). Individual characteristics includes one prior observation of relevant individual attribute, crown surface area (CSA), crown ratio (CR) or crown class (CC), basal area for tress larger than the subject tree (BAL), etc. (Lessard et al. 2001). Among these stand characteristics, mean basal area or quadratic mean diameter is the most important one because it contains information about BA and TPA and it is highly correlated to individual diameter. All the functions cited in Ritchie and Hann (1997b, p. 228) used one prior observation of the response as an explanatory variable. From the mixed model standpoint, the prior observation is used to identify the individual or predict the model parameter uniquely related to the individual. Since one prior measurement is used as an explanatory variable, reference

invariant property is desired or required. Although it is difficult to satisfy all the following desirable properties of disaggregation or stand table projection functions, insight into these properties will help construct consistent and sound projection equations. The following properties are desirable, though they may not constitute on exhaustive list.

The first property is invariance between the estimated aggregate of characteristics using stand level model and the aggregate of estimated individual characteristics using disaggregation model. In other words, if a certain characteristic is summable (i.e., TPA, basal area, and volume), the sum of the individual values projected by the estimated disaggregation function should be consistent with the corresponding values projected by the estimated stand level function. The invariance property is often maintained by adjusting the sum of individual values to equal the stand level estimate (Clutter & Jones, 1980; Pienaar & Harrison, 1988; Dahms, 1983; Nepal & Somers, 1992; McTague & Stansfield, 1994, 1995). Similarly, Somers and Nepal (1994) presented a general algorithm for maintaining invariance based on the assumption that the relative growth between individual trees remains constant and that individual predictions are subservient to the stand level model. Summarily, two main adjustment methods are proportional allocation (Clutter & Jones, 1980; Pienaar & Harrison, 1988; Dahms, 1983) and constant relative growth allocation (Somers and Nepal 1994). Proportional allocation can be summarized as $w_i = \frac{W}{\sum f(u_i)} f(u_i)$ (Ritchie & Hann, 1997a), where w_i is the adjusted individual

characteristic of interest, W is the estimated aggregate of w in the future, and $f(u_i)$ is a function of some estimated individual attribute u, often the same as w. It is easily seen that proportional allocation reallocates the estimated aggregate to every individual in proportion to its estimated contribution to the total. The allocation method of Somers and Nepal (1994) assumes that relative growth of any individual with respect to some selected reference individual remains the same after adjustments. Suppose basal areas are to be projected at both the individual level and stand level; solving $\left(\frac{b_{i2}-b_{i1}}{b_{r2}-b_{r1}}\right)=\left(\frac{k_i b_{i2}-b_{i1}}{k_r b_{r2}-b_{r1}}\right)$ for k_i and plugging the solution into $\hat{B}_2=\sum_{i=1}^{n_2} k_i \hat{b}_{i2}$ yields the estimator for k_r. Once k_r is estimated, k_i can be estimated, where B_2 is the stand basal area estimated by a stand basal area model, b_{i1} is the initial basal area of the ith tree, and b_{i2} is projected basal area of the ith tree, b_{r1} and b_{r2} are initial and projected basal areas for some selected tree r, respectively, and k_i and k_r are growth adjustments for ith tree and the reference tree r. It can be seen that the invariance property is satisfied with constraint such that $\hat{B}_2=\sum_{i=1}^{n_2} k_i \hat{b}_{i2}$.

In addition, when the total survival projection equation is available, the predicted total mortality must be distributed or mortality probabilities must be estimated. The allocation method of Somers and Nepal (1994) estimates k_r and survival adjustment simultaneously, whereas other authors distributed mortality before adjusting individual estimates to maintain the invariance property.

Since all disaggregation functions or stand table projections involve one observation of the response and explanatory variables, which can be thought to be functions of stand age, the reference invariance property is desired. Reference invariance is a very important property of a projection equation. A projection equation lacking this property is dependent on the choice of stand age and demonstrates different equation behavior if different reference ages are used; however stand age may not explicitly be included in the equation since most explanatory variables in a forest model system are functions of stand age. One simple criterion to determine whether a projection equation is reference invariant is to see whether the equation is the same as its inverse with respect to y_1, where y_1 denotes observation of the response. Since Clutter (1963) introduced the idea of compatibility between growth and yield functions, the idea has been accepted as an important property for such functions. Ramirez-Maldonado et al. (1987) demonstrated that such a relationship holds true for functions derived through the algebraic different approach (ADA) (Bailey & Clutter, 1974; Borders et al. 1984). In this sense, a projection equation derived either through derivative and integrals (Clutter, 1963) or ADA actually is a growth function. Consider the Schumacher function, $\ln y = \alpha + \left(\frac{\beta}{A}\right)$, if you would like to derive a growth function from A_1 to A_2, integrating $\frac{\partial(\ln y)}{\partial A} = \left(\frac{-\beta}{A^2}\right)$ from A_1 to A_2 yields $\ln y \left|\begin{pmatrix} y_2 \\ y_1 \end{pmatrix} = \left(\frac{\beta}{A} + c\right)\right|\begin{pmatrix} A_2 \\ A_1 \end{pmatrix}$ and $\Delta(\ln y) = \ln y_2 - \ln y_1 = \beta\left(A_2^{-1} - A_1^{-1}\right)$, with the initial condition

that $\ln y_2 = \ln y_1$ when A_1=A_2. The growth function is equivalent to $\ln y_2 = \ln y_1 + \beta\left(A_2^{-1} - A_1^{-1}\right)$. In a similar way, it can be shown that $\ln y_2 = \alpha + \left(\ln y_1 - \alpha\right)\left(A_1 A_2^{-1}\right)$ or, equivalently, $\Delta\left(\ln y\right) = \left(\ln y_1 - \alpha\right)\left(A_1 A_2^{-1} - 1\right)$, with parameter β replaced with the other parameter. Although various interpretations of projection equations have been made (Clutter 1963; Bailey & Clutter, 1974; Ratkowsky, 1990; Cieszewski, 2001, 2002), one important interpretation is that a projection function can be viewed as either a growth function or a yield function.

Disaggregation functions should be able to account for at least two sources of variation: stand level variation and individual variation. Stand level variation is usually accounted for by incorporating stand level attributes such as site index, quadratic mean diameter, basal area, and volume, whereas individual level variation is often accounted for by one prior observation of the response and other individual attributes. Individuals in the same stand should have different curve shapes and variable asymptotes, as are desired for site index models. This property requires that at least two parameters vary from individual to individual within the same stand and is extremely difficult to achieve since one observation can account for only variation of one parameter. If other individual attributes are included and the function is reference invariant, the model system must have additional components to estimate the projected values of such involved attributes as diameter, height, crown ratio, etc., and these projected values require other individual models that may not necessarily help improve the performance of whole system.

A disaggregation method by means of a specific growth function is intuitively appealing because the disaggregation function is an implied individual growth function (Ritchie & Hann, 1997a). Examples are McTague and Stansfield (1995, 1996), Clutter and Jones (1980), and Pienaar and Harrison (1988), and Nepal and Somers (1992). The Pienaar and Harrison function actually is a revised version of the Clutter and Jones function (1980). The disaggregation function proposed by McTague and Stansfield (1994) is as follows:

$$d_{i2}^2 = d_{i1}^2 + \beta_1 d_{i1}^2 \left(\frac{BA_2}{BA_1} \right) (A_2 - A_1) + \beta_2 d_{i1}^2 \left(\frac{N_2}{N_1} \right) (A_2 - A_1) + \beta_3 d_q (A_2 - A_1) \quad (2.1)$$

$$ba_{i2} = BA_2 \frac{d_{i2}}{K \sum d_{i2}} \quad (2.2)$$

where d_{i2}^2 and d_{i1}^2 are squared diameters at breast height of ith tree at stand age A_2 and A_1, respectively; d_q is the quadratic mean diameter at A_1; BA_2 and BA_1 are the estimated stand basal area corresponding A_2 and the observed stand basal area corresponding to A_1; N is trees per acre (TPA) K is conversion constant from diameter squared to basal area; ba_{i2} is the adjusted projected basal area for ith tree. Function 2.1 was used to project individual diameter squared forward over time, whereas function 2.2 was used to ensure the invariance between the estimated basal area at stand level and the sum of individual basal areas. Function 2.1 has all the desirable properties except reference invariance: solving function 2.1 for d_{i1}^2 does not yield an equation

that is identical to equation 2.1. The curve shape of function 2.1 depends on the choice of A_1, though it possesses logical behavior in that $d_{i2}^2=d_{i1}^2$ when $A_2= A_1$.

Only a few of the 15 functions cited in Ritchie and Hann (1997b) have reference invariance across the stand level and individual level, and were derived from an individual growth function. Examples include Clutter and Allison (1974, cited in Ritchie and Hann 1997b), Clutter and Jones (1980), and Pienaar and Harrison (1988). Actually, these functions are essentially the same. More detailed analysis of the Pienaar-Harrison function (1988) is given in the following section and a new stand table projection function is derived based on their assumptions.

2.2. Analysis of the Pienaar-Harrison Projection Equation

The called Pienaar-Harrison projection equation was first developed by Clutter and Allison in 1974 (Ritchie and Hann, 1997). Clutter and Jones (1980) used it for stand table projection of slash pine plantations after thinning. Subsequently revised by Pienaar and Harrison (1988). The equation is able to reproduce multimodal distribution and mathematically simple. Tagged diameter remeasurement data is a special requirement when this equation is fitted, and an initial stand table is required for application. Borders and Patterson (1990) showed it to be superior to the parameter recovery method and the percentile-based projection method. Other applications of the Pienaar-Harrison equation can be seen in Knowe and Hibbs (1996), Knowe et

al. (1997), Knowe (1994), Dyer (1997), and Nepal and Somers (1992). Dyer (1997) examined the Harrison-Daniels equation (1988, cited in Dyer 1997), the Pienaar-Harrison equation (1988), and a new disaggregation function. The Pienaar-Harrison equation was found to perform slightly better than others in terms of mean absolute residual based on 12-year projections.

As shown below, the Pienaar-Harrison equation can be derived from the Schumacher function based on the assumption that only β_2 is an individual-specific parameter in the Schumacher function, $d = \beta_1 Exp\left(\frac{-\beta_2}{A^{\beta_3}}\right)$. Zeide (1989, 1997) argued that the relative growth rate of diameter is a power rather than an exponential function of age. He evaluated the Chapman & Richards function, the Logistic function, the Weibull function, and the Schumacher function, which was called power decline function in his article and found the Schumacher function to be twice as accurate as the next best (Chapman-Richards function) and about five times as accurate as the Logistic function. Assuming that mortality was nonexistent or evenly distributed across diameter, Bailey (1980) showed that equation (2.3) is the diameter equation for some diameter distributions such as Weibull, Lognormal, and Generalized Gamma that will preserve the functional form of the distribution, where β_k (k=0, 1, 2, or 3) is parameters, and d_2 and d_1 are diameters at A_1 and A_2, respectively. Bailey (1980) derived the Schumacher function based on the further assumption such that $\beta_0 = \beta_3 = 0$ and that $\beta_2 = \left(\frac{A_2}{A_1}\right)^{k-1}$, using differentiation of both

sides of equation 2.3 with respect to A and integrating the result. Certainly, other functions, such as the Chapman-Richard, Logistic, and Gompertz functions, can be derived from $\beta_2 = \left(\frac{A_2}{A_1}\right)^{k-1}$ based on different assumptions.

$$(d_2 - \beta_0) = \beta_1 (d_1 - \beta_3)^{\beta_2} \tag{2.3}$$

A variation of equation 2.3, $d_2 = a_2 + b_2 \left(\frac{d_1 - a_1}{b_1}\right)^{\left(\frac{c_1}{c_2}\right)}$, derived from the Weibull distribution cumulative density function (CDF), was used by Nepal and Somers (1992) and Cao and Baldwin (1999) to project stand tables; a_i, b_i, and c_i (i=1, 2) are all Weibull distribution parameters corresponding to location, scale, and shape parameter, respectively. The parameter recovery method was employed to estimate Weibull distribution parameters in their studies. Algorithms of Nepal and Somers (1992) and Cao and Baldwin (1999) make use of Weibull distribution and prior observations for projection; obviously because their algorithm are a combination of distribution-based method and disaggregation function, they are not applicable to multimodal diameter distribution.

Deriving Pienaar-Harrison Equation From Schumacher Function

The simplicity of Pienaar-Harrison projection equation (equation 2.4) is an attractive feature. The function is reference invariant and derived from the Schumacher function. The invariance between the sum of estimated individual basal areas and projected total basal area can

be maintained by equation 2.5

$$\left(\frac{b_{i2}}{\overline{b}_2}\right)=\left(\frac{b_{i1}}{\overline{b}_1}\right)^{\left(\frac{A_1}{A_2}\right)^{\beta_3}} \tag{2.4}$$

$$b_{i2}=BA_2\frac{\left(\frac{b_{i1}}{\overline{b}_1}\right)^{\phi}}{\sum_{i=1}^{n_2}\left(\frac{b_{i1}}{\overline{b}_1}\right)^{\phi}} \tag{2.5}$$

where $\phi=\left(\frac{A_1}{A_2}\right)^{\beta_3}$, b_{i1} and b_{i2} are the *i*th tree basal area at stand age A_1 and A_2, $\overline{b}_1$ is the mean basal area of the subject stand at stand age A_1, β_3 is the only parameter to be estimated, n_2 is the number of survivals at age A_2, and BA_2 is the projected stand basal area at age A_2. When total survival projection model is available, the stand table projection model requires that the estimated total mortality be distributed over the estimated stand table at A_2.

If the diameter distribution is unimodal, the quadratic mean has a similar growth pattern to most of the individual diameters because it is the second moment of the diameter distribution. Accordingly, it is reasonable to assume that an individual diameter function can be used for quadratic mean. Suppose that growth functions for the *i*th individual and the quadratic mean are equations 2.6 and 2.7, respectively

$$d_i=\beta_{1i}Exp\left(\frac{-\beta_{2i}}{A^{\beta_3}}\right) \tag{2.6}$$

$$d_q = \beta_{1q} Exp\left(\frac{-\beta_{2q}}{A^{\beta_3}}\right) \tag{2.7}$$

where all individuals have the same parameter β_3, and d and d_q are the ith diameter and quadratic mean diameter, respectively. Dividing equation 2.6 by 2.7 yields equation 2.8

$$\frac{d_i}{d_q} = \alpha_{1i} Exp\left(\frac{\alpha_{2i}}{A^{\alpha_3}}\right) \tag{2.8}$$

where $\alpha_{1i} = \frac{\beta_{1i}}{\beta_{1q}}, \alpha_{2i} = \beta_{2i} - \beta_{2q}$, and $\alpha_3 = \beta_3$. The ratio of the ith diameter to the quadratic mean diameter also follows the Schumacher function. Furthermore, if both sides of equation 2.8 are squared, $\frac{kd_i^2}{kd_q^2} = \alpha_{1i}^2 Exp\left(\frac{2\alpha_{2i}}{A^{\alpha_3}}\right)$, it follows that $\frac{b_i}{\bar{b}} = \phi_{1i} Exp\left(\frac{\phi_{2i}}{A^{\phi_3}}\right)$, where k =0.005454 is the conversion from diameter square to basal area in square units if diameter units are inches and basal area units are square feet; b_i is the basal area of the ith tree, and $\bar{b}$ is the arithmetic basal area mean of all survivals; $\frac{b_i}{\bar{b}}$ is the relative size, which was defined by Pienaar and Harrison (1988). If only one initial stand table is available for prediction, either ϕ_{1i} or ϕ_{2i} has to be assumed to be a global parameter. The best value for ϕ_{1i} would be 1 for a polymorphic function with a single asymptote (ϕ_{1i} =1 means all individual diameters have the same asymptote). Replacing the parameter ϕ_{2i} with one prior observation of the relative size $\frac{b_{i1}}{\bar{b}_1}$ by ADA (Bailey & Clutter, 1974), equation 2.4 follows.

The derived equation is exactly the same as the one proposed by Pienaar and Harrison in 1988. It is noteworthy that the procedure I used to derive Pienaar and Harrison projection equation is only one possible way. Due to the assumption only one initial stand table is available for prediction, two parameters are specified as global parameters. Conversely, from the Pienaar-Harrison equation, the Schumacher function can also be derived.

It should be noted that other stand table projection equations could be derived following the approach described above. For example, if you start with the Chapman-Richard function, $y_i = \beta_{1i}\left(1-e^{(-\beta_2 A)}\right)^{\beta_{3i}}$ and $y_r = \beta_{1r}\left(1-e^{(-\beta_2 A)}\right)^{\beta_{3r}}$, then a stand table projection equation $y_{i2} = y_{r2}\left(\frac{y_{i1}}{y_{r1}}\right)\left(\frac{1-Exp(-\alpha_1 A_2)}{1-Exp(-\alpha_1 A_2)}\right)^{\alpha_2}$ can be obtained with the asymptote parameter replaced, where $\alpha_1 = \beta_2$, $\alpha_2 = \beta_{3i} = \beta_{3r}$, and y_r is reference stand parameter (e.g. quadratic mean diameter) with which individual diameters are to be compared.

Deriving Schumacher Function from the Pienaar-Harrison Projection Equation

It is straightforward to derive equation 2.9 conversely from equation 2.4. Equation 2.4 is equivalent to $\left(\frac{b_{i2}}{\bar{b}}\right)^{A_2{}^{\beta_3}} = \left(\frac{b_{i1}}{\bar{b}_1}\right)^{A_1{}^{\beta_3}} = 1$, which means that $\left(\frac{b_i}{\bar{b}}\right)^{A^{\beta_3}} = k(.) > 0$ holds for any stand age because A_1 and A_2 are two arbitrary stand ages, where k (.) must be a constant or any function of stand or individual attributes that can be viewed as constants. Accordingly, equation 2.9 holds for any individual.

If one further assume that equation 2.4 was derived based on the assumption that the mean basal area and individual basal area follow the same growth function with one local parameter, from equation 2.9 (β_2 in place of ln k(.)), it is concluded that the individual basal area must take the form of $b_i = u(.)Exp\left(\frac{\beta_{2i}}{A^{\beta_3}} + h(.)\right)$, which is equivalent to $b_i = \beta_1\ Exp\left(\frac{\beta_{2i}}{A^{\beta_3}}\right)$, where, u(.) and h (.) are constants or functions like k(.).

Compatibility Between the Aggregation of Individuals and Estimated Aggregate

In order to make the sum of projected individual basal areas compatible with predicted basal area using a stand level basal area equation, equation 2.4 can be adjusted by multiplying RHS with an adjustment coefficient w, and equation 2.10 follows:

$$\left(\frac{b_i}{\bar{b}}\right) = Exp\left(\ln k(.)A^{-\beta_3}\right) = Exp\left(\frac{\beta_2}{A^{\beta_3}}\right) \tag{2.9}$$

$$\tilde{b}_{i2} = w\bar{b}_2\left(\frac{b_{i1}}{\bar{b}_1}\right)^{\left(\frac{A_1}{A_2}\right)^{\beta_3}} \tag{2.10}$$

where $\tilde{b}_{i2}$ is the adjusted predicted basal area of the ith diameter at A_2, and w is the adjustment coefficient that ensures the summation of all individual diameters at A_2 equal to the projected total basal area A_2 from a separate basal area model (i.e. $\sum_{i=1}^{n_2}\hat{b}_{i2} = BA_2$). Summing both sides of

equation 2.10 and solving for *w* yields $BA_2 = \sum_{i=1}^{n_2} \tilde{b}_{i2} = w \sum_{i=1}^{n_2} \bar{b}_2 \left(\frac{b_{i1}}{\bar{b}_1} \right)^{\left(\frac{A_1}{A_2} \right)^{\beta_3}}$ and

$w = \frac{BA_2}{\sum_{i=1}^{n_2} \bar{b}_2 \left(\frac{b_{i1}}{\bar{b}_1} \right)^{\left(\frac{A_1}{A_2} \right)^{\beta_3}}} = \frac{BA_2}{\bar{b}_2 \sum_{i=1}^{n_2} \left(\frac{b_{i1}}{\bar{b}_1} \right)^{\phi}}$, where $\phi = \left(\frac{A_1}{A_2} \right)^{\beta_3}$. Finally, equation 2.11 and 2.12 are

derived by replacing *w* in equation 2.10 with $w = \frac{BA_2}{\bar{b}_2 \sum_{i=1}^{n_2} \left(\frac{b_{i1}}{\bar{b}_1} \right)^{\phi}}$.

$$\tilde{b}_{2i} = \left(\frac{BA_2 \left(\frac{b_{i1}}{\bar{b}_1} \right)^{\phi}}{\sum_{i=1}^{n_2} \left(\frac{b_{i1}}{\bar{b}_1} \right)^{\phi}} \right) \tag{2.11}$$

$$\tilde{b}_{2i} = \frac{n_2}{\sum_{i=1}^{n_2} \left(\frac{b_{i1}}{\bar{b}_1} \right)^{\phi}} \bar{b}_2 \left(\frac{b_{i1}}{\bar{b}_1} \right)^{\phi} \tag{2.12}$$

If individual diameters are used in equation 2.12 instead of individual basal areas, then an equation identical to the Clutter-Jones equation (Clutter & Jones, 1980) can be obtained.

$$\tilde{d}_{2i} = \left(\frac{BA_2 \left(\frac{b_{i1}}{\bar{b}_1} \right)^{\phi}}{k \sum_{i=1}^{n_2} \left(\frac{b_{i1}}{\bar{b}_1} \right)^{\phi}} \right)^{\frac{1}{2}} \tag{2.13}$$

From equation 2.11, it is easy to see that the equation reallocates the observed or predicted stand basal area to every individual in proportion to its predicted contribution to the

total. Any stand table projection equation may be reformulated in such a way to maintain invariance between observed or predicted stand yield and the aggregation of individual yields. The following generalization of adjusted individual prediction ensures that the sum of individual yields is equal to the stand level observation or prediction.

$$\tilde{b}_{2i} = \frac{f(x_1, x_2, b_{i1})}{\sum_{i=1}^{n_2} f(x_1, x_2, b_{i1})} \hat{B}_2 \tag{2.14}$$

where $f(x_1, x_2, y_{i1})$ is the function projecting an attribute of the ith individual from A_1 to A_2 using y_{1i} and x_i as explanatory variables; $\hat{B}_2$ is the predicted aggregate of b; n_2 is survivals at A_2, and b is the individual characteristic of interest, individual diameter or basal area in this case. It is apparent that $\sum_{i=1}^{n_2} \tilde{b}_{2i} = \hat{B}_2$ always holds true. It should be noted that this adjustment method requires that mortality allocation be done first. Somers and Nepal (1994) presented a complicated algorithm to allocate mortality and adjust diameter estimates simultaneously when there is no mortality probability model available. Their adjustment algorithm is based on the assumption that relative growth between individuals remains constant.

2.3. A New Stand Table Projection Model

Pienaar and Harrison (1988) claimed that comparing the relative size of individual survivors over long projection intervals indicated that the relative contribution of smaller than

average-sized survivors to the total size decreased over time, whereas the largest relative size increased over time. However, empirical evidence from CAPPS data disagrees with their results, showing that relative contribution is not significantly increasing or decreasing (see Figure 2.1). There is no doubt that the quadratic mean is highly correlated to individual diameters since it always has a value close to most of the individual diameter if the diameter distribution is unimodal. It might be the best stand parameter to account for variation among stands. Figure 2.2 shows a strong relation between individual diameter and quadratic mean diameter. Based upon the same assumption as the Pienaar and Harrison equation, the following power function is available for modeling the relationship between an arbitrary diameter and quadratic mean diameter. Equation 2.6 and 2.7 are as follows:

$$d_i = \beta_{1i} Exp\left(\frac{-\beta_{2i}}{A^{\beta_3}}\right) \tag{2.6}$$

$$d_q = \beta_{1q} Exp\left(\frac{-\beta_{2q}}{A^{\beta_3}}\right) \tag{2.7}$$

Through simple algebraic rearrangement, equation (2.15) and (2.16) result

$$A^{\beta_3} = (-\beta_{2i})\left\{\ln\left(\frac{d_i}{\beta_{1i}}\right)\right\}^{(-1)} \tag{2.15}$$

$$A^{\beta_3} = (-\beta_{2q})\left\{\ln\left(\frac{d_q}{\beta_{1q}}\right)\right\}^{(-1)} \tag{2.16}$$

Equating equation 2.15 and 2.16 yields equation 2.17. Equation 2.18 follows with further algebraic rearrangement. Reparameterization yields equation 2.19 and 2.20, where $\beta = \left(\frac{\beta_{2i}}{\beta_{2q}} \right)$ and $\alpha = \frac{\beta_{1i}}{\beta_{1q}{}^{\beta}}$, d_q is stand quadratic mean, and two parameters uniquely specify the relationship between the quadratic mean and the *DBH* of a given individual.

$$\left(\frac{\beta_{2i}}{\beta_{2q}} \right) = \ln\left(\frac{d_i}{\beta_{1i}} \right) \left(\ln\left(\frac{d_q}{\beta_{1q}} \right) \right)^{-1} \tag{2.17}$$

$$d_i = \beta_{1i} \left(\frac{d_q}{\beta_{1q}} \right)^{\left(\frac{\beta_{2i}}{\beta_{2q}} \right)} \tag{2.18}$$

$$d_i = \beta_{1i} \left(\frac{d_q}{\beta_{1q}} \right)^{\beta_i} \tag{2.19}$$

$$d_i = \alpha_i d_q{}^{\beta_i} \tag{2.20}$$

In addition to the Schumacher function, the power function can also be derived from other growth equations listed in Table 2.1. A diameter distribution percentile function for PMRC loblolly pine data, $\ln P_x = \beta_0 + \beta_1 \ln\left(\frac{BA}{TPA} \right)$, similar to model 2.20, was proposed by Harrison and Borders (1996), where P_x is xth percentile. This function achieved reasonable fit while preventing illogical crossover of adjacent percentiles. All functions in Table 2.1 have two

parameters vary across from individuals to another and other parameters common to all. They can be generalized in the form of $d_i = a_i h(A)^{b_i}$, where $h(A)$ is a function excluding parameters *a* and *b*. Furthermore, if you assume that quadratic mean diameter growth follows the same growth pattern with different values for parameters *a* and *b* and then relate each single diameter to the quadratic mean diameter, then equation 2.20 can be obtained.

Data and Models

Data used in this study is from the Consortium for Accelerate Pine Plantation Studies (CAPPS) initiated in 1987 and maintained by the Daniel B. Warnell School of Forest Resources, University of Georgia. Each installation had six blocks in which four 0.15 ha treatment plots were assigned one of four treatments. On each treatment plot, a 0.05ha measurement plot was centered and approximately 80 loblolly pine seedlings were planted at 2.4 by 2.4m spacing. The treatments were (1) Herbicide, (2) fertilization, (3) a combination of herbicide and fertilization, and (4) control. For detailed information, refer to Zhang et al. (2002) and Borders and Bailey (2001).

Out of CAPPS 148 plots, 100 sample plots were randomly selected to construct fit data, and the rest compiled the validation dataset. First, I fit the fit dataset to model 2.21, incorporating plot level random effects and individual diameter level random effects:

$$d_{ijk} = \phi_{1,ij}\left(d_{q,ik}\right)^{\phi_{2,ij}} + e_{ijk} \tag{2.21}$$

where

d_{ijk} = Diameter of the *j*th tree in the *i*th plot on *k*th measurement occasion

$d_{q,ik}$ = Quadratic mean diameter of the *i*th plot on *k*th measurement occasion

e_{ijk} = Random error term assumed to follow normal distribution with mean 0. The subscript *k* indicates that *e* might depend on the measurement occasion.

$\phi_{1,ij} = \beta_1 + P_{1,i} + t_{1,ij}$

$\phi_{2,ij} = \beta_2 + P_{2,i} + t_{2,ij}$

where β_is' are fixed parameters, $P_{h,i}$ (*h*=1, 2) is the plot random effect, and $t_{h,ij}$ is the tree random effect which is nested in plot. I assume bivariate normal distributions for both the tree level random effects and plot level effects such as $p = \begin{pmatrix} p_{1,i} \\ p_{2,i} \end{pmatrix} \sim N\left(0, \Sigma_1\right)$ and $t = \begin{pmatrix} t_{1,ij} \\ t_{2,ij} \end{pmatrix} \sim N\left(0, \Sigma_2\right)$. Random variables *p*, *t*, and *e* are mutually independent. Parameters were estimated with Splus NLME. Maximum likelihood estimates are presented in Table 2.2.

The fixed parameter estimates are both close 1: $\hat{\beta}_1$ =1.020464 and $\hat{\beta}_2$=0.991821. These estimates imply most individual diameters have a growth pattern similar to their corresponding quadratic mean and that the relative size growth pattern is almost a nearly horizontal line as shown by Figure 2.1. By comparing $V(P_1)$=0.0022 to $V(t_1)$=0.0986, and $V(p_2)$=0.0014 to V

(t_2)=0.0220, I notice that t_1 and t_2 account for a large proportion of variations among individual trees. It naturally follows that incorporating quadratic mean significantly decrease variation among stands. High correlation coefficient (0.999) between p_1 and p_2 (plot level) indicates that model 2.21 is overfitted. In addition, one initial stand stable provides insufficient information to give accurate estimates for four random effects. Therefore, I reduced model 2.21 to 2.22 by dropping two plot level random effects.

$$d_{ijk} = \phi_{1,ij}\left(d_{q,ik}\right)^{\phi_{2,ij}} + e_{ijk} \tag{2.22}$$

where, $\alpha_{ij} = \phi_1 + t_{1,ij}$, $\beta_{ij} = \phi_2 + t_{2,ij}$, and $t_{h,ij}$ is the tree level random effect which is nested in plot. I still assume bivariate normal distributions for the two random effects $t = \begin{pmatrix} t_{1,ij} \\ t_{2,ij} \end{pmatrix} \sim N\left(0, \Sigma_2\right)$ and that the random effect *t* and *e* are mutually independent. Parameter estimates for model 2.22 is presented in Table 2.3. Residuals vs. fitted dbh plot and observed dbh vs. fitted plot are presented in Figure 2.3 and 2.4, respectively. These visual inspections show that normality assumptions are not violated by model 2.22.

A general method for comparing nested model fit by maximum likelihood is the likelihood ratio test (LRT). If L_2 is the likelihood of the more general model and L_1 is the likelihood of the reduced model, $2\log\left(\frac{L_2}{L_1}\right) = 2\left(\log\left(L_2\right) - \log\left(L_1\right)\right)$, the likelihood ratio test statistic, can be used to test which model is more appropriate. The asymptotic distribution of the LRT

statistic is $\chi^2(k_2 - k_1)$ under the null hypothesis that the reduced model is adequate, where k_2 is the number of parameters to be estimated in the full model, and k_1 is the number of parameters to be estimated in reduced model (see Pinheiro and Bates 2000). Table 2.4 shows the likelihood ratio test results under the hypothesis that model 2.22 is as adequate as model 2.21. There is a significant increase in AIC and BIC, as evidenced by the larger value for the likelihood ratio test, indicating that model (2.21) gives a better fit. However, constrained to the assumption I made that one initial stand table is available for projection, plot level random effects are excluded. One solution is to incorporate installation, block, and treatment as covariate to account for parameter variations among stands. Since the estimated variance of plot level random effects is relatively small and I intend to validate the applicability of model 2.22 for general management situations, no covariate is used in the reduced model. As for individual tree diameters, I use an initial stand table to predict random effects $t_{h,ij}$ (h=1,2).

Model Comparisons

In this section, three models, the Pienaar-Harrison model 2.23, the new projection model 2.24, and the mixed model 2.22 are fitted to the fit data (100 randomly selected plots out 148). No convergence was obtained for $\frac{b_{ijk}}{\bar{b}_{ik}} = Exp\left(\frac{\phi_{2ij}}{A_{ijk}\phi_3}\right) + e_{ijk}$ by using Splus NLME, probably because $\frac{b_{ijk}}{\bar{b}_{ij}}$ does not show any significant trends over time (see Figure. 2.1), almost a

horizontal line after age 6, where $\frac{b_{ijk}}{\bar{b}_{ij}}$ is relative size of *j*th tree of *i*th plot on *k*th measurement occasion. Accordingly, these two projection models (2.23 and 2.24) were fitted using the ordinary least square for the convenience of comparison. The projection function of model 2.24 actually is the same as the function of mixed model 2.22 except it is reparameterized in a different way for convergence. By assuming that $\beta_{1i} = \beta_{1q} = a$ in equation 2.19, $\left(\frac{d_i}{a}\right) = \left(\frac{d_q}{a}\right)^{\beta_i}$ follows. The projection equation of model 2.24 is derived through algebraic difference approach.

$$d_{ijk_2} = d_{q(ik_2)}\left(\frac{d_{ijk_1}}{d_{q(ik_1)}}\right)^{\left(\frac{A_{ik_1}}{A_{ik_2}}\right)^{\beta}} + e_{ijk_2} \tag{2.23}$$

$$d_{ijk_2} = \beta\left(\frac{d_{ijk_1}}{\beta}\right)^{\left(\frac{\ln(d_{q(ik_2)}/\beta)}{\ln(d_{q(ik_1)}/\beta)}\right)} + e_{ijk_2} \tag{2.24}$$

where

d_{ijk_2} and d_{ijk_1} are the *j*th tree's diameter of *i*th plot on k_2th and k_1th measurement occasion; d_q is quadratic mean diameter, and A is stand age. The mixed model (2.22) uses EBLUP to estimate random effects $t = \begin{pmatrix} t_{1,ij} \\ t_{2,ij} \end{pmatrix} \sim N(0, \Sigma_2)$ as shown by equation 2.25.

$$\begin{pmatrix} \hat{\phi}_{ij} \\ \hat{\phi}_{ij} \end{pmatrix} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} + \begin{pmatrix} \hat{t}_{1,ij} \\ \hat{t}_{2,ij} \end{pmatrix} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} + \hat{\Sigma}_2 \hat{Z}_{ijk_1}{}^T \left(\hat{Z}_{ijk_1} \hat{\Sigma}_2 \hat{Z}_{ijk_1}{}^T + \hat{R}_{ijk_1}\right)^{-1} \left(d_{ijk} - \hat{\beta}_1 d_{q,ik}{}^{\hat{\beta}_2}\right) \tag{2.25}$$

where

$\hat{\phi}_1$ and $\hat{\phi}_2$=estimates for ϕ_1 or ϕ_2 β_1 and $\hat{\beta}_2$, each of which consists of two components: fixed parameter estimates $\hat{\beta}_1$ or $\hat{\beta}_2$, and predicted random effect $\hat{t}_{1,ij}$ or $\hat{t}_{2,ij}$, which is predicted using EBLUP based one prior observation d_{ijk} ;

$d_{q,ik}$ =Quadratic mean diameter of the *i*th plot on the *k*th measurement occasion;

$$\hat{Z}_{ijk_1}{}^T = \begin{pmatrix} \hat{Z}_{1,ijk_1} \\ \hat{Z}_{2,ik_2} \end{pmatrix} = \left| \begin{pmatrix} \left(\frac{\partial f(d_{q,ik})}{\partial t_{1,ij}} \right) \\ \left(\frac{\partial f(d_{q,ik})}{\partial t_{2,ij}} \right) \end{pmatrix} \right| \begin{pmatrix} t_{1,ij}, t_{2,ij} = 0 \\ d_{q,ik} = d_{q,ik_1} \end{pmatrix} = \begin{pmatrix} d_{q,ik_1}{}^{\hat{\phi}_2} \\ \hat{\phi}_1 d_{q,ik_1}{}^{\hat{\phi}_2} \ln\left(d_{q,ik_1}\right) \end{pmatrix};$$

$$f\left(d_{q,ik}\right) = \alpha_{ij}\left(d_{q,ik}\right)^{\beta_{ij}} + e_{ijk};$$

$$\hat{R}_{ijk_1} = \hat{V}\left(e_{ijk_1}\right);$$

$$\hat{\Sigma}_2 = \begin{pmatrix} \hat{\mathrm{cov}}(t_{1,ij}, t_{1,ij}) & \hat{\mathrm{cov}}(t_{1,ij}, t_{2,ij}) \\ \hat{\mathrm{cov}}(t_{1,ij}, t_{2,ij}) & \hat{\mathrm{cov}}(t_{2,ij}, t_{2,ij}) \end{pmatrix};$$

Most of stand table projection model comparisons in the research have been made at stand distribution level by using Kolmogrov-Smirnov two-sample test (Borders, 1984, 1990; Nepal & Somers, 1992; Knowe et al. 1997; Trincado. et al., 2003). To isolate the performance of the stand table projection models, observed stand attributes were used for each models, and comparisons were made only at the individual diameter growth level. The statistical criteria employed to compare models were Mean Residuals (MR), Mean Absolute Residuals (MAR)

(Vanclay and Skovsgaard, 1997), Mean Absolute Percentage Residuals (MAPR), and Mean Square Error (MSE) or its square root, RMSE, which might be the most widely used fit-statistics in the field of forestry.

The basic measure of how closely a model fits a dataset is MSE, which measures the average mismatch between each observation and the model. MSE is the statistic whose value is minimized with the parameter estimation for least squares. The expectation of MSE defined here is $E(MSE)=V(y_i-\hat{y}_i)+(E(y_i-\hat{y}_i))^2$, where $\hat{y}_i$ and y_i are the estimated response and observed response, respectively. MSE consists of two components: variance of the errors and the squared MR. It approximately holds that $MSE=V(e)+MR^2$. Often, RMSE is preferable to MSE because the former is measured in the same units as the data, and is representative of the size of an average error. Using *n-p* instead of *n* in equation 2.29 allows for a minor adjustment of the number of parameters estimated in order to make it an unbiased estimator, but for purposes of selecting models, other statistics that impose a heavier penalty on model complexity, such as the Akaike Information Criterion (AIC) or Schwarz' Bayesian Information Criterion (BIC), should be employed instead of RMSE (Rivas et al., 2004).

The MR is signed measures of error and indicates whether the model underestimates or overestimates the response. MR describes both the direction and magnitude of the error bias. The

MAR is also measured in the same units as the original data and is usually similar in magnitude to, but slightly smaller than, RMSE. MAR and RMSE indicate the magnitude of the average error, but fail to provide information on the relative size of the average difference between prediction and observation. In contrast, MAPR gives information on the relative size of residual. It is also viewed as a weighted version of MAR, using the reciprocal of observation as weight. The model evaluation criteria can be summarized as follows:

$$MAR = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.26}$$

$$MR = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{2.27}$$

$$MAPR = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|y_i - \hat{y}_i|}{y_i}\right) \tag{2.28}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.29}$$

First, I used age=6 (yr.) for model 2.23 and 2.24 as the reference age to project forward to ages 7 to 13. Observation at age 6 of each tree in each plot (100 out of 148) was used to estimate random effect for each tree with EBLUP. MR, MAR, RMSE, and MAPR were calculated for each projection interval from years 1 to 7, corresponding age 8 to 13. Table 2.6

presents fit statistics, and Figure 2.5 through 2.8 provides visual inspections that corresponds Table 2.6.

First of all, it is apparent that the accuracy of projection decreases as projection intervals increase for models 2.22, 2.23, and 2.24 (see Figure 2.5). Probably the decreasing accuracy is caused by heterogeneity and/or increasing deviations caused by the error associated with the prior observation with increasing projection intervals. All the models give reasonable representation of diameter growth curves, because the mean absolute percentage residual (MAPR) for seven-year projection interval is less than 10%, seeing that CAPPS plots are very fast-growing under intensive management.

It is clear from Figure 2.5 and Table 2.6 that the two random effects mixed model 2.22 is more accurate than model 2.23 and 2.24 in terms of all the fit statistics employed. For the one or tow-year projection interval, model 2.23 and 2.24 have preferable fit statistics because their regression functions are reference invariant such that the projected dbh and the observed dbh are the same when the projection interval is zero. Visual inspections of MR (Figure 2.5d) reveal that models 2.23 and 2.34 tend to overestimate dbh. RMSE shows that model 2.22 is preferable to the others for long projection intervals. For five, six, and seven-year projection intervals, which can be viewed as long intervals if one take such a fast-growing rate as a half inch per year on average

into consideration, model (2.22)'s RMSE is almost 10% lower than Pienaar-Harrison model's. The decrease in RMSE is significant.

Fit statistics from five-year projections using observations at eight year for both the fit dataset and the validation dataset are presented in Table 2.7. The same conclusions as those shown in Table 2.6 can be reached. Since the fit statistics from fit dataset are similar to those from the validation dataset, conclusion based on the fit dataset applies to the validation data.

So far, comparisons have been made on the basis of one selected reference age. To eliminate to the effect of a specific reference age on comparisons, I also compared the three models by projecting forward to every stand age available in the dataset using dbh observed at ages 6 to12 as reference. $RMSE_{k_1k_2} = \frac{1}{\left(\sum_{i=1}^{n} n_i\right)} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(d_{ijk_2} - \hat{d}_{(k_1),ijk_2}\right)^2$ summarizes the RMSE calculation, where $RMSE_{k_1k_2}$ =the average RMSE based on projection forward in time from age k_1 to k_2, k_1= 6, 7, …, m-1, k_2= k_1+1, k_1+2, …, m (m=13), n=100 (the number of sample plots in the fit dataset), and $\hat{d}_{(j_1),ijk_2}$ =the predicted diameter for the jth tree of the ith plot on the k_2th measurement occasion, based on the observation made on the k_1th occasion. Pairwise comparisons of $RMSE_{k_1k_2,m_1}$ vs. $RMSE_{k_1k_2,m_2}$ are presented in Figures 2.6 through 2.8, where m_1 and m_2 denote the models to be compared. If two models give exactly the same RMSE, all the circles lie in the slope line. If one model outperforms another, more circles are far way from the

axis that represents that model. Figures 2.7 to 2.8 show that model 2.22 provides the most accurate predictions because it has two random effects to represent diameter growth patterns better.

2.4.Conclusion and Discussion

Individual growth models are often classified into two categories: (a) polynomial linear models or multiple linear regression models, based on the hypothesized functional relation between the response and explanatory variables (Zhang et al., 1993; Zhao et al., 2004; Wycoff et al; 1982) and (b) models that can be decomposed into two components, potential growth models and modifier models. In the latter category, growth is described as a general growth function modified by individual tree attributes (Lessard et al., 2001, among others). Generally, individual growth models are for annual or fixed-time increments (e.g. five-year increments, Ritchie & Hann, 1997a, among others). Often, linear interpolation is required for model fitting or application. When an individual model is used for predictions longer than the predetermined interval, it is often used in a recursive manner.

An undesirable property is reference variance, which makes individual model curves depend on explanatory variables. One solution is to construct a growth model following approach proposed by Clutter (1963), Bailey and Clutter (1974), or Cieszewski (2001). Projection models derived using these approaches automatically have reference invariance

property such that model shapes are affected only by error terms in the observation of the response used as reference and the estimated values of explanatory variables at the projected age. One prior observation of the response might be the most important explanatory variable to account for variation among individuals. Either individual growth models or projection models take advantage of the prior observation. The prediction error $V(\hat{y}-y)$ of projection models comes from four sources of random errors, associated respectively with (a) the prior observations of the response used as reference, (b) estimated values at projected age of explanatory variables, (c) the response at projected age, and (d) parameter estimates; It is not clear which sort of functions, reference variant individual growth functions or reference invariant projection functions, is preferable. In addition to reference invariance, one advantage of projection models is that they can be viewed as either yield models or growth models. One disadvantage of projection models is that additional projection models are required to project other attributes, such as crown ratio, if individual attributes used as explanatory variables are functions of stand age. In addition, whether the estimated values of these attributes could improve predictions is determined by the accuracy of the estimated values, projection equation and projection intervals.

The Pienaar-Harrison projection model (1988) deals with all variations in a very simple way: the quadratic mean is used to account for variation among stands, and one prior observation is used to account for variation among individuals nested in the same stand. I

developed a new projection model (2.24) based on the same assumptions as those in the Pienaar-Harrison model. The function of the new model is reference invariant and derived from a hypothesized individual diameter growth functions. Like the Pienaar-Harrison model, it uses the quadratic mean diameter and one prior observation to account for variation among stands and variation among individual trees, respectively. It is easy to incorporate the new model into a model system for more detailed predictions.

Projection models actually can be viewed as mixed models with one random coefficient, which is to be predicted with one prior observation. Using EBLUP, two random coefficients can be predicted but predictions are not necessarily improved. The two random effects mixed model (2.22) in this study outperform model 2.23 and its counterpart 2.24, which might have resulted from the simple model form and moderate correlations between the two random effects. From studies in this chapter, if a two-random effects mixed model performs as well as its projection counterpart, it is preferable in the sense that it can be used to improve prediction substantially if one more observation becomes available for prediction.

It also seems reasonable to conclude that data, expectation function, parameter estimation and model application should be considered systematically in modeling. Any separate operation might decrease model applicability.

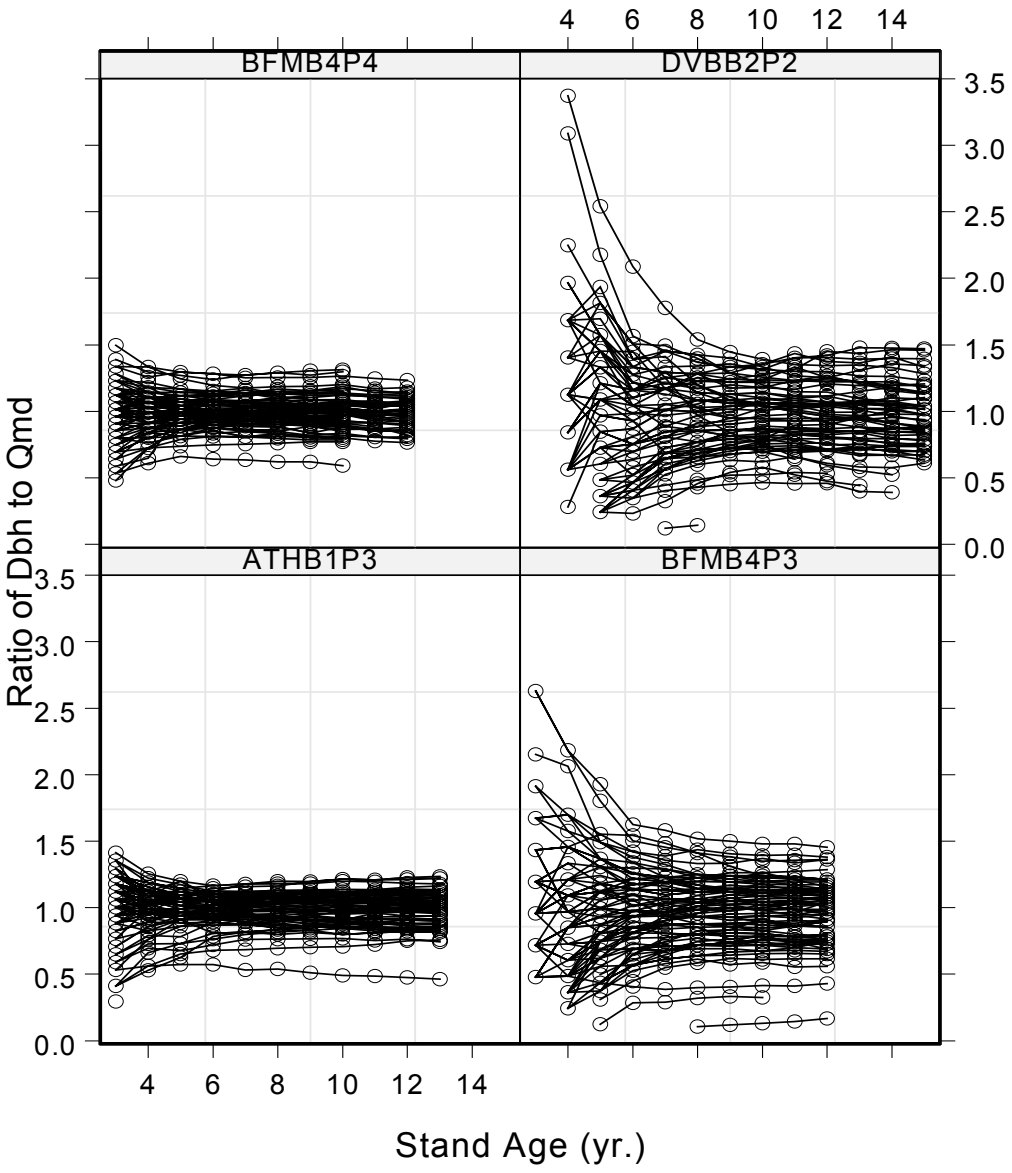

Figure 2.1. Profiles of relative diameter growth for 4 randomly selected CAPPS plots, showing that no significant trends over time exist

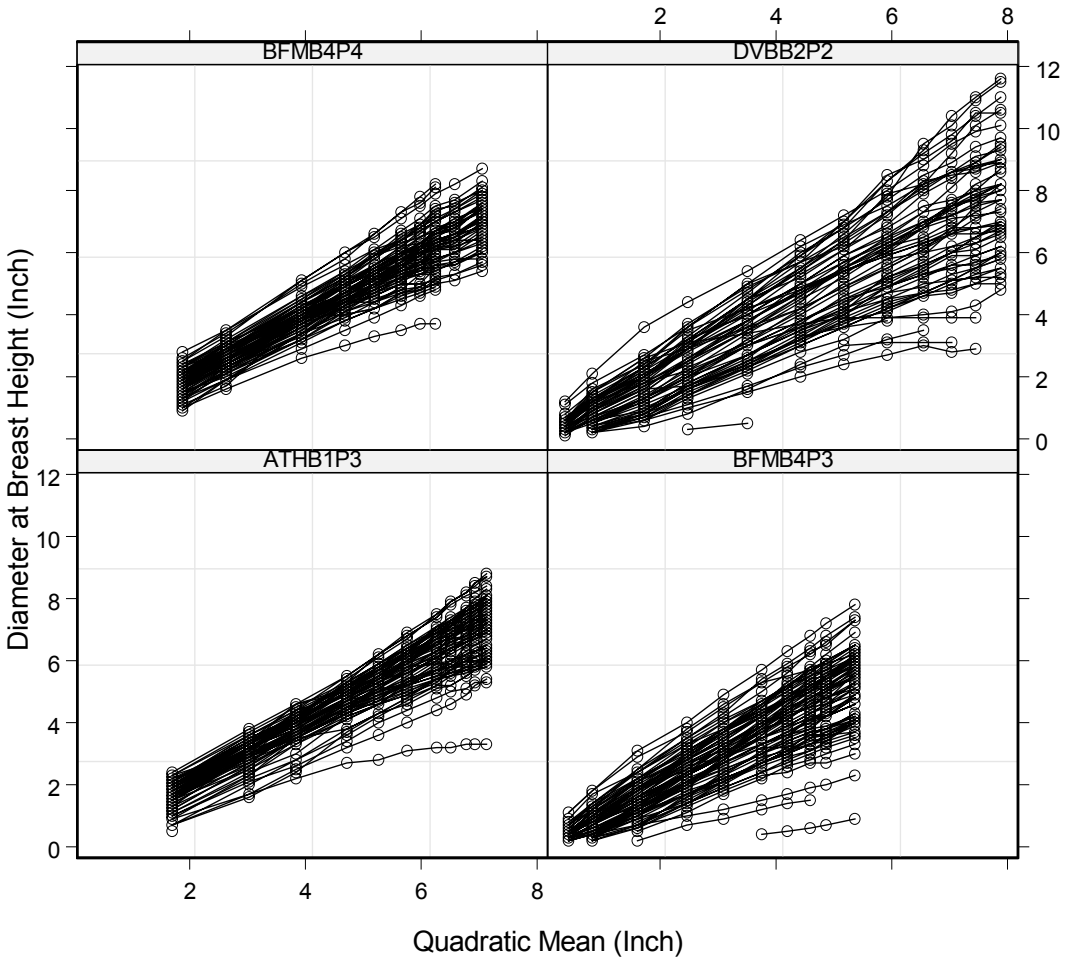

Figure 2.2. Diameter and Quadratic Mean Diameter for four randomly selected CAPPS plots, showing that a significant relation exists.

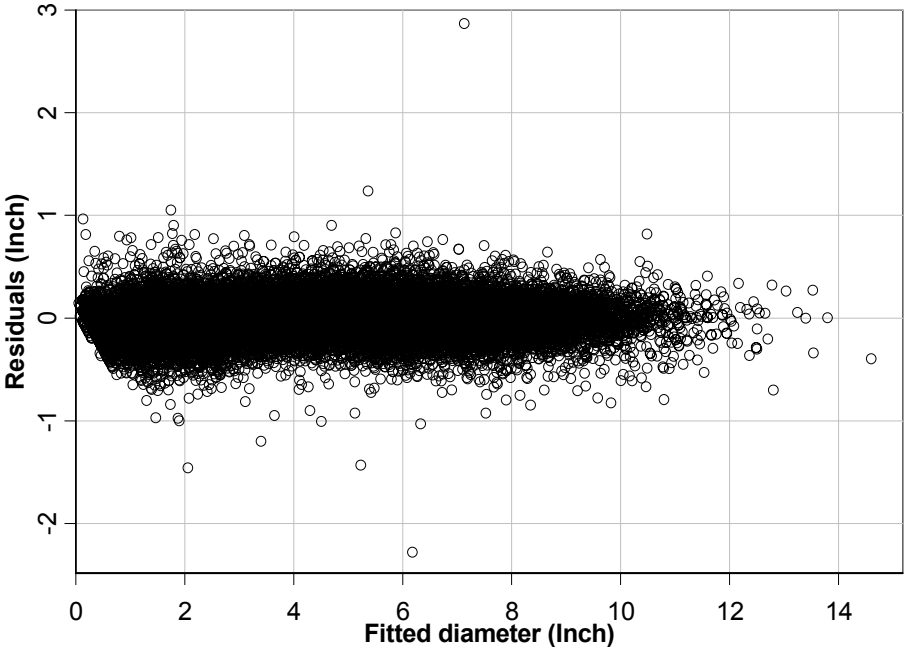

Figure 2.3. Residuals vs. Fitted diameters plot for model 2.22

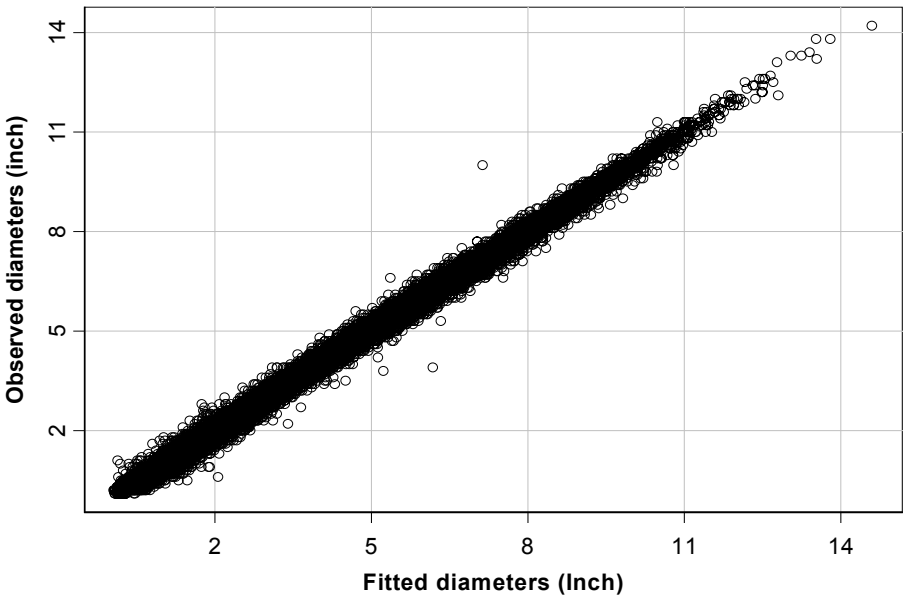

Figure 2.4. Observed dbh vs. Fitted dbh plot for model 2.22 (100 randomly selected plots)

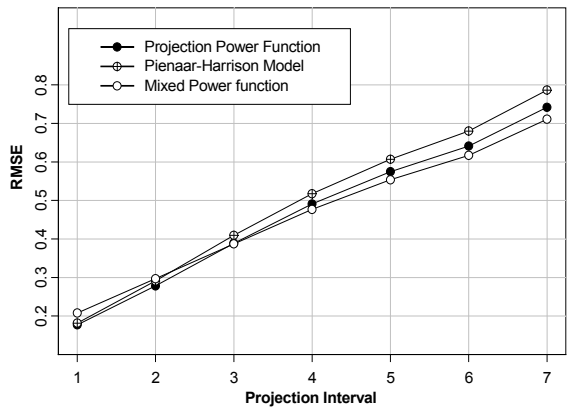

Figure 2.5a.

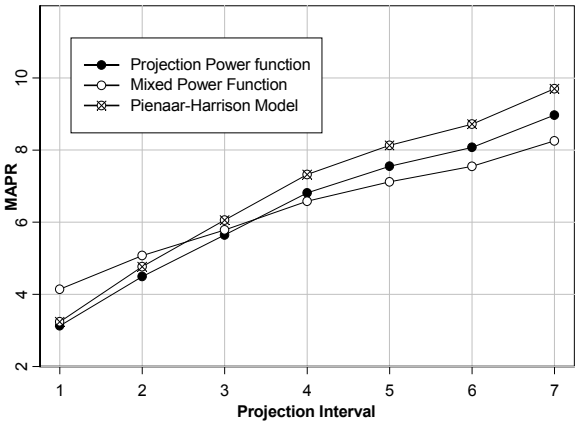

Figure 2.5b.

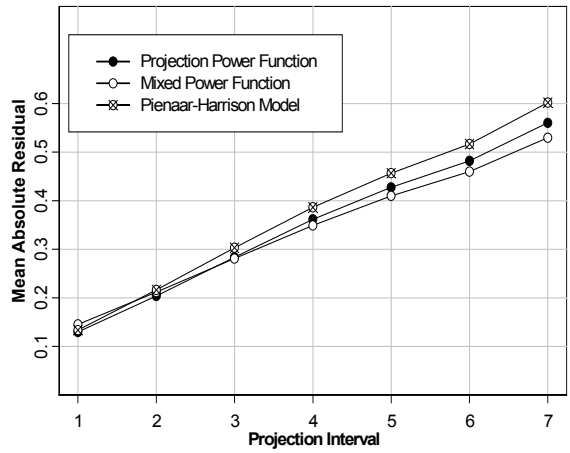

Figure 2.5c.

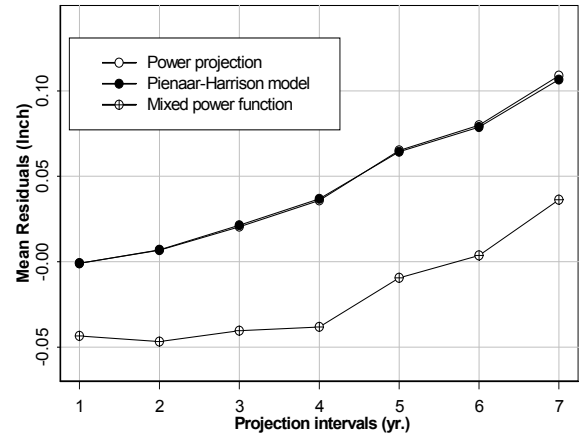

Figure 2.5d.

Figure 2.5. Comparisons of RMSE (a), MAPR (b), MAR (c), and MR (d) calculated with the mixed power function model 2.22, the Pienaar-Harrison model 2.23, the projection power model 2.24. Observations at age 6 were used to estimate random effects.

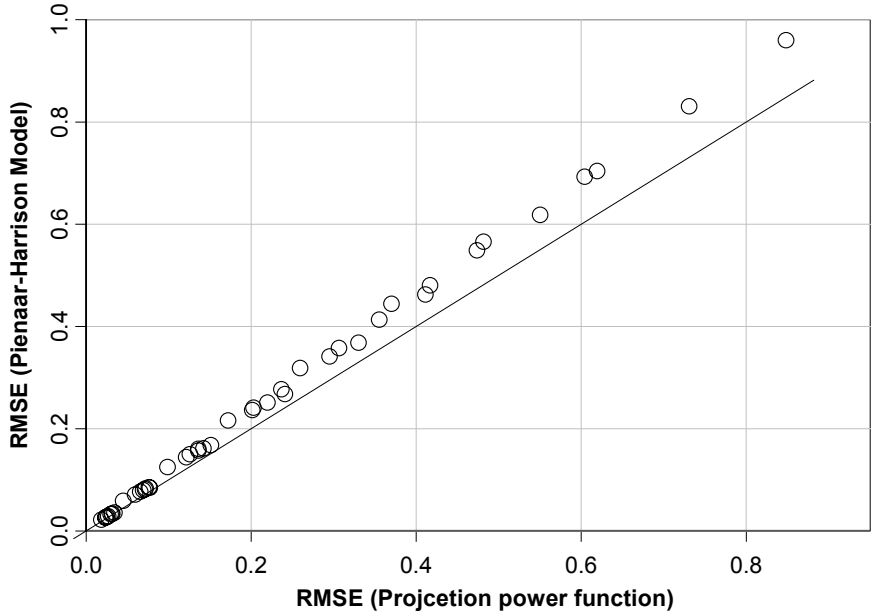

Figure 2.6. RMSE pairwise comparisons of the Pienaar-Harrison Model and Projection Power Function for 100 CAPPS sample plots.

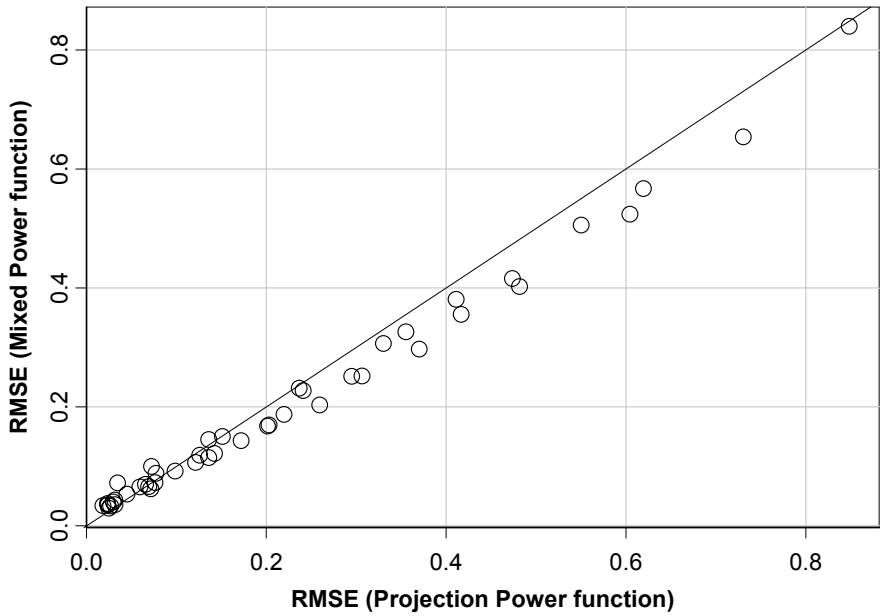

Figure 2.7. RMSE pairwise comparisons of the Mixed Power Function and Projection Power Function for 100 CAPPS sample plots.

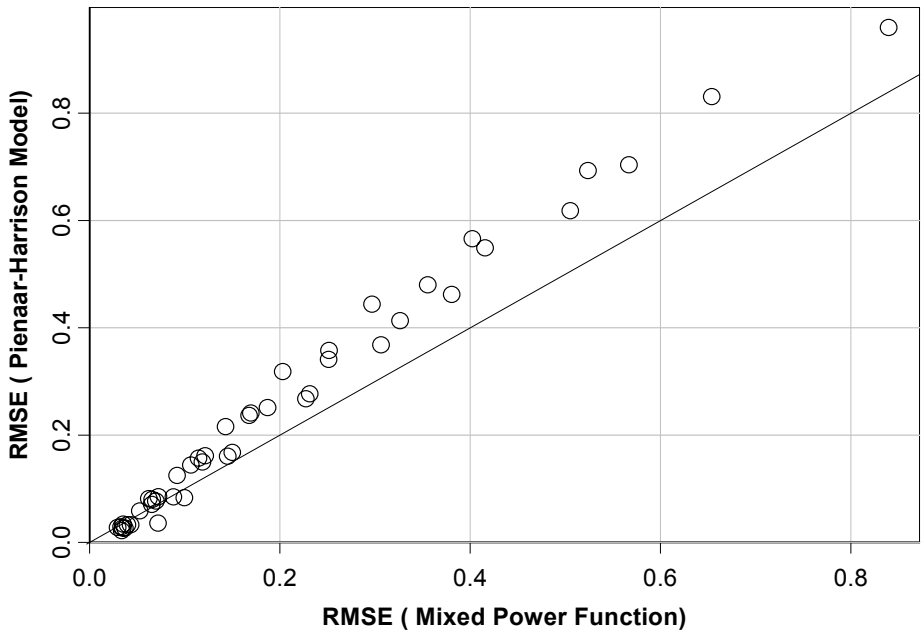

Figure 2.8. RMSE comparisons of the Projection Power Model and Mixed Model for 100 CAPPS sample plots.

Table. 2.1. Base functions that can be used to derive function 2.20 and corresponding individual-specific parameters.

functions	Math Form	Specific Parameters	
Chapman-Richards	$y = \beta_1\left(1 - e^{(-\beta_2 A)}\right)^{\beta_3}$	β_1	β_3
Gompertz	$y = \beta_1 e^{\left(-\beta_2 e^{(-\beta_3 A)}\right)}$	β_1	β_2
Levakovic III	$y = \beta_1\left(\frac{A^2}{\beta_2 + A^2}\right)^{\beta_3}$	β_1	β_3
Levakovic I	$y = \beta_1\left(\frac{A^{\beta_3}}{\beta_2 + A^{\beta_3}}\right)^{\beta_4}$	β_1	β_4
Sloboda	$y = \beta_1 e^{\left(-\beta_2 e^{\left(-\beta_3 A^{\beta_4}\right)}\right)}$	β_1	β_2

Table 2.2. Parameter estimates for a multi-level nonlinear mixed effects model (model 2.21) with Splus NLME.

	Parameters	Statistics				
		Value	StdE	DF	t-value	p-value
Fixed	β_1	1.020464	0.005953702	65505	171.3999	<.0001
	β_2	0.991821	0.004175526	65505	237.5319	<.0001
	Parameters	Plot level effect		Tree level effect		
		StdD	Corr	StdD	Corr	
Random	β_1	0.04652369	-0.999	0.3140669	-0.823	
	β_2	0.03758353	--	0.1483711	--	

Residual=0.1491881, 7604 trees nested in 100 plots

Table 2.3. Parameter estimates for a single-level nonlinear mixed effects model (model 2.22) with Splus NLME.

	Parameters	Statistics				
		Value	StdE	DF	t-value	p-value
Fixed	β_1	1.01957	0.003731	65505	273.2566	<.0001
	β_2	0.99281	0.001836	65505	540.6544	<.0001
	Parameters	Tree level effect				
		StdD		Corr		
Random	β_1	0.3173013		-0.827		
	β_2	0.1528874				

Residual=0.1492513, 7604 trees nested 100 plots

Table 2.4. Likelihood Ratio Test under the null hypothesis that model (2.22) is adequate

Model	df	AIC	BIC	Log-Likelihood	LRT	p-value
Full	9	-7372.617	-7289.82	3695.309		
Reduced	6	-7115.068	-7059.87	3563.534	263.5493	<.0001

Full=model (2.21), Reduced=model (2.22)

Table 2.5. Parameter estimate for projection power model and Pienaar-Harrison model

Model	Estimate	StdE	MSE	RMSE	R-Square
Projection Power	21.97673	0.1015	0.2780	0.5272	0.9094
Pienaar-Harrison	-0.71678	0.0019	0.3095	0.5563	0.8991

Table. 2.6 Fit statistics based on the fit dataset using observations at age 6 as reference

Interval (yr.)	Model	MAR (in)	MR (in)	RMSE (in)	MAPR (%)
1	1	0.1296	-0.0009	0.1772	3.1279
	2	0.1337	-0.0009	0.1816	3.2389
	3	0.1454	-0.0434	0.2079	4.1445
2	1	0.2038	0.0068	0.2781	4.4959
	2	0.2161	0.0070	0.2912	4.7663
	3	0.2115	-0.0468	0.2969	5.0760
3	1	0.2837	0.0205	0.3888	5.6481
	2	0.3032	0.0213	0.4098	6.0593
	3	0.2808	-0.0403	0.3874	5.7880
4	1	0.3616	0.0358	0.4909	6.8137
	2	0.3865	0.0368	0.5176	7.3227
	3	0.3492	-0.0382	0.4769	6.5828
5	1	0.4275	0.0652	0.5747	7.5518
	2	0.4568	0.0644	0.6069	8.1274
	3	0.4098	-0.0095	0.5536	7.1181
6	1	0.4820	0.0799	0.6413	8.0739
	2	0.5168	0.0788	0.6802	8.7186
	3	0.4599	0.0037	0.6170	7.5484
7	1	0.5604	0.1089	0.7418	8.9690
	2	0.6020	0.1066	0.7864	9.6997
	3	0.5298	0.0363	0.7108	8.2539

Model 1, 2, and 3 represents model 2.24, 2.23, and 2.22, respectively.

MAR =Mean Absolute Residual, MR =Mean Residual, RMSE=Root Mean Square Residual, and MAPR=Mean Absolute Percentage Residual

Table 2.7. Fit statistics for fit dataset and validation dataset using observations at age 8 as references

		Fit data (100 plots)		Validation (48)	
Fit Statistics	Model	Mean	StdE	Mean	StdE
AR	1	0.4228	0.3683	0.3943	0.3545
	2	0.4581	0.3829	0.4399	0.3676
	3	0.3725	0.3362	0.3585	0.3419
R	1	0.1237	0.5469	0.0886	0.5229
	2	0.1220	0.5845	0.0861	0.5668
	3	0.1204	0.4871	0.0679	0.4908
SE/RMSE	1	0.3144/0.5606	0.5716	0.2811/0.5301	0.5739
	2	0.3564/0.5969	0.5954	0.3286/0.5732	0.5889
	3	0.2517/0.5017	0.5067	0.2453/0.4953	0.5708
MAPR (%)	1	6.8059	7.8540	6.2520	6.7630
	2	7.4135	8.3883	7.0696	7.4237
	3	5.7607	6.0358	5.5977	6.3269

CHAPTER 3

QUANTILE REGRESSION APPROACH TO ESTIMATING PERCENTILE GROWTH MODEL

3.1. Introduction

Diameter distribution models can be broken into two categories: static models and dynamic diameter models. Here I refer to static models as models describing stand diameter distribution at a specific point in the stand growth trajectory with a chosen probability density function. I refer to dynamic models as those describing diameter distribution change through the entire growth trajectory. In other words, static models apply to a single stand at a given age whereas dynamic models can be used to describe a set of stands throughout their whole growth spans, assuming that all stands that have the same stand characteristics used as explanatory variables have the same diameter distribution. Many probability density functions (pdf) have been applied to model diameter distribution at a specific point in time during the stand growth process, including Normal, Lognormal, Gamma, Beta, Weibull (Nelson, 1964; Bliss & Reinker, 1964; Clutter & Bennett, 1965; all cited in Bailey & Dell, 1973), Johnson's S_b (Hafley & Schreuder, 1977) and a finite mixture of pdf functions (Liu et al., 2002).

The ultimate purpose of modeling diameter distributions is to estimate future stand tables conditional on stand age and other stand attributes since sometimes it is necessary and desirable to have a prediction system that provides estimates of the numbers of trees and volumes by diameter classes. In this sense, the parameters of interest for the pdf are functions of stand age and other attributes. In early works, pdf parameters were estimated directly as regression functions of stand attributes. Due to relationships among parameters and to some parameters varying inconsistently with stand attributes (Borders, 1987), a parameter recovery method (Bailey, 1981) and percentile-based method (Borders et al., 1987) have been devised to project stand tables instead of the parameter prediction method. Parameter recovery and percentile-based methods depend on predictions of percentiles to recover an empirical distribution directly from predicted percentiles or recover the parameters of chosen distribution, usually the Weibull distribution due to its closed form cumulative density function (CDF) and its flexibility. The usual process for estimating parameters of percentile growth models has been to: (a) acquire a dataset containing dbh data and other observed or calculated stand characteristics to be used as explanatory variables, (b) estimate selected percentiles for each plot by the order statistics, (c) run ordinary least square or seemingly unrelated regression on the estimated percentiles. Obviously, multiple observations of the response variable, dbh, at a given covariate are observed in this case.

The percentile-based method and parameter recovery method do not rely on the use of initial stand tables to reproduce future diameter distributions. Since they do not require the initial stand tables, these model can be more widely used than methods that require initial stand tables, such as the relative-size projection model proposed by Pienaar and Harrison (1988). Consequently they are likely to be less accurate than relative-size projection models according to Borders and Patterson (1990) and Knowe et al (1997). In a comparison of parameter recovery methods with percentile-based methods, percentile-based methods are more appropriate when the diameter distribution is multimodal. One reason that the parameter recovery is preferred is that it performs almost as well as percentile-based methods and is more mathematically simple in the unimodal case.

In this study, I proposed using quantile regression to estimate the parameters of the percentile regression models and compare this method with the traditional ordinary least squares method. I compare the efficiency of these two estimation methods through simulations. One advantage of quantile regression is its simplicity with mathematical computations. Instead of the above procedure, quantile regression employs a dataset in step (a) directly to estimate the percentile growth model. It is apparent that the application of quantile regression is quite simple compared with ordinary least squares or seemingly unrelated regression.

3.2. Quantile Regression

Mean regression is a widely used statistical method to investigate the relationship between the response variable and explanatory variables, usually taking the form of $h(y)=f(\mathrm{x})$, where $h(y)$ represents some transformation of the response variable. The mean regression method focuses on estimating the conditional mean of the response variable distribution as some function of a set of explanatory variables; in other words, the regression function is defined for the expected value of y conditional on x, $E(y|x)$, no matter whether the response variable distribution is homoscedastic or heteroscedastic. While the entire conditional distribution of y is of interest, rather than only the expected mean just as in the case of modeling a stand diameter distribution, the mean regression cannot provide a complete picture of the y distribution conditional on x, just as in the case of modeling a stand diameter distribution.

Conditional quantile $Q_{\mathrm{y}}(\theta|\mathrm{x})$, which is synonymous with percentile, is essentially a curve that consists of points at cumulative distribution function curves of the random variable, conditional on covariate x, where θ is the probability of observing a random variable $Y< Q_{\mathrm{y}}(\theta|\mathrm{x})$. Once a functional relation between conditional quantile and the covariate is specified, the conditional quantile can be estimated as a regression function of the covariate. Quantile regression is a statistical technique for estimating and conducting inference about conditional quantile functions, $Q_{\mathrm{y}}(\theta|\mathrm{x})=f(x)$, either in linear regression or nonlinear models, and was first

introduced by Koenker and Bassett (1978). It has become a widely used and accepted technique in many areas, especially in econometrics. Just as classical linear regression methods based on minimizing sums of squared residuals enable one to estimate models for conditional mean functions, the quantile regression method offers a mechanism for estimating models for the full range of conditional quantile functions by minimizing weighted absolute residuals. By supplementing the estimation of conditional mean functions with techniques for estimating an entire family of conditional quantile functions, quantile regression is capable of providing a more complete statistical analysis of the stochastic relationship among random variables.

A random variable is fully characterized by its cumulative distribution function (CDF), probability density function (PDF), and quantile function. Sometimes, it is necessary to assume that its distribution function parameters are functions of some explanatory variables. Hence, the estimation of distribution of Y conditional on X provides a means that can be used to investigate the influence of explanatory variable on the shape of the distribution. In forest biometrics, modeling diameter distribution as a function of other stand parameters is a perfect example that necessitates investigating how the random variable distribution is affected by a set of explanatory variables.

A classical linear regression with iid errors (independently and identically distributed) describes a linear functional relation between the mean of the response variable and a set of

explanatory variables, such as $y_i = x_i^T \beta + u_i$, where $E(y_i) = x_i^T \beta$ and u_i is IID. Accordingly, the corresponding conditional quantile function is $Q_{Y|X}(\theta | x_i) = x_i^T \beta + F^{(-1)}(\theta)$ since $P\left(y_i \le x_i^T \beta + F^{(-1)}(\theta)\right) = P\left(u_i \le F^{-1}(\theta)\right) = \theta$, where $(0 \le \theta \le 1)$ and $F^{(-1)}(\theta)$ is the inverse of cumulative density function of u_i. It is apparent all quantile curves are parallel to each other since $F^{(-1)}(\theta)$ is independent of x_i. If the model exhibits some kind of heteroscedasticity in a way such that $y_i = x_i^T \beta + v(x_i) u_i$, where $v(x_i) > 0$ is the variance function and u_i is iid with $E(u_i) = 0$, the corresponding quantile function is $Q_{Y|X}(\theta | x_i) = x_i^T \beta + v(x_i) F^{(-1)}(\theta)$.

Following representation was used to facilitate our simulation analysis since it is convenient to generate simulation dataset and to calculate the real quantile regression parameters for the sake of comparisons. The θth regression quantile $(0 \le \theta \le 1)$ for the heteroscedastic linear model $y_i = x_i^T \beta + x_i^T \rho u_i$ is defined as $Q_{Y|X}(\theta | x) = x^T \beta_\theta$ where $\beta_\theta = \beta + \rho F^{(-1)}(\theta)$. Other denotations are defined as follows:

y= the response variable corresponding to x

x=the explanatory variable vector

β=vector of unknown regression parameters

ρ= vector of unknown scale parameters

u= random errors that are independent and identically following a distribution that is unspecified

$F^{(-1)}(\theta)$=the inverse of the cumulative distribution of the errors

From $\beta_\theta = \beta + \rho F^{(-1)}(\theta)$, $y_i = x_i^T \beta + x_i^T \rho u_i$ can be rewritten as

$y_i = x_i^T \beta_\theta + x_i^T \rho u_i - x_i^T \rho F^{(-1)}(\theta)$ or $y_i = x_i^T \beta_\theta + w_i$ where $w_i = x_i^T \rho u_i - x_i^T \rho F^{(-1)}(\theta)$.

So I have

$$P\left(y_i \le x_i^T \beta_\theta\right)$$

$$= P\left(x_i^T \beta_\theta + x_i^T \rho u_i - x_i^T \rho F^{-1}(\theta) \le x_i^T \beta_\theta\right) = P\left(w_i \le 0\right)$$

$$= P\left(x_i^T \rho u_i \le x_i^T \rho F^{-1}(\theta)\right)$$

$$= P\left(u_i \le F^{-1}(\theta)\right)$$

Finally, $Q_{Y|X}(\theta \mid x) = x^T \beta_\theta$. It is important to note the term ρ allows the errors to change as a linear function of X and thus various heteroscedastic and homogeneous error model are accommodated with regression quantiles $Q_{Y|X}(w_i \mid x_i) = 0$ where $w_i = x_i^T \rho u_i - x_i^T \rho F^{(-1)}(\theta)$. Term ρ does not have to be estimated explicitly because it is automatically incorporated into the estimates of β_θ, say, $\hat{\beta}_\theta$. Homoscedastic regression models are a special case of the linear model with $\rho = (1,0,0,..,0)$, where all parameters other than the intercept are the same for all θ. $y_i = x_i^T \beta + x_i^T \rho u_i$ and $\beta_\theta = \beta + \rho F^{(-1)}(\theta)$ are not the necessary assumption of linear quantile regression as noticed by Buchinsky (1998). It is only assumed that w_i in $y_i = x_i^T \beta_\theta + w_i$ satisfies $Q_{Y|X}(w_i \mid x) = 0$(Buchinsky, 1998).

Assuming that $F_{Y|X}(\theta|x)$ is strictly increasing with its density $f_{Y|X}(\theta|x)$, the conditional quantile can be characterized as

$$Q_{Y|X}(\theta|x) = \arg\min_{a} E(\rho_\theta(y-a)|X=x) \tag{3.1}$$

where $\rho_\theta(u) = u(\theta - I(u<0))$ is the check function, and $I(u<0)$ is the indicator function, equal to 1 if the statement inside the brackets is true, and 0 otherwise. When $\theta = \frac{1}{2}$, then the check function is an absolute value function, i.e., $\rho_\theta(u) = |u|$. To verify that model 3.1 holds true, note that $E(\rho_\theta(y-a)|x) = (\theta-1)\int_{-\infty}^{a}(y-a)f_{Y|X}(y|x)dy + \theta\int_{a}^{\infty}(y-a)f_{Y|X}(y|x)dy$. After taking first order derivative with respect to *a*, I have $-(\theta-1)\int_{-\infty}^{a} f_{Y|X}(y|x)dy - \theta\int_{a}^{\infty} f_{Y|X}(y|x)dy = 0$ and $F(a|x) - \theta = 0$. Therefore, the solution to the first order condition is $\hat{a} = F^{(-1)}(\theta|x)$, which is the conditional quantile function. Under the assumption that $Q_{Y|X}(\theta|x) = x^T\beta_\theta$, equation 3.1 implies equation 3.2

$$\beta_\theta = \arg\min_{b} E\left(\rho_\theta\left(y - X^T b\right)|X=x\right) \tag{3.2}$$

Estimating function 3.3, the quantile regression estimator of β_θ, is defined as the sample analogue estimator based on equation 3.2.

$$\hat{\beta}_\theta = \arg\min_{b} \frac{1}{n}\sum_{i=1}^{n}\left(\rho_\theta\left(y - X^T b\right)|X=x\right) \tag{3.3}$$

$$\hat{\beta}_\theta = \min_{\beta}\left\{\sum_{y_i \geq x_i^T\beta}\theta\left|y_i - x_i^T\beta\right| + \sum_{y_i < x_i^T\beta}(1-\theta)\left|y_i - x_i^T\beta\right|\right\} \tag{3.4}$$

$$\hat{\beta} = \min_{\beta} \left\{ \sum_{i=1}^{n} \left| y_i - x_i^T \beta \right| \right\} \tag{3.5}$$

The minimization of estimating function 3.3 has a linear programming representation that makes estimation easy. Equation 3.3 can be solved by a modification of simplex linear program for any specified value of θ (Koenker and Bassett, 1978). By expanding the estimating function 3.3 to estimating function 3.4, it can be seen that function 3.4 minimizes the sum of weighted absolute residuals by giving weight θ to positive residuals and 1-θ to negative residuals. (see Koenker & Bassett, 1978; Buchinsky, 1998). If θ=0.5, function 3.4 collapses to function 3.5, which is the estimation function for median regression estimation. Median regression is a robust alternative to least squares when data is contaminated with outliers or error terms follow other distribution than normal distribution.

The estimator of β_θ is asymptotically normal under certain regularity conditions, $\sqrt{n}\left(\beta_\theta - \hat{\beta}_\theta\right) \to N(0, \Lambda_\theta)$. Various estimators for asymptotic covariance matrix have been proposed, each having some advantages and disadvantages (Buchinsky, 1998).

Regression quantiles have several important linear model properties that are common to least squares regression estimates of expected values. Once one denote the quantile regression estimate for a given $\theta \in (0,1)$ and observations (Y, X) by $\hat{\beta}(\theta, Y, X)$, then for any $P \times P$ nonsingular matrix A and a>0 the following properties holds (see Koenker & Bassett, 1978 for details):

1. $\hat{\beta}(\theta, aY, X) = a\hat{\beta}(\theta, Y, X)$	(Scale change)
2. $\hat{\beta}(\theta, -aY, X) = a\hat{\beta}(1-\theta, Y, X)$	(Scale change)
3. $\hat{\beta}(\theta, Y + X\phi, X) = \hat{\beta}(\theta, Y, X) + \phi$	(Location shift)
4. $\hat{\beta}(\theta, Y + X, XA) = A^{-1}\hat{\beta}(\theta, Y, X)$	(Design X reparameterization)

One important property of the quantile regression model is that, for any monotone function h (.), is invariance to monotonic transformations, $Q_{h(Y)}(\theta \mid x) = h(Q_Y(\theta \mid x))$. In other words, the quantiles of the transformed random variable h (y) are the transformed quantiles of the original variable Y.

Quantile regression does not require that multiple values of the response variable be observed for given covariates to estimate the conditional quantile. For the special case where multiple values of the response variable y are observed for given covariates x, two natural approaches to estimation of conditional quantile are quantile regression and ordinary least squares running on estimated quantiles. One question that arises in this case is which estimation method is more efficient in the first order and second order since both quantile regression and least squares are asymptotically consistent. I tried to use a simulation technique to investigate whether quantile regression has any other advantages over least squares estimates than the simplicity of mathematical computations. Since quantiles of random variable are not observable,

a simulation method is employed to evaluate these two estimators instead of using a dataset for estimation method assessment.

3.3.Simulation Model and Sample Quantile Estimation

The θth sample quantile of a data set is defined as that value where a θ fraction of the data is below that value and a 1-θ fraction of the data is above that value. I estimate the sample quantiles based on the order statistics in simulations. The formula is function 3.6.

$$q_\theta = \begin{cases} x_i & if\ \theta = \theta_i = \left(\dfrac{i-1}{n-1}\right) \\ (1-\alpha)x_i + (\alpha)x_{i+1} & if\ \theta_i \le \theta \le \theta_{i+1} \quad where\ \alpha = \theta - \theta_i \end{cases} \tag{3.6}$$

where x_i is the ith observations sorted in ascending order and n is sample size. The algorithm linearly interpolates between the order statistic of x_i, assuming that the ith order statistic is the $\dfrac{(i-1)}{(n-1)}th$ quantile. The estimation method is widely used in statistical packages (e.g. Splus).

I based the simulations on the linear model $y_i = x_i^T\beta + x_i^T\rho u_i$, where $x_i^T\rho$ is the linear combination of the explanatory variable; ρ is the unknown scale parameter; and $u_i \sim F$ are independent and identically distributed errors. If letting $\beta=\beta_\theta - F^{(-1)}(\theta)\,\rho$, one has $y_i = x_i^T\beta_\theta + (x_i^T\rho)u_i - (x_i^T\rho)\,F^{-1}(\theta)$, where $F^{-1}(\theta)$ is the inverse of the cumulative density function that does not have to be known. $\beta=\beta_\theta - F^{-1}(\theta)\,\rho$ implies that $Q_y(\theta\,|x_i) = x_i^T\beta_\theta$, as demonstrated in section 3.2.

Based on the model $y_i = x_i^T \beta + \left(x_i^T \rho\right) u_i$ I generated random samples, estimated regression coefficients of a specific conditional quantile regression using both the quantile regression method and ordinary least squares and compared them to the real values of $\beta_\theta = \beta + F^{(-1)}(\theta)\rho$. Apparently, ρ is incorporated into the estimate for β_θ and it does not have to be known or to be estimated. Quantile regression does not specify any sort of parametric distribution assumption but a parametric conditional quantile form. In our simulation study, I used parametric distribution for the convenience of generating random samples and calculating real regression coefficients. The algorithm given in Koenker & Bassett (1978) and Bassett & Koenker (1982) was used to find the solution to estimating function 3.4. SAS Proc IML was employed to perform all simulations.

3.4. Simulation Analysis

The purpose of our simulation study is to investigate the first and second order efficiency of two estimators for parameters in a linear quantile model in the case where the multiple observations of response are available at a fixed covariate. The first estimator is the quantile regression estimator while the second estimator is the ordinary least squares estimator on the sample quantile of the response at each covariate. Our attentions were focused on comparison of quantile regression with ordinary least square and no efforts were made to investigate the first order and second order behavior under model misspecification.

Extensive simulation experiments were performed for sub-sample size between 20 and 80, which covers most of the range of sampled trees in forest inventory sample plots. The sub-sample size here refers to the number of multiple observations of the response variable at a given covariate, and the sample size refers to the number of sample units in which multiple response observations respond to the same covariate. Simulation experiments included simple linear quantile models, multiple linear quantile models, homoscedastic and heteroscedastic linear models, symmetric distribution (normal), asymmetric distribution (weibull) and all combinations of these factors under the various sample sizes.

It is not surprising that the variance of the estimate differs among quantiles. Generally, the variance of the estimate increases as the value of θ approaches 0 or 1, but the specifics are dependent on the data distribution. Estimates further from the center of the distribution usually cannot be estimated as precisely as the median (Cade, 2003). It should be noted that the median is not necessarily the quantile that can be estimated most accurately for asymmetric distribution. Here I am going to refer the complement of the quantile $\theta=\alpha$ as $\theta=1-\alpha$ for the sake of simplicity. For example, the 10*th* quantile is the complement of the 90*th* percentile and vice versa. For symmetric distributions such as normal, one quantile has the same estimate variance as its complement does (see Tables 3.2 and 3.3), whereas for an asymmetric distribution with a finite lower bound such as Weibull, an estimate close to the finite lower bound can be estimated more

precisely than its complement, as shown in Table 3.1. It also can be seen that for this distribution the median is not the quantile that can be estimated most accurately. The Weibull distribution with shape parameter 3 and scale parameter 5 (see Table 3.1) is almost asymmetric. When a Weibull distribution is positively skewed, the estimate variance increases as θ increases. The conclusion above bears some significance for the parameter recovery method in selecting quantiles to recover parameters.

Simulation results indicated that many factors affect parameter estimates and their variances, including the number of parameters of a linear function, independent variable range, error term distribution in $y_i = x_i^T\beta+(x_i^T\rho)u_i$, sample size and sub-sample size, heteroscedasticity, and quantiles. As previously stated, the study's purpose is to investigate estimate efficiency of these two estimators in terms of the second order as well as the first order. From the viewpoint of comparing quantile regression with least squares, sub-sample size and error term heteroscedasticity are significantly associated with the difference of two estimators in terms of the first order and second order. The conclusion drawn from simple linear regression functions applies to multiple linear functions, but only a part of simulation results from simple linear function are presented.

In all situations involved in our simulation experiment, quantile regression estimation has a smaller bias than ordinary least squares for the same sample size and sub-sample size and

with smaller variance of the estimate in the case of heteroscedasticity. For a homoscedasticity situation, a quantile regression estimate sometimes has a slightly larger variance for some cases. In the case of homoscedasticity there is no need to estimate conditional quantiles since all conditional quantiles are parallel to the conditional mean and many regression techniques are available for estimating the conditional mean under various assumptions of error distributions.

Least squares estimates are much more sensitive to sub-sample size. When the number of multiple values of the response variable is relatively small, least square estimate bias is much larger than quantile regression, especially for a quantile far away from the median. Actually, quantile regression was not intended for the multiple response case, but for the single observation at a given covariate, so multiple response observation is not a necessary condition for quantile regression while it is for least squares. In the case where only a single observation of the response variable is available, quantile regression still is able to estimate the regression parameters for different quantiles while least square has the same parameter estimates for any quantiles. The single observation case helps illustrate why quantile regression is superior to least squares in the case of a small sub-sample size although it is unreasonable to apply least squares in this single observation case. As demonstrated by Table 3.2, quantile regression estimation is less biased than least squares with smaller variance when the number of multiple observations is relatively small and error terms are heteroscedastic. The scale parameter vector in Table 3.2 is (1,

0.5), which means the coefficient of variation is roughly 0.125. It should be noted that quantile regression did not show any advantage over ordinary least squares in terms of the second order in the case where error terms are homoscedastic or close to homoscedastic, as indicated by Table 3.4.

Simulation results indicated that quantile regression estimation compared very favorably with least squares also in terms of the second order in the case of heteroscedasticity where quantile regression estimates have smaller estimate bias and variance of the estimate (see Tables 3.2, 3.3, and 3.5). The advantage of quantile regression over least squares is proportional to heteroscedasticity (Tables 3.3 and 3.5). These three simulation models are identical except for that the scale parameter vectors, which are ρ=(1, 0), ρ=(1, 0.5), and ρ=(1, 1) respectively. Note that when error terms are homoscedastic the scale parameter vector ρ has all elements equal to 0 except the first one.

The following simulation experiment design is different from the ones above in the mechanism for generating simulation dataset and parameter estimation procedure. First of all, I assumed that the diameter distribution at any stand age of CAPPS stands is Weibull and that the percentile growth model of CAPPS stands is Schumacher, $Q_{Y|X}(\theta|x)=\alpha_\theta \exp\left(\frac{\beta_\theta}{A}\right)$, where A is stand age.

Secondly I estimated the parameters α_θ and β_θ for 10 percentiles from 0% to 90% with 10% increment using quantile regression, and then I can recover these two parameters for the diameter distribution at any arbitrary x_i by employing percentile 0 (location parameter) and any other two percentiles, $Q_{Y|X}(\theta_1|x)$ and $Q_{Y|X}(\theta_2|x)$, based on the Weibull cumulative function 3.7, where a_i is location parameter; b_i is scale parameter; and c_i is shape parameter. Parameters a_i, b_i, and c_i are conditional on x_i (stand age). Given a_i, $Q_{Y|X}(\theta_1|x)$ and $Q_{Y|X}(\theta_2|x)$ are known, c_i can be calculated with equation 3.9.

$$F\left(Q_{Y|X}(\theta|x)\right)=1-Exp\left(-\left(\frac{Q_{Y|X}(\theta|x)-a_i}{b_i}\right)^{c_i}\right) \tag{3.7}$$

$$a_i|(X=x_i)=\alpha_0\exp\left(\frac{\beta_0}{A}\right) \tag{3.8}$$

$$c_i|(X=x_i)=\ln\left(\frac{\ln(1-\theta_1)}{\ln(1-\theta_2)}\right)\left(\ln\left(\frac{Q_{Y|X}(\theta_1|x)-a}{Q_{Y|X}(\theta_2|x)-a}\right)^{c}\right)^{(-1)} \tag{3.9}$$

$$b_i|(X=x_i)=\left(Q_{Y|X}(\theta_k|x)-a_i\right)\left(-\ln(1-\theta_k)\right)^{-\left(\frac{1}{c_i}\right)} \tag{3.10}$$

Parameter b can be estimated using either of two percentiles given that parameters a and c have been known. Parameter a is a function of stand age as are b and c since they are functions of parameter a. Subscript i indicates parameters are associated with covariates, namely they are functions of explanatory variables, the stand age in this case. After obtaining the estimates for the parameters of the Weibull distribution at any point of time, I abstracted a stand

growth system with the Schumacher model as percentile growth model and Weibull as diameter distribution model from CAPPS data (see chapter 2). I simulated 100 sample plots that stand age uniformly distributes between [1,40] to construct a dataset for estimation and then estimated the parameters of percentile growth models with both quantile regression and least squares. It should be noted that there is only two quantiles strictly following Schumacher model, i.e., $Q_{Y|X}(\theta_1|x)$ and $Q_{Y|X}(\theta_2|x)$, in the constructed system. A part of simulation results, based on 1,000 simulated datasets, are presented in Table 3.4. The abstracted system did not take other stand attributes into account for the difference in percentile growth patterns from stand to stand. In addition, other percentiles other than 0*th* percentile and two percentiles used to recover the parameters could not be modeled with a function in this case.

All percentile growth curves are almost parallel lines after logarithm transformation since all slopes are almost the same. It is implied that the abstracted system is close to homoscedastic after transformation. Therefore, quantile regression is favorable only in the first order, not in the second order. I used simulated root mean square error (RMSE) to evaluate these two estimators. RMSE is a widely used statistical criterion in the comparison of estimators. The RMSE of an estimator $\hat{y}$ of a parameter y in a statistical model is defined as: $MSE(\hat{y}) = E\left((\hat{y}-y)^2\right)$. From the definition of the variance $V(X) = E\left(X^2\right) - E(X)^2$, one can express the MSE as

$MSE(\hat{y}) = V(\hat{y}) + (E(\hat{y} - y))^2$ by expanding the RHS of $MSE(\hat{y}) = E\left((\hat{y} - y)^2\right)$. It is easy to see the RMSE can give comprehensive comparisons of estimators by taking both estimate variance and bias into considerations. I used RMSE of the response, $Q_{Y|X}(\theta|x)$ (quantile of interest), to evaluate these two estimators. Simulation results presented in Table 3.6 shows that quantile regression still performs better than least squares.

I also changed values of parameter β for all selected percentiles to make the transformed system more heteroscedastic in order to see if quantile regression compared favorably with ordinary least squares in the second order in the case of heteroscedasticity. The simulation results showed the results I expected, which were presented in Table 3.5. One feature of the linear percentile growth model from the simulations is noteworthy. That is, intercept is less accurately estimated.

3.5. Conclusion and Discussion

In addition to the advantage that quantile regression is computationally simple, when multiple observations are available for a given covariate quantile regression produces less biased estimates than least squares for parameters of percentile growth models, especially where θ is far away from the median and the number of multiple observations is small. Compared with least squares, when error terms are homoscedastic or close to homoscedastic, quantile regression can

be is slightly unfavorable because the variance of the estimate may be slightly higher. In the case of heteroscedasticity, quantile regression is preferable not only in the first order but also in the second order, and the advantage is proportional to heteroscedasticity; or in other words, a more heteroscedastic response variable means a higher efficiency gain from using quantile regression. Increasing the number of multiple observations at a given covariate plays an important role in reducing the bias of least squares estimates. Although increasing the number of sample units (the number of sample plots) can reduce the variance of the estimate, simulation results indicated that it did not reduce least squares estimate bias substantially.

Based on the assumption that quantile estimation of the order statistics is approximately normal, in the case of heteroscedasticity weighted least squares is appropriate in order for the least squares estimates to have minimum variance. In the application of weighted least squares, weights have to be estimated before final estimates are obtainable. In addition, seemingly unrelated regression (Borders et al., 1987) is another alternative to reduce the variance of the estimate if the percentile models are related by the fact that the distributions are correlated across equations or a subset of explanatory variables (variables in model right-hand sides) are the same. One of the SUR assumptions is that for any given equation the error terms are homoscedastic and the assumption would be violated since the estimated quantile is also heteroscedastic in the case where the response variable distribution is heteroscedastic and the sub-sample size are the same

for any given covariate. One advantage of quantile regression is that it does not assume any parametric distribution for error terms.

Table 3.1. An example of simulation results indicating that the estimate variance of a quantile from an asymmetric distribution differs from that of its complement quantile

P	Real value		Estimate		Variance	
	α	β	α	β	α	β
10	4.361543	4.47230	4.355132	4.473149	0.098836	0.000647
20	5.032714	4.60654	5.035191	4.606570	0.079809	0.000518
30	5.545908	4.70918	5.545155	4.709355	0.072980	0.000477
40	5.996939	4.79938	5.995237	4.799533	0.070176	0.000460
50	6.424985	4.88499	6.421753	4.885290	0.071250	0.000463
60	6.856399	4.97128	6.856291	4.971445	0.073295	0.000486
70	7.319149	5.06383	7.319075	5.063796	0.079187	0.000513
80	7.859511	5.17190	7.864819	5.171173	0.092681	0.000593
90	8.602502	5.32050	8.604470	5.320257	0.130737	0.000847

Model, $y = \alpha + \beta x + \begin{pmatrix} 1 & 0.2 \end{pmatrix} \begin{pmatrix} 1 \\ x \end{pmatrix} e$ where $e \sim f(x) = \left(\frac{c}{b}\right)\left(\frac{x}{b}\right)^{c-1} \exp\left[-\left(\frac{x}{b}\right)^{c}\right]$ and α=2, β=4, a=3, and b=5. $x \in [1,\ 40]$. Sub-sample size, 40; sample size 40; and simulation replication for each percentile is 10,000. The estimator for parameters is quantile regression;

Table 3.2. Comparison of quantile estimate and least squares estimate in terms of the first order and the second order statistics in the case that error terms are heteroscedastic.

n	p	Real values		α				β			
				OLS		QR		OLS		QR	
		α	β	Est	Std	Est	Std	Est	Std	Est	Std
	10	0.7184	3.3592	0.8452	0.7662	0.7020	0.5395	3.4208	0.0994	3.3617	0.0849
	30	1.4756	3.7378	1.5093	0.6191	1.4790	0.4107	3.7545	0.0808	3.7375	0.0655
20	50	2.0000	4.0000	2.0022	0.5912	2.0050	0.3902	3.9996	0.0766	3.9993	0.0617
	70	2.5244	4.2622	2.4933	0.6152	2.5281	0.4075	4.2448	0.0802	4.2607	0.0656
	90	3.2816	4.6408	3.1445	0.7651	3.2959	0.5392	4.5807	0.0999	4.6375	0.0862
	10	0.7184	3.3592	0.7599	0.4064	0.7174	0.2670	3.3748	0.0531	3.3597	0.0426
	30	1.4756	3.7378	1.4836	0.3142	1.4740	0.2022	3.7419	0.0414	3.7378	0.0326
80	50	2.0000	4.0000	2.0038	0.3038	2.0014	0.1943	3.9998	0.0398	3.9999	0.0313
	70	2.5244	4.2622	2.5169	0.3175	2.5270	0.2036	4.2581	0.0413	4.2621	0.0324
	90	3.2816	4.6408	3.2449	0.4047	3.2820	0.2643	4.6254	0.0530	4.6411	0.0420

Model $y = \alpha + \beta x + \begin{pmatrix} 1 & 0.5 \end{pmatrix} \begin{pmatrix} 1 \\ x \end{pmatrix} e$, where e~ N (0, 1) and α =2, β =4 and $x \in [1,20]$. Simulation replication for each percentile is 10,000.

Table 3.3. An example of simulation results indicating the gain from the quantile estimate increases as error term heteroscedasticity increases.

ρ	p	Real value		α				β			
				OLS		QR		OLS		QR	
		α	β	Est	Std	Est	Std	Est	Std	Est	Std
ρ=(1, 0)	10	0.7184	4	0.7616	0.1005	0.7185	0.1039	4.0000	0.0083	4.0001	0.0086
	30	1.4756	4	1.4883	0.0772	1.4769	0.0792	4.0000	0.0065	4.0000	0.0066
	50	2.0000	4	1.9996	0.0741	1.9990	0.0751	4.0000	0.0061	4.0000	0.0062
	70	2.5244	4	2.5138	0.0776	2.5250	0.0785	4.0000	0.0065	4.0000	0.0066
	90	3.2816	4	3.2376	0.1001	3.2797	0.1028	4.0000	0.0083	4.0000	0.0085
ρ=(1, 0.5)	10	0.7184	3.3592	0.7612	0.4706	0.7128	0.3106	3.3809	0.0612	3.3603	0.0497
	30	1.4756	3.7378	1.4943	0.3686	1.4778	0.2375	3.7427	0.0482	3.7375	0.0376
	50	2.0000	4.0000	1.9963	0.3502	1.9989	0.2236	4.0006	0.0455	4.0002	0.0358
	70	2.5244	4.2622	2.5114	0.3671	2.5254	0.2369	4.2563	0.0484	4.2617	0.0383
	90	3.2816	4.6408	3.2332	0.4708	3.2853	0.3082	4.6195	0.0613	4.6399	0.0493
ρ=(1, 1)	10	0.7184	2.7184	0.7539	0.8537	0.7014	0.4865	2.7626	0.1154	2.7207	0.0883
	30	1.4756	3.4756	1.4814	0.6743	1.4764	0.3686	3.4878	0.0909	3.4758	0.0676
	50	2.0000	4.0000	2.0089	0.6390	2.0046	0.3514	3.9990	0.0856	3.9996	0.0639
	70	2.5244	4.5244	2.5186	0.6634	2.5288	0.3681	4.5135	0.0895	4.5248	0.0678
	90	3.2816	5.2816	3.2370	0.8400	3.2945	0.4738	5.2390	0.1131	5.2799	0.0874

The number of observations at a given covariate, 60; simulation models $y = \alpha + \beta x + x^T \rho\, e$; e~ N (0, 1) and α =2, β =4 and $x \in [1,20]$; simulation replication, 10,000

Table 3.4. Simulation results from the growth system abstracted from CAPPS data

P	α	β	n	α				β			
				OLS		QR		OLS		QR	
				Est	Std	Est	Std	Est	Std	Est	Std
0.1	8.8299	-6.2501	25	9.0696	0.0910	8.8288	0.0906	-6.2281	0.0963	-6.2519	0.0946
			35	8.9997	0.0805	8.8231	0.0815	-6.2254	0.0831	-6.2406	0.0827
			45	8.9667	0.0671	8.8315	0.0667	-6.2360	0.0712	-6.2508	0.0725
			55	8.9408	0.0637	8.8310	0.0632	-6.2358	0.0680	-6.2469	0.0679
0.2	9.6807	-6.1383	25	9.8457	0.0846	9.6796	0.0840	-6.1296	0.0754	-6.1398	0.0757
			35	9.7943	0.0744	9.6749	0.0733	-6.1249	0.0635	-6.1323	0.0652
			45	9.7728	0.0649	9.6812	0.0624	-6.1345	0.0575	-6.1401	0.0561
			55	9.7567	0.0566	9.6819	0.0560	-6.1332	0.0509	-6.1379	0.0502
0.3	10.369	-6.1463	25	10.4747	0.0634	10.370	0.0647	-6.1420	0.0516	-6.1472	0.0537
			35	10.4444	0.0540	10.370	0.0556	-6.1443	0.0439	-6.1485	0.0436
			45	10.4292	0.0485	10.371	0.0484	-6.1437	0.0382	-6.1464	0.0376
			55	10.4168	0.0419	10.368	0.0414	-6.1434	0.0353	-6.1455	0.0346
0.4	10.8544	-6.1138	25	10.9457	0.0629	10.854	0.0635	-6.1110	0.0472	-6.1146	0.0486
			35	10.9203	0.0532	10.854	0.0518	-6.1138	0.0396	-6.1157	0.0392
			45	10.9068	0.0466	10.856	0.0458	-6.1125	0.0329	-6.1148	0.0328
			55	10.8947	0.0402	10.854	0.0409	-6.1114	0.0310	-6.1131	0.0312

n representing the number of trees in one plot; OLS, least square; QR, quantile regression; The system is close to homoscedastic.

Table 3.5. Simulation results from the modified growth system abstracted from CAPPS data

P	α	β	n	α				β			
				OLS		QR		OLS		QR	
				Est	Std	Est	Std	Est	Std	Est	Std
0.1	8.8299	-6.2501	25	9.0774	0.1453	8.8280	0.1227	-6.1391	0.1939	-6.2466	0.1806
			35	9.0096	0.1240	8.8330	0.1035	-6.1687	0.1700	-6.2510	0.1475
			45	8.9668	0.1154	8.8252	0.0898	-6.1839	0.1595	-6.2400	0.1321
			55	8.9405	0.1097	8.8254	0.0865	-6.1931	0.1517	-6.2433	0.1264
0.2	9.6807	-5.8383	25	9.8630	0.1364	9.6839	0.1153	-5.7847	0.1592	-5.8409	0.1431
			35	9.8053	0.1172	9.6817	0.0972	-5.7958	0.1398	-5.8385	0.1203
			45	9.7754	0.1049	9.6750	0.0876	-5.8028	0.1308	-5.8332	0.1134
			55	9.7572	0.0959	9.6775	0.0781	-5.8069	0.1185	-5.8354	0.1010
0.3	10.369	-5.3463	25	10.4821	0.1230	10.372	0.0965	-5.3031	0.1430	-5.3497	0.1200
			35	10.4415	0.1040	10.366	0.0835	-5.3083	0.1230	-5.3436	0.1055
			45	10.4293	0.0929	10.368	0.0744	-5.3217	0.1110	-5.3465	0.0915
			55	10.415	0.0839	10.367	0.0658	-5.3254	0.1013	-5.3484	0.0836
0.4	10.8544	-5.1138	25	10.9526	0.1175	10.857	0.0955	-5.0767	0.1282	-5.1156	0.1103
			35	10.9210	0.1004	10.853	0.0811	-5.0880	0.1119	-5.1144	0.0966
			45	10.9085	0.0907	10.853	0.0720	-5.0968	0.1018	-5.1139	0.0846
			55	10.8953	0.0793	10.853	0.0642	-5.0967	0.0897	-5.1144	0.0763

n representing the number of trees in one plot; OLS, least square; QR, quantile regression; . The system is modified to be heteroscedastic.

Table 3.6. RMSE comparison of quantile regression and ordinary least squares

p	n	RMSE		RMSE
		OLS	QR	OLS/QR
	25	0.1416	0.0432	3.2778
	35	0.1048	0.0381	2.7507
	45	0.0825	0.0334	2.4701
0.1	55	0.0691	0.0301	2.2957
	25	0.0986	0.0401	2.4489
	35	0.0729	0.0345	2.1130
	45	0.0581	0.0298	1.9497
0.2	55	0.0491	0.0272	1.8051
	25	0.0834	0.0389	2.1362
	35	0.0628	0.0331	1.8973
	45	0.0488	0.0282	1.7305
0.3	55	0.0417	0.0249	1.6747
	25	0.0737	0.0373	1.9759
	35	0.0562	0.0316	1.7749
	45	0.0436	0.0269	1.6202
0.4	55	0.0377	0.0243	1.5514

CHAPTER 4

MODEL-BASED SAFIS SAMPLE PLOT UPDATING

4.1. Introduction

The forest inventory and analysis (FIA) program provides information on the status and trends of forest resources. The USDA forest service has developed an annual inventory system where 20% of each state's inventory is conducted every year, and all FIA plots are to be measured in a 5-year cycle. Since only 20% of the sample plots are measured in current year and other plots are 1 to 4 years old, updating techniques are required to eliminate this lag and improve population parameter estimates. One simple approach to annual estimates is to use only the current 20%. However, it is less accurate due to the small sample size. It has been suggested that estimates for current forest status should take advantage of previous data since they are only a few years old and contain a significant amount of information about the current status. Samples of research works on utilizing previous FIA data to improve estimates for the current forest conditions include Lessard et al. (2001), Johnson et al. (2003), and Van Deusen et al. (1999).

Many approaches to calculating annual FIA estimates have been considered. These approaches can be classified into 3 categories depending on the level of variable being updated,

namely, population level (Van Deusen, 1999), plot level (Reams & Van Deusen, 1999), and individual tree level (Lessard et al, 2001).

Population Level Models

Many approaches have been proposed to calculate annual FIA based on the sampled population. One of the simplest approaches is to use the five most recent panels of measurements to calculate annual estimates for current status and trends. FIA has selected a five-year moving average (MA) as the default estimator for the new annual inventory system (Van Deusen, 2002). MA is obtained by assuming that there is no time trend. Suppose that the numbers of plots in five panels are n_1, n_2, n_3, n_4 and n_5 and a total of N plots distributed in the five panels. A simple five-year moving average involving summing the means from five consecutive panels is

$MA_{t-4,t} = \sum_{i=t-4}^{t} w_i \bar{y}_i$, where $w_i = \frac{n_i}{N}$, t=5, and $\bar{y}_i$ is the mean of the panel measured in year i. The variance of the MA can be calculated as

$$V\left(MA_{t-4,t}\right) = \sum_{i=t-4}^{t} w_i^2 V\left(\bar{y}_i\right) = \sum_{i=t-4}^{t} w_i^2 \frac{\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)}{n_i(n_i - 1)}$$

The MA estimate for a population parameter is easy to calculate and has a variance estimator that is relatively easy to calculate from data as well (Johnson et al., 2003). The major problem with the MA is that it is not an unbiased estimator for any particular population

parameter (Van Deusen, 2002). Another problem is the selection of weights. The weight defined as $w_i = \frac{n_i}{N}$ might mask time trends that Southern Annual Forest Inventory System (SAFIS) was intended to evaluate. However, the equal weighting estimator can be viewed as an unbiased estimator at some time approximately in the middle of the rotation cycle (Johnson et al. 2003). Johnson et al. (2003) studied the performance of three classes of weighted average estimators for an annual inventory: ARIMA(0,1,1) time series model, ARIMA(0,2,2) time series model, and locally least squares regression. Simple moving average with equal weight is a special case of ARIMA(0,1,1). Johnson et al. claimed that the MA performed well in terms of mean square error in virtually every simulation situation. It is tended to be the best among the estimators tested if spatial variation was large and change was relatively small (Johnson, 2001). Their conclusion was consistent with Van Deusen's (2002). Van Deusen (2002) compared the MA approach with two other alternatives: simple one panel mean and a mixed estimator. He concluded, based on simulated data, that level trend (in which change is small, or horizontal) is a near optimal situation for the MA because the expected value of every year is almost the same. In this case, the MA follows the level trends well and has small variance relative to simple one panel mean and the mixed estimator. However, when there is a trend in FIA data over time, simulated comparisons showed that the mixed estimator outperforms the MA. The MA approach tends to lag evolving trends, which can result in very large bias (Van Deusen 2002).

Van Deusen's mixed estimator (Van Deusen, 1996, 1999, 2002; Roesch 1999; Theil 1971; Scott et al. 1999) can be viewed as a compromise between a frequentist and a Beyesian approach. Two models, the observation equation and the transition equation, are required to describe the ME formulation. The observation model describes the observations for time t=1, 2, 3,…T, $\overline{Y} = \beta + e$, where $\overline{Y} = (\overline{y}_1, \overline{y}_2 ..., \overline{y}_T)^T$ is the mean vector of plot observations in the panel measured at time t, $e = (e_1, e_2, ..., e_T)^T$ is an independent random error vector representing sampling error, and $\beta = (\beta_1, \beta_2, ..., \beta_T)^T$ is an unknown random coefficient representing the population mean at time t. The transition equation describes how the random coefficients change over time or constraints on the time trend, $R\beta = V$, where R is the constraint matrix on the parameter β, and V is an error vector. Combining $\overline{Y} = \beta + e$ and $R\beta = V$ yields

$$\begin{pmatrix} \overline{Y} \\ 0 \end{pmatrix} = \begin{pmatrix} I \\ R \end{pmatrix} \beta + \begin{pmatrix} e \\ V \end{pmatrix}, \text{ where } \overline{Y} \sim N(\beta, \Sigma), \; V \sim N(0, p\Omega)$$

Plot Level Models and Individual Tree Models

Plot level and individual level models provide the means to update the current year data for plots measured in previous years and then base estimates on the data for all plots. Two categories updating techniques, imputation (Reams et al., 1999; McRoberts, 1999, 2001) and modeling (McRoberts, 1999, 2001; Lessard et al. 1999, 2001), are of general interest (McRoberts, 1999).

The imputation approach to dealing with plots previously measured is to view their current observations as missing. Imputation can be done by either matching methods or modeling methods. Matching imputation methods seek plausible and consistent replacement observations by selecting from a pool of current observations that either match prescribed attributes of the missing observations or are only similar to the missing observations with respect to the prescribed attributes. Modeling imputation methods use a regression model to estimate the mean value and add an error term. Both the matching and modeling methods can be divided into two main categories: multiple imputation and single imputation. The multiple imputation differs from the single imputation in that it allows assessment of uncertainty in imputed variables and excludes extreme results by multiple completions of the imputed data set. Reams and McCollum (1999) evaluated multiple imputation models for SAFIS, and their research results indicated that modeling methods and matching methods gave nearly identical results for both means and variances. Furthermore, they found that there is no practical difference between inventory estimates when using a model with predictive capability and one that is relatively low. Gartner and Reams (2001) applied multiple imputation to update FIA data for Georgia at both the plot level and the tree level. They concluded that the tree level modeling imputation performed best overall.

The use of modeling to update FIA is not new in forestry. One famous example is STEMS (Belcher et al., 1982), by which the north central research station (NCRS) of the USDA updates FIA data for a proportion of well-established, undisturbed FIA plots. Other recent examples of using model-based techniques intended to update data for annual inventory are Lessard et al. (2001, 1999) and McRoberts (2001,1999). Lessard et al. (2001) constructed distance-independent individual models for species groups using FIA data from Minnesota. Models calibrated using the form and methodology presented in their article will be used by NCRS for updating information on plots collected under the annual inventory system in the north central region. McRoberts (2001,1999) compared one imputation with model-based updating for annual forest inventory with respect to basal area. The comparison indicated that simple plot level imputation and model-based updating techniques produced similar estimates, though the best model-based results were slightly superior to the best imputation results. McRoberts (2001) realized that model-based updating techniques would be facilitated as FIA measurements at a five-year cycle accumulate. Model-based updating may be further justified if models can be used for long-term predictions. Borders (1997) built a set of whole stand models for natural pine stand in Georgia to update individual FIA plot data. All models fitted in his study were algebraically differenced and calibrated to each individual FIA plot.

Although imputation is accurate for annual inventory statistics, one inherent disadvantage of imputation techniques is their dependence on five-year average annual growth as a surrogate for annual growth. The disadvantage prevents imputation techniques from being a good means of estimating change. The goal of FIA is to provide information on current status and trends of forest resources. Change is at least as important as current status to most users of FIA data (Van Deusen 2001). As the annual system proceeds, more and more measurements of FIA plots become available and will greatly facilitate model-based updating techniques. How to make best use of SAFIS data to estimate trends of forest resources on the basis of the annual inventory plots is of general interest.

Since the FIA program requires a sampling intensity of one plot per approximately 6,000 acres (Brand et al., 1999), plot parameters, such as volume, basal area, and trees per acre, etc., vary much more than permanent research sample plots. This feature will challenge modeling. Forest biometricians should achieve a compromise between use of growth information on similar plots and variation among FIA plots. Although some feasible models were developed for the intended purpose of annual forest statistics (Lessard, 1999, 2001; McRoberts, 2001,1999), unfortunately, they are not appropriate for long-term predictions. First, the functions of these models are annual growth functions, which can be viewed as integration of instantaneous growth functions. Therefore, it is apparent that prediction precision would decrease as prediction

intervals increase. Secondly, these models fail to incorporate observations of sample plots to improve predictions. Although covariates can be used to account for variations among FIA plots, observations can provide more information on plot growth. This can be justified by widely used applications of projection models in forestry, which take the advantage of one observation as a snapshot of the cross-section of subject growth trajectory. Thirdly, model-based updating for FIA data depends on two sources of information, information on the subject plot and information from similar plots from which strength can be borrowed to improve prediction. Since I do not intend to apply models built on based FIA to other stands or plots than the sampled, model fitting should not be constrained by model applications.

Lessard (2001) proposed the average dbh function $E(\Delta d) = \beta_1 \exp^{(-\beta_2 d)} d^{\beta_3}$ for Minnesota FIA updates. Suppose that the curve generated with $E(\Delta d) = \beta_1 \exp^{(-\beta_2 d)} d^{\beta_3}$ (parameter estimates are 0.1535, 0.0378, 0.3897 for β_1, β_2, and β_3 in Lessard (2001) for Jack Pine) represents the real instantaneous dbh growth rate curve. The other curves in Figure 4.1 were generated using the following method: a) calculating instantaneous dbh growth rates at dbh values, which were taken systematically from the hypothesized real curve; b) reproducing a dbh growth series such that $\tilde{d}_i = d_{i-1} + I * R_{i-1}$, where $\tilde{d}_i, d_{i-1}, R_{i-1}$, and I denote the new dbh, the hypothesized real dbh, instantaneous rate corresponding to d_{i-1}, and forward projection interval, respectively; c) calculating a new instantaneous dbh growth rate curve based on the newly

generated dbh series. It can be seen that a new growth rate curve deviates from the hypothesized real one approximately proportional to projection interval. Figure 4.1 is used to illustrate why a growth function like $E(\Delta d) = \beta_1 \exp^{(-\beta_2 d)} d^{\beta_3}$ is inappropriate for long-term prediction since different instantaneous growth rate curves imply different yield curves.

Clutter (1963) proposed in his groundbreaking article that a growth function should be compatible with the corresponding yield function in sense that integration of the growth function should produce the yield function. Based on a derivative-integral relationship, Clutter developed a set of stand level functions, which are the first projection function system. Sullivan and Clutter (1973) derived the same function system with the algebraic difference approach (ADA). Clutter et al. (1983) summarized that a projection function system either by derivative-integral or by ADA provided a logically consistent set of equations for prediction, i.e., current status, future status, and instantaneous growth rate. Projection functions are applicable to various projection intervals and do not require that data be annual increments. This property facilitates its application in FIA sample plot updates since forward projection intervals in time are 1, 2, 3, and 4 years, not a fixed interval. It should be noted that some desirable logical properties were also discussed (Clutter et al. 1983, p.123) and that a projection function is illogical if it lacks these properties.

Take the Clutter basal area equation, $\ln B=\beta_0+\beta_1 S+\beta_2\left(\frac{1}{A}\right)+\beta_3\left(\frac{\ln D}{A}\right)+\beta_4\left(\frac{S}{A}\right)$, for example, where D is basal area at age 20. Differentiation with respect to A and algebraic rearrangement gives $\frac{dB}{dA}=B\left(\frac{1}{A}\right)\left(\beta_0+\beta_1 S-\ln B\right)$ as the corresponding growth rate equation. Integration of the growth equation yields $\left[-\ln\left(B\left(\beta_0+\beta_1 S-\ln B\right)\right)\right]\left|\begin{matrix}B_2\\B_1\end{matrix}\right.=\left(\ln A\right)\left|\begin{matrix}A_2\\A_1\end{matrix}\right.$ by integrating $\frac{dB}{B\left(\beta_0+\beta_1 S-\ln B\right)}=\frac{dA}{A}$. It comes to $\ln B_2=\left(\frac{A_1}{A_2}\right)\ln B_1+\left(\beta_0+\beta_1 S\right)\left(1-\frac{A_1}{A_2}\right)$, which is identical to the equation derived with ADA replacing $\left(\beta_2+\beta_3\ln D+\beta_4 S\right)$ in $\ln B=\beta_0+\beta_1 S+\left(\frac{1}{A}\right)\left(\beta_2+\beta_3\ln D+\beta_4 S\right)$ with $\left(\ln B_1-\beta_0-\beta_1 S\right)A$.

Ramirez-Maldonado et al. (1987) noted that the derivative-integral relationship between growth and yield equations holds true for those derived by the algebraic difference approach. There is no distinction existing between the derivative-integral and ADA. ADA features identifying a parameter of the underlying equation to account for variation among individuals, solving the equation at a reference age, say A_1, for the identified parameter, replacing the parameter in the equation at a projection age, say A_2, with a prior observation. The parameter replaced corresponds to the constant of integration, say c, of the solution of a certain differential equation (Ramirez-Maldonado et al. 1987). Which parameter in a yield equation corresponds to c depends on the derivative of the yield equation. Again, I take the Schumacher function $\ln y=\beta_0+\beta_1\left(\frac{1}{A}\right)$ for example. Differentiating it with respect to A yields $\frac{dy}{dA}=\beta_1 y\left(-\frac{1}{A^2}\right)$ or further $\frac{dy}{dA}=y\left(-\frac{1}{A}\right)\left(\ln y-\beta_0\right)$ with $\beta_1\left(\frac{1}{A}\right)$ in $\frac{dy}{dA}=\beta_1 y\left(-\frac{1}{A^2}\right)$ replaced with $\ln y-\beta_0=\beta_1\left(\frac{1}{A}\right)$.

Integrating these two growth rate equations yields $\ln y = c + \beta_1\left(\frac{1}{A}\right)$, and $\ln(\ln y - \beta_0) = \ln\left(\frac{1}{A}\right) + c$, respectively. Imposing the constraint that $y_2=y_1$ when $A_1=A_2$ is equivalent to specifying the parameter corresponding to C to be the individual specific parameter. Projection models derived with the derivative-integral approach can be derived with ADA.

In this chapter, I focus on estimating current status and trends of forest resources in Georgia using modeling techniques. Analysis on published models in this context leads to application of projection equations with desired logical properties. In addition, realizing some drawbacks of projection models in the case that multiple prior observations are available for predictions, I discuss the applicability of EBLUP for long-term predictions, though it is not feasible to fit mixed model system at present. A whole stand growth and yield model system will be developed.

4.2. Model System for SAFIS Updates and Trend Evaluation

The data used for this study was provided by Southern Research Station (SRS), the USDA forest service, in Knoxville, Tennessee. The data are from 1997 periodic sampling and following annual inventory samplings 1998 through 2004. The plot design in these samplings was based on a cluster of four fixed plots spaced 120 feet apart. Each served as the center of a 1/24-acre circular subplot used to sample trees 5.0 inches dbh and larger. A 1/300-acre circular microplot, located at the center of the subplot, was used to sample trees 1.0 through 4.9 inches

dbh as well as seedlings. SRS provided a total of 3,790 plots measured both in 1997 and a following annual survey performed in Georgia.

Loblolly pine and Slash pine are the two most abundant species, accounting for 26.6% and 12.6% of all trees in the data, respectively. The following criteria were used to screen the data for model fitting: a) either loblolly or slash pine, c) only use records for live trees, and b) no signs for excessive damage, cutting or mortality. A total of 813 loblolly pine and 301 slash pine plots satisfied the screening criteria and were available for this study. These plots were combined to construct one dataset containing 1114 plots. Volume/plot (1/6 acre), basal area/plot, and trees per plot were calculated for all trees alive with five inches and larger. The plots were grouped according to species (loblolly pine and slash pine), forest origin (natural or artificial stands), and physiographic regions (piedmont, upper coastal plain, and lower coastal plain). Table 4.1 shows the plot distribution by physiographic regions, species, and stands origins.

In the following context, variables are defined as follows: H=dominant height (ft); V=cubic ft volume; BA=basal area of a give plot (ft^2); N=trees per plot; A=stand age of a given plot; others are parameters.

Dominant/Codominant Height Model

Dominant height was determined for each SAFIS plot by averaging the total height of all tally trees that were classified as dominant or codominant. Several model forms were

investigated, including the Chapman-Richards function, the Schumacher function, the Weibull function, and the McDill-Amateis function (McDill & Amateis, 1992). The best empirical model form by statistics of fit (RMSE) is the Chapman-Richards function (Equation 4.1) with ϕ_3 specified as the local parameter. Equation 4.2 was derived through the algebraic difference approach by solving for ϕ_3 and replacing it with a prior observation.

$$H = \phi_1 \left(1 - e^{\phi_2 A}\right)^{\phi_3} \tag{4.1}$$

$$H_2 = \phi_1 \left(\frac{H_1}{\phi_1}\right)^{\left(\frac{\ln(1-e^{-\phi_2 A_2})}{\ln(1-e^{-\phi_2 A_1})}\right)} \tag{4.2}$$

Equation 4.2 was fitted to the dataset, where $\phi_1 = \beta_{10} + \sum_{i=1}^{4} \beta_{1i} I_i$ and $\phi_2 = \beta_{20} + \sum_{i=1}^{4} \beta_{2i} I_i$.

Indicator variable I_1=1 if one plot is loblolly pine and I_1=0 otherwise. Indicator variables I_2, I_3, and I_4 take 1 for plantation, piedmont, and upper costal plain, respectively, and 0 otherwise.

There is no statistically significant difference between species or physiographic regions, but there is between natural stands and artificial stands. Plantation stands have a smaller ϕ_1 and larger ϕ_2 estimates. This implies that plantations grow faster but with smaller asymptotic height. The final parameter estimates and statistics of fit are summarized in Table 4.1. Parameter estimates by stands origin are presented in Table 4.3. Figure 4.2 shows that there is no systematic

pattern of residuals. The predicted dominant height vs. the observed is illustrated in Figure 4.4, which indicates that model form 4.2 underestimate dominant height when it is large.

Volume Projection Model

Analysis of the data showed that a strong relationship between volume and two commonly used measures of stand density, basal area and survivals. The final model form is 4.3 (Borders, 1997; Hall & Clutter, 2004). It was based on the basic relation described by equation 4.4.

$$\ln V_2 = \ln V_1 + \phi_1 \ln\left(\frac{H_2}{H_1}\right) + \phi_2 \ln\left(\frac{BA_2}{BA_1}\right) + \phi_3 \ln\left(\frac{N_2}{N_1}\right) \tag{4.3}$$

$$V = \phi_0 H^{\phi_1} BA^{\phi_2} N^{\phi_3} \tag{4.4}$$

where $\phi_k = \beta_{k0} + \sum_{i=1}^{4} \beta_{ki} I_i$ (k=1, 2, 3), I_1, I_2, I_3, and I_4 equals 1 for loblolly pine, plantation, piedmont, and upper coastal plain, respectively, 0 otherwise.

OLS Parameter estimates are listed in Table 4.4. Partial F-test shows that there is no significant difference among physiographic region. It is evidenced by statistical tests that a statistical significant difference exists between loblolly pine and slash pine by stands origin. Accordingly, a total of four models are fitted simultaneously. Parameter estimates by species, and stands origin are summarized in Table 4.5. Plots of residual vs. fitted value (in natural logarithm scale) and predicted value vs. observed value are presented in Figures 4.3 and 4.5, respectively.

Figure 4.5 shows that model form 4.3 will provide accurate predictions of volume per plot, provided that basal area and dominant height can be predicted accurately. It is also revealed that model form 4.4 slightly underestimates volume per plot when it is large.

Survival Model

The most challenging parts of this model system are basal area and trees per plot. Modeling survivals of trees is the most difficult due to relatively high variation in mortality pattern (Borders, 1997). The variation was exacerbated in SAFIS sample plots by variation in environment, management, stands origin, physiographic region, and measurement error. Several commonly used nonlinear survival functions, $N_2 = \left(N_1^{\phi_1} + \phi_2 \ln\left(A_2^{\beta_3} - A_1^{\beta_3}\right)\right)^{\frac{1}{\phi_3}}$ (Clutter & Jones, 1980), $\ln N_2 = \ln N_1 + \phi_1\left(A_2^{\beta_3} - A_1^{\beta_3}\right)$ (Pienaar & Shiver, 1981), and $N_2 = N_1\left(\frac{A_2}{A_1}\right)^{\phi_1} Exp\left(\phi_2\left(A_2 - A_1\right)\right)$ (Clutter et al., 1983) were examined. The Pienaar-Shiver function, for which significant parameter estimates were obtained, produced nonsensical survival trends. Parameter estimates for others failed to converge. Close examination of the dataset indicated that there was a very strong relationship between basal area and survivals. The only explanatory variable that explains most of the variation in trees per plot is basal area. Variations of the Schumacher function were evaluated for reasonable trends and statistics of fit. Statistics of fit and inspection of trends suggest that the final model form be equation 4.5.

$$\ln N_2 = \ln N_1 + \phi_1 \ln\left(\frac{H_2}{H_1}\right) + \phi_2 \ln\left(\frac{BA_2}{BA_1}\right) + \phi_3\left(\frac{1}{A_2} - \frac{1}{A_1}\right) \quad (4.5)$$

where all parameters are as defined previously.

Parameter estimates are presented in Table 4.6. The table shows significant differences between natural stands and artificial stands by physiographic region. Survivals of natural stands are not statistically dependent on stands age. This is most likely because stand age measurement errors mask the relationship between stand age and survivals. The measurement errors also might explain why models based on data for permanent research plots do not apply to FIA plots. Stand age of natural stands is the most difficult parameter to measure. It was determined based on the age of two or three dominant or codominant trees. Parameter estimates by physiographic region and stands origin are given in Table 4.7. The plot of predicted vs. observed survivals is presented in Figure 4.7.

Basal Area Projection Model

The only available independent variable for basal area projection model is stand age. Statistics of fit and inspection of trends led to the Clutter basal area model (Clutter, 1963; Clutter & Jones 1980). The model form is indexed as equation 4.6 in this chapter. Parameter estimates for equation 4.6 and estimates by origin are given in Table 4.8 and 4.9, respectively. Figure 4.6 presents the plot of the predicted basal area vs. the observed.

$$\ln BA_2 = \phi_1\left(1-\left(\frac{A_1}{A_2}\right)^{\phi_2}\right)+\ln BA_1\left(\frac{A_1}{A_2}\right)^{\phi_2} \quad (4.6)$$

Applicability of the Model System

The model system will provide per plot estimates of dominant height, volume, basal and trees per plot. Because projection intervals in the dataset range from 1 to 8 years, statistics of fit are very favorable. This indicates that these models should work well for updating SAFIS sample plots (1 through 4 years). Basal area and survival of trees are less accurate components of the model system due to relatively high variation in mortality pattern among sample plots and measurement errors in stand age. These less accurate components likely decrease the applicability of the system for a long projection interval. In addition, the accuracy of projection decreases as projection interval increases. As a result, the models might not effectively achieve the objective of evaluating forest trends. However, as SAFIS proceeds, more annual inventory data accumulates for individual plots as well as the population. Incorporating the new observations of each plot will certainly improve individual plot updating. How to incorporate new observations of each plot deserves further research. Mixed effects models are the most efficient approach to make use of multiple observations. A theoretical analysis of applicability of mixed models is given in the following section.

It should be noted that the model system applies only to sample plots not disturbed excessively by management and damage. Therefore, an additional model is required to estimate the probability of disturbance. In addition, a separate model is needed for sample plots that are expected to have many ingrowths. All these additional models can be fitted only after sufficient individual plot data become available. As for species or species groups that make up a small proportion of the population, the matching imputation method should be employed.

4.3. Using EBLUP to Update SAFIS Sample Plots-A Theoretical Analysis

Suppose that for the *i*th of *m* individual plots, n_i responses have been observed so that a total of $N = \sum_{i=1}^{m} n_i$ data values are available. The data vector, y_i, for the *i*th individual plot, satisfy linear model 4.7

$$y_i = X_i\beta + Z_i b_i + e_i \tag{4.7}$$

$$E(y_i|b_i) = X_i\beta + Z_i b_i \tag{4.8}$$

$$Cov(y_i|b_i) = R_i \tag{4.9}$$

where y_i is n_i by 1 response vector for observations on the *i*th individual, X_i is an n_i by p model matrix for the fixed effects, β is p by 1 vector of fixed effects coefficient, b_i is a r by 1 vector of random effects coefficient, Z_i is an n_i by r model matrix linking y_i to random effects; $e_i \sim N_{n_i}(0, R_i)$ and R_i is n_i by n_i covariance matrix. All except β are specific to the *i*th individual.

Suppose that b_i is from a normal distribution such that $b_i \sim N_r(0,D)$, where D is a r by r covariance matrix, independent of b_j or e_i. The marginal mean and covariance for y_i are expressed by equations 4.10 and 4.11, respectively.

$$E(y_i) = E(E(y_i|b_i)) = X_i\beta \tag{4.10}$$

$$Cov(y_i) = Cov(E(y_i|b_i)) + E(Cov(y_i|b_i)) = Z_i D Z_i^T + R_i \tag{4.11}$$

Stacking all data for all m individuals gives model 4.12

$$y = X\beta + Zb + e \tag{4.12}$$

where $y = \left(\left(y_1^T, ..., y_m^T \right)^T \right)_{N\times 1}$; $y \sim N(X\beta, V(\theta))$, $V(\theta)$ indicates that V is assumed to be known up to a parameter vector that is supposed to be q by 1; $b = \left(\left(b_1^T, ..., b_m^T \right)^T \right)_{rm\times 1}$,

$X = \left(\left(X_1^T, ..., X_m^T \right)^T \right)_{N\times p}$, $e = \left(\left(e_1^T, ..., e_m^T \right)^T \right)_{N\times 1}$, $Z = DM\left(Z_{n_1\times r}, ..., Z_{n_m\times r} \right)_{N\times rm}$,

$\tilde{D} = DM\left(D_{r\times r}, ..., D_{r\times r} \right)_{rm\times rm}$, and $R = DM\left(R_{n_1\times n_1}, ..., R_{n_m\times n_m} \right)_{N\times N}$, and

$V = DM(V_1, ..., V_m)_{N\times N} = R + Z\tilde{D}Z^T$, $V_i = R_i + Z_i\tilde{D}Z_i^T$. The acronym DM stands for diagonal matrix here.

Given that V is known (D and R are known), it can be shown that the estimator for β is the generalized least squares estimator $\hat{\beta} = \left(X^T V^{-1} X\right)^{-1} X^T V^{-1} y$, and that estimator for b is $\hat{b} = DZ^T V^{-1}\left(y - X\hat{\beta}\right)$ (BLUP) so that the individual estimator for ith individual is given by

$\hat{b}_i = DZ_i^T V_i^{-1}\left(y_i - X_i\hat{\beta}\right)$. If a point estimate of V_i (D and R_i) is available, it is to be used to replace D and R and yields EBLUP such that $\hat{b}_i = \hat{D}Z_i^T \hat{V}_i^{-1}\left(y_i - X_i\hat{\beta}\right)$.

When D and R_i are unknown, and y is assumed to be normal, the log likelihood is equation 4.13. Partial derivatives with respect to β and θ yield equations 4.14 and 4.15, respectively. In the context of unbalanced longitudinal data, a closed form solution is not available. They can be solved simultaneously with a numerical algorithm (e.g., Newton-Raphson) to estimate β and θ (for details see McCulloch and Searle 2001).

$$\log L(\beta,\theta;y) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log\left(|V(\theta)|\right) - \frac{1}{2}(y - X\beta)^T V(\theta)^{-1}(y - X\beta) \tag{4.13}$$

$$\frac{\partial L}{\partial \beta} = -\beta^T X^T V(\theta)^{-1} X + y^T V(\theta)^{-1} X \tag{4.14}$$

$$\frac{\partial L}{\partial \theta_k} = -\frac{1}{2}\left(\mathrm{tr}\left(V(\theta)\frac{\partial V(\theta)}{\partial \theta_k}\right) - (y - X\beta)^T V(\theta)^{-1}\frac{\partial V(\theta)}{\partial \theta_k}V(\theta)^{-1}(y - X\beta)\right) \tag{4.15}$$

Nonlinear mixed model estimation is based on a first order Taylor expansion around b_i=0 (or other values). A nonlinear model such as model 4.16 can be linearized through the expansion in a way such that $y_{ij} = f\left(A_i\beta, x_{ij}\right) + Z_i b_i + e_{ij}$, where $Z_i = \frac{\partial f\left(A_i\beta + B_i b_i, x_{ij}\right)}{\partial b_i^T}\Big| b_i = 0$. Note that $f\left(A_i\beta, x_{ij}\right)$ plays the role of $X\beta$ in a linear model.

I follow Pinheiro and Bates (2000) notations to present mixed model and EBLUP. The jth observation on ith plot is modeled as model (4.16)

$$y_{ij} = f\left(\phi_i, x_{ij}\right) + e_{ij} \tag{4.16}$$

where $\phi_i = A_i\beta + B_i b_i$, i=1,…, m; j=1,…,n_i, $b_i \sim N(0,D)$; m is the number of sample plots; n_i is the number of observations on ith plot; ϕ_i is plot specific parameter vector; the matrices A_i and B_i depend on plot (also possibly depend on the values of some covariates at the jth observation); β is a p dimensional vector of fixed effects. b_i is r dimensional vector of random effects associated with ith plot; e_{ij} does not have be iid. Davidian and Giltinan (1995) used variance function $Cov(e) = \sigma^2 G^{1/2}(\beta,\theta)H(\alpha)G^{1/2}(\beta,\theta)$ to account for heterogeneity and correlation within individual, where $H(\alpha)$ is correlation function and $G^{1/2}(\beta,\theta)$ is the diagonal matrix with elements of the square root of $G(\beta,\theta)$, which is used to model heterogeneity. The variance function $Cov(e)$ is in place of R in following formulations if e_{ij} is not iid. Model (4.17) is the mixed effects Schumacher model with all parameters being mixed. Suppose one has a certain factor f, with 3 values, say f_1, f_2, and f_3, affects ϕ_{i1} and ϕ_{i2} through linear models $\phi_{1i} = \beta_1 + \beta_2 I_1 + \beta_3 I_2 + b_{1i}$ and $\phi_{2i} = \beta_4 + \beta_5 I_3 + \beta_6 I_4 + b_{2i}$, where I_1=I_3=1 if the factor $f = f_2$, 0 otherwise, and I_2=I_4=1 if the factor $f = f_3$, and 0 otherwise. It is apparent that β_2, β_3, β_5, and β_6 represent the differences from f_1. The jth response of the ith plot at level f_2 can be modeled with model 4.17:

$$y_{ij} = \phi_{1i} Exp\left(\phi_{2i} A^{\phi_{3i}}\right) + e_{ij} \qquad (4.17)$$

$$\underbrace{\begin{pmatrix}\phi_{1i}\\ \phi_{2i}\\ \phi_{3i}\end{pmatrix}}_{\phi_i} = \underbrace{\begin{pmatrix}1 & 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 1\end{pmatrix}}_{A_i} \underbrace{\begin{pmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \beta_4\\ \beta_5\\ \beta_6\\ \beta_7\end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}}_{B_i} \underbrace{\begin{pmatrix}b_{1i}\\ b_{2i}\\ b_{3i}\end{pmatrix}}_{b_i} \tag{4.18}$$

where β_1 =the intercept of ϕ_1 , β_4 =the intercept of ϕ_2 , β_7 =fixed parameter of ϕ_3 .

Suppose a vector of n_{k1} prior observations on plot *k*, say $Y_{k_1} = \begin{pmatrix} y_{k_1 1} \\ \vdots \\ y_{k_1 n_k} \end{pmatrix}$, corresponding to $X_{k_1} = \begin{pmatrix} x_{k_1 1} \\ \vdots \\ x_{k_1 n_k} \end{pmatrix}$, A_k, and B_k (X_k could be a matrix).

One can predict b_k with equation 4.19, where, b_k is *r* by 1; $\hat{Z}_k = \begin{pmatrix} \hat{Z}_{k1} \\ \vdots \\ \hat{Z}_{kn_k} \end{pmatrix}$ is n_{k1} by *r*;

$\hat{Z}_{ks} = \left.\frac{\partial f\left(A_k\beta + B_k b_k, x_{ks}\right)}{\partial {b_k}^T}\right|_{\beta=\hat{\beta}}$ (S =1,…,n_{k1}); $\left(Y_k - f\left(A_k\hat{\beta}, X_k\right)\right)$ is n_{k1} by 1; R is n_{k1} by n_{k1}; D is *r* by *r*; Z_k is n_{k1} by *r*.

$$\hat{b}_k = \hat{D}_{k_1}{\hat{Z}_{k_1}}^T\left(\hat{Z}_{k_1}\hat{D}_{k_1}{\hat{Z}_{k_1}}^T + \hat{R}_{k_1}\right)^{-1}\left(Y_{k_1} - f\left(A_k\hat{\beta}, X_{k_1}\right)\right) \tag{4.19}$$

$$\hat{Y}_{k_2} = f\left(A_k\hat{\beta} + B_i\hat{b}_i, X_{k_2}\right) \tag{4.20}$$

After b_k is predicted, equation 4.20 can be used to predict n_{k2} unobserved responses of *k*th plot at covariate X_{k2}, say Y_{k2}.

The dependence between e_{k1} and e_{k2}, error vectors associated with Y_{k1} and Y_{k2}, respectively, is ignored by equation 4.19. If the dependence exists, the following model is preferable since the dependence is accounted for through covariance matrix V_{k1k2} between Y_{k1} and Y_{k2}. The best (minimum variance) linear unbiased estimator of Y_{k2} given Y_{k1} known is equation 4.21, which is the expectation of Y_{k2} conditional on Y_{k1} if Y_{k1} and Y_{k2} follow a joint distribution such that $\begin{pmatrix} Y_{k_1} \\ Y_{k_2} \end{pmatrix} \sim N\left(\begin{pmatrix} E(Y_{k_1}) \\ E(Y_{k_2}) \end{pmatrix}, \begin{pmatrix} V_{k_1} & V_{k_1k_2} \\ V_{k_1k_2} & V_{k_2} \end{pmatrix} \right)$ (see Hall & Bailey, 2001; Hall & Clutter, 2004).

$$\hat{Y}_{k_2} = E(Y_{k_2}) + V_{k_1k_2} V_{k_1}^{-1} \left(Y_{k_1} - E(Y_{k_1}) \right) \tag{4.21}$$

Replacing all unknown quantities in RHS of equation 4.22 with their estimates yields $\hat{Y}_{k_2} = E(\hat{Y}_{k_2}) + \hat{V}_{k_1k_2} \hat{V}_{k_1}^{-1} \left(\hat{Y}_{k_1} - \hat{E}(Y_{k_1}) \right)$, where $\hat{E}(Y_{k_u}) = f\left(A_k \hat{\beta} + B_i \hat{b}_i, X_{k_u} \right)$ (u=1 or 2) corresponding to $\hat{V}_{k_1k_2} = Cov(e_{k_1}, e_{k_2})$, and $\hat{V}_{k_1k_2} \hat{V}_{k_1}^{-1} \left(Y_{k_1} - \hat{E}(Y_{k_1}) \right)$ is the corrector taking covariance into account. Another version of $\hat{E}(Y_{k_u})$ will be $\hat{E}(Y_{k_u}) = f\left(A_k \hat{\beta}, X_{k_u} \right)$ (u=1 or 2) corresponding to $\hat{V}_{k_1k_2} = Z_{k_2} D Z_{k_1}^T + Cov(e_{k_1}, e_{k_2})$.

As equation 4.22 shows, $\hat{Y}_k = f\left(A_k \hat{\beta} + B_k \hat{b}_k, X_k \right)$ can be rewritten in the form of a weighted average combining information from the ith individual only and information from the population. Equation 4.22 explains why it is often said that EBLUP estimator shrinks the individual predicted response towards the population-averaged mean response profile. The

weighting scheme is quite reasonable, since more weight should be given to the individual observations if within individual variation is relatively small, whereas less weight should be given to the individual observations when the population is relatively homogenous. Equation 4.22 may be viewed as " borrowing strength" across individuals to get the best prediction for *i*th individual. The amount of shrinkage towards population mean depends also on the number of observations. In general, there is more shrinkage toward the population mean curve when n_i is small. This is reasonable since less weight should be given to observed responses when fewer data are available. So far, it is clear that EBLUP is the best approach to predicting FIA sample plots as shown by equation 4.22 and following derivation.

$$\hat{Y}_k = \hat{R}\left(\hat{Z}_k \hat{D} \hat{Z}_k{}^T + \hat{R}\right)^{-1} f\left(A_k \hat{\beta}, X_k\right) + \left(I_{n_k} - \hat{R}\left(\hat{Z}_k \hat{D} \hat{Z}_k{}^T + \hat{R}\right)^{-1} \right) Y_k \tag{4.22}$$

$$Y_k \approx f\left(A_k \beta, X_k\right) + Z_k b_k + e_k ,$$

$$\hat{Y}_k \approx f\left(A_k \hat{\beta}, X_k\right) + Z_k \hat{b}_k$$

$$\hat{Y}_k = f\left(A_k \hat{\beta}, X_k\right) + Z_k \hat{D} \hat{Z}_k{}^T \left(\hat{Z}_k \hat{D} \hat{Z}_k{}^T + \hat{R}_k\right)^{-1} \left(Y_k - f\left(A_k \hat{\beta}, X_k\right)\right)$$

$$\hat{Y}_k = \left(I_{n_k} - Z_k \hat{D} \hat{Z}_k{}^T \left(\hat{Z}_k \hat{D} \hat{Z}_k{}^T + \hat{R}_k\right)^{-1} \right) f\left(A_k \hat{\beta}, X_k\right) + Z_k \hat{D} \hat{Z}_k{}^T \left(\hat{Z}_k \hat{D} \hat{Z}_k{}^T + \hat{R}_k\right)^{-1} Y_k$$

$$\hat{Y}_k = \underbrace{\hat{R}\left(\hat{Z}_k \hat{D} \hat{Z}_k{}^T + \hat{R}_k\right)^{-1}}_{Weight} \underbrace{f\left(A_k \hat{\beta}, X_k\right)}_{Population\ \ Average} + \left(I_{n_k} - \hat{R}\left(\hat{Z}_k \hat{D} \hat{Z}_k{}^T + \hat{R}_k\right)^{-1} \right) \underbrace{Y_k}_{Individual\ \ Obs}$$

Since

$$\left(I_{n_k} - Z_k \hat{D} \hat{Z}_k{}^T \left(\hat{Z}_k \hat{D} \hat{Z}_k{}^T + \hat{R}_k\right)^{-1} \right)$$

$$= \left(\left(\hat{Z}_k \hat{D} \hat{Z}_k{}^T + \hat{R} \right) \left(\hat{Z}_k \hat{D} \hat{Z}_k{}^T + \hat{R}_k \right)^{-1} - Z_k \hat{D} \hat{Z}_k{}^T \left(\hat{Z}_k \hat{D} \hat{Z}_k{}^T + \hat{R}_k \right)^{-1} \right)$$

$$= \hat{R}_k \left(\hat{Z}_k \hat{D} \hat{Z}_k{}^T + \hat{R}_k \right)^{-1}$$

4.4 Conclusion and Discussion

In this chapter, I reviewed methodologies to update SAFIS sample plots and focused on model-based updating. Compared to the matching imputation method, model-based updating provides information about trends of forest resources, which is another primary objective of forest inventory. Obviously, as SAFIS proceeds, accumulated data will greatly facilitate and justify the applicability of model-based updating.

Our model system is based on SAFIS data measured in 1997 and following annual inventories for Georgia. The model system is to be used in conjunction with other models such as disturbance probability models and imputation models for other species or species groups. Statistics of fit show it is appropriate for one through four year updates. However, these models most likely will provide less accurate information about trends of forest resources. All models in our model system are projection models due to the availability of data. Data with only two remeasurements is not sufficient to fit more appropriate statistical models. One characteristic of projection models is that they use only one coefficient parameter to account for variation among sample plots. Obviously, one coefficient parameter is insufficient to account for variation in SAFIS plots, each of which represents about 6,000-acre forestland. Accurate projection requires

more information about individual plots. This will be satisfied as more annual inventory data accumulates.

When more SAFIS data becomes available, more appropriate models should be fitted for more accurate projections. Mixed effects models and the empirical best linear unbiased predictor (EBLUP) should be employed to update SAFIS plots because they allow more random coefficients to account for variation among plots, make efficient use of prior observations, and borrow information from similar plots.

It should be noted that model-based updating applies only to well established and undisturbed dominant species plots in conjunction with disturbance probability models. Imputation methods are also required for species or species groups that take only a small proportion. Updating SAFIS plots is the most challenging due to the involves of almost all aspects of forest biometrics and definitely requires collaborative research work of multiple partners.

Table 4.1 Number of sample plots by physiographic region, species, and stand origin

	Loblolly			Slash			Total
Region	Natural	Artificial	Subtotal	Natural	Artificial	Subtotal	
UCP	83	219	302	41	88	129	431
LCP	21	65	86	49	123	172	258
Piedmont	243	182	425	-	-	-	425
Subtotal	347	466	813	90	211	301	

Table 4.2 Nonlinear OLS parameter estimates and statistics of fit for dominant height model (Model 4.2)

		Parameter	Estimate	StdE	t-Value	Pr > |t|	RMSE	R^2
$\hat{\phi}_1$	Intercept	$\hat{\beta}_{10}$	90.40703	2.3754	38.06	<.0001	3.9560	0.9320
	Plantation	$\hat{\beta}_{12}$	-6.30645	2.9099	-2.17	0.0305		
$\hat{\phi}_2$	Intercept	$\hat{\beta}_{20}$	0.058787	0.00475	12.37	<.0001		
	Plantation	$\hat{\beta}_{22}$	0.021746	0.00615	3.54	0.0004		

Table 4.3 Parameter estimates for dominant height projection model by stands origin.

Natural Stands		Artificial Stands	
$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_1$	$\hat{\phi}_2$
90.40703	0.058787	84.10058	0.08053

Table 4.4 Nonlinear OLS parameter estimates and statistics of fit for volume projection model (Equation 4.3)

$\hat{\phi}_k$		Parameter	Estimate	StdE	t-Value	Pr >|t|	RMSE	R^2
$\hat{\phi}_1$	Plantation	$\hat{\beta}_{12}$	0.376501	0.0232	16.25	<.0001	0.1435	0.9755
	Intercept	$\hat{\beta}_{20}$	1.689769	0.0431	39.17	<.0001		
$\hat{\phi}_2$	Loblolly	$\hat{\beta}_{21}$	-0.1837	0.0336	-5.46	<.0001		
	Plantation	$\hat{\beta}_{22}$	0.272136	0.0398	6.84	<.0001		
	Intercept	$\hat{\beta}_{30}$	-0.60069	0.0456	-13.16	<.0001		
$\hat{\phi}_3$	Loblolly	$\hat{\beta}_{31}$	0.16233	0.0351	4.63	<.0001		
	Plantation	$\hat{\beta}_{32}$	-0.33448	0.0412	-8.12	<.0001		

Table 4.5 Parameter estimates for volume projection model by species and stands origins.

Species	Origin	Estimates		
		$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$
Loblolly	N	0	1.50607	-0.43836
	A	0.376501	1.77825	-0.77284
Slash	N	0	1.68977	-0.60069
	A	0.376501	1.96191	-0.93517

Table 4.6. Nonlinear OLS parameter estimates and statistics of fit for survival projection model (Equation 4.5)

$\hat{\phi}_k$		Parameter	Estimate	StdE	t Value	Pr > |t|	RMSE	R^2
	Intercept	$\hat{\beta}_{10}$	0.07209	0.0358	2.01	0.0446		
$\hat{\phi}_1$	Plantation	$\hat{\beta}_{12}$	-0.43331	0.0602	-7.20	<.0001		
	UCP	$\hat{\beta}_{14}$	-0.20791	0.0492	-4.22	<.0001	0.2120	0.9249
$\hat{\phi}_2$	Intercept	$\hat{\beta}_{20}$	0.78614	0.0295	26.66	<.0001		
	Plantation	$\hat{\beta}_{21}$	0.11158	0.0427	2.61	0.0091		
$\hat{\phi}_3$	Plantation	$\hat{\beta}_{32}$	-0.63767	0.1254	-5.08	<.0001		

Table 4.7. Parameter estimates for survival projection model by physiographic regions and stands origins.

Physiographic Region	Origin	Estimates		
		$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$
UCP	N	-0.13581	0.78614	0
	A	-0.56912	0.89772	-0.63767
Others	N	0.07209	0.78614	0
	A	-0.36122	0.89772	-0.63767

Table 4.8. Nonlinear OLS parameter estimates and statistics of fit for basal area projection model (Equation 4.6)

$\hat{\phi}_k$		Parameter	Estimate	StdE	t Value	Pr > |t|	RMSE	R^2
$\hat{\phi}_1$	Intercept	$\hat{\beta}_{10}$	2.666299	0.1714	15.55	<.0001	0.3094	0.8327
	Plantation	$\hat{\beta}_{12}$	2.098867	0.5567	3.77	0.0002		
$\hat{\phi}_2$	Intercept	$\hat{\beta}_{20}$	0.643684	0.0754	8.54	<.0001		
	Plantation	$\hat{\beta}_{22}$	-0.35897	0.0886	-4.05	<.0001		

Table 4.9. Parameter estimates for basal area projection model by stands origins.

Natural Stands		Artificial Stands	
$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_1$	$\hat{\phi}_2$
2.666299	0.643684	4.76516	0.28471

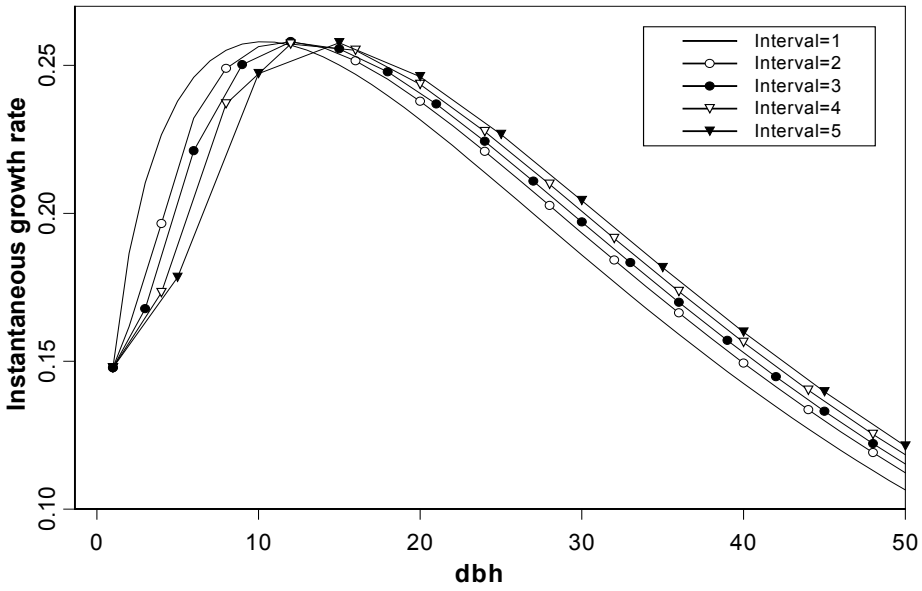

Figure 4.1. Simulation results showing that different dbh values generate difference instantaneous dbh growth rate curves.

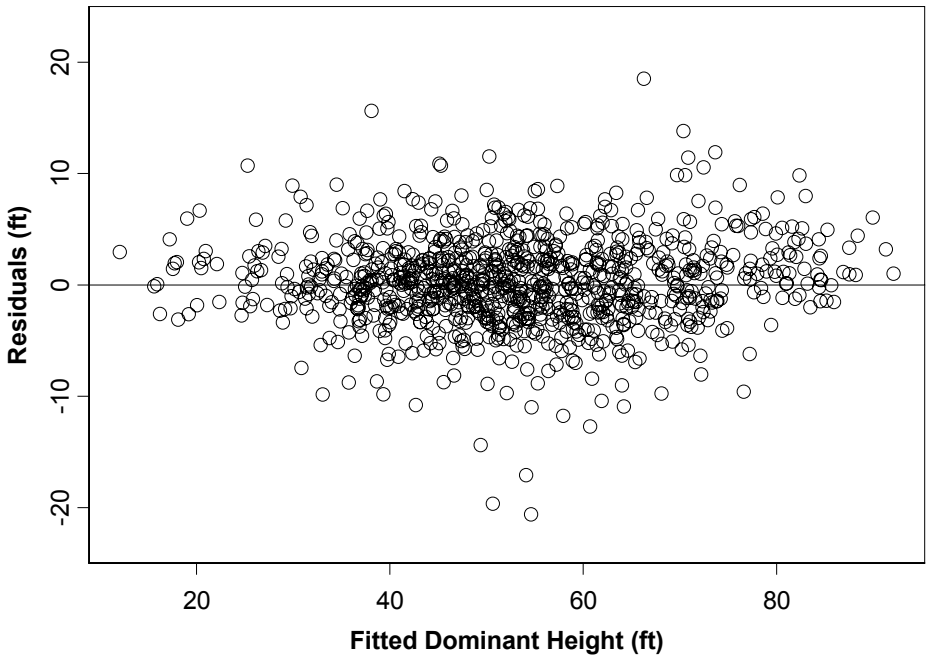

Figure 4.2. Plot of residuals vs. fitted dominant heights, showing that no significant systematic pattern exists

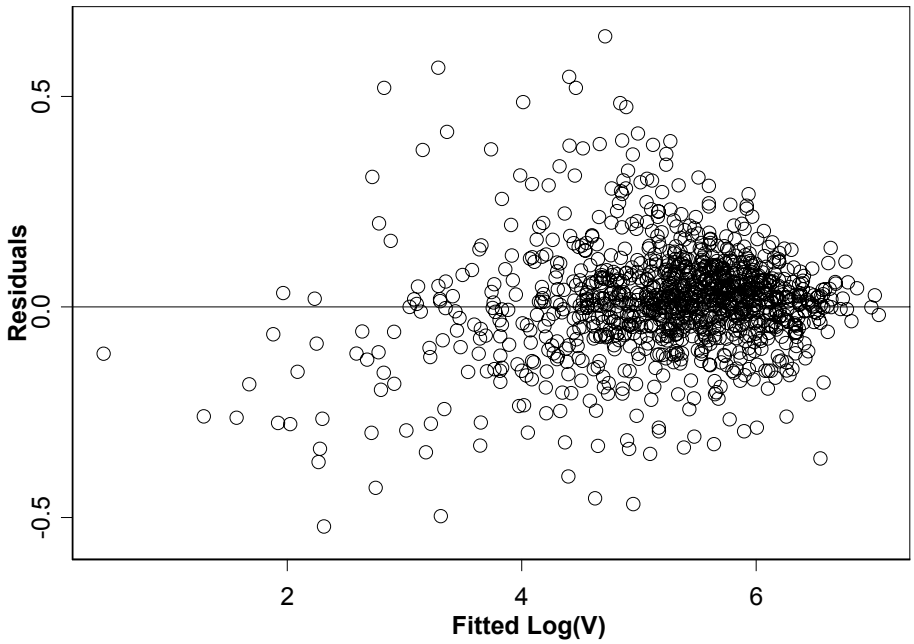

Figure 4.3. Plot of residuals vs. fitted log (volume)

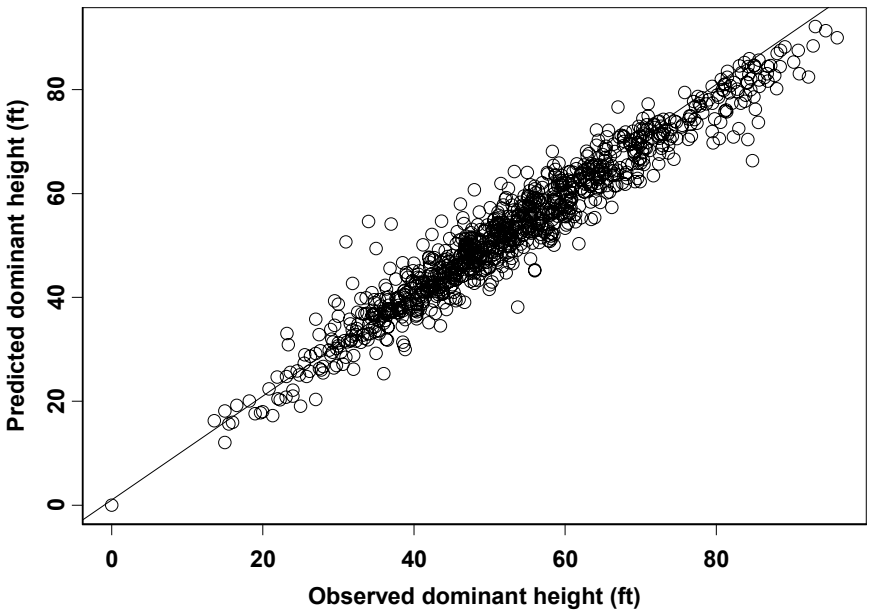

Figure 4.4. Plot of predicted vs. observed dominant height

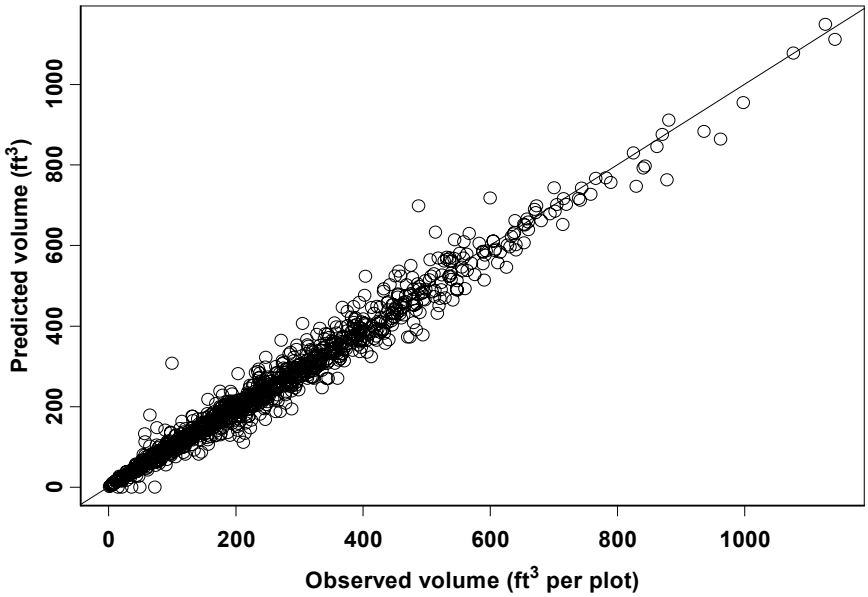

Figure 4.5. Plot of predicted vs. observed volume per plot

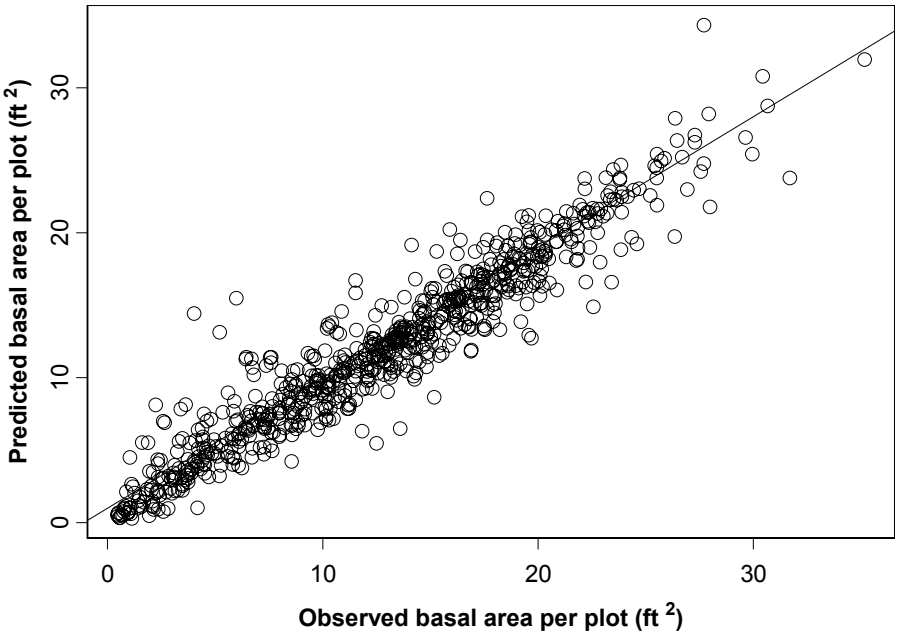

Figure 4.6. Plot of predicted vs. observed basal are per plot

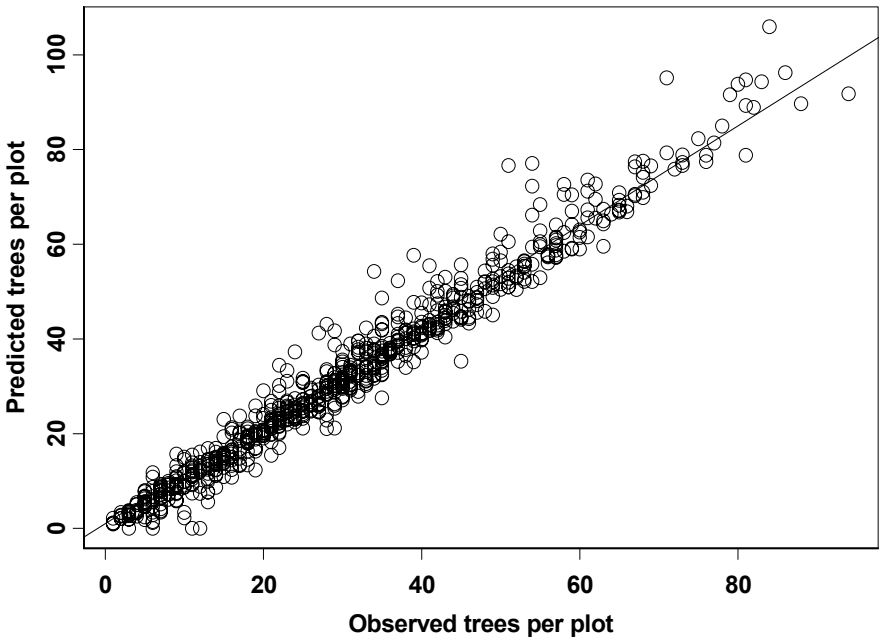

Figure 4.7. Plot of predicted vs. observed trees per plot

CHAPTER 5

SUMMARY AND CONCLUSION

Data for forest growth and yield modeling is longitudinal data. Two sources of variation are often presented in growth and yields models. One source is within observational units and the other is among observational units. It is apparent that mixed effects models provide a powerful and flexible tool for the analysis of longitudinal data.

Projection models are employed to account for the variation among individuals in forestry. A projection model can be viewed as a combination of a mixed model with a single random coefficient and an estimator for it. Mixed effects models are not always appropriate for the analysis of forest data since no sufficient prior information is available to predict random effects for a given individual. Our studies distinguish two situations, one of which is where there is only one prior observation available for predictions, and the other is where there are multiple observations. The latter situation is exemplified by SAFIS sample plot update.

Through our studies of projection models, which have been most widely applied in forest biometrics since 1963, I conclude that projection models are very useful as long as no multiple observations can be afforded for projecting forest inventory. Projection models can

behave consistently only if their model forms have properties such as reference invariance or path invariance. A simple method to justify whether or not a projection model is consistent is proposed. Theoretical and empirical analyses suggest that projection models should be estimated with maximum likelihood or generalized least squares. Comparison of EBLUP and the estimation method implied by projection models for the random coefficient indicates that EBLUP is not always superior to the other in the case where only one prior observation is available for predictions.

Studies were also given to stand table projection stand models, which are an indispensable component of a model system, providing detailed information on the underlying stands for management. A new stand table model was developed for the situation where one initial stand tables are available. Empirical studies support the conclusion that the new model outperforms the well-known Pienaar-Harrison model.

Extensive simulations were performed to compare the traditional estimation method for percentile growth model. Simulation results show that quantile regression gives more accurate estimates for percentile growth models in terms of the first order and second order statistics.

The southern annual forest inventory system for Georgia was initiated in 1998. The new FIA system requires that sample plots should be updated to current condition for annual

statistics. Since multiple observations are becoming available, mixed effects models are the most appropriate means for updating sample plots. Due to a lack of multiple observations at the present stage, a projection model system was built for Georgia. The projection error analyses show that projection models are accurate for short projection intervals, but not for long intervals. Accordingly, when over three measurements are obtained, a model system should be constructed using mixed effects models for more accurate projections.

REFERENCES

1. Alerich, C. L., L. Klevgard, C. Liff, and P. D. Miles. The forest inventory and analysis database: Database description and users guide Version 1.7

2. Avery, T. E., and H. E. Burkhart. 1994. Forest Measurements. McGraw-Hill, New York.

3. Bailey, R. L. and C. J. Cieszewski. 2000. Development of a well-behaved site-index equation: jack pine in north-central Ontario: comments. Can. J. For. Res. 30: 1167-1168.

4. Bailey, R. L. 1980. Individual tree growth derived from diameter distribution models. For. Sci. 26(4); 626-632.

5. Bailey, R. L., and N. C. Abernathy, and E. P. Jones, Jr. 1981. Diameter distribution models for repeatedly thinned slash pine plantations. USDA For. Serv. Tech. Rep. SO-34

6. Bailey, R. L. and J. L. Clutter.1974. Base-age invariant polymorphic site curves. For. Sci. 20(2):155-159.

7. Bailey, R. L., and T. R. Dell. 1973. Quantifying diameter distributions with the weibull function. For. Sci. 19(2): 97-104.

8. Bassett, C. and R. Koenker. 1982. An empirical quantile function for linear models with iid errors. JASA. 37(378): 407-415.

9. Belcher, D.W., M. R. Holdaway, and G. J. Brand, 1982. A description of STEMS, the stand and tree evaluation and modeling system. USDA For. Serv. Gen. Tech. Rep. NC-279, 19p.

10. Biging, G. S. 1985. Improved estimates of site index curves using a varying parameter model. For. Sci. 31: 248-259.

11. Brand, G.J., M.D. Nelson, D. G. Wendt, and K.K. Nimerfro, 1999, Proceedings of the first annual forest inventory and analysis symposium, San Antonio, Texas, 1999.

12. Bredenkamp, B. V., and Gregoire, T. G. 1988. A forestry application of Schnute's generalized growth function. For. Sci. 34(3): 790-797.

13. Borders, B. E., and R. L. Bailey. 2001. Loblolly pine-pushing the limits of growth. SJAF 25(2): 69-74.

14. Borders, B. E. 1997. Natural Slash pine and Loblolly pine projection models for Timber Supply Projection algorithms. SOFAC report, 1997, The University of Georgia.

15. Borders, B. E. and W. D. Patterson. 1990. Projecting stand tables: a comparison of the weibull diameter distribution method, a percentile-based projection method, and a basal area growth projection method. For. Sci. 36(2): 413-424.

16. Borders, B. E., R. A. Souter, R. L. Bailey, and K. D. Ware. 1987. Percentile-based distributions characterize forest stand tables. For. Sci. 33(3): 570-576.

17. Borders, B. E., R. L. Bailey, and K. D. Ware. 1984. Slash pine site index from a polymorphic model by joining (splining) non-polynomial segments with an algebraic difference method. For. Sci. 30: 411-423.

18. Buchinsky, M. 1998. Recent advances in quantile regression models: A practical guideline for empirical research. J. Human Res. 33: 88-126

19. Cade, B. S. and B. R. Noon. 2003. A gentle introduction to quantile regression for ecologies. Front Ecol. Environ. 1(8): 412-420.

20. Calegario, N., and R. F. Daniels, R. Maestri, and R. Neiva. 2004. Modeling dominant height growth based on nonlinear mixed-effects model: a clonal Eucalyptus plantation case study. For. Ecol. Manage. 204: 11-20.

21. Cade, B. S. 2003. Quantile regression models of animal habitat relationships (PhD dissertation). Fort Collins, CO: Colorado State University.

22. Cao, Q. V. and V. C. Jr. Baldwin. 1999. A new algorithm for stand table projection models. For. Sci. 45(4): 506-511.

23. Cieszewski, C. J. 2002. Comparing fixed- and variable-base-age site equation having single versus multiple asymptotes. For. Sci. 48(1): 7-28.

24. Cieszewski, C. J. 2001. Three methods of deriving advanced dynamic site equations demonstrated on inland Douglas-fir site curves. Can. J. For. Res. 31:165-173.

25. Cieszewski, C. J., M. Harrison, and S. T. Martin. 2000. Practical methods for estimating non-biased parameters in self-referencing growth and yield models. PMRC technical report, February 22, 2000. The University of Georgia.

26. Cieszewski, C. J., and R. L. Bailey. 1999. Generalized algebraic difference approach: Theory based derivation of dynamic site equations with polymorphism and variable asymptotes. For. Sci. 46(1): 116-126.

27. Cieszewski, C. J., and I.E. Bella. 1989. Polymorphic height and site index curves for lodgepole pine in Alberta. Can. J. For. Res. 19: 1151-1160.

28. Clutter, J. L. 1963. Compatible growth and yield models for loblolly pine. For. Sci. 9: 354-371.

29. Clutter, J.L., J. C. Fortson, L.V. Pienaar, G.H. Brister, and R.L. Bailey. Timber management: A quantitative approach. Wiley, New York, 1983.

30. Clutter, J. L., and E. P. Jones. 1980. Prediction of growth after thinning old-field slash pine plantations. USDA For. Serv. Res. Pap. SE-217. 14pp.

31. Daniels, R. F., and H. E. Burkhart. 1988. An integrated system of forest stand models. For. Ecol. Manage. 23: 159-177.

32. Davidian, M., and D. M. Giltinan 1995. Nonlinear models for repeated measurement data. London: Chapman & Hall.

33. Demidenko, E. 2004. Mixed models: theory and application, Wiley.

34. Dhote, J. F. 1994. Hypothesis about competition for light and water in even-aged common beech. For. Ecol. Manage. 69: 219-232.

35. Dyer, M. E. 1997. Dominance/Suppression competitive relationships in loblolly pine (Pinus taeda L.) plantations. Virginia Tech dissertation, 1997.

36. Eerikainen, K. and M. Maltamo. 2003. A percentile based basal area diameter distribution model for predicting the stand development of Pinus Kesiya plantations in Zambia and Zimbabwe. For. Ecol. Manage. 172: 109-124.

37. Fang, Z. and R.L. Bailey. 2001. Nonlinear mixed effects modeling for slash pine dominant height growth following intensive silvicultural treatments. For. Sci. 47(3) 287-300.

38. Fang, Z., R. L. Bailey, and B. D. Shiver. 2001. A multivariate simultaneous prediction system for stand growth and yield with fixed and random effects. For. Sci. 47(4): 550-562.

39. Fang, Z. 1999, A simultaneous system of linear and nonlinear mixed effects models for forest growth and yield prediction. The University of Georgia, PhD dissertation.

40. Fitzmaurice, G. M. 2004. Applied longitudinal analysis. John Wiley & sons, Inc.

41. Gartner, D. and G. Reams. 2001. A comparison of several techniques for imputing tree level data. Proceedings of the third annual forest inventory and analysis symposium, Traverse city, Michigan, October 17-19, 2001.

42. Goelz, J.C.G. 2001. A diameter distribution perspective on forest growth and yield models. Math. Model. Sci. Comp. 13(34): 177-189.

43. Hafley, W. L. and H. T. Schreuder. 1977. Statistical distributions for fitting diameter and height data in even-age stands. Can. J. For. Res. 7: 481-459.

44. Hall, D. B. and M. Clutter. 2004. Multivariate multilevel nonlinear mixed effects models for timber yield predictions. Biometrics 60:16-24.

45. Hall, D.B. and R. L. Bailey. 2001. Modeling and Prediction of Forest Growth Variables based on multilevel nonlinear mixed models. For. Sci. 47(3): 311-321.

46. Harrison, W. M., and B. E. Borders. Yield prediction and growth projection for site-prepared loblolly pine plantations in the Carolinas, Georgia, Alabama, and Florida. PMRC technical reports, The university of Georgia, 1996.

47. Johnson, D.S., M.S. Williams, and R.L. Czaplewski. 2003. Comparison of estimators for

rolling samples using forest inventory and analysis data. For. Sci. 49(1): 50-63.

48. Kangas, A., and M. Maltamo. 2000. Percentile based basal area diameter distribution models for Scot pine, Norway spruce and Birch Species. Silva Fennica, 34(4): 371-380.

49. Knowe, S. A., G. R. Ahrens, and D. S. DeBell. 1997. Comparison of diameter-distribution-prediction, stand-table-projection, and individual-tree-growth modeling approaches for you red alder plantations. For. Ecol. Manage. 98: 49-60.

50. Knowe, S.A. and D. E. Hibbs. 1996. Stand structure and dynamics of young red alder as affected by planting density. For. Ecol. Manage. 82: 69-85.

51. Knowe, S. A. 1994. Incorporating the effects of inter-specific competition and vegetation management treatments in stand table projection model for Douglas-fir saplings. For. Ecol. Manage. 67: 87-99.

52. Koenker, R. and G. Bassett. 1978. Regression quantiles. Econometrica 46: 107-112.

53. Kudus, K. B., M. I. Ahmad, and J. Lapongan. 1999. Nonlinear regression approach to estimating Johnson SB parameters for diameter data. Can. J. For. Res. 29: 310-314.

54. Lappi, J., and R. L. Bailey. 1988. A height prediction model with random stand and tree parameters: An alternative to traditional site index methods. For. Sci. 34(4): 907-927.

55. Lessard, V.C., R. E. McRoberts, and M.R. Holdaway. 2001. Diameter growth models using Minnesota forest inventory and analysis data. For. Sci. 47(3): 301-310.

56. Lessard, V.C., R. E. McRoberts, and M.R. Holdaway. 1999. Diameter growth models using FIA data from the Northeastern, Southern, and North Central Research stations. Proceedings of the first annual forest inventory and analysis symposium. San Antonio, Texas, 1999.

57. Liu, C., L. Zhang, C.J. Davis, D.S. Solomon, and J.H. Grove. 2002. A finite mixture model for characterizing the diameter distribution of mixed-species forest stands. For. Sci. 48:653-661.

58. Lynch, T. B., K. L. Hitch, M.M. Huebschmann, and P. A. Murphy. 1999. An individual-tree

growth and yield prediction system for even-aged natural shortleaf pine forests. SJAF 23(4): 203-211.

59. Lynch, T. B., P. A. Murphy. 1995. A compatible height prediction and projection system for individual trees in natural, even-aged short leaf pine stands. For. Sci. 41(1): 194-209.

60. Maltamo, M., A. Kangas, J. Uuttera, T. Torniainen, and J. Saramaki. 2000. Comparison of percentile based prediction methods and the weibull distribution in describing the diameter distribution of heterogeneous Scots pine stands. For. Ecol. Manage. 133: 263-274.

61. McCulloch, C.E. and S. R. Searle (2001). Generalized, linear, and mixed models, Wiley.

62. McDill, M. E. and R. L. Amateis. 1992. Measuring Forest site quality using the parameters of a dimensionally compatible height growth function. For. Sci. 38(2): 409-429.

63. McRoberts, R.E. 2001. Imputation and model-based updating techniques for annual forest inventories. For. Sci. 47: 322-330

64. McRoberts, R.E. 1999. Evaluating imputation and modeling in the North central region. Proceedings of the first annual forest inventory and analysis symposium, San Antonio, Texas, 1999.

65. McTague, J.P. and W. F. Stansfield. 1994. Stand and tree dynamics of uneven-aged ponderosa pine. For. Sci. 40(2): 289-302.

66. Nepal, S. K. and G. L. Somers.1992. A generalized approach to stand table projection. For. Sci. 38(1) 120-133.

67. Pienaar, L. V. and W. M. Harrison. 1988. A stand table projection approach to yield prediction in unthinned even-aged stands. For. Sci. 34(3): 804-808.

68. Pienaar, L.V., and B. D. Shiver. 1981. Survival functions for site prepared slash pine plantations in the flatwoods of Georgia and northern Florida. SJAF. 5: 59-62.

69. Pinheiro, J. C., and D. M. Bates. 2000. Mixed-effects models in S and S-plus. New York: Springer.

70. Ramirez-Maldonado, H., R. L. Bailey, and B. E. Borders, 1987. Some implications of the algebraic difference approach for developing growth models presented at the IUFRO forest growth modeling and prediction conference, Minneapolis, MN, August 24-28, 1987.

71. Ratkowsky, D.A. Handbook of nonlinear regression models 1990 Marcel Dekker, Inc.

72. Reams, G.R. and J. McCollum. 1999. Evaluating multiple imputation models for the southern annual forest inventory. Proceedings of the section on statistics and the environment of the American Statistical Association.

73. Ritchie, M. W. and D. W. Hann. 1997a. Evaluation of individual-tree and disaggregative prediction methods for Douglas-fir stands in western Oregon. Can. J. For. Res. 27: 207-216.

74. Ritchie, M. W. and D. W. Hann. 1997b. Implications of disaggregation in forest growth and yield modeling. For. Sci. 43(2): 223-233.

75. Rivas, J. J. C, J. G. A. Gonzalez, A. D. R. Gonzalez, and K. Von Gadow. 2004.Compatible height and site index models for five pine species in E1 Salto, Durango (Mexico). For. Ecol. Manage. 201: 145-160.

76. Robison, G.K. 1991. That BLUP is a good thing: the estimation of random effects. Stat. Sci. 6(1): 15-51.

77. Roesch, F. A. 1999. Mixed estimation for a forest survey sample design. Proceedings of the section on statistics and the environment of the American Statistical Association.

78. Schumacher, F.X. 1939. A new growth curve and its application to timber-yield studies. J. For. 37:819 - 820.

79. Schnute, J. 1981. A versatile growth model with statistically stable parameters. Can. J. Fish. Aquat. Sci. 38: 1128-1140.

80. Scott, C.T., M. Kohl, and H. J. Schnellbacher. 1999. A comparing of periodic and annual forest surveys. For. Sci. 45(3): 433-451.

81. Somers, G.L., S. K. Nepal. 1994. Liking individual-tree and stand-level growth models. For. Ecol. Manage. 69: 233-243.

82. Sullivan, A.D., and J. L. Clutter. 1972. A simultaneous growth and yield model for Loblolly pine. For. Sci. 18: 76-86.

83. Theil, H. 1971. Principle of econometrics. Wiley, New York.

84. Thompson, M. T., and L. W. Thompson. 2002. Georgia's forests, 1997. USDA For. Serv. SRS, Res. Bull. SRS-72.

85. Trincado, G. V., R. P. Quezada, and K. Von Gadow. 2003. A comparison of two stand table projection methods for young Eucalyptus nitens (Maiden) plantations in Chile. For. Ecol. Manage. 180: 443-451.

86. Van Deusen, P.C. 2002. Comparison of some annual forest inventory estimators. Can. J. For. Res. 32: 1992-1995.

87. Van Deusen, P.C. 2001. Issues related to panel creep. Proceedings of the third annual forest inventory and analysis symposium, Traverse city, Michigan, October 17-19, 2001.

88. Van Deusen, P.C. 1999. Modeling trends with annual survey data. Can. J. For. Res. 29: 1824-1828.

89. Van Deusen, P.C. 1997. Annual forest inventory statistical concepts with emphasis on multiple imputation. Can. J. For. Res. 27: 379-384.

90. Van Deusen, P.C. 1996. Incorporating predictions into an annual forest inventory. Can. J. For. Res. 26: 1709-1713.

91. Wycoff, W. R., N.L. Crookston, and A.R. Stange. 1982. User's guide to the stand prognosis model. USDA For. Serv. Gen. Tech. Rep. INT-133.

92. Zhang, L., J. A. Moore, and J. D. Newberry. 1993. Disaggregating stand volume growth to individual trees. For. Sci. 39(2): 295-309.

93. Zhang, Y., B. E. Borders, and R. L. Bailey. 2002. Derivation, fitting, and implication of a compatible stem taper-volume-weight system for intensively managed, fast growing loblolly pine. For. Sci. 48(3); 595-607.

94. Zhao, D, B. Borders, and M. Wilson. 2004. Individual-tree diameter growth and mortality models for bottomland mixed-species hardwood stands in the lower Mississippi alluvial valley. For. Ecol. Manage. 199: 307-322.

95. Zeide, B. 1999. Pattern of height growth for southern pine species. For. Ecol. Manage. 118: 183-196.

96. Zeide, B. 1993. Analysis of growth equations. For. Sci. 39(3): 594-616.

97. Zeide, B. 1989. Accuracy of equations describing diameter growth. Can. J. For. Res. 19: 1283-1286.