Predicting Stand Volume for Intensively Site Prepared Slash Pine Plantations: Levels of Prediction Precision from a Mixed-model System

Plantation Management Research Cooperative

Daniel B. Warnell School of Forest Resources
University of Georgia
Athens, Georgia 30602

PMRC TECHNICAL REPORT 2000-3

March, 2000

Compiled by:

Robert L. Bailey

Zixing Fang[1]

Executive Summary

Conventional modeling approaches based on “difference equations” use only the most recent measurement on a plot in predicting expected future yields and ignore any other measurements on that plot. This paper explains a new modeling approach that allows all the information included in an inventory, no matter how many repeat measurements exist, to be used in making predictions of expected future yields. The result is a better “localization” of predictions. Because of some particular features in this approach, namely the mixed-parameter formulation (i.e. fixed and random parameters in the same equation), valid prediction intervals (i.e. 95% confidence intervals) can be determined for the predictions.

The model was fitted with data through age 17 from the PMRC Coastal Plain Site Preparation Study. Five examples of predictions for an intensively site-prepared slash pine plantation illustrate the gains in prediction precision when using repeated measurements of dominant height in order to “localize” predictions for a given stand.

---

[1] This report is based on a part of Dr. Fang’s doctoral research that he completed while at The Daniel B. Warnell School of Forest Resources, The University of Georgia .

Introduction

For predicting future stand values, foresters often rely on a simultaneously inter-dependent model system that includes common stand characteristics such as dominant height, basal area, and total volume. Permanent plots with some specific design, such as the PMRC Coastal Plain Site Preparation Study, are usually the main source of data in developing such a model system to evaluate stand growth and yield in conjunction with different silvicultural treatments or site conditions. Two basic sources of variation in stand growth and yield models are very common for such permanent plot data with a special experimental design. One is with-in plot error and the other is the variation from plot to plot. Within-plot variation is usually modeled by a reasonable variance function that accounts for within-plot heteroscedasticity and correlation. The between plot variation can be modeled by random effects that allow certain parameters in the model to vary from plot to plot. The more available information on these sources of variation, the more precise the predictions of the stand characteristics for a new plot. The interdependency among components, such as height and basal area, within the system is another key to successful predictions. Values for observed components in the system could be used to improve predictions for unobserved components by application of the contemporaneous correlation among the components.

We used the data for ages 5, 8, 11, 14, and 17 from the PMRC Coastal Plain Site Preparation Study (Pienaar *et al*. 1998). The prediction models are a simultaneous system containing component models for dominant height growth, basal area, and total volume for slash pine (*Pinus elliottii* Engelm.) plantations with different silvicultural treatments. The system models and error structure account for both contemporaneous correlation and random effects by using a mixed model approach. The system allows prediction (confidence) intervals to be determined for yield predictions and projections. Thus, gains in the precision of the prediction for new cases can be evaluated with detailed examples for distinct situations.

Dominant height is the fundamental component in this three-component system. With contemporaneous correlation and random effects for dominant height considered, including random effects in the basal area and total volume models proved to be unnecessary. Between-plot variation (stand-to-stand) is the main source of error for dominant height prediction, and dominant height itself contributes greatly to the precision of basal area prediction. Basal area is the main source of error for total volume prediction.

The Model System

Extensive details of the model's development are reported in Fang (1999) and in three journal article manuscripts (Fang and Bailey 2000a, Fang and Bailey 2000b, Fang and Bailey 2000c). The system consists of models for dominant height, basal area, and stand volume:

$$\left.\begin{aligned}
H &= \left(b_{1,0} + b_{1,1}zburn + b_{1,2}zf\right) \\
&\quad \times \left(\frac{1-e^{-\left(b_{1,3}+b_{1,4}zf+b_{1,5}zh\right)t}}{1-e^{-\left(b_{1,3}+b_{1,4}zf+b_{1,5}zh\right)t_0}}\right)^{\left(b_{1,6}+b_{1,7}zf+b_{1,8}zbed+b_{1,9}zh\right)} + e_H \\
\ln[B] &= b_{2,0} + b_{2,1}zf + b_{2,2}zburn \times zsoil + b_{2,3}zchop \times t \\
&\quad + \left(b_{2,4} + b_{2,5}zh + b_{2,6}zburn + b_{2,7}zchop\right)/t \\
&\quad + \left(b_{2,8} + b_{2,9}zchop\right)\ln(H) + \left(b_{2,10} + b_{2,11}zbed\right)\ln(T) + e_B \\
\ln[V] &= b_{3,0} + b_{3,1}\ln(H) + b_{3,2}\ln(B) + b_{3,3}zchop \\
&\quad + b_{3,4}zchop \times t + b_{3,5}zh \times zsoil + e_V
\end{aligned}\right\} \qquad (1)$$

where $b_{1,0}$ through $b_{3,5}$ are the parameters to be estimated, *zchop*, *zburn*, *zf*, *zbed* and *zh* are respective dummy variables indicating if chop, burn, fertilizer, bed or herbicide treatments were applied on a specified treatment plot (=1 for yes, =0 if no). To identify soils, nonspodosols have *zsoil*=1 and spodosols have *zsoil*=0. The main variables are as follows:

H = average height of dominants and codominants (m)

B = basal area (m^2/ha)

T = trees/ha

V = total stem outside-bark volume (m^3/ha)

t = plantation age (years) at which H, B, T, and V occur

t_0 = a reference age (25 years in this work)

System (1) is a modified version of a common simultaneous growth and yield system in forestry in that three components (dominant height, basal area and volume) are the focus with silvilcultural treatments and soil groups included as exogenous variables. Endogenous variables *H* and *B* are also used as predictors in the system and they are usually correlated with the response on the left-hand side of the equation. For example, dominant height is correlated with basal area and both dominant height and basal area are correlated with volume.

To avoid the “simultaneous bias” in estimation, various parameter estimation methods (such as three-stage least squares, full information likelihood, the moment of methods) have been recommended in the forestry and econometrics literature (Amemiya 1985, Borders 1989, LeMay 1990).

We believe that a mixed-effects model will better explain the variations in forest growth and thus be more efficient in prediction. Based on results from earlier work, and a comparison of model performance on the PMRC Coastal Plain Site Prep Study data (likelihood ratio test, *AIC* criterion), a first step was to treat the “intercept” terms in the three components as mixed effects. These “intercept” parameters are $b_{1,0}$, $b_{1,3}$, and $b_{1,6}$ in the dominant height equation, $b_{2,0}$ in the basal area equation and $b_{3,0}$ in the total volume equation. These were assumed to be composed of two parts, a “fixed” or non-varying population average part and another part that varies randomly from plot to plot. That is, in (1),

$$b_{1,0,i} = b_{1,0} + b_{1,0,i} \qquad (2a)$$

$$b_{1,3,i} = b_{1,3} + b_{1,3,i} \qquad (2b)$$

$$b_{1,6,i} = b_{1,6} + b_{1,6,i} \qquad (2c)$$

$$b_{2,0,i} = b_{2,0} + b_{2,0,i} \qquad (2d)$$

$$b_{3,0,i} = b_{3,0} + b_{3,0,i} \qquad (2e)$$

In (2), the b_{ijk} are the fixed or non-varying parts and the $\boldsymbol{b}_{ijk}$ are random effects with an expected average of zero over the entire population of possible plots and with variances and covariances to be estimated.

The variance covariance structure for the random effects of the domain height model was assumed to be positive and symmetric. The random intercepts of the basal area model and the volume model are correlated, but both are independent of the random effects of dominant height.

That is,

$$\begin{pmatrix} b_{1,0,i} \\ b_{1,3,i} \\ b_{1,6,i} \end{pmatrix} \sim N(\mathbf{0}, \mathbf{D}_H), \quad \text{where} \quad \mathbf{D}_H = \begin{pmatrix} j_{1,11} & j_{1,12} & j_{1,13} \\ j_{1,12} & j_{1,22} & j_{1,23} \\ j_{1,13} & j_{1,23} & j_{1,33} \end{pmatrix}$$

$$\begin{pmatrix} b_{2,0,i} \\ b_{3,0,i} \end{pmatrix} \sim N(\mathbf{0}, \mathbf{D}_{BV}), \quad \text{where} \quad \mathbf{D}_{BV} = \begin{pmatrix} j_{2,11} & j_{2,12} \\ j_{2,12} & j_{2,22} \end{pmatrix}$$

The endogenous variables that appear on the right-hand-side (RHS) of an equation were assumed to be independent of the error term of the left-hand-side (LHS) responses for given values of the random effects. For example, the term "ln(H)" is no longer to be correlated with the error term of the basal area equation given the random effects. Thus the "simultaneous error" problem is overcome by including random effects in the model and the three-component simultaneous equation system is specified as a seemingly unrelated equation system (Srivastava and Giles 1987). Given the random effects, however, the error terms e_H, e_B, and e_V were assumed to be correlated with each other. These are the so-called contemporaneous correlations in a simultaneous or seemingly unrelated equation system (Borders and Bailey 1986, LeMay 1990). For simplicity, errors within plot for the same response are assumed to be independent. The variances of the models are assumed to be proportional to a power of the mean response and the proportionality parameters of the power are distinct for the different responses. Let $e_{H_{ij}}$, $e_{B_{ij}}$, and $e_{V_{ij}}$ denote the error of model dominant height, basal area and volume for plot *i* at occasion *j*. Then,

$$\begin{pmatrix} e_{H_{ij}} \\ e_{B_{ij}} \\ e_{V_{ij}} \end{pmatrix} \Big| \mathbf{b}_i \sim N(\mathbf{0}, \mathbf{R}_i)\,, \qquad (3)$$

where $\mathbf{R}_i = s^2 G_i^{1/2}(b_i,,q)\Gamma_i(r)G_i^{1/2}(b_i,q)$ (4)

$$G_i(b_i,q) = diag(f_{HD_{ij}}^{q_1}, f_{BA_{ij}}^{q_2}, f_{V_{ij}}^{q_3})\;, \qquad (5)$$

which is the variance function for the three models (Fang and Bailey 2000a); and

$$\Gamma_i(r) = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{pmatrix} \qquad (6)$$

which is the contemporaneous correlation matrix.

To implement the above assumptions in our model formulation, we define the response vector as a general form, in which three vector components [*H*, *ln(B)* and *ln(V)*] are combined into one long vector and the components are identified by the function factor: Level 1 of the function factor *H*, Level 2 indicates *ln(BA)*, and Level 3 indicates *ln(V)*. Different levels have different mean functions and error variances. In fitting the model system we mainly relied on *NLME*, an S-plus library by Bates and Pinheiro (1999).

Model Fitting

Using the PMRC Coastal Plain Site Preparation Study data for slash pine, the Restricted Maximum Likelihood (REML) estimates (Bates and Pinheiro 1999) for the above covariance parameters for the random effects are as follows.

$$\hat{\mathbf{D}}_H = \operatorname{cov}\left(\begin{pmatrix} b_{1,0,i} \\ b_{1,3,i} \\ b_{1,6,i} \end{pmatrix} \begin{pmatrix} b_{1,0,i} & b_{1,3,i} & b_{1,6,i} \end{pmatrix} \right) = \begin{pmatrix} 4.92350063\,5 & -0.00196176\,707 & 0.15775075\,30 \\ -0.00196176\,7 & 0.00006827\,368 & -0.00045761\,25 \\ 0.15775075\,3 & -0.00045761\,250 & 0.00753530\,92 \end{pmatrix}$$

$$\hat{\mathbf{D}}_{BV} = \operatorname{cov}\left(\begin{pmatrix} b_{2,0,i} \\ b_{3,0,i} \end{pmatrix} \begin{pmatrix} b_{2,0,i} & b_{3,0,i} \end{pmatrix} \right) = \begin{pmatrix} 0.00016101\,623 & 0.00001092\,752 \\ 0.00001092\,752 & 0.00001245\,645 \end{pmatrix}$$

The relatively small variance-covariance of the random effects in basal area and volume indicates that the random components in the "intercept" terms for those models have only a small effect and are essentially negligible. So, the random effects in the basal area and volume models were dropped thus reducing the system to one where only the three "intercept" parameters in dominant height are taken as random. A likelihood ratio test supported such a model reduction. The log likelihood values of the two model systems are almost the same (3032.141 vs. 3032.143), but the full system has three more parameters. So, preference should be given to the reduced model.

It is interesting that random effects in basal area and volume models are not important any more as long as the parameters in the dominant height model are considered to be random and contemporaneous correlation among the three components is considered. Therefore, in the final model, only the three parameters in the dominant height model are taken to be mixed and the errors of the components in the system are correlated to each other, with variance proportional to the power function of its mean. Since the mean function and the power parameters are different for different components, the variances for the three components are different from each other. After considering the random effects and contemporaneous correlation, the term *zbed´ln(T)* was no longer significant (p-value = 0.3599), so it was dropped from the system. This produced a log-likelihood value of 3031.799. Again, a likelihood ratio test supported the change. Therefore, the final model system is as follows:

$$
\left.
\begin{aligned}
y_{1,i}(t_k) &= H_i(t_k) \\
&= \left(\beta_{1,0} + \beta_{1,i} + \beta_{1,1} zburn + \beta_{1,2} zf\right) \\
&\quad \times \left(\frac{1 - e^{-\left(\beta_{1,3}+\beta_{2,i}+\beta_{1,4}zf+\beta_{1,5}zh\right)t_k}}{1 - e^{-\left(\beta_{1,3}+\beta_{2,i}+\beta_{1,4}zf+\beta_{1,5}zh\right)t_0}}\right)^{\left(\beta_{1,6}+\beta_{3,i}+\beta_{1,7}zf+\beta_{1,8}zbed+\beta_{1,9}zh\right)} + \varepsilon_H \\
y_{2,i}(t_k) &= \ln[B_i(t_k) + 1] \\
&= \beta_{2,0} + \beta_{2,1} zf + \beta_{2,2} zburn \times zsoil + \beta_{2,3} zchop \times t_k \\
&\quad + (\beta_{2,4} + \beta_{2,5} zh + \beta_{2,6} zburn + \beta_{2,7} zchop)/t_k \\
&\quad + (\beta_{2,8} + \beta_{2,9} zchop)\ln(H) + \beta_{2,10}\ln(T) + \varepsilon_B \\
y_{3,i}(t_k) &= \ln[V_i(t_k) + 1] \\
&= \beta_{3,0} + \beta_{3,1}\ln(H) + \beta_{3,2}\ln(B) + \beta_{3,3} zchop \\
&\quad + \beta_{3,4} zchop \times t_k + \beta_{3,5} zh \times zsoil + \varepsilon_V
\end{aligned}
\right\} \tag{7a}
$$

$$
\begin{pmatrix} \beta_{1,i} \\ \beta_{2,i} \\ \beta_{2,i} \end{pmatrix} \sim N(\mathbf{0},\mathbf{D}), \quad \text{where} \quad \mathbf{D} = \begin{pmatrix} \varphi_{1,11} & \varphi_{1,12} & \varphi_{1,13} \\ \varphi_{1,12} & \varphi_{1,22} & \varphi_{1,23} \\ \varphi_{1,13} & \varphi_{1,23} & \varphi_{1,33} \end{pmatrix} \tag{7b}
$$

$$
\left(\varepsilon_{H_i} \quad \varepsilon_{B_i} \quad \varepsilon_{V_i}\right)^T | \mathbf{b}_i \sim N(\mathbf{0},\mathbf{R}_i), \tag{7c}
$$

$$
\text{where} \quad \mathbf{R}_i = \sigma^2 \mathbf{G}_i^{1/2}(\beta_i,,\theta)\Gamma_i(\rho)\mathbf{G}_i^{1/2}(\beta_i,\theta) \tag{7d}
$$

$$
\mathbf{G}_i(\beta_i,\theta) = diag(f_{HD_{ij}}^{\theta_1}, f_{BA_{ij}}^{\theta_2}, f_{V_{ij}}^{\theta_3}) \tag{7e}
$$

$$
\Gamma_i(\rho) = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix} \tag{7f}
$$

, with i = 1, …,191, indexing individual plots and t_k= $(5,8,11,14,17)^T$ being the ages of the stand. Matrix ***R*** is the variance covariance of the three response components in the system given the random effects and ***G*** is the power variance function to explain the within-plot heteroscedasticity. Matrix Γ is the contemporaneous correlation matrix.

The estimated parameters for the revised system using REML are as follows:

$$\hat{\Delta} = \mathrm{cov}\left(\begin{pmatrix} \beta_{1,i} \\ \beta_{2,i} \\ \beta_{3,i} \end{pmatrix} \begin{pmatrix} \beta_{1,i} & \beta_{2,i} & \beta_{3,i} \end{pmatrix}\right)$$

$$= \begin{pmatrix} 4.92152728 & -0.001929953 & 0.158306766 \\ -0.00192995 & 0.000067357 & -0.000460514 \\ 0.15830677 & -0.0004605135 & 0.007591564 \end{pmatrix} \qquad (8a)$$

$$\hat{\Gamma}_i = \begin{pmatrix} 1 & 0.441 & 0.330 \\ 0.441 & 1 & 0.568 \\ 0.330 & 0.568 & 1 \end{pmatrix} \qquad (8b)$$

$$\hat{\theta}_1 = 0.4695512, \quad \hat{\theta}_2 = -0.2272813, \quad \hat{\theta}_3 = -1.228734, \quad \hat{\sigma} = 0.096753489 \qquad (8c)$$

The parameter estimates for the fixed effects are all significant (Table 1).

For the youngest age (i.e. 5), small changes in the observed values resulted in large residuals due to log transformations on small basal area and total volume. Therefore, we replaced *ln(B)* and *ln(V)* with *ln(B+1)* and *ln(V+1)* respectively in model fitting.

Table 1. -- Parameter estimates for fixed effects in the dominant height, basal area and total volume simultaneous model system

Parameter	Value	Std.Error	t-value	p-value
$b_{1,0}$	18.36704	0.3627850	50.6279	<.0001
$b_{1,1}$	1.19048	0.3713349	3.2060	0.0014
$b_{1,2}$	2.38687	0.3908644	6.1066	<.0001
$b_{1,3}$	0.08115	0.0030647	26.4791	<.0001
$b_{1,4}$	-0.02192	0.0041760	-5.2491	<.0001
$b_{1,5}$	0.01540	0.0038456	4.0040	0.0001
$b_{1,6}$	1.98367	0.0393179	50.4522	<.0001
$b_{1,7}$	-0.26349	0.0452981	-5.8168	<.0001
$b_{1,8}$	-0.16069	0.0246817	-6.5103	<.0001
$b_{1,9}$	-0.16502	0.0446935	-3.6922	0.0002
$b_{2,0}$	-4.61006	0.1547915	-29.7824	<.0001
$b_{2,1}$	0.03259	0.0101457	3.2117	0.0013
$b_{2,2}$	0.03880	0.0111913	3.4666	0.0005
$b_{2,3}$	-0.01806	0.0014064	-12.8408	<.0001
$b_{2,4}$	-0.77866	0.2881011	-2.7027	0.0069
$b_{2,5}$	1.51645	0.0827950	18.3158	<.0001
$b_{2,6}$	0.27581	0.1046136	2.6365	0.0084
$b_{2,7}$	-0.58888	0.1700667	-3.4626	0.0005
$b_{2,8}$	1.21727	0.0222883	54.6145	<.0001
$b_{2,9}$	0.14209	0.0120492	11.7922	<.0001
$b_{2,10}$	0.63554	0.0177388	35.8276	<.0001
$b_{3,0}$	-0.47398	0.0084069	-56.3803	<.0001
$b_{3,1}$	0.79417	0.0067315	117.9793	<.0001
$b_{3,2}$	1.06569	0.0044231	240.9372	<.0001
$b_{3,3}$	-0.01620	0.0054913	-2.9492	0.0032
$b_{3,4}$	0.00097	0.0003228	3.0030	0.0027
$b_{3,5}$	0.00543	0.0023153	2.3432	0.0192

Predictions

The typical conventional growth and yield system based on regression approaches would use only the coefficients displayed in Table 1, the estimates for fixed parameters. Two problems arise that are not easily overcome with such an approach. The first is that estimated variances for predicted growth and yield values at the stand level are either impossible or very difficult to obtain, especially for predictions of future stand conditions (i.e. projections). Estimates of variances and covariances related to the time series nature of the prediction problem are not an output from the usual fixed-effects only regression approach. The second problem in the conventional approach is that a prediction for a future condition only utilizes information from one prior event. In a situation where inventory data for ages 12 and 17 exists and we want a prediction of expected yield at age 20, the conventional approach only uses the age-17 inventory information. The usual regression model structure does not have the capability of extracting whatever information exists in the age-12 data and using it along with the age-17 information to predict age-20 yield. As will be illustrated later, both of these shortcomings are overcome with a mixed-model system such as we present above.

Yet another shortcoming of the conventional regression approach is the failure to “correct” volume predictions to account for deviations in observations of input variables, such as height and basal area, from “average” stand conditions. Without estimated variance and covariance information for the within and between-plot error structure from model to model, this is impossible. For example, suppose we have a stand on a non-spodosol soil (*zsoil*=1) site prepared by an application of herbicide along with burning, chopping and fertilizing (i.e. the FCBH treatment). Currently there are 720 trees/ha at age 20. Inventory records for this stand shows it to have a dominant height of 19.6 m and basal area of 26.0 m^2/ha. With the typical regression approach, the last 6 coefficient estimates in Table 1 along with the third equation in (7a) would be used to obtain a predicted volume of 213.8 m^3/ha (3,056 ft^3/ac). If we apply the mixed-model prediction system represented by (7) and the estimated variance-covariance structure to this particular stand, the expected dominant height and basal area are 19.2 m and 27.6 m^2/ha, respectively. So, the inventory records of 19.6 m and 26.0 m^2/ha are higher for dominant height and lower for basal area. But, these deviations from the modeled expected values were not taken into account in obtaining the estimated volume of 3,056 ft^3/ac. In fact, if the full mixed-model system is used, the predicted volume is 3,045 ft^3/ac. The result of having higher than “normal” height and lower than “normal” basal area is a net reduction in predicted volume from that predicted with a traditional fixed-parameter regression model approach. In this case, there is a balance struck between reducing the prediction due to lower basal area and increasing it due to the higher height. Since the basal area is a more potent predictor of volume, the effect is a net reduction in predicted volume by comparison to the fixed-parameter prediction.

The motivation for our current work is prediction. However, several somewhat complex mathematical derivations are necessary in order to develop predictions for the various situations that may arise. We present these in other manuscripts (Fang 1999, Fang and Bailey 2000c), but due to their complexity will not repeat them here. Instead, we will focus on examples illustrating the various situations regarding available input data.

For predictions based on the simultaneous growth and yield system with random effects, we classify the situations into several cases:

Case 1 -- prediction is for an individual unit (plot) on which no previous observations have ever been made

Case 2 -- prediction is for an individual unit on which observations at prior times are available.

For each of these situations there are two sub-cases to consider:

Sub-case a -- both the endogenous variables (H and B) and all exogenous variables (t, T, and all the "z" values) on the RHS of the system are observed on the occasion for which prediction is wanted.

Sub-case b -- only partial variables on the RHS are observed on the occasion for which prediction is wanted.

For *Case 2* there may also be two other special cases:

Sub-case 2a1 and 2b1 -- all the responses (H, B, and V), as well as all the exogenous variables, were measured on those past occasions for which observations are available.

Sub-case 2a2 & 2b2 -- for some past occasions, complete data are available but on others only partial data are recorded.

A summary of these situations follows:

	Information Available for			
	Current Condition?		Past Conditions?	
Case	Exogenous	Endogenous	Exogenous	Response + Endogenous
1a	Yes	Yes	No	No
1b	Yes	No	No	No
2a1	Yes	Yes	Yes	Yes
2a2	Yes	Yes	Yes	No
2b1	Yes	No	Yes	Yes
2b2	Yes	No	Yes	No

Exogenous: age, trees/ha, site preparation treatment
Endogenous: dominant height, basal area
Response: volume/ha

Case 2b* is perhaps the most common situation in forestry practice. For example, it is not uncommon in forestry that projections are required for future volume or basal area on a stand for which inventory records contain volume as well as age, height and basal area for some past occasions. For other past occasions only dominant height and age, or maybe only age, will be known. The results of the individual predictions in these different cases will usually be different. Generally speaking, the observed components of endogenous variables in the simultaneous system can be used to improve the precision of prediction of the unobserved components in the system because of the contemporaneous correlation in the simultaneous system. For a stand on which some or all variable values are known for prior ages, localized parameters can be estimated and applied to improve the precision of predictions.

Examples

Through the use of several examples, we will illustrate the various cases and sub-cases summarized above.

Example 1

Suppose we want to predict the slash pine dominant height (H), basal area (B), and total volume (V) per hectare at age 20 on a new plot with silvicultural treatments *FCBH* (chop, burn, herbicide, fertilize). We also know the soil type is nonspodosol and the stand density is 720 trees/ha at age 20. Trees/ha could be measured or projected by some other models but, since it is an exogenous variable, it is assumed know without error. There is no information available on this plot for prior ages. So, this is *Case 1b* as discussed above. Using the appropriate equations, coefficients, and estimated variance-covariance matrices (Fang and Bailey 2000c), we determined the predicted values, approximate variances, and approximate 95% prediction intervals. For basal area and volume, these were in logarithms and were converted back to the original measures. We then calculated the error bound (one-half of the 95% prediction interval), converted it to the original measure and expressed it as a percent of the predicted value:

Variable	Predicted value	Error Bound (%)
Height (m)	19.2	18.9
Basal Area (m^2/ha)	27.6	30.3
Volume (m^3/ha)	224.2	37.3

The proper interpretation is that for this *Case 1b* situation, we can only predict the volume on an individual plot basis to within ±37% with 95% confidence.

Example 2

Suppose for the example shown above, we also know that the dominant heights were 12.5 *m*, 14.8 *m* and 17.6 *m* at ages 11, 14 and 17 years, respectively, for the slash pine stand on which predictions of stand characteristics at age 20 are required. Thus, we have an example of *Case 2b2* with only the prior endogenous variable of height (H) known. The dominant height observations for prior ages can be used to estimate random effects associated with this individual stand and the precision of predictions in the simultaneous system will therefore be improved:

Variable	Predicted value	Error Bound (%)
Height (m)	19.4	4.4
Basal Area (m^2/ha)	27.9	18.5
Volume (m^3/ha)	228.7	20.2

For this example, on a "localized" basis (i.e. for a particular plot in a particular stand) a 17% reduction in the Error Bound results when heights from three prior ages are used. Generally speaking, with information on more than one plot in a given stand the estimates could be combined to further reduce the prediction error for the stand.

Example 3

Suppose we know that the basal area at age 20 is 26 m^2/ha in addition to knowing the dominant heights at ages 11, 14 and 17 as given above in Example 2 and the exogenous variables identified in Example 1. Thus, we have an example of *Case 2a1* with the endogenous variable height (H) known for three prior ages and the endogenous variable basal area (B) know at the "current" age of 20 years. The dominant height observations for prior ages are used to estimate random effects associated with this individual stand (i.e. to localize the predictions). The known value for basal area at age 20 and the estimated contemporaneous correlations from the fitting of the model help to further refine the predictions age 20:

Variable	Predicted value	Error Bound (%)
Height (m)	19.4	4.0
Basal Area (m^2/ha)	26.0	---
Volume (m^3/ha)	211.7	6.0

Example 4

In addition to all the information available in *Example 3*, suppose we know the height at age 20 is 19.6 meters. This is another variation on of *Case 2a1* and the additional information provided by current height refines the localization of the models but adds very little to precision over *Example 3*:

Variable	Predicted Value	Error Bound (%)
Height (m)	19.6	---
Basal Area (m^2/ha)	26.0	---
Volume (m^3/ha)	197.5	5.8

Table 2. -- Five examples of predicted values and confidence intervals with the simultaneous mixed-model system for a plot in an intensively site prepared slash pine plantation at age 20.

Available Information	*Dominant Height Prediction (m)* $\hat{HD}$	*95%CI*[*]	*Basal Area Prediction (m^2/ha.)* $\ln(\hat{BA}+1)$	$\hat{BA}$	*95% CI*	*Total Volume Prediction (m^3/ha.)* $\ln(\hat{V}+1)$	$\hat{V}$	*95% CI*
Example 1: *Exogenous variables only*	19.18 (1.85)	[15.55, 22.81]	3.3184 (0.1558)	26.62	[20.35, 36.48]	5.4170 (0.1854)	224.2	[155.6, 322.8]
Example 2: *Example 1 + Dom. Ht. at ages 11, 14 and 17*	19.42 (0.439)	[18.56, 20.28]	3.3349 (0.09053)	27.08	[22.51, 32.53]	5.4443 (0.10311)	230.4	[188.1, 282.3]
Example 3: *Example 2 +BA @ age 20*	19.42 (0.431)	[18.57, 20.26]	-	-	-	5.3624 (0.03426)	212.2	[198.4, 227.1]
Example 4: *Example 3 + H @ age 20*	-	-	-	-	-	5.3035 (0.02923)	200.0	[188.8, 211.9]
Example 5: *Example 1 +H & BA @ age 20*	-	-	-	-	-	5.3663 (0.03868)	213.1	[197.4, 230.9]

[*]95%CI is the approximated confidence intervals at α=95% level, CI = Mean $\pm$ 1.96$\times$std. For basal area and total volume, the CI's are obtained by transforming back from the CI in logarithm.

Example 5

Suppose we have no prior height measurements but know the current height and basal area at age 20 as well as all the exogenous variables. This is an example of *Case 1a*:

Variable	Predicted Value	Error Bound (%)
Height (m)	19.6	---
Basal Area (m^2/ha)	26.0	---
Volume (m^3/ha)	213.1	7.6

Table 3.-- Comparisons of the predictions (at age 20) with the simultaneous model system*
Based on different combinations of former HD measurements

Example	Former HD available	HD at 20 pred.	HD at 20 var.	log(BA+1) pred.	log(BA+1) var.	log(Vol+1) pred.	log(Vol+1) var.	Vol pred.	95% C.I. low	95% C.I. upper	RBE
		meters		ln(sq m/ha)		ln(cu m/ha)			------ cu m/ha ------		
1	none	19.18355	3.43523	3.318365	0.024260	5.417018	0.034361	224.21	155.60	322.87	37.30%
2	11+14+17	19.35231	0.13130	3.330271	0.007906	5.436663	0.010204	228.67	187.42	278.96	20.02%
2a	age 11	20.39052	1.25779	3.401309	0.012676	5.553869	0.017240	257.23	198.64	333.03	26.12%
2b	age 14	19.52261	0.54499	3.342181	0.009825	5.456313	0.013037	233.23	186.26	291.98	22.66%
2c	age 17	19.79072	0.17258	3.360722	0.008047	5.486905	0.010409	240.51	196.74	293.97	20.21%
2d	age 11+14	18.58111	0.45428	3.274991	0.009679	5.345455	0.012827	208.65	166.92	260.76	22.49%
2e	age 11+17	19.50910	0.13697	3.341240	0.007919	5.454760	0.010222	232.87	190.83	284.12	20.03%
2f	age 14+17	19.84788	0.14236	3.364643	0.007909	5.493374	0.010205	242.08	198.41	295.30	20.01%
3	11+14+17	19.35181	0.12431	known	-	5.359700	0.000941	211.66	199.25	224.84	6.04%
4	11+14+17	known	-	known	-	5.290514	0.000853	197.45	186.41	209.14	5.76%
5	none	known	-	known	-	5.366263	0.001496	213.06	197.43	229.92	7.62%

* Bailey and Fang, PMRC TECHNICAL REPORT 2000-3

****NOTE: Some of these results (i.e. RBE) are slightly different than those in the text. These are the correct values.**

By comparing Examples 4 and 5 we see that a reduction of 2% results from “localization” with prior height data even when current height and basal area are known (Table 2, Figure 1).

The precision of dominant height prediction mainly depends on the repeated measurement factor (Figure 2) and dominant height itself is the main factor for precise basal area predicting (Figure 3). Basal area and the repeated measurement factor in dominant height both heavily affect total stem volume prediction (Figure 4).

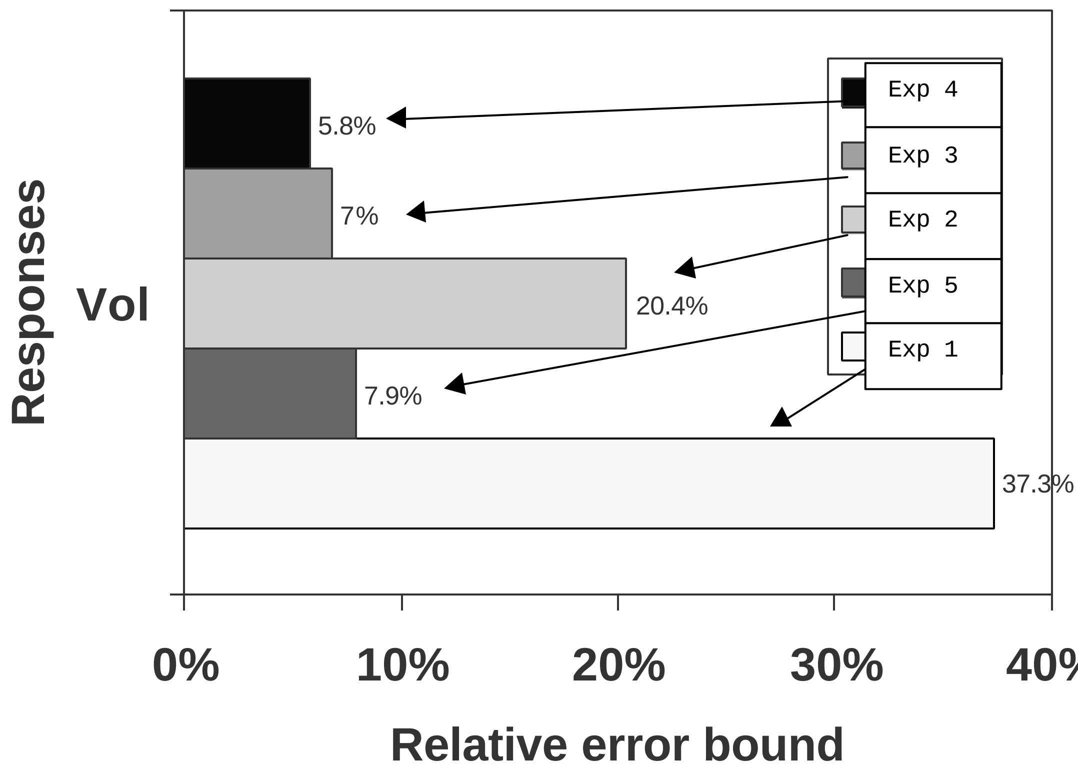

Figure 1.-- A comparison of "Bound on Error" for different prediction situations (5 examples) in the simultaneous growth and yield model system. "Bound on Error" is defined as half the approximate 95% confidence interval divided by the predicted value.

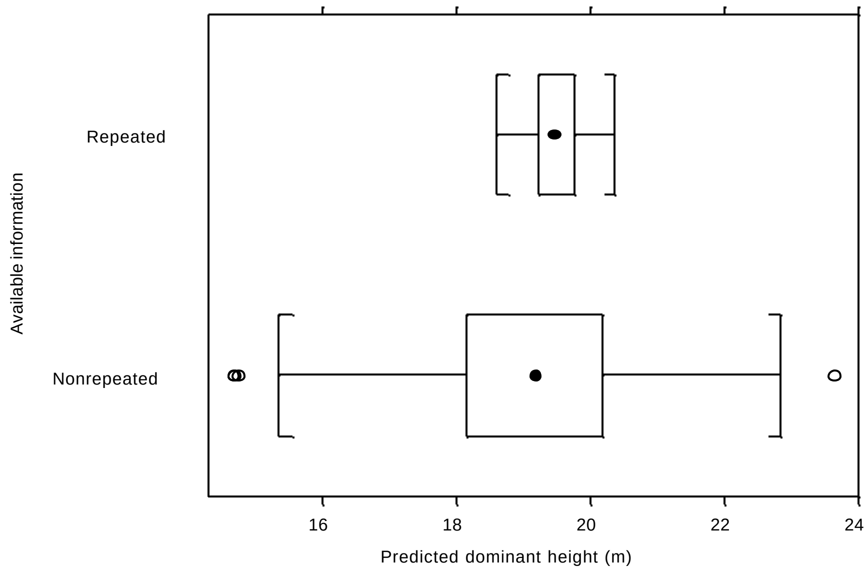

Figure 2.-- A comparison of the predictions and corresponding confidence intervals for slash pine dominant height in the simultaneous growth and yield model system with or without the information on repeated measurements.

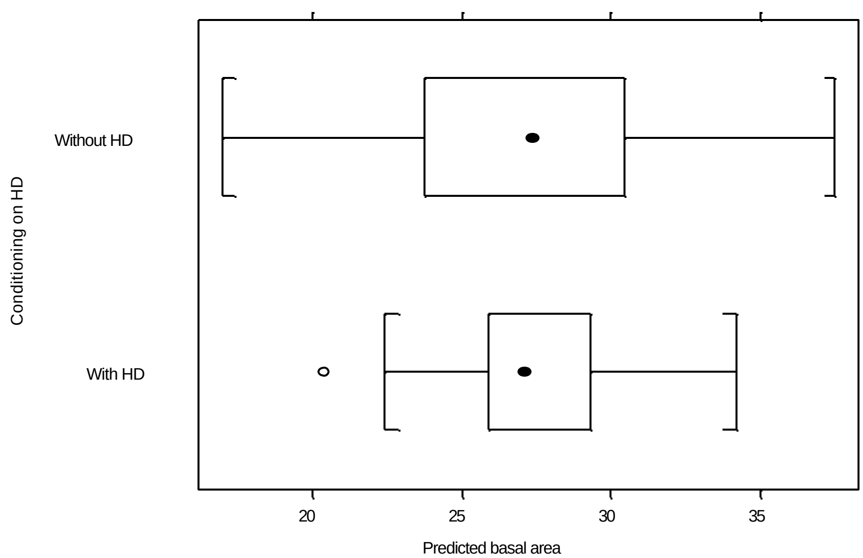

Figure 3.-- A comparison of the predictions and corresponding confidence intervals for slash pine basal area (in logarithm) in the simultaneous growth and yield model system with or without the contemporary dominant height measured.

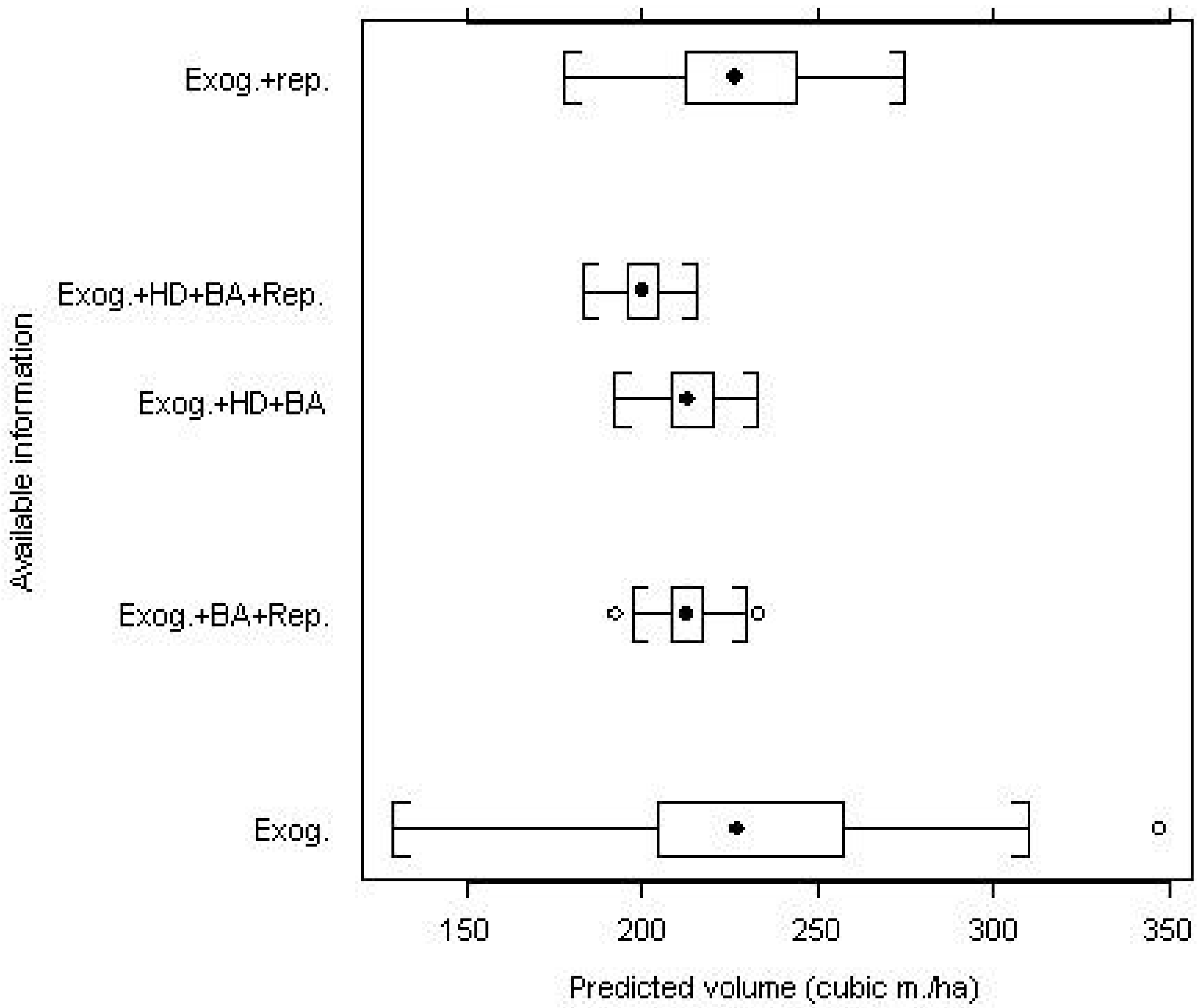

Figure 4.-- A comparison of the predictions and corresponding confidence intervals for slash pine total stem volume (in logarithm) in the simultaneous growth and yield model system based on 5 distinct situations with different information available. Note: *Exog* = exogenous variables known; *BA* = current basal area known; *HD* = current dominant height known; *rep* = dominant height know at three prior ages.

Discussion and Conclusions

Two basic sources of variation in predicted stand volume are very common for survey data, such as those from permanent plots in a forest inventory. One is the within-plot error and the other is the variation from plot to plot. Within-plot variation is usually modeled by a reasonable variance function that accounts for the lack of a constant variance within-plot and correlation of measurements. The between plot variation can be modeled by random effects that allow the parameters in the model to be varied from plot to plot. The more available information on these sources of variation, the more precise the predictions of the stand characteristics for a new plot. This is the intuitive logic behind the improvements in prediction of stand growth and yield for a new plot with the use of a simultaneous system that includes random effects. The interdependency among the components in the system (dominant height, basal area and volume) is another key to improved predictions. The observed components in the system can be used to improve the prediction of the unobserved components because of the contemporaneous correlation among these components (i.e. height, basal area, and volume).

Dominant height is the fundamental component in our simultaneous system. It is a predictor in both the basal area model and the volume model. In addition, the prediction error for dominant height is one of the main sources of error in both the basal area and the total volume predictions when observed dominant height is unavailable. Thus, precise prediction of dominant height is critically important in this simultaneous system. For example, when dominant height is known, the coefficients of variation (CV) for predicted values (logarithms) in the models decreases by about 46% in the case of basal area and by 48% in the case of volume. When observations for dominant height at three prior ages are available but current dominant height is unknown, the CV values for the basal area and total volume response variables (logarithms) still decrease by 42%, and 45%, respectively. This is because the estimated random effects from previously observed dominant heights dramatically decreased the prediction error for dominant height (from 1.85 *m* to 0.44 *m*). When observed basal area is unavailable, the prediction error of basal area is the main source of the error in volume prediction. So, the three components in this simultaneous system form a pyramid with dominant height at the base and volume at the top.

Once the random effects in dominant height and the contemporaneous correlations among the components are reasonable modeled, accounting for random effects in basal area and total volume proved to be unnecessary in our analysis. This result is supported by statistical criteria, such as *AIC* and the likelihood ratio test. Intuitively, this makes logical sense in that dominant height is a predictor in both the basal area model and the volume model. Random effects are reasonably included in the dominant height model. Therefore, it is no surprise that random effects in the basal area and volume models are not significant in the simultaneous fit.

From the error structure of dominant height, random effects (i.e., between-plot variation) is the main source of error. For example, in the marginal mean response random effects account for 97.4% of the variation in the total error. For simplicity, within-plot correlation for the same component of the response in the system has not been included in the model. This does not mean that autocorrelation is generally unimportant. Actually, the random effects are often capable of representing the serial correlation (Jones 1990). Random effects and autocorrelation can compete with each other in covariance modeling (Pinheiro and Bates, 1998). In order to avoid this complexity, modeling serial correlation has given way to approaches using random effects.

References

Amemiya, T. 1985. Advanced econometrics. Cambridge, MA: Harvard University Press.

Bates, D. M. and J. C. Pinheiro 1999. Mixed effects models in S. Springer, New York.

Borders B. E. 1989. Systems of equations in forest stand modeling. For. Sci. 35:548-556.

Borders B. E. and R. L. Bailey, 1986. A compatible system of growth and yield equations for Slash pine fitted with restricted three-stage least squares. For. Sci. 32:185-201.

Burkhart, H. E. and P. E. Sprinz 1984. Compatible cubic volume and basal area projection equations for thinned old-field loblolly pine plantations. For. Sci. 30:86-93.

Castleberry, J. P. 1998. Evaluation of models developed on a designed study data set applied to operational stands. The University of Geogia, M.S. thesis.

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.

Clutter J. L. 1963. Compatible growth and yield models for loblolly pine. For. Sci. 9(3):354-370.

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

Fang, Zixing. 1999. A Simultaneous System of Linear and Nonlinear Mixed Effects Models for Forest Growth and Yield Prediction. PhD dissertation, The Univ of Ga, Athens, GA, USA 30602, 177 p.

Fang, Zixing and Robert L. Bailey 2000a A linear multilevel mixed effects model for slash pine basal area growth following intensive silvicultural treatments. Can. Jour. of For. Res. [in review]

Fang, Zixing and Robert L. Bailey 2000b A nonlinear mixed effects model for slash pine dominant height growth following intensive silvicultural treatments. For. Sci. [in review]

Fang, Zixing and Robert L. Bailey 2000c A multivariate simultaneous prediction system for stand growth and yield with fixed and random effects. For. Sci. [in review]

Furnival, G. M. and R. W. Wilson, JR. 1971. System of equations for predicting forest growth and yield. Statist. Ecol. 3: 43-55.

Hasenauer, H., R.A. Monserud, and T.G. Gregoire. 1998. Using simultaneous regression techniques with individual-tree growth models. For. Sci. 44(1): 87-95.

Jones, R.H. 1990. Serial correlation or random subject effects? Commu. in Stat. A 19:1105-1123.

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. 38(2): 409-429.

LeMay, M. V. 1990. MSLS: a linear least squares technique for fitting a simultaneous system of equations with a generalized error structure. Can. J. For. Res. 20: 1830-1839.

Liski, E. P. and T. Nummi 1990. Prediction in growth curve models using the EM algorithm. Computational Statistics & Data Analysis 10: 99-108.

Liski, E. P. and T. Nummi 1995. Prediction of tree stems to improve efficiency in automatized harvesting of Forests. Scandinavian Journal of Statistics 22: 255-269.

Liski, E. P. and T. Nummi 1996. Prediction in repeated-measures models with engineering applications. Technometrics 38(1):25-36.

Pienaar, L.V., B.D. Shiver, and W. M. Harrison. 1998. Slash pine site preparation study: age 17 results. Sch. Forest Resour., Univ. Ga., Athens, PMRC Tech. Rep. 1998-4.

Pinheiro, J.C. and D.M. Bates 1998. Model building for nonlinear mixed effects models. Technical report, Department of statistics, University of Wisconsin.

Rao, C. R. 1976. Prediction of future observations with special reference to linear models. In Multivariate Analysis IV (P. R. Krishnaiah, ed.) 193-208. North-Holland, Amsterdam.

Ratkowsky, D. A. 1990. Handbook of nonlinear regression models. Marcel Dekker, New York.

Richards, F.J. 1959. A flexible growth function for empirical use. J.Exp.Bot.10:290-300.

Robinson, G. K. 1991. That BLUP is a good thing: the estimation of random effects. Statist. Sci. 6(1): 15-51. (with discussions)

Schnute, J. 1981. A versatile growth model with statistically stable parameters. Can. J. Fish. Aquat. Sci. 38(9):1128-1140.

Srivastava, V. K. and D. E. A. Giles. 1987. Seemingly unrelated regression equations models. New York: Dekker.

Sullivan, A. D. and J. L. Clutter 1972. A simultaneous Growth and yield model for loblolly pine. For. Sci. 18(1):76-86.

Tait D.E., C.J. Cieszewski, I. E. Bella. 1988. The stand dynamics of lodgepole pine. Can. J. For. Res. 18(10) 1255-1260.

Valliant, R. 1985. Nonlinear prediction theory and the estimation of proportions in a finite population. J. of the Amer. Statist. Assoc. 80(391): 631-641.

Verbeke, G., and G. Molenberghs. 1997. Linear mixed effects model in practice: a SAS-oriented approach. Springer-verlag, New York.

Vonesh, E.F., and V.M. Chinchilli, 1997. Linear and nonlinear models for the analysis of repeated measurements. Marcel Dekker, New York.