Journal of Applied Statistics, Vol. 23, Nos 2&3, 1996, 257–271

CARFAX

A non-linear mixed-effects model to predict cumulative bole volume of standing trees

TIMOTHY G. GREGOIRE & OLIVER SCHABENBERGER, *College of Forestry and Wildlife Resources, Virginia Polytechnic Institute and State University, USA*

SUMMARY *For purposes of forest inventory and eventual management of the forest resource, it is essential to be able to predict the cumulative bole volume to any stipulated point on the standing tree bole, while requiring measurements of tree size that can be made easily, quickly and accurately. Equations for this purpose are typically non-linear and are fitted to data garnered from a sample of felled trees. Because the cumulative bole volume of each tree is measured to numerous upper-bole locations, correlations between measurements within a tree are likely. A mixed-effects model is fitted to account for this within-subject (tree) correlation structure, while also portraying the sigmoidal shape of the cumulative bole volume profile.*

1 Introduction

Since the introduction of regression methods into forestry more than 60 years ago, it has been common to fit a regression model to predict the woody volume in the bole of a standing tree. Morphological differences among species and even intra-specific differences caused by varying physiographic, climatic and other environmental effects generally require that different equations be used for—or at least that a particular equation be fitted separately to data from—each regional population of tree species to which it eventually will be applied for the purpose of volume prediction. The volume of interest may include the bark (outer-bark volume) or not (under-bark volume), and it may comprise the entire bole or only the portion of it between stump level and a stipulated point on the upper bole.

Correspondence: T. Gregoire, Department of Forestry and Wildlife Resources, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0324, USA.

0266-4763/96/020257-15 $6.00

Equations of the last sort are known as merchantable volume equations, because the upper-bole point is often determined by a minimum diameter that establishes a merchantability threshold, above which there is too little volume in the tip of the bole to convert to a merchantable product economically. Regardless of the type of volume (outer- or under-bark; total-bole or merchantable) that serves as the response variable in the model, the set of covariates included in the model must be restricted to the relatively small set of overall tree dimensions that can be measured quickly, yet accurately in the forest—otherwise, the fitted model will never be applied. Almost universally, the bole diameter at breast height (D) and total tree height (H) comprise this set. Breast height, which is typically 1.3 or 1.37 m above ground, is the customary height at which to measure a tree's reference diameter, i.e. D, to avoid the buttressing in the lower reaches of the stem. In some cases, an alternative to H is the height H_m to the merchantable diameter limit.

Although volume equations are developed to provide a prediction of the volume of standing trees, these equations are fitted to measurements conducted on felled trees, so as to minimize measurement error and its consequent effect on parameter estimation. For example, a sample of trees is selected that spans the range of tree sizes for which the fitted equation is intended to be applicable. Each sample tree is felled, its bole delimbed and cut into short sections of possibly unequal lengths. The volume of each section is determined and accumulated with the volumes of lower sections. This is repeated to the top of the tree bole if total-bole volume is the response variable in the regression model, or to the point where the bole's diameter has tapered to the merchantable diameter, if merchantable volume is the response variable. Almost always, the number of observations of cumulative bole volume will vary among trees, according to tree size; short trees may have as few as two, whereas tall trees will have 30 or more. Therefore, the observed sample will be unbalanced, in the sense of having unequal numbers of observations per subject.

As implied above, a multitude of different bole-volume equations have been developed over the years and, with little effort, one can easily identify hundreds for just temperate-zone tree species. Part of the reason for the multiplicity of volume equations derives from the on-going change in merchantability standards over time. While this change is dictated partly by the anticipated end-use product, it is also a result of technological advances in milling and fluctuations in the economic value of the raw material. It has been commonplace to fit a new bole-volume equation, as required, in response to changes in the upper-bole merchantability diameter limit, which is a costly and perhaps duplicative endeavor.

Burkhart (1977), however, suggested an alternative strategy in which the upper-bole diameter appears as a pseudo-covariate in the volume equation. Given an appropriately fitted volume equation of this sort, one can then evaluate it with measured values of D and H on a standing tree, in order to predict the merchantable volume, say V_d, to an upper-bole diameter d. When $d = 0$, V_0 constitutes the total-bole volume. For a given species, therefore, a single equation can be used to predict the merchantable volume to a diameter limit d on one occasion, and to a different upper-bole diameter on another occasion. The forestry literature features a limited but growing number of applications of this type of model; see, for example, Golden *et al.* (1982) and Knoebel *et al.* (1984) with yellow poplar (*Liriodendron tulipifera* L.); Van Deusen *et al.* (1981), Newberry and Burk (1985) and Amateis and Burkhart (1987) with loblolly pine (*Pinus taeda* L.); Bailey (1994) with slash pine (*Pinus elliotii* Engelm.); and Gregoire and Schabenberger (1995)

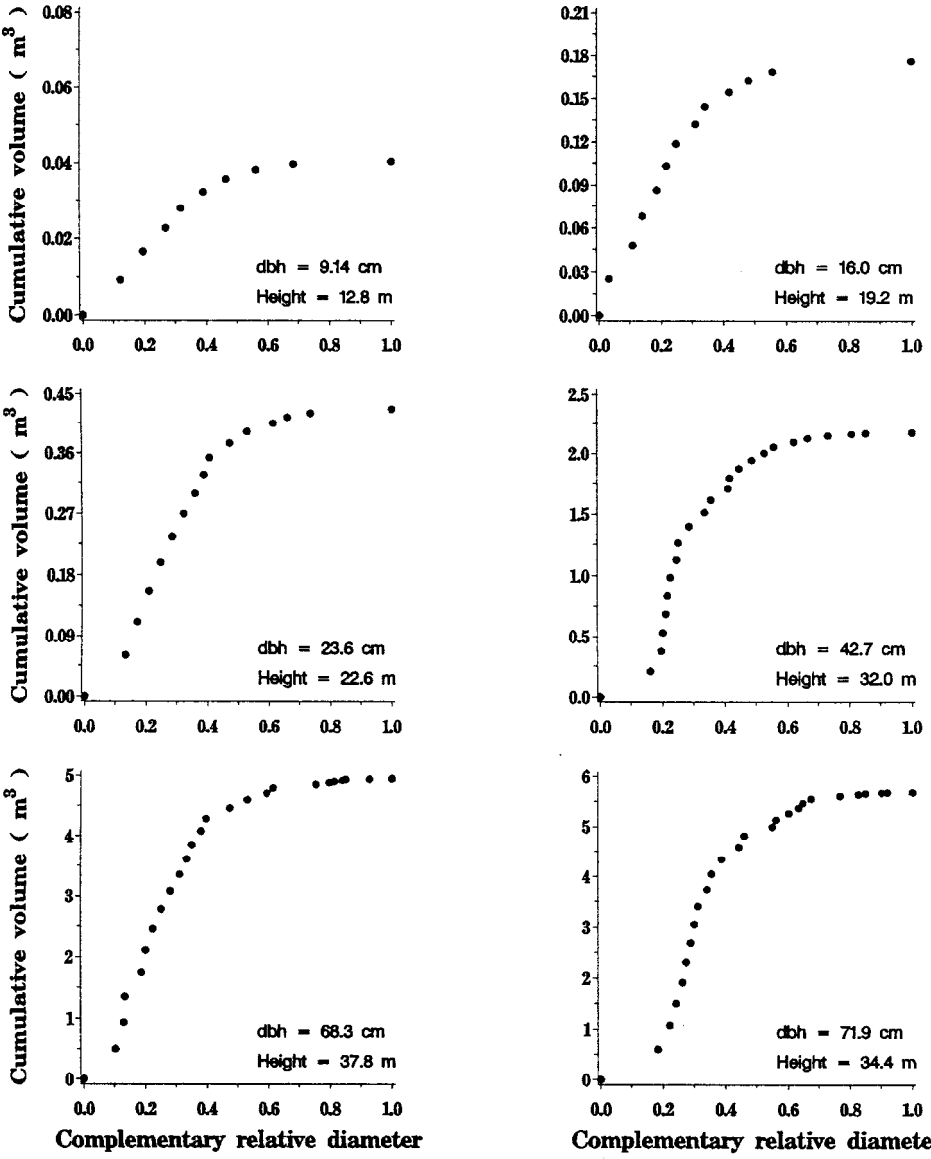

FIG. 1. Yellow poplar cumulative outer-bark bole volume: empirical profiles.

with sweetgum (*Liquidambar styraciflua* L.). The customary approach has been to express V_d as the product $V_0 R_d$, where R_d represents the ratio of the merchantable volume to the total-bole volume. Newberry and Burk (1985) and Avery and Burkhart (1994) refer to this type of model as a 'volume-ratio equation' and we also adopt this lexicon.

Figure 1 portrays the empirical bole-volume function for each of six yellow poplar trees of various sizes, where the cumulative bole volume is plotted against $r = 1 - d/d_s$, where d_s is the stump diameter of the tree. At the base of the tree, where $d = d_s$, the cumulative bole volume obviously is $V_{d_s} = 0$; at the tip of the tree, $d = 0$ and the cumulative volume V_0 is the total-bole volume. Gregoire and

Schabenberger (1995) remarked on the similarity of cumulative bole volume profiles to cumulative distribution functions, and to many biological growth functions. Previously, Van Deusen *et al.* (1981) had fitted R_d as an exponential function in d/D, and Newberry and Burk (1985) fitted the S_B distribution function, both of which captured the essential sigmoidal shape of the response curve.

Because the multiple measurements on each tree are likely to be correlated, Gregoire and Schabenberger (1995) proposed a mixed-effects model to account for the within-subject correlation. Their work in this area was novel within forestry, because all other modelling efforts had ignored the within-subject correlations. In the present work, a modified type I extreme value distribution function is incorporated into the ratio (R_d) term, and we exemplify a model-fitting approach based on Gaussian maximum likelihood and on generalized estimating equations (GEEs) for continuous responses.

2 Model development

The customary approach to developing volume-ratio equations (cf. Avery & Burkhart, 1994) has been to fit a model for the total-bole volume separately from that of the ratio term. However, it seems more reasonable to pursue a joint estimation of both terms in the composite model.

Since Spurr (1952), an extensively used expression for V_0 has been the simple linear regression

$$V_0 = \beta_1 + \beta_2 X$$

where $X = D^2H/1000$. Our introduction of the scaling factor 1/1000 puts β_2 on the same scale as β_1.

For the ratio term, we chose $R_d = \exp(-\beta_3 t' \exp(\beta_4 t))$, where $t = d/D$, and $t' = t/1000$. This function resembles a type I extreme value function. R_d is always positive and tends to unity as $d \to 0$; thus, it ensures the logical constraint that V_d cannot exceed V_0 and cannot assume a negative value. It exhibits an interior inflection point, and is parsimonious. Moreover, we have found it to be very flexible in adapting to a wide variety of bole-volume profiles, such as those shown in Fig. 1. Finally, when imbedded in a mixed-model framework as described below, the above expression for R_d is considerably more straightforward than the S_B distribution function used by Newberry and Burk (1985) with its attendant difficulties in percentile prediction.

Let V_{id_j} denote the observation of the cumulative bole volume on the ith tree ($i = 1, \ldots, n$; $j = 1, \ldots, m_i$) to the upper-bole diameter d_j at the jth location on the bole. Let $V_{i0} = \beta_1 + \beta_2 X_i$, where $X_i = D_i^2 H_i/1000$, and let $R_{id_j} = \exp(-\beta_3 t'_{ij} e^{\beta_4 t_{ij}})$, where $t_{ij} = d_j/D_i$ and $t'_{ij} = t_{ij}/1000$. The fixed-effects version of our volume-ratio model is

$$V_{id_j} = (\beta_1 + \beta_2 X_i) \exp(-\beta_3 t'_{ij} e^{\beta_4 t_{ij}}) + \varepsilon_{ij}$$

To account for the intra-individual variation that arises from the multiple measurements of the bole volume on each tree, we opted to add a random element to $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)'$, in preference to modelling the within-subject covariance directly through the joint distribution of $\varepsilon_{ij}, \varepsilon_{ik}, j \neq k$ (cf. Jones, 1990; Gregoire *et al.*, 1995). Both Davidian and Giltinan (1993) and Pinheiro *et al.* (1994) have suggested a model-building strategy for mixed-effects models that begins with all

effects as random. Accordingly, we regard $\mathbf{b}_i = (b_{1i}, b_{2i}, b_{3i}, b_{4i})'$ as a random vector, i.e. $\mathbf{b}_i = \boldsymbol{\beta} + \boldsymbol{\gamma}_i$, where $\boldsymbol{\gamma}_i = (\gamma_{1i}, \gamma_{2i}, \gamma_{3i}, \gamma_{4i})'$, and

$$E[\mathbf{b}_i] = \boldsymbol{\beta}$$

$$\text{var}[\mathbf{b}_i] = E[(\mathbf{b}_i - \boldsymbol{\beta})(\mathbf{b}_i - \boldsymbol{\beta})'] = E[\boldsymbol{\gamma}_i \boldsymbol{\gamma}_i'] = \sigma^2 \Delta, \qquad \forall i$$

and

$$E[\boldsymbol{\gamma}_i \varepsilon_{ij}] = \mathbf{0}, \ \forall i, j$$

The resulting full mixed-effects version of our volume-ratio model is

$$\begin{aligned} V_{id_j} &= (b_{1i} + b_{2i} X_i) \exp(-b_3 t_{ij}' e^{b_4 t_{ij}}) + \varepsilon_{ij} \\ &= f(Q_{ij}; \mathbf{b}_i) + \varepsilon_{ij} \end{aligned} \tag{1}$$

where Q_{ij} represents the set of covariates $\{X_i, t_{ij}, t_{ij}'\}$.

Sheiner and Beal (1980) pioneered procedures for fitting mixed-effects non-linear models. In recent years, there has been a flurry of work in this area (see, for example, Lindstrom & Bates, 1990; Vonesh & Carter, 1992; Davidian & Gallant, 1993; Davidian & Giltinan, 1993, 1995; Wolfinger, 1993; Schabenberger, 1995; Gregoire & Schabenberger, 1995). Our approach is to approximate the marginal distribution of the response vector by expanding $f(\cdot)$ in a first-order Taylor series, as did Sheiner and Beal, and Lindstrom and Bates. One can then derive maximum likelihood or restricted maximum likelihood estimators, based on the approximate marginal density of the linearized response. Typically, a Gaussian distribution is assumed.

A first-order Taylor expansion of our model in equation (1), around the values $\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}_i$, gives the approximating linear function as

$$V_{id_j} \doteq f(Q_{ij}; \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}_i) + \mathbf{z}_{ij}'(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) + \mathbf{w}_{ij}'(\boldsymbol{\gamma}_i - \tilde{\boldsymbol{\gamma}}_i) + \varepsilon_{ij}$$

where

$$\mathbf{z}_{ij}' = \left. \frac{\partial f(Q_{ij}; \boldsymbol{\beta}, \boldsymbol{\gamma}_i)}{\partial \boldsymbol{\beta}'} \right|_{\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}_i}$$

$$\mathbf{w}_{ij}' = \left. \frac{\partial f(Q_{ij}; \boldsymbol{\beta}, \boldsymbol{\gamma}_i)}{\partial \boldsymbol{\gamma}_i'} \right|_{\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}_i}$$

Rearranging yields

$$Y_{id_j} \doteq \mathbf{z}_{ij}' \boldsymbol{\beta} + \mathbf{w}_{ij}' \boldsymbol{\gamma}_i + \varepsilon_{ij} \tag{2}$$

where

$$Y_{id_j} = V_{id_j} - f(Q_{ij}; \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}_i) + \mathbf{z}_{ij}' \tilde{\boldsymbol{\beta}} + \mathbf{w}_{ij}' \tilde{\boldsymbol{\gamma}}_i$$

Equation (2) is a linear mixed model where the observed cumulative bole volume V_{id_j} has been replaced by what Gregoire and Schabenberger (1996) labelled a 'pseudo-response' Y_{id_j}.

Let $\mathbf{Y}_i = (Y_{id_1}, \ldots, Y_{id_{m_i}})'$ denote the vector of pseudo-observations for the ith subject tree, and let $\mathbf{Z}_i = (\mathbf{z}_{i1}', \ldots, \mathbf{z}_{im_i}')'$, $\mathbf{W}_i = (\mathbf{w}_{i1}', \ldots, \mathbf{w}_{im_i}')'$ and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{im_i})'$. For the ith tree, the approximated linear model of cumulative bole-volume is

$$\mathbf{Y}_i = \mathbf{Z}_i \boldsymbol{\beta} + \mathbf{W}_i \boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i \tag{3}$$

where the conditional variance is $\text{var}[\mathbf{Y}_i | \mathbf{Z}_i, \mathbf{W}_i, \boldsymbol{\gamma}_i] = \sigma^2 \mathbf{I}_{m_i}$, and the marginal variance is

$$\text{var}[\mathbf{Y}_i | \mathbf{Z}_i] = \sigma^2 (\mathbf{W}_i \Delta \mathbf{W}_i' + \mathbf{I}_{m_i}) = \sigma^2 \boldsymbol{\phi}_i \tag{4}$$

3 Likelihood estimation under a Gaussian model

When both ε_i and γ_i are independently Gaussian distributed, minus twice the logarithm of the approximate marginal likelihood is

$$\tilde{L} = \sum_{i=1}^{n} (m_i \ln(2\pi) + m_i \ln(\sigma^2) + \ln|\phi_i| + \sigma^{-2}\mathbf{r}_i\phi_i^{-1}\mathbf{r}_i')$$

where $\mathbf{r}_i = \mathbf{Y}_i - \mathbf{Z}_i\beta$. It is straightforward (cf. Jones, 1993; Wolfinger & O'Connell, 1993; Diggle *et al.*, 1994, p. 63) to maximize $\tilde{L}$ analytically for β and σ^2, and then estimate the unique parametric components of Δ that jointly maximize the resulting concentrated (or profile) likelihood. Letting $\hat{\Delta}$ denote this estimate, one can estimate

$$\hat{\beta} = \hat{\Omega}^{-1}\left(\sum_{i=1}^{n} \mathbf{Z}_i'\hat{\phi}_i^{-1}\mathbf{Y}_i\right) \tag{5}$$

where

$$\hat{\Omega} = \sum_{i=1}^{n} \mathbf{Z}_i'\hat{\phi}_i^{-1}\mathbf{Z}_i \tag{6}$$

and

$$\hat{\text{var}}[\hat{\beta}] = \hat{\sigma}^2\hat{\Omega}^{-1}$$

is the estimated variance of $\hat{\beta}$. The random effects can be predicted by

$$\hat{\gamma}_i = \hat{\Delta}\mathbf{W}_i'\hat{\phi}_i^{-1}\hat{\mathbf{r}}_i \tag{7}$$

where $\hat{\mathbf{r}}_i = \mathbf{Y}_i - \mathbf{Z}_i\hat{\beta}$. In a linear model context, equation (7) has been termed an empirical best linear unbiased predictor (Harville & Carriquiry, 1992), or EBLUP for short.

Because the pseudo-responses and the derivative matrices in equation (3) depend on the current estimates, the linear mixed model of equation (3) is fitted repeatedly until successive changes in the estimates or the likelihood are inconsequential. At this point, the scale parameter σ^2 can be estimated by

$$\hat{\sigma}^2 = M^{-1}\sum_{i=1}^{n} \hat{\mathbf{r}}_i\hat{\phi}_i^{-1}\hat{\mathbf{r}}_i', \qquad \text{where } M = \sum_{i=1}^{n} m_i \tag{8}$$

In their approach, Sheiner and Beal (1980) opted to expand the non-linear response around $\tilde{\beta}$ and $\tilde{\gamma}_i = E[\gamma_i] = 0$. In contrast, Lindstrom and Bates (1990) chose to expand around $\hat{\beta}$ and the current solutions of the random effects.

Minus twice the restricted Gaussian likelihood of equation (2) is

$$\tilde{L}_R = \sum_{i=1}^{n} [m_i \ln(2\pi) + \lambda m_i \ln(\sigma^2) + \ln|\phi_i| + \sigma^{-2}\mathbf{r}_i\phi_i^{-1}\mathbf{r}_i' + \ln|\mathbf{Z}_i\phi_i^{-1}\mathbf{Z}_i'|]$$

where $\lambda = M^{-1}(M-p)$, and p is the dimension of β. Because restricted maximum likelihood estimators (REMLs) of variance components are less biased than corresponding maximum likelihood estimators, they have come to be preferred, in general. The restricted likelihood estimators developed by Lindstrom and Bates (1990) were shown by Wolfinger (1993) to be the solutions to the linear system

$$\begin{pmatrix} \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{W} \\ \mathbf{W}'\mathbf{Z} & \mathbf{W}'\mathbf{W} + \mathbf{I}_n \otimes \Delta^{-1} \end{pmatrix}\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} \mathbf{Z}'\mathbf{Y} \\ \mathbf{W}'\mathbf{Y} \end{pmatrix}$$

where $\mathbf{Z}' = (\mathbf{Z}_1', \ldots, \mathbf{Z}_n')$, $\mathbf{W}' = \text{diag}(\mathbf{W}_i')$, $\mathbf{Y}' = (\mathbf{Y}_1', \ldots, \mathbf{Y}_n')$, $\hat{\gamma}' = (\hat{\gamma}_1', \ldots, \hat{\gamma}_n')$. The solutions to this system are equivalent to equations (5) and (7).

4 Generalized estimating equations

Schabenberger (1995) discusses how GEEs can be utilized for non-linear continuous response models. It is based on an estimating function (Godambe, 1960) that involves the first two marginal moments of the response distribution only, so it is semi-parametric in that higher-order moments are unspecified.

The key idea (cf. Zeger *et al.*, 1988) is to approximate the first two marginal moments from conditional moments, taking a first-order Taylor series expansion of equation (1) around $E[\gamma_i] = 0$. This yields

$$V^{\star}_{id_j} \doteq f(Q_{ij}; \beta, 0) + \mathbf{w}'_{ij}\gamma_i + \varepsilon_{ij}$$

and, consequently, we have

$$E[V^{\star}_{id_j}] \doteq f(Q_{ij}; \beta, 0)$$

$$\text{var}[V^{\star}_{id_j}] \doteq \sigma^2(1 + \mathbf{w}'_{ij}\Delta\mathbf{w}_{ij})$$

Let

$$\mathbf{V}_i = (f(\mathbf{Q}_{i1}; \beta; 0), \ldots, f(\mathbf{Q}_{im_i}; \beta, 0))' + \varepsilon_i$$

$\mu_i = E[\mathbf{V}_i]$ and $\mathbf{Z}_i = \partial\mu_i/\partial\beta$. The GEEs for $\beta \mid \hat{\Delta}$ become

$$U(\beta; \hat{\Delta}, \mathbf{V}_i, \forall i) = \sum_{i=1}^{n} \mathbf{Z}'_i \hat{\phi}_i^{-1}(\mathbf{V}_i - \mu_i) = 0 \tag{9}$$

One then solves equation (9) iteratively, such as by a Newton–Raphson algorithm with Fisher scoring, which leads to the following update of the current estimate, $\hat{\beta}^{(u)}$ say:

$$\hat{\beta}^{(u+1)} = \hat{\beta}^{(u)} + \left(\sum_{i=1}^{n} \mathbf{Z}'_i \hat{\phi}_i^{-1}\mathbf{Z}_i\right)^{-1} \sum_{i=1}^{n} \mathbf{Z}'_i \hat{\phi}_i^{-1}(\mathbf{V}_i - \mu_i)$$

This is as shown in Schabenberger and Gregoire (1996). Since the estimates $\hat{\beta}^{(u)}$ depend on Δ, a moment estimator can be used to update $\hat{\Delta}$ after each iteration. From $\hat{\phi}_i = \mathbf{W}_i\hat{\Delta}\mathbf{W}'_i + \mathbf{I}_{m_i}$, the following consistent estimators are suggested:

$$\hat{\Delta} = n^{-1}\sum_{i=1}^{n} (\mathbf{W}'_i\mathbf{W}_i)^{-1}\mathbf{W}'_i[\hat{\sigma}^{-2}(\mathbf{V}_i - \hat{\mu}_i)(\mathbf{V}_i - \hat{\mu}_i)' - \mathbf{I}_{m_i}]\mathbf{W}_i(\mathbf{W}'_i\mathbf{W}_i)^{-1} \tag{10}$$

and

$$\hat{\sigma}^2 = M^{-1}\sum_{i=1}^{n} (\mathbf{V}_i - \hat{\mu}_i)'\hat{\phi}_i^{-1}(\mathbf{V}_i - \hat{\mu}_i)$$

Following the main result in Liang and Zeger (1986), $\hat{\beta}$ will be asymptotically unbiased and Gaussian distributed, provided that Δ and σ^2 are estimated consistently. At convergence of the algorithm, EBLUPs for the random effects (equation (8)) are obtained as in the fully parametric implementation, and an asymptotically unbiased estimator of var$[\hat{\beta}]$ is as shown in equation (7).

Irrespective of whether or not a parametric likelihood approach or a semi-parametric GEE approach to estimation is adopted, an asymptotically unbiased estimator of the variance of $\hat{\gamma}_i - \gamma_i$ is (cf. Laird & Ware, 1982; Gregoire *et al.*, 1995):

$$\hat{\text{var}}[\hat{\gamma}_i - \gamma_i] = \hat{\Delta}[\mathbf{I}_q - \mathbf{W}'_i\hat{\phi}_i^{-1}(\mathbf{I}_{m_i} - \sigma^2\mathbf{Z}_i\hat{\Omega}^{-1}\mathbf{Z}'_i\hat{\phi}_i^{-1})\mathbf{W}_i\hat{\Delta}]$$

where q is the dimension of Δ.

Let $\mathring{V} = f(\mathring{Q}; \hat{\beta}, E[\gamma])$ denote the marginal estimate of $E[V_d \mid \mathring{Q}]$ for some stipulated set of covariate values indicated by $\mathring{Q}$. An asymptotic $(1-\alpha)100\%$ confidence interval is

$$\mathring{V} \pm \zeta_\alpha\hat{\sigma}\{\mathring{\mathbf{z}}'\hat{\Omega}^{-1}\mathring{\mathbf{z}}\}^{1/2} \tag{11}$$

where $\mathring{\mathbf{z}}' = \partial f(\mathring{Q}; \hat{\boldsymbol{\beta}}, \mathbf{0})/\partial \hat{\boldsymbol{\beta}}'$, and where ζ_α is the $(1-\alpha/2)$ quantile of the standard Gaussian distribution. If $\mathring{V} = f(\mathring{Q}; \hat{\boldsymbol{\beta}}, \hat{\gamma})$ instead denotes the prediction of V_d, then one will normally be compelled to stipulate $\hat{\gamma} = E[\gamma]$, unless prediction is being made to a subject in the data to which the model was fitted. Regardless of the value of $\hat{\gamma}$ used to evaluate $\mathring{V}$, an asymptotic $(1-\alpha)100\%$ prediction interval is

$$\mathring{V} \pm \zeta_\alpha \hat{\sigma} \{\mathring{\mathbf{z}}' \hat{\Omega}^{-1} \mathring{\mathbf{z}} + \mathring{\mathbf{w}}' \hat{\Delta} \mathring{\mathbf{w}} + 1\}^{1/2}$$

5 Data

The trees profiled in Fig. 1 were six of 336 trees that were felled and measured for the purpose of developing a bole-volume prediction equation for yellow poplars in the southern Appalachian region of the southeastern US. Tree heights H ranged from 3.7 to 42.1 m, averaging 27.7 m; tree breast-height diameters D ranged from 1.8 to 76.2 cm, averaging 33.6 cm; and their total bole volumes V_0 ranged from 0.001 to 7.362 m^3, averaging 1.544 m^3. The outside-bark diameter of each stem was measured at intervals of 1.2 m along the felled stem and the volume of each 1.2 m-section was computed as the product of its length with its average cross-sectional area. The diameters were measured to the nearest 0.25 cm and the heights were measured to within ± 3 cm. There was an average of 21 bole-diameter measurements per tree, yielding a total of 6972 observations of the cumulative bole volume and these were used to fit equation (1).

The results that we report were obtained with a program written for the task in the GAUSS programming language. All the REML results were checked with the NLINMIX macro available from SAS of Cary, NC.

6 Results

Although the linearized version of equation (1) was fitted by both REML and GEEs, we were unsuccessful in fitting it by either method with all effects random. Evidently, for these yellow poplar data, a completely random structure cannot be supported, i.e. the model is over-parameterized. While the GEE approach implictly uses a linearization around $E[\gamma_i]$, the REML approach can be implemented by expanding around $\hat{\gamma}_i$ or $E[\gamma_i]$. We tried both types of expansion, to no avail. Subsequent investigation indicated that even versions of equation (1) with only three random effects were similarly over-parameterized, when fitted both to the yellow poplar data and to similar sets of data from other tree species.

Consequently, we examined models with at most two effects random, with the results shown in Table 1 for both expansions. For models with more than one random effect, one can choose to restrict Δ to a simple diagonal structure (uncorrelated random effects), or not. We fitted all models both ways. In some cases, when Δ was unrestricted, the resulting model could not be fitted, presumably as a result of over-parameterization. In cases where the model could be fitted with Δ unrestricted, the difference in the observed $-2 \ln \tilde{L}_R$ from that obtained with a simple diagonal structure was very minor. We opted to present results for the more parsimonious model only.

From the results in Table 1, we concluded that (a) any model with one or more random effect is a major improvement over a purely fixed-effects model; and (b) models with two random effects are substantially better than those with a single

TABLE 1. Observed $-2 \ln \tilde{L}_R$ when equation (1) was fitted under various combinations of random effects

	Expansion locus	
Random effect(s)	$\hat{\gamma}_i$	$E[\gamma_i]$
None	45 490	45 490
γ_1	38 070	38 082
γ_2	37 594	37 608
γ_3	41 311	41 374
γ_4	41 885	42 082
γ_1, γ_3	32 890	33 081
γ_2, γ_3	32 393	32 532
γ_1, γ_4	33 194	33 446
γ_2, γ_4	32 697	32 952

Note: In all cases, Δ was fitted with a simple diagonal structure.

random effect. The observed value of $-2 \ln \tilde{L}_R$ for the model linearized around $\hat{\gamma}_i$ is always slightly smaller than that for the model linearized around its expectation, although it is questionable whether or not there is any meaningful difference.

Among the models with two random effects, there is only a minor difference in the observed likelihoods. Because the model with b_2 and b_3 both random is arguably superior to the others, we explore its performance in more detail. To be explicit, the following results pertain to

$$V_{id_j} = (\beta_1 + (\beta_2 + \gamma_{2i})X_i)(\exp\{-(\beta_3 + \gamma_{3i})t'_{ij}e^{\beta_4 t_{ij}}\} + \varepsilon_{ij} \tag{12}$$

Table 2 shows the REML-based parameter estimates and, for the sake of comparison, we also include the GEE estimates. There is little apparent difference between the three alternative estimates of β_2 and β_4 and their estimated standard errors. The GEE estimate of β_1 is about 20% larger than that of either REML estimate, whereas the REML estimate of β_3 obtained by expanding around $\hat{\gamma}_i$ is about 20% smaller than that of either the GEE or the REML estimate obtained by expanding around $E[\gamma_i]$. The importance of these differences is questionable, however, when one looks at Fig. 2, in which are shown the GEE and REML (expansion around $\hat{\gamma}_i$) fitted profiles for the six trees for which empirical profiles were exhibited in Fig. 1. The full line that represents the REML fit overlays the dash–dot line that represents the GEE fit nearly perfectly for all but the very smallest tree. We found

TABLE 2. Parameter estimates of equation (10)

	Expansion locus		
	$\hat{\gamma}_i$	$E[\gamma_i]$	GEE
β_1	0.25 (0.118)	0.26 (0.120)	0.31 (0.115)
β_2	2.30 (0.012)	2.30 (0.012)	2.30 (0.011)
β_3	2.63 (0.057)	3.21 (0.067)	3.23 (0.066)
β_4	6.80 (0.020)	6.56 (0.021)	6.55 (0.021)
$\sigma^2_{b_2}$	0.023	0.023	0.016
$\sigma^2_{b_3}$	0.218	0.235	0.140
σ^2	4.8	4.9	5.1

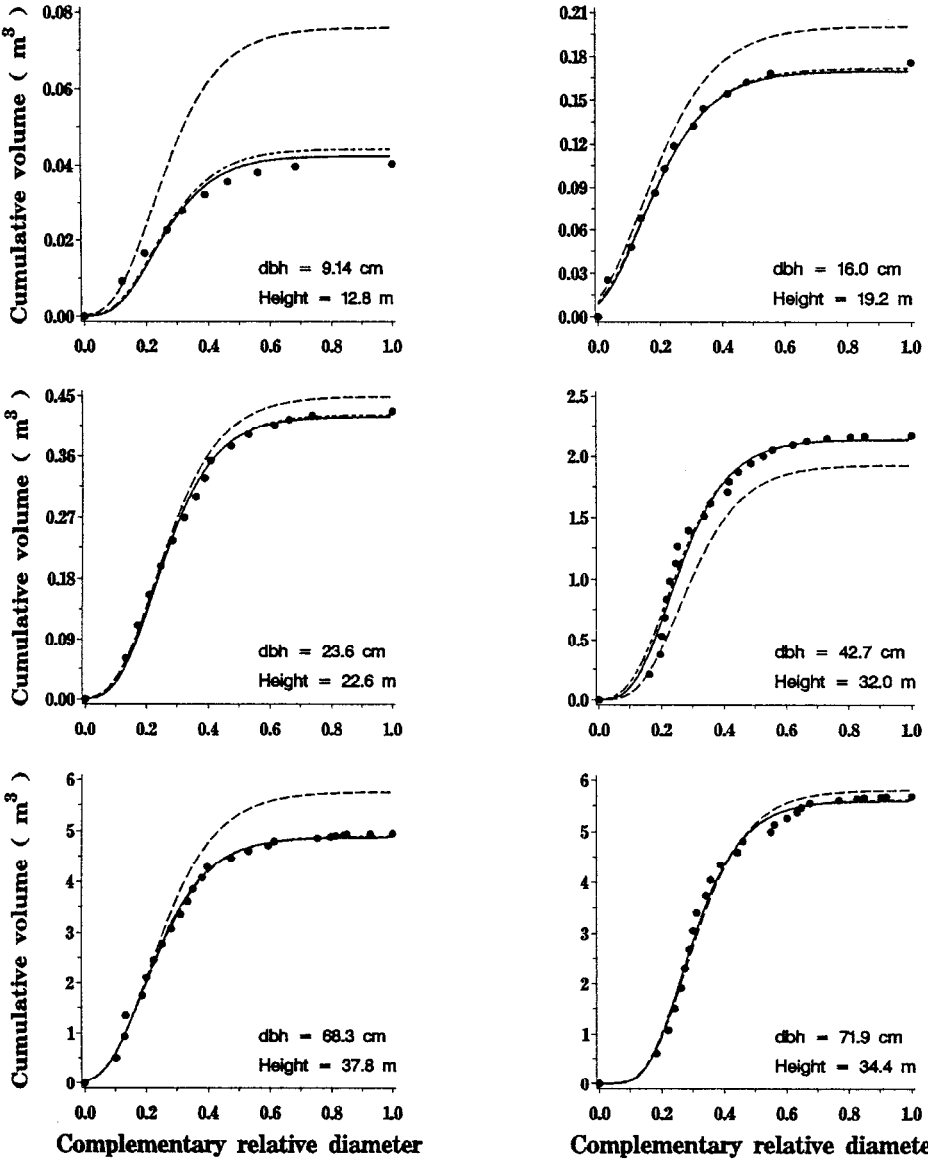

FIG. 2. Yellow poplar cumulative outer-bark bole volume: fitted profiles.

this to be true generally: as we scanned the fitted profiles, plotted individually for each of the 336 trees, the model fitted the very smallest trees slighly less well than it did the large and intermediate-sized trees, and it is only with these very small trees that there is a noticeable departure of the GEE fit from the REML fit. In no case was this departure so sizeable to raise concern about a systematic lack of fit.

For the sake of comparison, we have shown the fitted profile from the purely fixed effects model as the dashed line in Fig. 2. While the fixed effects model retains the sigmoidal shape of the cumulative bole volume profile, it fails to trace an individual tree's form, unless it coincides with the average trend in the population.

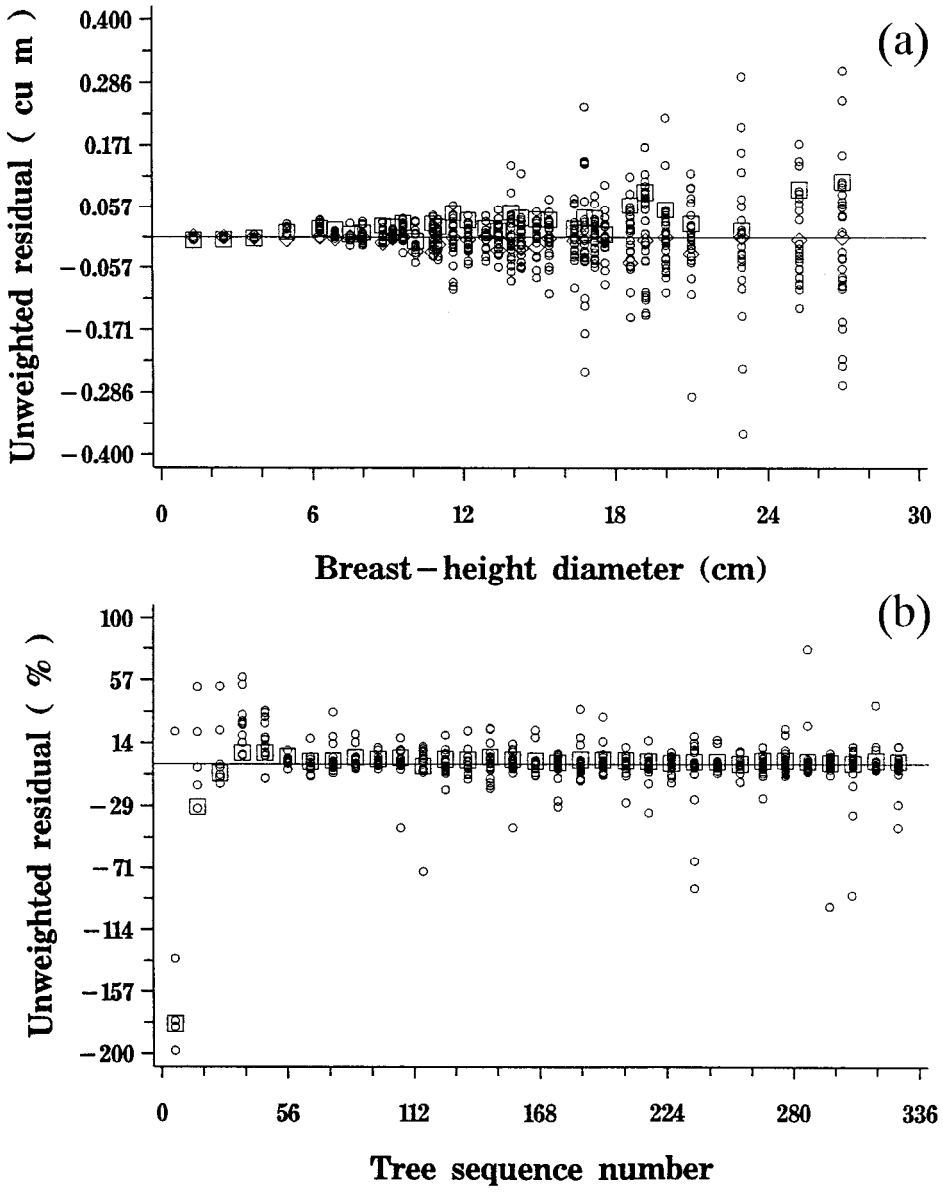

FIG. 3. Yellow poplar cumulative outer-bark bole volume residuals. Trees sequenced in order of breast height diameter D. REML fit of mixed model with b_2 and b_3 random, using

$$V = (b_1 + b_2 X)\exp(-b_3 t e^{b_4 t}/1000)$$

To date, diagnostic tools for non-linear mixed-effects models are generally lacking. Pinheiro *et al.* (1993) looked at box plots of raw residuals by subject. There is some question concerning the informativeness of raw residuals in a non-linear setting (cf. Seber & Wild, 1989, p. 174), because intrinsic curvature and parameter effects will affect the magnitude of residuals in a way that generally cannot be discerned. None the less, residual plots are an appealing diagnostic tool. In Fig. 3(a), we show the raw residual $(V_{id_j} - \hat{V}_{id_j})$ on the vertical axis versus $\hat{V}_{i0}$

on the horizontal axis for every 10th tree, after having sequenced the trees in order of increasing diameter at breast height D. The square symbol denotes the residual at the tip of the tree, i.e. the residual $V_{i0} - \hat{V}_{i0}$, which is highlighted because of the importance of predicting the total-bole volume well. The diamond symbol represents the residual at the base of the tree. The circles are the residuals at the intermediate points on the bole. A pattern of increasing dispersion with increasing tree size is evident. When the residuals are put on a relative basis, i.e. $100\%(V_{id_j} - \hat{V}_{id_j})/V_{id_j}$, as in Fig. 3(b), the dispersion pattern reverses. The relative residuals display the slightly poorer fit of the model to the very smallest trees.

Compared with differences among the alternative estimates of $\hat{\beta}_j$, the GEE estimates of the covariance parameters differed much more from the corresponding REML estimates. To examine the effect of these differences, we estimated

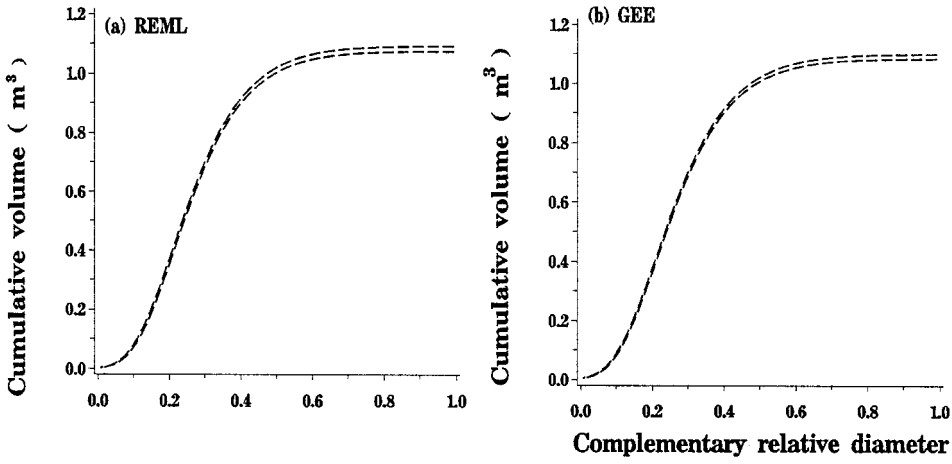

FIG. 4. Yellow poplar cumulative outer-bark bole volume for a tree with $D = 33.5$ cm and $H = 29$ m: 95% confidence bands for marginal response; (a) REML; (b) GEE.

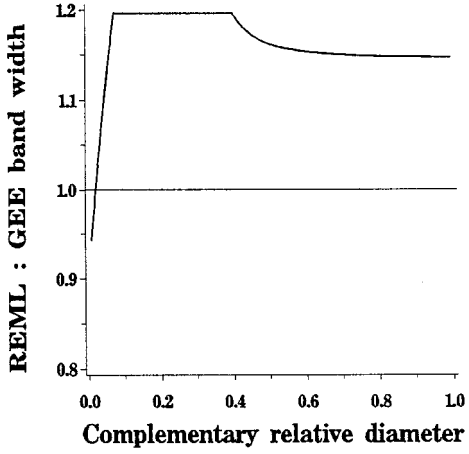

FIG. 5. Yellow poplar cumulative outer-bark bole volume for a tree with $D = 33.5$ cm and $H = 20$ m: ratio of 95% confidence band widths.

95% confidence interval (equation (11)) bands using both the REML and GEE parameter estimates for a hypothetical tree of size $D = 33.5$ cm and $H = 29$ m, which corresponds closely to the average D and H values among the 336 yellow poplar trees. As seen in Fig. 4, the corresponding confidence interval bands are nearly indistinguishable. The ratio of the REML interval width to that of the GEE interval is plotted in Fig. 5, from which we conclude that asymptotic inference about the estimated marginal response is affected inconsequentially by the choice of estimation procedure.

7 Discussion

As is evident from a comparison of the alternative estimates in Table 2 and from the graphical comparisons in Figs 2 and 4, there is little difference between REML and GEE estimation in the present setting. The GEE fit is always closer to the REML fit of the model that is linearized around $E[\gamma_i]$, presumably because our GEE implementation uses the same expansion locus.

The semi-parametric GEE approach requires only minimal assumptions: the correct specification of a mean model and consistent estimation of the covariance parameters. It is less restrictive in this sense than Gaussian-based likelihood estimation. The lack of distributional assumptions, however, practically restricts the choice of covariance parameter estimation to the method of moments. Moment estimators are neither unique nor necessarily useful. As seen from equation (10), $\hat{\Delta}$ can be non-positive definite. We have found such occurrences to be a useful indication of over-parameterized covariance structures, rather than a hindrance.

However, semi-parametric estimation is less computationally demanding, because the covariance parameters can be estimated in closed form after updates of the fixed effects are obtained, whereas they are obtained iteratively in the parametric implementation. One may opt to use GEE estimates as starting values for subsequent parametric estimation. As far as predictions are concerned, our results clearly show that little or nothing is gained in predictive capability by such an approach. The asymptotic basis for inference is also the same in either approach. We believe that semi-parametric estimation in non-linear mixed models constitutes a valid methodology in its own right, and is not a mere front-end vehicle for likelihood inference, although this usage is sensible at times.

The volume-ratio equation developed here is a substantial improvement over its precursors that have appeared in the foresty literature, because the random effects serve to individualize the fit of the model to each subject tree and account for the inter-tree variation, through the marginal covariance structure Δ. Moreover, because equation (1) mimics the inflection of the empirical cumulative bole volume profile (Fig. 1), it provides a superior fit than models that are uninflected.

In principle, the conditional variance could be modelled more generally as

$$\text{var}[\mathbf{Y}_i | \mathbf{Z}_i, \mathbf{W}_i, \gamma_i] = \sigma^2 \mathbf{R}_i$$

where $\mathbf{R}_i$ is specified in a manner that accounts for residual intra-individual correlation around $E[\mathbf{Y}_i | \mathbf{Z}_i, \mathbf{W}_i, \gamma_i]$. In our experience, the mixed-effects model effectively annihilates the within-subject correlation, in agreement with the observation by Jones (1990) that random subject effects may account for within-subject covariances, and vice versa. Gregoire *et al.* (1995) concluded similarly. However,

$\mathbf{R}_i$ could also be specified to account for the interindividual heteroscedasticity, such as is evident in Fig. 3, if warranted. In this application, the heteroscedasticity was not deemed severe enough to justify the added complexity that would be introduced.

The scientist (Beck, 1963) who felled, sectioned and measured the sectional diameters (outer-bark) and heights of the 336 yellow poplar trees also measured the corresponding under-bark diameters. As is frequently the case, there was an interest and need to fit a volume equation for the under-bark volume as well as for the outer-bark volume. Under-bark volume equations have always been fitted separately from outer-bark volume equations. In view of the strong correlation between under-bark and outer-bark volume, Gregoire *et al.* (1994) fitted the two volume-ratio equations jointly, using the pooled two-stage procedure proposed by Davidian and Giltinan (1993). A similar tactic could be employed with the approach pursued in the present paper, but such work remains to be done.

REFERENCES

Amateis, R. L. & Burkhart, H. E. (1987) Cubic-foot volume equations for loblolly pine trees in cutover, site-prepared plantations, *Southern Journal of Applied Forestry*, 11, pp. 190–192.

Avery, T. E. & Burkhart, T. E. (1994) *Forest Measurements*, 4th edn (New York, McGraw-Hill).

Bailey, R. L. (1994) A compatible volume-taper model based on the Schumacher and Hall generalized form factor volume equation, *Forest Science*, 40, pp. 303–313.

Beck, D. E. (1963) Cubic-foot volume tables for yellow poplar in the southern Appalachians, *USDA Forest Service, Research Note SE-16*.

Burkhart, H. E. (1977) Cubic-foot volume of loblolly pine to any merchantable top limit, *Southern Journal of Applied Forestry*, 1, pp. 7–9.

Davidian, M. & Gallant, A. R. (1993) The nonlinear mixed effects model with a smooth random effects density, *Biometrika*, 80, pp. 475–488.

Davidian, M. & Giltinan, D. M. (1993) Some general estimation methods for nonlinear mixed-effects models, *Journal of Biopharmaceutical Statistics*, 3, pp. 23–55.

Davidian, M. & Giltinan, D. M. (1995) *Nonlinear Models for Repeated Measurement Data* (New York, Chapman & Hall).

Diggle, P. J., Liang, K.-Y. & Zeger, S. L. (1994) *Analysis of Longitudinal Data* (Oxford, Oxford University Press).

Godambe, V. P. (1960) An optimum property of regular maximum likelihood estimation, *The Annals of Mathematical Statistics*, 31, pp. 1208–1211.

Golden, M. S., Knowe, S. A. & Tuttle, C. L. (1982) Cubic-foot volume for yellow-poplar in the hilly coastal plain of Alabama, *Southern Journal of Applied Forestry*, 6, pp. 167–171.

Gregoire, T. G. & Schabenberger, O. (1996) Nonlinear mixed-effects modeling of cumulative bole volume with spatially correlated within-tree data, *Journal of Agricultural, Biological, and Environmental Statistics*, 1, p. xx.

Gregoire, T. G., Schabenberger, O. & Rennolls, K. (1994) Tree utilization models with spatially correlated errors and random effects, paper presented at *Joint Statistical Meeting (session 264)*, 14–18 August 1994, Toronto.

Gregoire, T. G., Schabenberger, O. & Barrett, J. P. (1995) Linear modelling of irregularly spaced, unbalanced, longitudinal data from permanent plot measurements, *Canadian Journal of Forest Research*, 25, pp. 137–156.

Harville, D. A. & Carriquiry, A. L. (1992) Classical and Bayesian production as applied to an unbalanced mixed linear model, *Biometrics*, 48, pp. 987–1003.

Jones, R. H. (1990) Serial correlation or random subject effects?, *Communications in Statistics—Simulation*, 19, pp. 1105–1123.

Jones, R. H. (1993) *Longitudinal Data with Serial Correlation: A State–space Approach* (New York, Chapman & Hall).

Knoebel, B. R., Burkhart, H. E. & Beck, D. E. (1984) Stem volume and taper functions for yellow-poplar in the southern Appalachians, *Southern Journal of Applied Forestry*, 8, pp. 185–188.

Laird, N. M. & Ware, J. H. (1982) Random-effects models for longitudinal data, *Biometrics*, 38, pp. 963–974.

LIANG, K.-Y. & ZEGER, S. L. (1986) Longitudinal analysis using generalized linear models, *Biometrika*, 73, pp. 13–22.

LINDSTROM, M. J. & BATES, D. M. (1990) Nonlinear mixed effects models for repeated measures data, *Biometrics*, 46, pp. 673–687.

NEWBERRY, J. D. & BURK, T. E. (1985) S_B distribution-based models for individual tree merchantable volume–total volume ratios, *Forest Science*, 31, pp. 389–398.

PINHEIRO, J. C., BATES, D. M. & LINDSTROM, M. J. (1993) Nonlinear mixed effects classes and methods for *S*, Department of Statistics, *Technical Report 906*, University of Wisconsin, Madison.

PINHEIRO, J. C., BATES, D. M. & LINDSTROM, M. J. (1994) Model building for nonlinear mixed effects models, Department of Statistics, *Technical Report 931*, University of Wisconsin, Madison.

SCHABENBERGER, O. (1995) Nonlinear mixed effects growth models for repeated measures in ecology, *Proceedings of the 1994 Joint Statistical Meetings, Section on Statistics and the Environment*, 14–18 August, Toronto.

SCHABENBERGER, O. & GREGOIRE, T. G. (1995) A conspectus of estimating function theory and its applicability to recurrent modeling issues in forest Biometry, *Silva Fennica*, 29, pp. 49–70.

SEBER, G. A. F. & WILD, C. J. (1989) *Nonlinear Regression* (New York, Wiley).

SHEINER, L. B. & BEAL, S. L. (1980) Evaluation of methods for estimating population pharmacokinetic parameters. I. Michaelis–Menton model: routine clinical pharmacokinetic data, *Journal of Pharmacokinetics and Biopharmaceutics*, 8, pp. 553–571.

SPURR, S. H. (1962) *Forest Inventory* (New York, Ronald Press).

VAN DEUSEN, P. C., SULLIVAN, A. D. & MATNEY, T. G. (1981) A prediction system for cubic foot volume of loblolly pine applicable through much of its range, *Southern Journal of Applied Forestry*, 5, pp. 186–189.

VONESH, E. F. & CARTER, R. L. (1992) Mixed-effects nonlinear regression for unbalanced repeated measures, *Biometrics*, 48, pp. 1–17.

WOLFINGER, R. (1993) Laplace's approximation for nonlinear mixed models, *Biometrika*, 80, pp. 791–795.

WOLFINGER, R. & O'CONNELL, M. (1993) Generalized linear mixed models: a pseudo-likelihood approach, *Journal of Statistical Computing and Simulation*, 48, pp. 233–243.

ZEGER, S. L., LIANG, K.-Y. & ALBERT, P. S. (1988) Models for longitudinal data: a generalized estimating equation approach, *Biometrics*, 44, pp. 1049–1060.