A stochastic differential equation model for the height growth of forest stands*

Oscar García

Forest Research Institute
Rotorua, New Zealand

Abstract

A model and estimation procedure for predicting the height growth of even-aged forest stands was developed as part of a methodology for modelling stand growth in forest plantations (García 1979). The data consists of heights measured at several ages in a number of sample plots. The ages and number of measurements may differ among plots and the measurements may not be evenly spaced in time. The height growth model is assumed to have some parameters which are common to all plots and others which are specific to each plot. In addition to random environmental variation affecting the growth, there are random measurement errors.

The height growth is modelled by a stochastic differential equation in which the deterministic part is equivalent to the Bertalanffy-Richards model (von Berfealanffy 1949, 1957, Richards 1959). The model also includes a component representing the measurement errors. Explicit expressions for the likelihood function are obtained.

All the parameters are estimated simultaneously by maximum likelihood. A modified Newton method which exploits the special structure of the problem is used.

Some experience with the model and estimation procedure is discussed.

Keywords: Stochastic differential equations, growth, von Bertalanffy model, Richards model, maximum likelihood, estimation, forestry.

1 Introduction

A model and an estimation procedure for predicting the height growth of even-aged forest stands has been developed, as part of a methodology for

*Presented at the 5th Australian Statistical Conference. Sydney, August 1980

modelling growth in forest plantations. The techniques developed may also be useful in other applications.

The essential characteristics of the problem may be described as follows. Heights are measured at several ages in a number of sample plots. The ages and number of measurements may differ among plots and the measurements may not be evenly spaced in time. The height growth model is assumed to have some parameters which are common to all plots and others which are specific to each plot. In addition to environmental fluctuations affecting the growth, random measurement errors may be present.

After a brief review of the uses and methods of height growth prediction in forestry, the proposed growth model is presented in section 3. The model consists of a stochastic differential equation related to the Bertalanffy-Richards growth model, and a measurement error component. Explicit expressions and an efficient computational procedure for the likelihood function are obtained.

In section 4 a method for the simultaneous maximum likelihood estimation of common and specific parameters is presented. The log-likelihood function is maximized using a modified Newton method. The special structure of the problem is exploited in order to handle the very large number of variables involved in the optimization.

The approach has been successfully implemented, and some computational experience is reported in section 5.

2 Height growth and site index

The prediction of height growth in even-aged forest stands is used for two related purposes: as a component of stand growth models and for assessing site quality. By the height of a stand we will understand some measure of "top height", such as the average height of the 100 largest trees per hectare. Top height has the advantage over mean height of being little affected (within limits) by manipulation of the stand density through low thinnings, where mainly small and malformed trees are extracted.

Stand growth models are used to predict the development of forest stands subjected to different silvicultural regimes. In addition to equations for predicting height growth, stand models include relationships for predicting other variables such as mean diameter and natural mortality. Since top height can be considered as approximately independent of these other variables, the height prediction component constitutes a self-contained submodel, and can be developed separately (García 1979).

As mentioned before, height growth is also used for assessing the potential productivity of forest land (site quality). Top height is used in preference to more direct measures, such as volume production, because it is more easily measured and is relatively independent of variations in stand density and thinning treatments[1]. The Site Index is defined as the height that a stand would have at a specified age (20 years for radiata pine in New Zealand). A family of height over age curves (site index curves) is used to estimate the site index given the age and height of a stand (see figures 1 and 2). See Spurr (1952), Jones (1969), and Carmean (1975) for a review of this and other approaches to site quality evaluation.

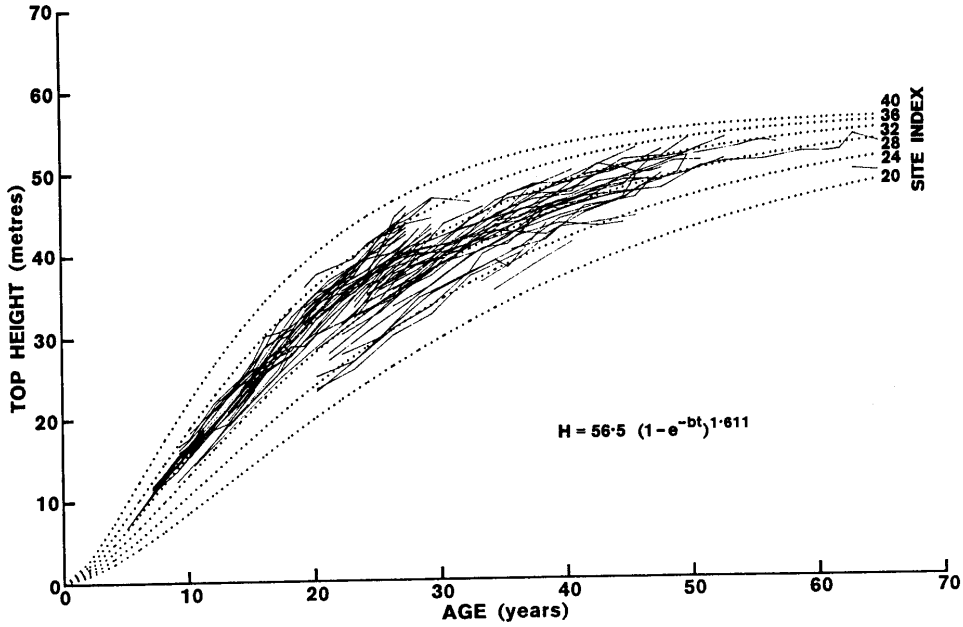

Figure 1:

Although there has been some discussion about the need for different

[1]In New Zealand the top height in forest plantations is usually computed as follows. In a sample plot, typically of 0.05 to 0.2 hectares, all trees are measured for diameter at breast height (d.b.h.) and heights are measured in a sample of 10–15 trees. A curve of height over d.b.h. is fitted using the height sample trees. Then the quadratic mean d.b.h. of the proportion of trees in the plot which corresponds to the 100 largest-diameter trees per hectare is computed. The height given by the height-d.b.h. curve for this "mean top d.b.h." is the "mean top height" or "top height". It may be mentioned that there are some standing questions about the effect of different plot sizes with procedures of this type (Fries 1974, Matérn 1976, Rennolls 1978).

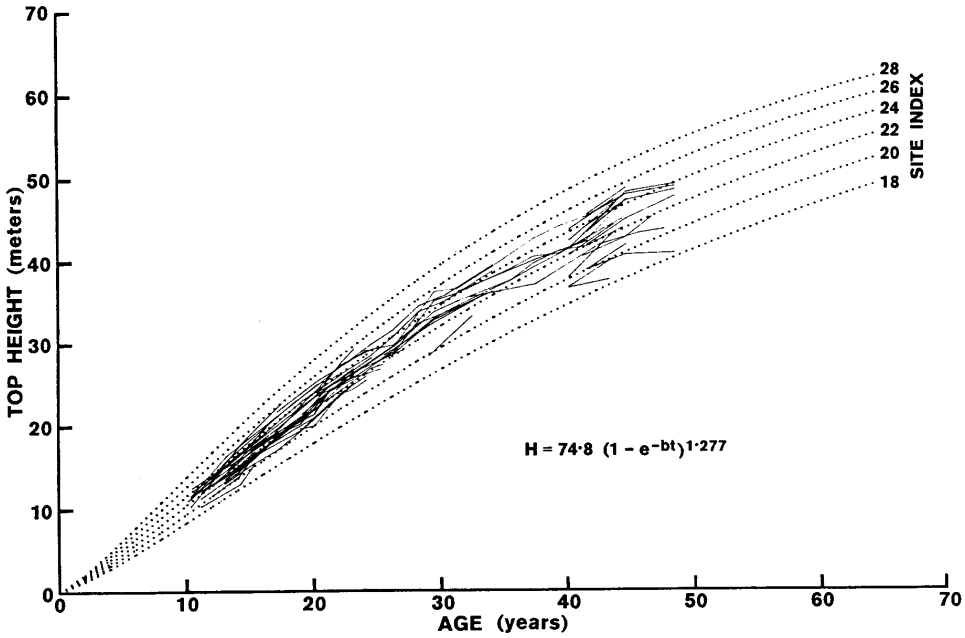

Figure 2:

estimation procedures for height prediction and for site classification (Curtis, De Mars, Herman 1974), the problem is essentially the same. Site index curves are height growth curves. The site index concept, however, depends on the assumption that variations in the height growth pattern, for a given species and region, can be described by a one-parameter family of curves.

A large variety of procedures for developing site index (height-age) curves has been used, and a review of these will not be attempted here. Site index curves can be obtained by cross-sectional analysis using a large number of single height-age measurements on different sample plots. This approach has severe limitations and, whenever possible, curves based on a number of consecutive measurements on each plot are preferred (Spurr 1952). These sequences of measurements may be obtained by repeated measurements on permanent sample plots, or by stem analysis, where the past growth of trees is reconstructed from the annual growth ring patterns up the tree. In some tree species the data can also be obtained from the position of branch whorls marking the course of annual height growth.

Some methods for deriving site index curves do not use height-age equations. An example of this is an interesting non-parametric nefchod developed by Tveite (1969). Among those using equations the procedure used

by Burkhart and Tennent (1977) is typical. They first estimate .the site index for each plot by interpolation or extrapolation (only plots with measurements close to the index age are used). Then an equation expressing the height as a function of age and site index is fitted to the data using non-linear least-squares. Bailey and Clutter (1974) use an approach based on the idea of a one-parameter family of curves, which does not depend on the arbitrary index age and which allows the use of all the data available. Some concern has also been expressed about using least-squares with repeated measurements (Sullivan and Clutter 1972, Sullivan and Reynolds 1976, Ferguson and Leech 1978).

Here a model and estimation procedure for height growth based on a stochastic differential equation and maximum likelihood estimation is proposed. Some advantages over existing methodsd are: (a) the error structure generated by repeated measurements is recognized; (b) different parametrizations of the height-age curves can be tried and compared, including multi-parametric families of curves; (c) atypical variation in early growth caused by frosts, weed competition, establishing techniques, etc., can be handled by shifting or leaving free the origin of the curves.

3 Model

3.1 Linear differential equations, power transformations, and the Bertalanffy-Richards model

There are advantages in taking differential equations as the basis for growth models, specifically equations of the form:

$$\frac{\mathrm{d}H}{\mathrm{d}t} = f(H) \,, \tag{1}$$

where H is the height (or any other size variable) and t is time (Hottelling 1927, García 1979). Deterministic models are discussed first.

One of the simplest differential equations is the linear equation

$$\frac{\mathrm{d}H}{\mathrm{d}t} = b(a - H) \,. \tag{2}$$

This is the well-known Mitscherlich or monomolecular model, and integrating it produces the growth function

$$H = a[1 - (1 - \frac{H_0}{a})\mathrm{e}^{-b(t-t_0)}] \,, \tag{3}$$

where H_0 is the height at time t. The height tends to an upper asymptote a, and there is no inflection point.

Much greater flexibility is attained by substituting a power transformation H^c for H:

$$\frac{\mathrm{d}H^c}{\mathrm{d}t} = b(a^c - H^c) \; . \tag{4}$$

Calculating the derivative on the left-hand side, this can be written as

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \frac{b}{c} H[(\frac{a}{H})^c - 1]$$

or

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \eta H^m - \kappa H \; , \tag{5}$$

a model proposed by von Bertalanffy (1949, 1957) and studied by Richards (1959). (Conversely, it may be observed that von Bertalanffy's model is a Bernoulli differential equation, for which the standard method of integration involves its transformation to a linear differential equation through a power transformation of H). The integrated form is (cf. equation 3)

$$H = a[1 - (1 - \frac{H_0^c}{a^c})\mathrm{e}^{-b(t-t_0)}]^{\frac{1}{c}} \; . \tag{6}$$

This function has generally a sigmoid shape, with upper asymptote a and an inflection point at $H = a(l - c)^{1/c}$. In most applications $t_0 = H_0 = 0$.

The Bertalanffy-Richards model is very flexible, including as special cases several well-known growth functions such as the Mitscherlich ($c = 1$), logistic ($c = -1$), exponential ($c = 1$, $a = 0$), and Gompertz (limit when $c \to 0$, see below) (Richards 1959). It has been frequently used for site index curves (e.g., Burkhart and Tennent 1977) and for modelling the development of other forest variables (Pienaar and Turnbull 1973). It has also been used for describing animal growth by von Bertalanffy (1949, 1957), and in fisheries research by Beverton and Holt (1957) and Chapman (1961).

Note that (4) – (6) breaks down for $c = 0$. This value can be included if instead of H^c we use the transformation

$$y = \begin{cases} \frac{1}{c}(H^c - a^c) & \text{if } c \neq 0 \\ \ln(H/a) & \text{if } c = 0 \; , \end{cases} \; , \tag{7}$$

and

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -by \; . \tag{8}$$

Equation (7), as a function of c, is continuous at $c = 0$, and it is a modification of the Box-Cox transformation (Box and Cox 1964) suggested by

Schlesselman (1971). Equations (7) and (8) are equivalent to (4)–(6) for $c \neq O$, and to the Gompertz model for $c = 0$. We will have no need to consider the case $c = 0$ because for height-age curves c is normally between 0.3 and 1, but the form (7)–(8) is slightly more convenient in some of the developments that follow.

It may be mentioned that, if needed, a model even more flexible may be obtained by adding to (4) a term in H^{2c}. A Riccati-type differential equation in H^c is obtained, which still can be integrated to obtain an explicit form for the $H - t$ equation. Levenbach and Reuter (1976) use a Riccati equation in H for the forecasting of time series.

3.2 Stochastic components

We adopt the Bertalanffy-Richards model defined by (4), (5), or (7) and (8) to describe the most probable course of the top height development of a forest stand. Even if we are interested only in point estimates of height, and not in the stochastic aspects of height growth, some assumptions about the nature of the random deviations from the model are necessary for a rational selection of parameter estimation procedures. For example, the usual approach of fitting the integrated equation by non-linear least-squares can be shown to have some optimality properties if the deviations from the curve are independently distributed with zero mean and common variance. It is widely acknowledged, however, that repeated measurements on an individual or sample plot are correlated, and that the deviations tend to increase with time (e.g., Hottelling 1927, Sullivan and Clutter 1972, Sullivan and Reynolds 1976, Ferguson and Leech 1978, Sandland and McGilchrist 1979).

As pointed out by Hottelling (1927), apart from possible measurement errors which may be considered as independent, deviations from the most probable growth curve may be seen as caused by the accumulative effect of numerous random disturbances operating for brief periods. It is then natural to attempt modelling the process through stochastic differential equations (Sandland and McGilchrist 1979, García 1979).

We modify then (4) by adding a Brownian motion or Wiener process which represents the effect of the fluctuating environment (to simplify the notation we assume $c \neq 0$):

$$\mathrm{d}H^c(t) = b[a^c - H^c(t)]\,\mathrm{d}t + \sigma(t)\,\mathrm{d}w(t)\ . \tag{9}$$

This is a stochastic differential equation (SDE), where w is a Wiener process and σ is some function of age, possibly containing unknown parameters. Essentially, this means that the variation or error in H^c accumulated over

a short time interval is normally distributed with zero mean and variance increasing with the interval length, and that errors for non-overlapping time intervals are independent (Karlin 1966, Gihman and Skorohod 1972). In the present application we will assume that a is a constant, except possibly for a few years after planting (and before the first measurement) where we might expect a larger variation to occur. A multivariate generalization of this model is discussed by García (1979).

In addition to the environmental variation, we allow for measurement or observation errors. It is mathematically convenient to assume that for a given sample plot the observed heights h_i at ages t_i are such that

$$h_i^c = H^c(t_i) + \epsilon_i \; ; \quad i = 1, \ldots, n \; , \tag{10}$$

where the ϵ_i are independent normal variables with zero means and variance η^2. This implies that the variance of h_i is approximately $\eta^2/(\frac{\mathrm{d}h_i^c}{\mathrm{d}h_i})^2$, that is, $(\eta h_i^{1-c})^2$. In our application c is typically around 0.7, which would make the standard deviation of h_i proportional to $h_i^{1-c} = h_i^{0.3}$, which seems reasonable.

Any of the parameters, possibly after reparametrization, may have the same value for all plots or different values for different plots.

We will also assume statistical independence between plots.

The stochastic aspects of the model are a compromise between realism and mathematical tractability. Several over-simplifications and assumptions that are not completely satisfactory may be mentioned. The error term in (9) could conceivably cause the height to decrease with time or even cause H^c to become negative. The additive effect of environmental fluctuations in (9) is questionable; perhaps a multiplicative effect would be more realistic[2]. Environmental fluctuations are not necessarily serially independent (Tomlinson 1976). More important, height increments for the same year in different plots are not independent because they are affected by similar weather conditions. Nevertheless, the stochastic structure of the model is intended to be used only for the development of estimation procedures. The widespread and successful use of linear models in statistics suggests that

[2]An alternative to (9) which is also tractable but has not been tried is (using the notation of (7) and (8)):

$$\mathrm{d}y(t) = -by(t)\,\mathrm{d}t + \sigma(t)y(t)\,\mathrm{d}w(t) \; ,$$

or

$$\mathrm{d}\ln|y(t)| = -b\,\mathrm{d}t + \sigma(t)\,\mathrm{d}w(t) \; .$$

in most cases the performance of estimators is not too badly affected by moderate deviations from the distributional assumptions in which they are based.

3.3 The likelihood function

Many estimation methods require the knowledge of the likelihood function, i.e., the probability density of the observations, considered as a function of the parameters. Because of the assumed independence between plots, the likelihood will be the product of the densities for each of the sample plots. Then we first find the probability distribution of the observations for one plot.

The model is:

$$\begin{aligned} \mathrm{d}H^c(t) &= b[a^c - H^c(t)]\,\mathrm{d}t + \sigma(t)\,\mathrm{d}w(t) && (9') \\ H(t_0) &= H_0 \\ h_i^c &= H^c(t_i) + \epsilon_i && (10') \\ \epsilon_i &\sim \mathrm{N}(0, \eta^2)\,, \quad \mathrm{cov}(\epsilon_i, \epsilon_j) = 0 \text{ if } i \neq j\ , \end{aligned}$$

where w is a Wiener process and h_i is the height observation at age t_i, $t_1 < t_2 < \ldots < t_n$[3]. The linear SDE (9) is a special case of those considered by Erickson (1971) and by Gihman and Skorohod (1972, p. 36–38). It may be simplified using (7). Integrating between t_0 and t_i and using (10) we get

$$h_i^c = a^c - (a^c - H_0^c)\mathrm{e}^{-b(t_i - t_0)} + \delta_i + \epsilon_i\ , \tag{11}$$

where

$$\delta_i = \int_{t_0}^{t_i} \mathrm{e}^{-b(t_i - s)}\sigma(s)\,\mathrm{d}w(s) \tag{12}$$

is a normal random variable with zero mean (compare with (6)). Heuristically, this result may be obtained by integrating (9) as an ordinary linear differential equation, taking w as a fixed function of t. The covariances for the δ_i are

$$\mathrm{cov}(\delta_i, \delta_j) = \mathrm{E}(\delta_i, \delta_j) = \int_{t_0}^{\min\{t_i, t_j\}} \exp[-b(t_i + t_j - 2s)]\sigma^2(s)\ \mathrm{d}s\ . \tag{13}$$

Equations (10), (11), and (13) completely define the joint (normal) distribution of the h_i^c The joint density for the observations h_i (and hence the

[3]The measurements are not necessarily evenly spaced in time. Moreover, the t_i may not be integers, as when seasonal adjustments (García 1979) or artificial time scales (Nelder, Austin, Bleasdale, Salter, 1960) are used.

likelihood function) can be obtained by multiplying the density of the h_i^c and the Jacobian of the transformation. A simpler expression can be found, however.

Let

$$z_i = (a^c - h_i^c) - \mathrm{e}^{-b(t_i - t_{i-1})}(a^c - h_{i-1}^c) \ , \qquad (14)$$
$$i = 1, \ldots, n, \text{ with } h_0 = H_0 \ .$$

Then, from (11) and (12), defining $\epsilon_0 = 0$,

$$\begin{aligned} z_i &= \delta_i + \epsilon_i - \mathrm{e}^{-b(t_i - t_{i-1})}(\delta_{i-1} - \epsilon_{i-1}) \\ &= \int_{t_{i-1}}^{t_i} \mathrm{e}^{-b(t_i - s)}\sigma(s)\ \mathrm{d}w(s) + \epsilon_i - \mathrm{e}^{-b(t_i - t_{i-1})}\epsilon_{i-1} \ . \end{aligned}$$

It follows that the z_i are normal random variables with

$$\mathrm{E}(z_i) = 0$$

and

$$\mathrm{cov}(z_i, z_j) = \begin{cases} \int_{t_0}^{t_1} \mathrm{e}^{-2b(t_1 - s)}\sigma^2(s)\ \mathrm{d}s + \eta^2 \ , & i = j = 1 \\ \int_{t_{i-1}}^{t_i} \mathrm{e}^{-2b(t_i - s)}\sigma^2(s)\ \mathrm{d}s + [1 + \mathrm{e}^{-2b(t_i - t_{i-1})}]\eta^2 \\ \quad - \mathrm{e}^{-b}|t_i - t_j|\eta^2 \ , & |i - j| = 1 \\ 0 \ , & \text{otherwise} \end{cases} \qquad (15)$$

In the special case where $\sigma(t)$ is a constant for $t > \tau$ and some $\tau \leq t_1$, $\sigma^2(t) = 2b\sigma^2$ say, and $\sigma^2(t) = 2b\sigma^2 + \xi^2(t)$ for $t \leq \tau$, the environmental component in (15) is given by

$$\int_{t_{i-1}}^{t_i} \mathrm{e}^{-2b(t_i - s)}\sigma^2(s)\,\mathrm{d}s = \begin{cases} [1 - \mathrm{e}^{-2b(t_i - t_{i-1})}]\sigma^2 + \sigma_0^2 & \text{for } i = 1 \\ [1 - \mathrm{e}^{-2b(t_i - t_{i-1})}]\sigma^2 & \text{for } i > 1 \ , \end{cases} \qquad (16)$$

where $\sigma_0^2 = \int_{t_0}^{\tau} \mathrm{e}^{-2b(t_1 - s)}\xi^2(s)\ \mathrm{d}s$.

The joint density of the observations is then

$$f(h_1, \ldots, h_n) = (2\pi)^{-\frac{1}{2}n}|\mathrm{C}|^{-\frac{1}{2}} \exp(-\frac{1}{2}\boldsymbol{z}'\mathrm{C}^{-1}\boldsymbol{z})\mathrm{J} \ , \qquad (17)$$

were $\boldsymbol{z} = (z_1, \ldots, z_j)'$, C is the covariance matrix with elements $c_{ij} = \mathrm{cov}(z_i, z_j)$ given by (15), and J is the Jacobian of the transformation (14):

$$\mathrm{J} = \left|\det[\frac{\partial z_i}{\partial h_j}]\right| = |c|^n(\prod_{i=1}^{n} h_i)^{c-1} \ . \qquad (18)$$

It is often convenient to work with the negative log-likelihood:

$$-\ln L = \frac{1}{2}\sum[n\ln(2\pi) + \ln|\mathrm{C}| + \boldsymbol{z}'\mathrm{C}^{-1}\boldsymbol{z} - 2n\ln|c| + 2(1-c)\sum_{i=1}^{n}\ln h_i] \ . \qquad (19)$$

The sum is over the N plots.

3.4 Computing the likelihood

An efficient procedure for evaluating $\ln|\mathrm{C}| + \boldsymbol{z}'\mathrm{C}^{-1}\boldsymbol{z}$ in (19) has been obtained by using Cholesky factorization and exploiting the fact that the matrix C is tridiagonal.

The symmetric positive-definite matrix C can be factorized as

$$\mathrm{C} = \mathrm{LDL}' \,, \tag{20}$$

where D is a diagonal matrix and L is lower triangular with ones on the main diagonal (Martin, Peters, Wilkinson,1971). Moreover, due to C being tridiagonal, (20) takes the form

$$\begin{bmatrix} q_1 & p_2 & 0 & \cdots & 0 \\ p_2 & q_2 & p_3 & \cdots & 0 \\ 0 & p_3 & q_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & q_n \end{bmatrix} = \tag{21}$$

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ r_2 & 1 & 0 & \cdots & 0 \\ 0 & r_3 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} s_1 & 0 & 0 & \cdots & 0 \\ 0 & s_2 & 0 & \cdots & 0 \\ 0 & 0 & s_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & s_n \end{bmatrix} \begin{bmatrix} 1 & r_2 & 0 & \cdots & 0 \\ 0 & 1 & r_3 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

where $q_i = \mathrm{cov}(z_i, z_i)$ and $p_i = \mathrm{cov}(z_i, z_{i-1})$ are given in (15),

The determinant of C can be written as

$$|\mathrm{C}| = |\mathrm{LDL}'| = |\mathrm{L}||\mathrm{D}||\mathrm{L}| = |\mathrm{D}| = \prod_{i=1}^{n} s_i \,,$$

and

$$\ln|\mathrm{C}| = \sum_{i=1}^{n} \ln s_i \,.$$

Also,

$$\boldsymbol{z}'\mathrm{C}^{-1}\boldsymbol{z} = \boldsymbol{z}'(\mathrm{LDL}')^{-1}\boldsymbol{z} = (\mathrm{L}^{-1}\boldsymbol{z})'\mathrm{D}^{-1}(\mathrm{L}^{-1}\boldsymbol{z}) \,.$$

Let $\mathrm{L}^{-1}\boldsymbol{z} = \boldsymbol{u}$. Then

$$\boldsymbol{z}'\mathrm{C}^{-1}\boldsymbol{z} = \boldsymbol{u}'\mathrm{D}^{-1}\boldsymbol{u} = \sum_{i=1}^{n} u_i^2/s_i \,,$$

We can then write

$$Q = \ln|\mathrm{C}| + \boldsymbol{z}'\mathrm{C}^{-1}\boldsymbol{z} = \sum_{i=1}^{n}(\ln s_i + u_i^2/s_i)\ . \tag{22}$$

From (21),

$$\left\{\begin{array}{lcl} q_1 &=& s_1 \\ q_i &=& s_{i-1}r_i^2 + s_i \\ p_i &=& s_{i-1}r_i\ ; \quad i = 2,\ldots,n\ . \end{array}\right.$$

It is easy to see then that the s_i can be computed recursively from the elements of C:

$$\left\{\begin{array}{lcl} s_1 &=& q_1 \\ r_i &=& p_i/s_{i-1} \\ s_i &=& q_i - r_i p_i\ ; \quad i = 2,\ldots,n\ . \end{array}\right. \tag{23}$$

Analogously, from $\boldsymbol{z} = \mathrm{L}\boldsymbol{u}$,

$$\left\{\begin{array}{lcl} u_1 &=& z_1 \\ u_i &=& z_i - r_i u_{i-1}\ ; \quad i = 2,\ldots,n\ . \end{array}\right. \tag{24}$$

Summarizing, $Q = \ln|\mathrm{C}| + \boldsymbol{z}'\mathrm{C}^{-1}\boldsymbol{z}$ can be computed recursively as follows:[4]

$$\begin{array}{rcl} s &\leftarrow& q_1 \\ u &\leftarrow& z_1 \\ Q &\leftarrow& \ln s + u^2/s \\ & \text{For} & i = 1 \text{ to } n: \\ & & r \leftarrow p_i/s \\ & & s \leftarrow q_i - rp_i \\ & & u \leftarrow z_i - ru \\ & & Q \leftarrow Q + \ln s + u^2/s. \end{array} \tag{25}$$

3.5 Derivatives

The method used to estimate the parameters of the model requires the computation of first and second derivatives of the log-likelihood with respect to the parameters. That is, for each plot we need the derivatives of

$$F = n\ln(2\pi) + \ln|\mathrm{C}| + \boldsymbol{z}'\mathrm{C}^{-1}\boldsymbol{z} - 2\ln \mathrm{J}\ . \tag{26}$$

[4] Although it requires extra computing effort, the accumulation of the sum of $\ln s$ instead of the product of s seems preferable in order to avoid potential numerical problems.

Explicit formulas can be obtained using matrix derivatives (Dwyer 1967, Neudecker 1969). Using subindices to denote derivatives with respect to parameters θ and μ, we get

$$F_\theta = \mathrm{tr}(\mathrm{C}^{-1}\mathrm{C}_\theta) - \boldsymbol{z}'\mathrm{C}^{-1}\mathrm{C}_\theta\mathrm{C}^{-1}\boldsymbol{z} + 2\boldsymbol{z}'\mathrm{C}^{-1}\boldsymbol{z}_\theta - 2(\ln \mathrm{J})_\theta \,, \tag{27}$$

$$\begin{aligned} F_{\theta\mu} = {} & \mathrm{tr}(\mathrm{C}^{-1}\mathrm{C}_{\theta\mu}) - \mathrm{tr}(\mathrm{C}^{-1}\mathrm{C}_\theta\mathrm{C}^{-1}\mathrm{C}_\mu) - \boldsymbol{z}'\mathrm{C}^{-1}\mathrm{C}_{\theta\mu}\mathrm{C}^{-1}\boldsymbol{z} \\ & + 2\boldsymbol{z}'\mathrm{C}^{-1}\mathrm{C}_\theta\mathrm{C}^{-1}\mathrm{C}_\mu\mathrm{C}^{-1}\boldsymbol{z} - 2\boldsymbol{z}'\mathrm{C}^{-1}\mathrm{C}_\theta\mathrm{C}^{-1}\boldsymbol{z}_\mu \\ & - 2\boldsymbol{z}'\mathrm{C}^{-1}\mathrm{C}_\mu\mathrm{C}^{-1}\boldsymbol{z}_\theta - 2\boldsymbol{z}'_\theta\mathrm{C}^{-1}\boldsymbol{z}_\mu - 2\boldsymbol{z}'\mathrm{C}^{-1}\boldsymbol{z}_{\theta\mu} - 2(\ln \mathrm{J})_{\theta\mu} \,. \end{aligned} \tag{28}$$

These formulas, however, are computationally inefficient, except perhaps in programs written in interpretative languages with fast built-in matrix operations, such as APL and some versions of BASIC. They have been used during program development for checking the coding of the methods described below.

A second possibility, used in an earlier version of the computer program, is to derive from (23) and (24) recursive expressions for the derivatives. This resulted in a reasonably efficient program, but the expressions were somewhat cumbersome to use. The formulas are rather complex, and most terms are not used in the computation of the derivatives with respect to individual parameters. Program changes require extensive recoding.

The method used in the current version of the program is as follows. First, a subroutine for computing (26), based on (25), was written. This subroutine is needed in the program. Then this subroutine was used as a framework for coding the subroutine which computes the derivatives. This was achieved by inserting, after each assignment statement, code for computing the derivatives of the variable on the left-hand side with respect to all the parameters involved in the statement. These derivatives are generally functions of other derivatives computed earlier in the subroutine.

For example, suppose that we are considering a statement

```
C = A * B
```

and we have previously computed `A3`, the derivative of A with respect to parameter 3, `B1`, `B3`, `B11` and `B33`, the first and second derivatives of B with respect to parameters 1 and 3, and `B13`, the cross-derivative of B with respect to parameters 1 and 3. Then, after `C = A * B` we insert:

```
C1 = A * B1
C3= A3 * B + A * B3
  C11 = A * B11
  C13 = A3 * B1 + A * B13
  C33 = 2 * A3 * B3 + A * B33
```

This is essentially a manual version of an automatic technique for computing derivatives devised by Wengert (1964, see also Wilkins 1964 and Lesk 1967). The procedure, although tedious, is straightforward and reasonably flexible.

4 Parameter estimation

4.1 Maximum likelihood estimation

The proposed model, if (16) is used, contains eight parameters for each sample plot: a, b, c, σ_0, σ, η, t_0, and H_0. In specific "versions" of the-model some of these are assumed known, others may be common to all plots ("global" parameters), and others may be specific to each plot ("local" parameters). Functional relationships between parameters may also be imposed through reparametrization of the model. The data consist of pairs of age-height observations for each sample plot,

$$t_1, h_1, \ldots, t_n, h_n \ ,$$

where n may differ among plots. We are interested in estimating the parameters for different model versions.

The parameters will be estimated by the method of maximum likelihood (ML). The ML estimates are the values of the parameters fcr which the likelihood function reaches a maximum, for the given data. These estimates will be computed through minimization of the negative log-likelihood (19), using a modification of Newton's method. This function has the general form

$$F(\boldsymbol{\theta}) = \sum_{k=1}^{N} F_k(\boldsymbol{\theta}_0, \boldsymbol{\theta}_k) \ , \tag{29}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$, $\boldsymbol{\theta}_0$ is the vector of global parameters, and $\boldsymbol{\theta}_k$ is the vector of local parameters for plot k.

The statistical properties of the ML estimates in this case are not clear (see 4.4). Nevertheless, besides usually producing reasonable estimates, the ML method has two attractive characteristics: It specifies an objective, well-defined procedure for estimating parameters, no matter how complicate the model might be, and the method is invariant under parameter transformations. The invariance property means that any quantity which is a function of the parameters is estimated by substituting the parameter ML estimates.

The methods developed here will be applicable to any model containing global and local parameters, such that the function to be minimized is of the form (29). The same procedures could also be used with criteria other

than ML. For example, Bayesian estimates would involve minimizing (19) minus the logarithm of the prior distribution, a function which is also of the form (29).

4.2 Newton's method and some variants

There are a number of algorithms which could, in principle, be used for minimizing (29) (Jacoby, Kowalik, Pizzo,1972, Fletcher 1972, Chambers 1973). Algorithms based on Newton's method have the disadvantage of requiring the computation of second derivatives but, on the other hand, tend to be more reliable and to take significantly fewer iterations than alternative methods. This is important in the present application, which involves a complicate function of a large number of variables.

Newton's method for finding the minimum of a function $F(\boldsymbol{\theta})$ is based on approximating the function by the first three terms of its Taylor expansion around an initial estimate $\boldsymbol{\theta}^0$:

$$F(\boldsymbol{\theta}) \approx F(\boldsymbol{\theta}^0) + \boldsymbol{g}'(\boldsymbol{\theta} - \boldsymbol{\theta}^0) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^0)'\mathrm{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^0) \tag{30}$$

where $\boldsymbol{g}$ is the gradient

$$\boldsymbol{g} = (\frac{\partial F}{\partial \theta_1}, \ldots, \frac{\partial F}{\partial \theta_p}) \ ,$$

and H is the Hessian matrix

$$\mathrm{H} = \begin{bmatrix} \frac{\partial^2 F}{\partial {\theta_1}^2} & \cdots & \frac{\partial^2 F}{\partial \theta_1 \partial \theta_p} \\ \vdots & & \vdots \\ \frac{\partial^2 F}{\partial \theta_p \partial \theta_1} & \cdots & \frac{\partial^2 F}{\partial {\theta_p}^2} \end{bmatrix} \ ,$$

both evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^0$. If (30) were an equality, the optimum $\boldsymbol{\theta}$ could be obtained immediately by equating to zero the derivative of (15) with respect to $\boldsymbol{\theta}$:

$$\begin{aligned} \boldsymbol{g} + \mathrm{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^0) &= \mathbf{0} \ , \\ \boldsymbol{\theta} &= \boldsymbol{\theta}^0 - \mathrm{H}^{-1}\boldsymbol{g} \ . \end{aligned} \tag{31}$$

Since (30) is only an approximation, the value given by (31) will not in general be the optimum, but usually it will be an improvement over $\boldsymbol{\theta}^0$. Formula (31) can then be iterated using the new value as $\boldsymbol{\theta}^0$, until no further improvement is possible. The repeated application of (31) is the basic Newton method.

Two problems may cause convergence to a minimum to fail:

(a) If at $\boldsymbol{\theta}^0$ the Hessian is not positive-definite (the function is not convex at $\boldsymbol{\theta}^0$), F may not decrease in the direction given by $-\mathrm{H}^{-1}\boldsymbol{g}$. A solution is to temporarily substitute for H some other postive-definite matrix. This is discussed below.

(b) Even if H is positive-definite, taking a step $-\mathrm{H}^{-1}\boldsymbol{g}$ may overshoot the minimum in that direction and fail to decrease the value of F. The usual way around this is to use, instead of (31),

$$\boldsymbol{\theta} = \boldsymbol{\theta}^0 - \alpha \mathrm{H}^{-1}\boldsymbol{g} \ , \tag{32}$$

where α is selected at each iteration in such a way that F decreases. In the current version of the program, for each iteration initially $\alpha = 1$, and if it fails in reducing F it is halved until $F(\boldsymbol{\theta}) \leq F(\boldsymbol{\theta}^0)$.

Several ways of solving problem (a) are discussed by Murray (1972) and Fletcher and Freeman (1977). A slight variation of Murray's approach has been used here. To explain the method, the application of (32) can be described as follows. In each iteration, if H is positive-definite, we compute a vector $\boldsymbol{\delta}$ defining the direction of movement by solving

$$\mathrm{H}\boldsymbol{\delta} = -\boldsymbol{g} \ , \tag{33}$$

and then we move by a multiple α of this vector:

$$\boldsymbol{\theta} = \boldsymbol{\theta}^0 + \alpha\boldsymbol{\delta} \ .$$

If H is positive definite, an efficient way of solving (33) is by using the Cholesky factorization $\mathrm{H} = \mathrm{LL}'$, where L is lower triangular (Martin et al. 1971). Making $\mathrm{L}'\boldsymbol{\delta} = \boldsymbol{y}$ it is easy to see that (33) can then be solved in two stages:

$$\mathrm{L}\boldsymbol{y} = -\boldsymbol{g} \ , \qquad \mathrm{L}'\boldsymbol{\delta} = \boldsymbol{y} \ .$$

These equations are easy to solve due to L being triangular. For a general H, not necessarily positive-definite, Murray (1972) uses a modified Cholesky factorization that is equivalent to applying Cholesky's method to H + D, where D is a diagonal matrix. D is zero if H is sufficiently positive-definite. The modification is such that the method is numerically stable. A similar method based on the three-factor Cholesky decomposition (20) saves seme computing time, at the cost of additional programming effort (Gill and Murray 1974).

The algorithm may now be summarized as follows:

(i) Compute $\boldsymbol{g}$ and H at $\boldsymbol{\theta}$

(ii) Find the modified Cholesky factor L of H

(iii) Solve $\mathrm{L}\boldsymbol{y} = -\boldsymbol{g}$ for $\boldsymbol{y}$ and $\mathrm{L}'\boldsymbol{\delta} = \boldsymbol{y}$ for $\boldsymbol{\delta}$

(iv) If $\boldsymbol{\delta}$ and $\boldsymbol{g}$ are small enough, stop.
Otherwise set $\alpha \leftarrow 1$ (34)

(v) Compute $F(\boldsymbol{\theta} + \alpha\boldsymbol{\delta})$

(vi) If $F(\boldsymbol{\theta} + \alpha\boldsymbol{\delta}) > F(\boldsymbol{\theta})$ set $\alpha \leftarrow \alpha/2$ and go to step (v)

(vii) Otherwise set $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\boldsymbol{\delta}$ and go to step (i).

It is possible for the algorithm to stop at a saddle point instead of at a minimum, a case which can be recognised because at that point the Hessian is not positive-definite. Special techniques may be included in the algorithm to handle this contingency (Gill and Murray 1974). However, convergence to a saddle point seems unlikely, and if it happens we can simply make a perturbation of $\boldsymbol{\theta}$ and restart the program from the new point (Murray 1972). More troublesome is the possibility of convergence to local minima different from the absolute minimum. The only way of guarding against this possibility is to rerun the program several times with different starting points.

A direct application of algorithm (34) for minimizing (29) is impractical, due to the large size of the Hessian H. It is possible, however, to exploit the special structure of (29) and arrange the computations in a sequential form with only modest memory requirements.

4.3 Exploiting special structure

It is convenient to redefine $\boldsymbol{\theta}$ in (29) as a column vector partitioned as:

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_1 \\ \vdots \\ \boldsymbol{\theta}_N \\ \boldsymbol{\theta}_0 \end{bmatrix},$$

where $\boldsymbol{\theta}_k$, $k \neq 0$, is a column vector of local parameters for plot k and $\boldsymbol{\theta}_0$ is a column vector of global parameters. $\boldsymbol{\delta}$, the gradient and the Hessian may

be partitioned in the same way:

$$\boldsymbol{\delta} = \begin{bmatrix} \boldsymbol{\delta}_1 \\ \vdots \\ \boldsymbol{\delta}_N \\ \boldsymbol{\delta}_0 \end{bmatrix}, \qquad \boldsymbol{g} = \begin{bmatrix} \boldsymbol{g}_1 \\ \vdots \\ \boldsymbol{g}_N \\ \boldsymbol{g}_0 \end{bmatrix}, \qquad \mathrm{H} = \begin{bmatrix} \mathrm{H}_{11} & \cdots & \mathrm{H}_{1N} & \mathrm{H}_{10} \\ \vdots & \ddots & \vdots & \vdots \\ \mathrm{H}_{N1} & \cdots & \mathrm{H}_{NN} & \mathrm{H}_{N0} \\ \mathrm{H}_{01} & \cdots & \mathrm{H}_{0N} & \mathrm{H}_{00} \end{bmatrix}.$$

We note that

$$\begin{aligned}
\boldsymbol{g}_0 &= \frac{\partial F}{\partial \boldsymbol{\theta}_0} = \sum_{k=1}^{N} \frac{\partial F_k}{\partial \boldsymbol{\theta}_0} \equiv \sum_{k=1}^{N} \boldsymbol{g}_{k0} \\
\boldsymbol{g}_k &= \frac{\partial F}{\partial \boldsymbol{\theta}_k} = \frac{\partial F_k}{\partial \boldsymbol{\theta}_k}\,; \quad k = 1, \ldots, N \\
\mathrm{H}_{00} &= \frac{\partial^2 F}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0} = \sum_{k=1}^{N} \frac{\partial^2 F_k}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0} \equiv \sum_{k=1}^{N} \mathrm{H}_{k00} \\
\mathrm{H}_{0k} &= \mathrm{H}'_{k0} = \frac{\partial^2 F}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_k} = \frac{\partial^2 F_k}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_k}\,; \quad k = 1, \ldots, N \\
\mathrm{H}_{kk} &= \frac{\partial^2 F}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_k} = \frac{\partial^2 F_k}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_k}\,; \quad k = 1, \ldots, N \\
\mathrm{H}_{km} &= \frac{\partial^2 F}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_m} = 0\,; \quad k, m = 1, \ldots, N\,,\ k \neq m\,.
\end{aligned}$$

If H is positive-definite, the Cholesky factorization of H takes the form

$$\begin{bmatrix} \mathrm{H}_{11} & & & & \mathrm{H}_{10} \\ & H_{22} & & & \mathrm{H}_{20} \\ & & \ddots & & \vdots \\ & & & \mathrm{H}_{NN} & \mathrm{H}_{N0} \\ \mathrm{H}_{01} & \mathrm{H}_{02} & \cdots & \mathrm{H}_{0N} & \mathrm{H}_{00} \end{bmatrix} =$$

$$\begin{bmatrix} \mathrm{L}_{11} & & & & \\ & L_{22} & & & \\ & & \ddots & & \\ & & & \mathrm{L}_{NN} & \\ \mathrm{L}_{01} & \mathrm{L}_{02} & \cdots & \mathrm{L}_{0N} & \mathrm{L}_{00} \end{bmatrix} \begin{bmatrix} \mathrm{L}'_{11} & & & & \mathrm{L}'_{10} \\ & L'_{22} & & & \mathrm{L}'_{20} \\ & & \ddots & & \vdots \\ & & & \mathrm{L}'_{NN} & \mathrm{L}'_{N0} \\ & & & & \mathrm{L}'_{00} \end{bmatrix} \tag{35}$$

Working with the submatrices analogously to the matrix elements in Cholesky's method (Martin *et al.* 1971) it is found that the non-zero submatrices of the

Cholesky factor can be obtained as follows:

$$\begin{cases} \mathrm{L}_{kk} &= \text{Cholesky factor of } \mathrm{H}_{kk} \\ \mathrm{L}_{0k} &= \mathrm{L}_{kk}^{-1}\mathrm{H}_{0k}\ ; \quad k = 1, \ldots, N \\ \mathrm{L}_{00} &= \text{Cholesky factor of } \mathrm{H}_{00} - \sum_{k=1}^{N} \mathrm{L}_{0k}\mathrm{L}_{0k}' \\ & \quad \text{or } \sum_{k=1}^{N}(\mathrm{H}_{k00} - \mathrm{L}_{0k}\mathrm{L}_{0k}')\ . \end{cases} \tag{36}$$

For non-positive-definite H, if the Cholesky factorization in (36) are replaced by Murray's modified Colesky factorizations the resultant factorization of H is not the same as the one which would be obtained from Murray's method applied to H. Nevertheless, it is easy to see that the result is also equivalent to a Cholesky factorization of H plus some diagonal matrix. It can be shown that this diagonal matrix is in general closer to zero than Murray's and that Murray's conditions for numerical stability still apply.

Using this approach, the computations for steps (i) to (iii) in (34) can be arranged sequentially:

(a) For $k = 1$ to N:

  (i) Read in the data and current value of $\boldsymbol{\theta}$ for plot k, and compute $\boldsymbol{g}_k$, $\boldsymbol{g}_{k0}$, H_{kk}, H_{0k}, and H_{k00}

  (ii) Find the modified Cholesky factor L_{kk} of H_{kk}

  (iii) Solve $\mathrm{L}_{kk}\boldsymbol{y}_k = -\boldsymbol{g}_k$ for $\boldsymbol{y}_k$ and $\mathrm{L}_{kk}\mathrm{L}_{0k} = \mathrm{H}_{0k}$ for L_{0k}

  (iv) Accumulate $\sum \boldsymbol{g}_{k0} = \boldsymbol{g}_0$, $\sum \mathrm{L}_{0k}\boldsymbol{y}_k$, and $\sum(\mathrm{H}_{k00} - \mathrm{L}_{0k}\mathrm{L}_{0k}')$

  (v) Store in a file $\boldsymbol{y}_k$, L_{kk} and L_{0k}

(b) Find the modified Cholesky factor L_{00} of $\sum(\mathrm{H}_{k00} - \mathrm{L}_{0k}\mathrm{L}_{0k}')$

(c) Solve $\mathrm{L}_{00}\boldsymbol{y}_0 = -\boldsymbol{g}_0 - \sum \mathrm{L}_{0k}\boldsymbol{y}_k$ for $\boldsymbol{y}_0$ and $\mathrm{L}_{00}'\boldsymbol{\delta}_0 = \boldsymbol{y}_0$ for $\boldsymbol{\delta}_0$

(d) For $k = 1$ to N:

  (i) Read from file $\boldsymbol{y}_k$, L_{kk} and L_{0k} and solve $\mathrm{L}_{kk}'\boldsymbol{\delta}_k = \boldsymbol{y}_k - \mathrm{L}_{0k}'\boldsymbol{\delta}_0$ for $\boldsymbol{\delta}_k$

  (ii) Store in file $\boldsymbol{\delta}_k$

Steps for the computation of the norms of $\boldsymbol{g}$ and $\boldsymbol{\delta}$ to be used in the convergence tests can be added. It is clear that steps (v) and (vii) of (34) can also be implemented sequentially, processing one plot at a time.

4.4 Statistical inference

Under certain conditions, it has been proved that ML estimators are consistent, asymptotically efficient, and the estimates are asymptotically normal with covariance matrix equal to the inverse of the matrix of second derivatives of the negative log-likelihood. In addition, logarithms of likelihood ratios are asymptotically distributed as Chi-squared, a fact that can be used for testing hypotheses and comparing different models. However, the standard conditions are not satisfied in this model, so that the properties of the estimates are uncertain.

For inference about the global parameters, the situation is similar to a class of problems studied by Kiefer and Wolfowitz (1956) and Kalbfleisch and Sprott (1970) (see also Cox and Hinkley 1974, p.292 and 298). For the local parameters, ignoring the global parameters we have the case of dependent observations discussed by Weiss (1971), Cox and Hinkley (1974, p.293 and 299), and Crowder (1976). The small number of observations in each sample plot, however, means that any results about asymptotic properties for the local parameters would be of doubtful value.

On the other hand, if one accepts the ideas of Likelihood Inference, it is possible to make use of the ML estimates, likelihood ratios, and second derivatives of the log-likelihood in a very direct way (Barnett 1973 section 8.2, Edwards 1972). Likelihood Inference is a controversial subject (as most other approaches are, see, e.g., Barnett 1973), but it seems to provide at least some useful qualitative guidelines for comparing models and for making other inferences.

Whatever the interpretation, it seems useful to compute the inverse of the matrix of second derivatives of the negative log-likelihood, that is, the inverse of our Hessian H evaluated at the point of convergence. This inverse can be computed from the stored values of L_{kk}, and L_{0k} using formulas derived from (35).

The computed minimum of the negative log-likelihood may be used as an aid in comparing different versions of the model. Edwards (1972) indicates that a difference of about two units in the negative log-likelihood might be taken as "significant". When comparing models with different numbers of parameters, Edwards suggests adding half a unit for each additional parameter; other approaches suggest adding one unit (Akaike 1973, Stone 1977).

5 Implementation and computational experience

5.1 Computer program

The parameter estimation procedure has been implemented in Fortran for the ICL 2980 computer. The program contains approximately 800 lines and is available from the author. Except for the direct access file handling statements, only standard ANS 66 Fortran has been used so that conversion to other computers should be relatively easy.

The log-likelihood and derivatives for each sample plot are computed as shown in 3.4 and 3.5, using the eight parameters a, b, c, σ_0, σ, $c\eta$, t_0, and H_0^c. Particular versions of the model are specified through two user-supplied subroutines. One produces the eight basic parameters, a to H_0^c, given the global and local parameters in $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_k$. The second subroutine transforms the derivatives with respect to the basic parameters to derivatives with respect to $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_k$. Some care is needed to ensure that there are no errors in these subroutines, and that the model is identifiable. For example, at least one of t_0 and H_0 must be fixed, and the possible number of local parameters is limited by the smallest number of observations in any sample plot.

The minimization is done essentially as described in 4.3. The height-age data are held in a sequential file. Heights and ages are scaled down by 10 in order to cause the parameters to be more nearly of the same magnitude, and reduce the likelihood of over- underflow and roundoff problems. All computations are done in double precision. It was found necessary to limit the step size in order to guard against overflow, specially in the initial iterations. A bound of 0.5 for the largest component of $\boldsymbol{\delta}$ is used. The current values of $\boldsymbol{\theta}_k$, L_{kk}, L_{0k}, and other plot information that needs to be updated are stored in a direct access file.

If the program converges within the pre-specified number of iterations, approximate variances and covariances for the parameters, based on the inverse Hessian as described in 4.4, are computed. Only a weighted average over all plots is printed for the cross-correlations of local parameters.

5.2 Results

Experience with the program to date has been limited to two sets of data. One consists of 91 sample plots with a total of 543 measurements of radiata pine in Kaingaroa Forest, New Zealand (figure 1). The other contains 58 plots with 247 measurements of radiata pine in Southland Conservancy (figure 2). Simulated data with similar characteristics have also been used (see

below). Only plots with three or more measurements were used for some model versions.

Convergence is satisfactory, provided that reasonable starting points are used. Most of the time is spent in regions of the parameter space where the Hessian is not positive-definite. In these regions the step length often has to be reduced several times in each iteration. Once the Hessian becomes positive-definite the rate of convergence tends to be much faster. The use of good starting points seems important. The best approach appears to be to start with a model version having a small number of parameters, and use the estimates for this model as starting point for the next more complex version.

No accurate timing has been carried out. The number of iterations needed varies widely, depending on the starting point and number of parameters. As a general guide, the number of iterations is typically between 10 and 50, and a run with the Southland data using three global and two local parameters and requiring 18 iterations took less than 20 seconds on the ICL 2980.

The most complete analysis was carried out with the Southland data, and results have been reported by García (1979). The main tests involved model versions with $t_0 = H_0 = \sigma_0 = 0$, c and η global, σ local, and three alternatives for parameters a and b: (i) a global and b local, (ii) b global and a local, and (iii) S local and $a = \alpha + \beta S$, where α and β are global parameters and $S = a[1 - \exp(-2b)]^{1/c}$ is the site index. Case (iii) covers (i) and (ii) as special cases ($\beta = 0$ and $\alpha = 0$, respectively), at the cost of one additional parameter. The maximum log-likelihoods were found to be 374.6, 377.3, and 377.9, for (i), (ii), and (iii), respectively. No appreciable improvement was observed when t_0 was allowed to take values different from zero. No alternative local minima have been found, other than the trivial ones obtained by changing the sign of the standard deviations.

An unexpected result was that in all cases, including runs with the Kaingaroa data, the estimator for σ was zero. This prompted a thorough checking of the whole system, and eventually the writing of a completely new version of the program including σ_0 as an additional parameter. In order to test if the effect was due to the data, artificial data sets were generated by taking the ages in the Southland and Kaingaroa data sets, and computing heights according to the model (with $\sigma \neq 0$) using pseudo-random numbers. In all cases the estimte for σ was found to be zero.

Finally, a simplified discrete version of the model, with only one sample plot, has been studied through simulation and likelihod plotting (see Appendix). It was found that, depending on the true values of the parameters,

the maximum likelihod estimate of either σ or η is usually zero. Difficulties in estimating the variances occur then even in cases where the data are generated according to the model, the stochastic process is discrete on time, and there are no local parameters. This indicates that the effect is probably not caused by the use of the Wiener process, by the presence of both global and local parameters, by the minimization algorithm, or by lack of fit of the model to the data. The most likely explanation appears to be simply that the data do not provide enough information for a reliable separation of the random variation into the "environmental" and the "measurement" components. It is also possible that the ML estimator performs poorly for this model and that other estimators might be more appropriate.

5.3 Discussion

The use of stochastic differential equations for approximating the stochastic structure of growth data should conduce to better estimators than an indiscriminate use of least-squares. It would be interesting to study also other SDEs, such as the one in the footnote on page 8.

The failure in the simultaneous estimation of σ and η has been somewhat disappointing. One possible way of aproaching this problem is to assume some value for σ or η) . The standard error of the height-diameter regressions used for computing the top height (see footnote on page 3) provides information about the magnitude of η . In data from stem analysis η may be negligible. The other parameters may then be estimated with η fixed at a given value, although this might still give $\sigma = 0$. Different values of η could be tried to examine the sensitivity of the parameters of interest to changes in σ and η. Limited experience with the simplified model of the Appendix indicates that the parameters of interest may be rather insensitive to changes in the variances. Another possibility is a Bayesian approach, adding the logarithm of a prior density for σ and η to the log-likelihood.

In the development of site index curves the local parameters are nuisance parameters, and techniques based on partial likelihoods might be worth investigating (Kalbfleisch and Sprott 1970, Cox 1975). In any case, some research on the properties of estimators for this class of models, starting with the simple model in the Appendix, would be desirable.

The procedure proposed in 4.3 is a feasible and reasonably efficient estimation method for models containing both "global" and "local" parameters. Processing time could be reduced by using a more sophisticated line search method instead of the halving of α (Gill and Murray 1974). Some recent Newton-type methods perform better than Murray's in non-convex regions

of the parameter space (Fletcher and Freeman 1977, More and Sorensen 1979), but it is not clear if they could be modified in order to exploit the special structure of the Hessian.

Appendix

In order to investigate the difficulties experienced in trying to separate the "environmental" and "measurement" components of the random variation, a simplified model using only one "sample plot" was studied. This model is obtained by applying (8) to evenly spaced observations. The model assumes that $x_1, \ldots, x_n$ is a set of n observations such that

$$\begin{aligned} x_i &= y_i + \epsilon_i \qquad \text{and} \\ y_i &= b y_{i-1} + \delta_i \; ; \quad i = 1, \ldots, n \; , \end{aligned} \tag{37}$$

where the δ_i and ϵ_i are all independent normal random variables with zero means, and $\mathrm{var}(\delta_i) = \sigma^2$, $\mathrm{var}(\epsilon_i) = \eta^2$. ($b$ and δ here are related to, but not the same as the b and δ in the body of the paper). We assume that $y_0 < 0$ is known, and we want to estimate b, σ and η given $x_1, \ldots, x_n$.

Defining

$$\begin{aligned} z_i &= x_i - b x_{i-1} = \epsilon_i - b\epsilon_{i-1} + \delta_i \; ; \quad i = 2, \ldots, n \; , \\ z_1 &= x_1 - b y_0 = \epsilon_1 + \delta_1 \; , \end{aligned} \tag{38}$$

it is easy to see that the log-likelihood is

$$-\frac{1}{2}[n \ln(2\pi) + \ln|\mathrm{C}| + \boldsymbol{z}' \mathrm{C}^{-1} \boldsymbol{z}] \; , \tag{39}$$

where $\boldsymbol{z} = (z_1, \ldots, z_n)'$, and the elements of the covariance matrix C are

$$c_{ij} = \begin{cases} \sigma^2 + \eta^2 & , \quad i = j = 1 \\ \sigma^2 + (1 + b^2)\eta^2 & , \quad i = j \neq 1 \\ -b\eta^2 & , \quad |i - j| = 1 \\ 0 & , \quad \text{otherwise} \; . \end{cases} \tag{40}$$

To study the properties of this model, sets of five observations were generated according to (37) using pseudo-random numbers, with $y_0 = -3$, $b = 0.95$, and various values for σ and η. The ML estimates for b, σ and η, or for σ and η with b fixed at 0.95, were computed using a version of the Nelder-Mead simplex algorithm on a microcomputer (Nelder and Mead

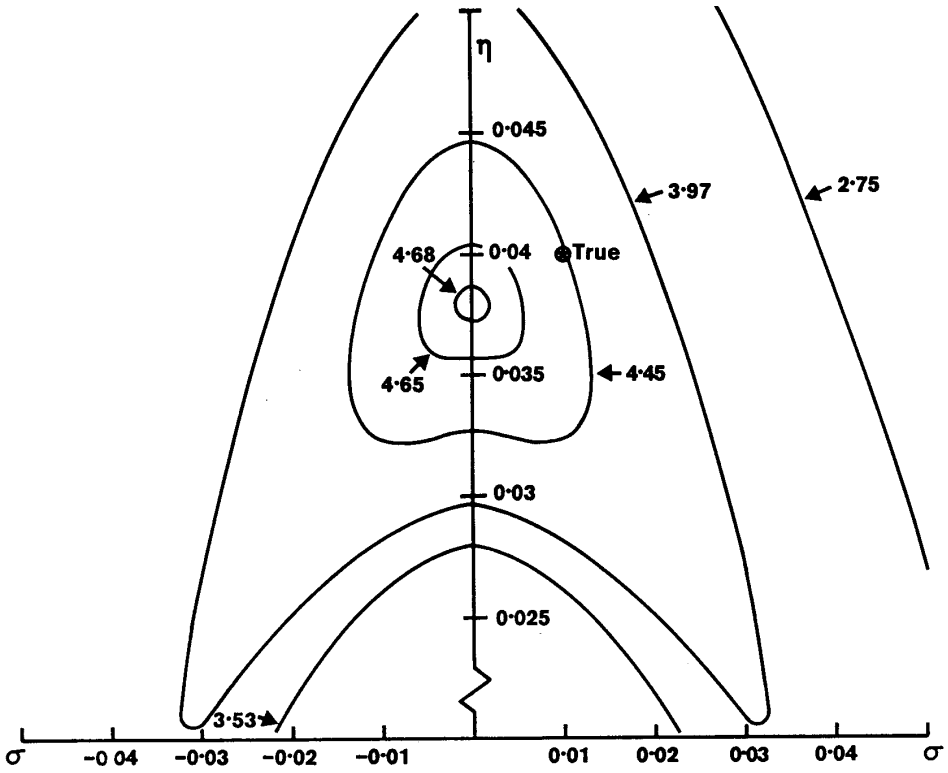
η
0·045
3·97
2·75
4·68
0·04
True
4·65
0·035
4·45
0·03
0·025
3·53
σ
-0 04
-0·03
-0·02
-0·01
0·01
0·02
0·03
0·04
σ

Figure 3:

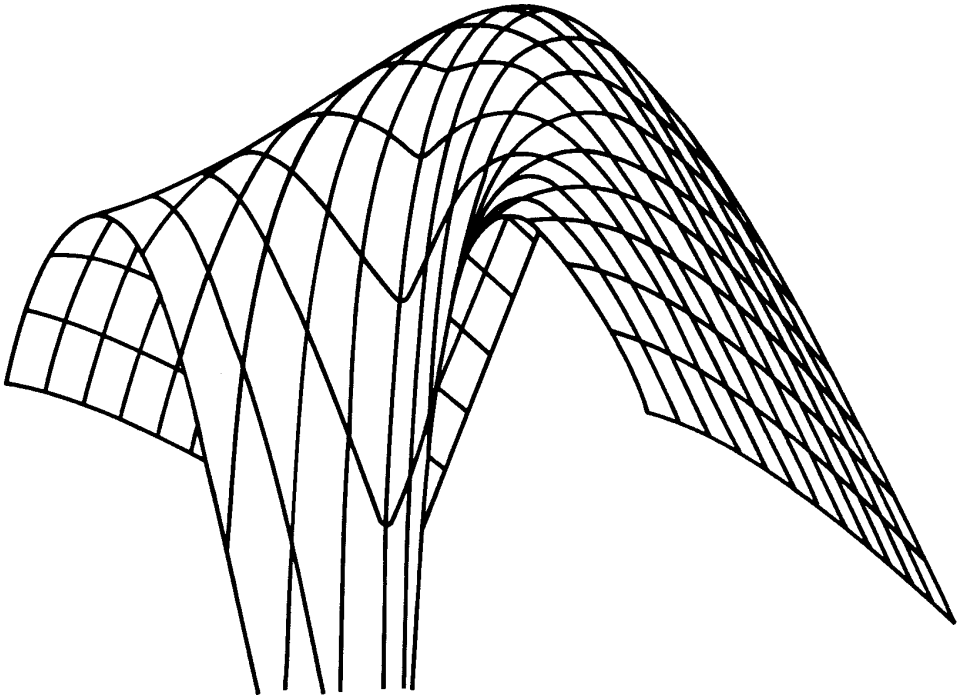

Figure 4:

1965). In some cases contours of the log-likelihood obtained by fixing one of the parameters were plotted. A typical form of the log-likelihood function for $b = 0.95$ is shown in figures 3 and 4.

Based on a limited number of simulations, the following preliminary observations may be made: (a) With $b = 0.95$ (fixed), most simulations result in either $\sigma = 0$ or $\eta = 0$, depending on the true values of the parameters. A simulation with $a = 0.02$ and $\eta = 0.03$ resulted in non-zero estimates for both parameters (0.013 and 0.021, respectively), (b) All simulations with b free have resulted in zero estimates for σ. (c) No singularities or other anomalies in the log-likelihood have been observed in the contour plots.

References

Akaike, H. (1973). Information Theory and an extension of the maximum likelihood principle. In: Petrov, B.N. and Czaki, G. (eds.), 2nd International Symposium on Information Theory. Akademia Kiadó, Budapest.

Bailey, R.L. and Clutter, J.L. (1974). Base-age invariant polymorphic site curves. Forest Science, 20, 155-159.

Barnett, V. (1973). Comparative Statistical Inference. Wiley, London.

Bertalanffy, L. von (1949). Problems of organic growth. Nature, 163, 156-158.

Bertalanffy, L. von (1957). Quantitative laws in metabolism and growth. Quart. Rev. Biol., 32, 217-231.

Beverton, R.J.H. and Holt, S.J. (1957). On the dynamics of exploited fish populations. Min. Agri. Fish. Food (Great Brit.). Fish. Invest. Ser.2, Vol.19, 533p.

Fletcher, R. and Freeman, T.L. (1977). A modified Newton method for minimization. J. Optim. Theory and Appl., 23, 357-372.

Fries, H. (1974). Praktisk bonitering. Sveriges Skogsvardsforbunds Tidskrift, 72. 551-563.

García, 0. (1979). Modelling stand development with stochastic differential equations. In: Elliofct, D.A. (ed.), Mensuration fcr management planning of exotic forest plantations, FRI Symposium 20, October 1978. N.Z. Forest Serv.

Gihman, I.J. and Skorohod, A.V. (1972). Stochastic Differential Equations. Springer-Verlag, New York.

Gill, P.E. and Murray, W. (1974). Newton-type methods for unconstrained and linearly constrained optimization. Mathematical Programming, 7, 311-350.

Hottelling, H. (1927). Differential equations subject to error, and population estimates. J. Am. Stat. Assoc., 22, 283-314.

Jacoby, S.L.S., Kowalik, J.S. and Pizzo, J.T. (1972). Iterative Methods for Nonlinear Optimization Problems. Prentice-Hall, Englewood Cliffs, N.J.

Jones, J.R. (1969). Review and comparison of site evaluating methods. USDA For. Serv., Research Paper RM-51.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. J.R. Statist.Soc. B. 26. 211-252.

Burkhart, H.E. and Tennent, R.B. (1977). Site index equations for radiata pine in New Zealand. N.Z.J. Forest Sci., 7, 408-416.

Carmean, W.H. (1975). Forest site quality evaluation in the United States. In: Brady, N.C.(ed.), Advances in Agronomy, Vol.27.

Chambers, J.M. (1973). Fitting nonlinear models: Numerical techniques. Biometrika, 60, 1-13.

Chapman, D.G. (1961). Statistical problems in dynamics of exploited fisheries populations. In; Proc. Fourth Berkeley Symp. Math. Statist, and Prob. University of California Press, Berkeley.

Cox, D.R. (1975). Partial likelihood. Biometrika, 62, 269-276.

Cox,,D.R. and Hinkley, D.V. (1974). Theoretical Statistics. Chapman and Hall, London.

Crowder, M.J. (1976). Maximum likelihood estimation for dependent observations. J.R. Statist. Soc. B, 38, 45-53.

Curtis, P.O., DeMars, D.J. and Herman, F.R. (1974). Which dependent variable in site index-height regressions? For. Sci. 20, 74-90

Dwyer, P.S. (1967). Some applications of matrix derivatives in multivariate analysis. J.Am. Stat. Assoc., 62. 607-625.

Edwards, A.W.F. (1972). Likelihood. Cambridge Univ. Press

Erickson, R.V. (1971). Constant coefficient linear differential equations driven by white noise. Ann. Math. Stat., 42. 820-823.

Ferguson, I.S. and Leech, J.W. (1978). Generalized least squares estimation of yield functions. Forest Science, 24, 27-42.

Fletcher, R. (1972). A survey of algorithms for unconstrained optimization. In: Murray, W. (ed.), Numerical methods for unconstrained optimization. Academic Press, London.

Kalbfleisch, J.D. and Sprott, D.A. (1970). Application of likelihood methods to models involving large numbers of parameters (with discussion). J.R. Stat. Soc. B, 32, 175-208.

Karlin, S. (1966). A First Course in Stochastic Processes. Academic Press, New York.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann. Math. Stat., 27, 887-906.

Lesk, A.M. (1967). Dynamic computation of derivatives. Comm. A.C.M., 10, 571-572.

Levenbach, H. and Reuter, B.E. (1976). Forecasting trending time series with relative growth rate models. Technometrics, 18, 261-271.

Matérn, B. (1976). Om skattning av ovre hojden. Sveriges Skogsvardsforbunds Tidskrift, 74, 51-53.

Martin, R.S., Peters, G. and Wilkinson, J.H. (1971). Symmetric decomposition of a positive definite matrix. In: Wilkinson, J.H. and Reinsch, C. Linear Algebra - Handbook for Automatic Computation, Vol. II. Springer-Verlag, Berlin.

Moré, J.J. and Sorensen, D.C. (1979). On the use of directions of negative curvature in a modified Newton method. Mathematical Programming, 16, 1-20.

Murray, W. (1972). Second derivative methods. In: Murray, W. (ed.), Numerical Methods for Unconstrained Optimization. Academic Press, London.

Nelder, J.A. and Mead, R. (1965). A simplex method for function minimization. Computer J., 7, 308-313.

Nelder, J.A., Austin, R.B., Bleasdale, J.K.A. and Salter, P.J. (1960). An approach to the study of yearly and other variation in crop yields. J. Hort. Sci., 35, 73-82.

Neudecker, H. (1969). Some theorems on matrix differentiation with special reference to Kronecker matrix products. J. Am. Stat. Assoc., 64. 953-963.

Pienaar, L.V. and Turnbull, K.J. (1973). The Chapman-Richards generalization of von Bertalanffy's growth model for basal area growth and yield in even-aged stands. Forest Science, 19, 2-22.

Rennolls, K. (1978). "Top height"; Its definition and estimation. Commonw. For. Rev.,57, 215-219.

Richards, F.J. (1959). A flexible growth function for empirical use J. Expt. Botany, 10, 290-300.

Sandland, R.L. and McGilchrist, C.A. (1979). Stochastic growth curve analysis. Biometrics, 35, 255-271.

Schlesselman, J. (1971). Power families: A note on the Box and Cox transformation. J. R. Stat. Soc. B, 33, 307-311.

Spurr, S.H. (1952). Forest Inventory. Ronald Press, New York.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. J.R. Stat. Soc. B, 39, 114-47.

Sullivan, A.D. and Clutter, J.L. (1972). A simultaneous growth and yield model for Loblolly pine. Forest Science, 18, 76-86.

Sullivan, A.D. and Reynolds, M.R. i,1976). Regression problems from repeated measurements. Forest Science, 22, 382-385.

Tomlinson, A.I. (1976). Short period fluctuations of New Zealand rainfall. N.Z.J. Science, 19, 149-161.

Tveite, B. (1969). A method for construction of site-index curves. Medd. Norske Skogforsoksvesen, 28, 131-159.

Weiss, L. (1971). Asymptotic properties of maximum likelihood estimators in some non-standard cases. J. Am. Stat. Assoc., 66, 345-350.

Wengert, R.E. (1964). A simple automatic derivative evaluation program Comm. A.C.M.. 7. 463-464.

Wilkins, R.D.(1964). Investigation of a new analytical method for numerical derivative evaluation. Comm. A.C.M., 7, 465-471.