

Interregional nonlinear height–diameter model with random coefficients for stone pine in Spain

Rafael Calama and Gregorio Montero

Abstract: An individual-tree height–diameter model was developed for stone pine (*Pinus pinea* L.) in Spain. Five biparametric nonlinear equations were fitted and evaluated based on a data set consisting of 8614 trees from 455 plots located in the four most important regions where the species occurs in Spain. Because of the problem of high correlation among observations taken from the same sampling unit, a mixed-model approach, including random coefficients, is proposed. Several stand variables, such as density, dominant height, or diametric distribution percentiles, were included in the model as covariates to explain among plot variability. To determine interregional variability among the regions studied, regional effects were included in the model using fixed dummy variables. Two models, one for inland regions and one for coastal regions, were found to be sufficient to explain regional variability in the height–diameter relationship for the species in Spain. Mixed models allow predictive role in two ways: a typical response, including only fixed effects, and a calibrated response, where random effects are predicted and included in the model from the prior measurement of the height in a subsample of trees. Different alternatives were tested to determine optimum subsample size. Measurement of the height of the 20% largest trees in the plot has been shown to be a useful approach.

Résumé : Les auteurs ont développé un modèle reliant la hauteur et le diamètre des arbres individuels pour le pin pignon (*Pinus pinea* L.) en Espagne. Cinq équations non linéaires biparamétriques ont été ajustées et évaluées en utilisant une base de données portant sur 8614 arbres. Ces arbres viennent de 455 parcelles réparties dans quatre régions où l'espèce est la plus représentée. À cause de la forte corrélation entre les observations, les auteurs proposent l'emploi d'un modèle mixte qui contient des coefficients aléatoires pour chaque parcelle. Plusieurs variables relatives au peuplement telles que la densité, la hauteur dominante ou les percentiles de la distribution diamétrale ont été incluses comme covariables dans le modèle pour tenir compte de la variation entre les parcelles. Des variables nominales fixes ont aussi été incluses dans le modèle pour tenir compte de l'effet dû aux régions et déterminer la variation entre les régions étudiées. Deux modèles se sont avérés suffisants pour expliquer la relation entre la hauteur et le diamètre chez le pin pignon en Espagne : un pour les régions de l'intérieur et l'autre pour la zone côtière. Les modèles mixtes permettent de faire deux types de prédictions : on peut obtenir d'une part une réponse typique en incluant seulement des effets fixes et d'autre part, une réponse calibrée lorsque les effets aléatoires sont prédits et inclus dans le modèle en ayant mesuré au préalable les hauteurs dans un sous-échantillon d'arbres. Plusieurs alternatives ont été testées pour déterminer la taille optimale de ce sous-échantillon. La meilleure solution consiste à mesurer les 20 % plus grosses tiges dans la parcelle.

[Traduit par la Rédaction]

Introduction

Knowledge of the breast-height diameter and the total tree height is fundamental for developing growth and yield models in forest stands. Measuring breast-height diameter is simpler, more accurate, and cheaper than measuring tree height. As a consequence, in forest inventories, diameter is measured for all the sampled trees, but height is measured only for a subsample of trees. From these measurements, graphic or statistical models are fitted to define the relation between these two variables.

Development of simple and accurate models that allow forest managers to determine with reliability the height of the trees in a stand from diameter data is a prime objective in forest management. Knowledge of the relation between these variables permits managers to obtain, without investing large amounts of money in height measurement, the input values needed to estimate single-tree volume, dominant height of the stands, competition indices for individual tree growth, height/diameter ratio, and structural diversity indexes.

The relation between the diameter of a tree and its height varies from one stand to another. Within the same stand, this relation is not constant over time (Curtis 1967). Development of general models that for a wide range of geographical validity connect tree height and diameter has been approached from different perspectives:

- (1) Two-stage approach (Ferguson and Leech 1978). The height–diameter relationship is first fitted individually for each sampling unit (plot, stand). In a second stage, parameters are explained using stand variables such as

Received 21 February 2003. Accepted 13 August 2003.
Published on the NRC Research Press Web site at
<http://cjfr.nrc.ca> on 22 January 2004.

R. Calama¹ and G. Montero. Centro de Investigación Forestal – Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, Apdo. 8111, 28080 Madrid, Spain.

¹Corresponding author (e-mail: rcalama@inia.es).

density, basal area, dominant height, age, and dominant diameter as covariates ([Krumland and Wensel 1978](#); [Larsen and Hann 1987](#); [Parresol 1992](#)).

- (2) Development of single models for different geographical or ecological regions, or inclusion of regional effects in the model as categorical variables ([Fulton 1999](#); [Huang et al. 2000](#)).
- (3) Analysis of the spatial pattern of variability in height/diameter ratio and geostatistical modelling of the parameters ([Nanos 2001](#)).
- (4) Use of models with random coefficients, which define a fixed population average response, including varying random parameters for each sampling unit ([Hokka 1997](#); [Lappi 1997](#)).

When modelling the height–diameter relationship, measurements of both variables taken from trees growing in sample plots located in different stands or regions are usually used. This hierarchical structure (i.e., trees within plots within stands) in the data means that there is a lack of independence among measurements, since observations from the same sampling unit are highly correlated ([West et al. 1984](#); [Gregoire 1987](#)). Using ordinary least squares regression techniques when data are correlated leads to bias in the confidence interval for the mean value of the parameters. The problem of how to deal with these correlated data has been tackled from several different approaches ([Henderson et al. 1959](#); [Ferguson and Leech 1978](#); [West et al. 1984](#); [Fox et al. 2001](#)).

Mixed models estimate both fixed and random parameters simultaneously for the same model. The introduction of random parameters into the model, specific for every sampling unit, enables us to model the variability detected for given phenomena among different locations, after defining a common fixed functional structure ([Lindstrom and Bates 1990](#)).

Random-parameter models have been widely used in forest research to model variables such as dominant height ([Lappi and Bailey 1988](#)), single-tree volume ([Lappi 1991](#); [Eerikäinen 2001](#)), and taper functions ([Lappi 1986](#); [Gregoire and Schabenberger 1996](#)). The height–diameter relationship has also been calibrated following these methods, using both repeated measurements from permanent plots ([Lappi 1997](#)) as well as cross-temporal measurements taken simultaneously in different temporary plots ([Jayaraman and Lappi 2001](#)).

Mixed models give an unbiased and efficient estimation of the fixed parameters of the model. Furthermore, mixed models improve predictive ability if we are able to predict the value of the random parameters for an unsampled location. This is possible if complementary observations of the dependent variable are available. Measuring height in a subsample of trees is a very common practice in forest inventories. Determining the minimum size of this subsample to get the reliability required by forest managers is a basic decision to be confronted in inventory design ([Houghton and Gregoire 1993](#); [Bégin and Raulier 1995](#)), since height measurement involves high costs.

The main objective of this study is to construct a model relating breast-height diameter and total height for stone pine (*Pinus pinea* L.) trees, growing in even-aged stands. In Spain, height–diameter models for stone pine have been con-

structed by [García Güemes \(1999\)](#) and [Cañadas \(2000\)](#), using ordinary least squares linear and nonlinear regression techniques, respectively. Taking into account the nested structure of the data set, we use a model that included random parameters. To diminish the level of variance among sampling units (plots), stand variables are included as covariates. In the same way, we study whether a single model for the whole area where the species occurs in Spain explains the relationship between height and diameter, or, to the contrary, whether inclusion of regional parameters is necessary. Finally, we also study the impact on predictability of the model of having a subsample of complementary trees where both variables (diameter and height) are measured, since this enables the random parameters to be predicted.

Materials and methods

Data

Since 1992, a research line of the Department of Mediterranean Silviculture, Centro de Investigación Forestal – Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, has been devoted to the sustainable management and multiple uses of stone pine stands in Spain. As a result of this work, a network of 470 permanent plots was installed in even-aged stands in the four principal stone pine regions in Spain: Northern Plateau, Catalonia, Central Range, and West Andalusia. Plots are circular, with variable sizes, and all plots have 20 trees, except for fifty 10-tree plots located in the Central Range. The design used to select the plots tried to include, for each region, a balanced representation of the possible range of age, density, and site quality ([García Güemes 1999](#)). Around each plot, a buffer area with similar conditions was also defined.

For every tree, breast-height diameter over bark (measured at 1.30 m above ground), total tree height, height to crown base, and crown diameter were measured. The position of each tree with respect to the centre of the plot was recorded. In the five nearest-to-the-centre trees, total age and radial growth for the last 5 years were also measured.

A multivariate analysis based on the distribution of the Mahalanobis distance between the observations and their expectations was carried out to detect outlier data. Among all the plots falling out from the multivariate normal distribution, 15 plots were rejected because they represented special cases, namely very low dense stands, overmature stands, or stands growing on moist or rocky lands. The remaining 455 plots were divided into two groups. Ten percent of them, 45 plots, were selected randomly within the four regions studied as a validation data set, while the rest were used as a fitting data set to develop the model. The total number of trees available where both variables, height and diameter, were measured was 7754 for the fitting data set and 860 for the validation data set. Summary statistics for both fitting and validation data sets are shown in Table 1.

Methods

Function selection

Several functions to model the relationship between breast-height diameter and total height of the trees in a stand

Table 1. Summary statistics of breast-height diameter and tree height for model fitting data set and model validation data set.

Data set	Region	N*	Diameter (cm)		Height (m)	
			Mean	Range	Mean	Range
Fitting	Central Range	816	34.4	11.2–81.4	10.47	3.82–23.00
	Catalonia	1280	26.8	1.0–52.4	9.91	1.82–20.14
	West Andalusia	3320	22.5	2.5–57.6	8.32	1.60–22.00
	Northern Plateau	2338	28.4	2.5–69.3	8.79	1.35–22.50
Validation	Central Range	100	35.5	23.9–49.9	11.13	6.75–17.30
	Catalonia	140	31.9	7.5–56.8	11.50	3.89–17.75
	West Andalusia	420	21.0	3.0–44.2	7.95	2.00–18.00
	Northern Plateau	200	32.9	6.0–66.3	10.04	1.68–19.00

*Number of trees.

Table 2. Height–diameter functions tested.

Function No.	Function	Source
II.1	$H = aD^b$	Stage 1975
II.2	$H = \exp\left(a + \frac{b}{D+1}\right)$	Wykoff et al. 1982
II.3	$H = a \exp\left(\frac{b}{D}\right)$	Burkhart and Strub 1974
II.4	$H = a[1 - \exp(-bD)]$	Meyer 1940
II.5	$H = \frac{aD}{b+D}$	Bates and Watts 1980

Note: D is breast-height diameter, H is total tree height, and a and b are fitting parameters.

are available (Curtis 1967; Huang et al. 1992; Fang and Bailey 1998). In this study, five functions (Table 2) were compared to determine which showed the best predictive ability and fit to the height–diameter data. All functions tested have two parameters, are asymptotic, and are nonlinear.

Functions were fitted independently to the data from each of the 410 plots in the fitting data set (Zakrzewski and Bella 1988; Fang and Bailey 1998). Nonlinear regressions using ordinary nonlinear least squares (ONLS) techniques were carried out with the SAS NLIN procedure (SAS Institute Inc. 2000). To determine which function describes the relationship between height and diameter for the individuals of stone pine more accurately, the following statistics, proposed by Arabatzis and Burkhart (1992), were calculated:

(1) Mean residual, $\bar{E}_i$, for plot i :

$$\bar{E}_i = \frac{\sum_{k=1}^{n_i} e_{ik}}{n_i}$$

(2) Model precision, V_i , for plot i :

$$V_i = \frac{\sum_{k=1}^{n_i} e_{ik}^2 - \left(\sum_{k=1}^{n_i} e_{ik}\right)^2 / n_i}{(n_i - 1)n_i}$$

(3) Mean square error, MS_i , for plot i :

$$MS_i = \bar{E}_i^2 + V_i$$

(4) Mean square error of prediction, MSE_i , for plot i :

$$MSE_i = \frac{\sum_{k=1}^{n_i} (e_{ik})^2}{n_i - 1}$$

where e_{ik} is the residual from the measurement of the k th tree in the i th plot (observed minus predicted height), and n_i is the number of trees in plot i .

Once value of these statistics was calculated for each plot, mean values of $\bar{E}_i$, V_i , MS_i , and MSE_i were computed over the 410 plots. These mean values, defined as $\bar{E}$, V , MS , and MSE , served as criteria to compare proposed models. To detect any pattern of nonconstancy in residual variances, residual plots showing means of squared residuals in 5-cm diameter classes were developed for each function (Hokka 1997; Lappi 1997).

Nonlinear mixed modelling

Available data were based on a sample of multiple measurements (diameter at breast height – height) taken from different plots located in different regions. As a consequence of this nested structure, we detected high correlation among measurements taken from the same plot. To alleviate this, a mixed-model approach has been proposed by other authors (Laird and Ware 1982; Palmer et al. 1991; Gregoire et al. 1995; Wolfinger 1997b). A general expression for a nonlinear mixed-effects model can be defined as (Lindstrom and Bates 1990; Vonesh and Chinchilli 1997; Pinheiro and Bates 1998)

$$[1] \quad y_{ij} = f(\Phi_i, x_{ij}) + e_{ij}$$

where y_{ij} is the j th observation (tree) of the response variable taken from the i th sampling unit (plot in our study case), x_{ij} is the j th measurement from the predictor variable taken from the i th plot, Φ_i is a parameter vector, $r \times 1$ (where r is the number of parameters in the model), specific for each sampling unit, f is a nonlinear function of the predictor variables and the parameter vector, and e_{ij} is the residual noise term. In vector form:

$$[2] \quad \mathbf{y}_i = f(\Phi_i, \mathbf{x}_i) + \mathbf{e}_i$$

where $\mathbf{y}_i$ is the $n_i \times 1$ vector including the n_i complete observations from the i th plot $[y_{i1}, y_{i2}, \dots, y_{in_i}]^T$, $\mathbf{x}_i$ is the $n_i \times 1$ predictor vector for the n_i observations of the predic-

tor variable x taken from the i th plot $[x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{in_i}]^T$, and $\mathbf{e}_i$ is a $n_i \times 1$ vector for the residual terms $[e_{i1}, e_{i2}, \dots, e_{ij}, \dots, e_{in_i}]^T$.

The main feature of mixed-effects models is that they allow parameter vectors to vary from plot to plot; regression coefficients are broken down into a fixed part, common to the population, and random components, specific to each plot. The parameter vector, Φ_i , can then be defined as (Pinheiro and Bates 1998):

$$[3] \quad \Phi_i = \mathbf{A}_i \boldsymbol{\lambda} + \mathbf{B}_i \mathbf{b}_i$$

where $\boldsymbol{\lambda}$ is the $p \times 1$ vector of fixed population parameters (where p is the number of fixed parameters in the model), $\mathbf{b}_i$ is the $q \times 1$ vector of random effects associated with the i th plot (where q is the number of random parameters in the model), $\mathbf{A}_i$ and $\mathbf{B}_i$ are design matrices of size $r \times p$ and $r \times q$, for the fixed and random effects specific to each plot, respectively. Elements of these design matrices are usually 0, 1, or the value of the covariates associated to the fixed and (or) random effects.

Basic assumptions for the nonlinear mixed-model theory include the asymptotic multivariate normal distribution for both the random-effects vector and the residual-terms vector:

$$\begin{aligned} \mathbf{b}_i &\sim N(0, \mathbf{D}) \\ \mathbf{e}_i &\sim N[0, \mathbf{R}_i(\boldsymbol{\lambda}, \mathbf{b}_i, \boldsymbol{\rho})] \end{aligned}$$

where $\mathbf{D}$ is a positive-definite variance-covariance matrix $q \times q$, for the random effects, representing the among-plot variability, assumed to be common to every plot, and $\mathbf{R}_i(\boldsymbol{\lambda}, \mathbf{b}_i, \boldsymbol{\rho})$ is the $n_i \times n_i$ intraindividual variance-covariance matrix, which defines within-plot variability. $\mathbf{R}_i$ is allowed to depend on both random and fixed effects, as well as on a set of common but unknown parameters $\boldsymbol{\rho}$.

The usual aim of an analysis of this type is the estimation of $\boldsymbol{\lambda}$, $\mathbf{D}$, and $\boldsymbol{\rho}$. To make inference over these parameters according to maximum-likelihood methods, the joint density function is maximized (Lindstrom and Bates 1990; Palmer et al. 1991; Pinheiro and Bates 1995; Vonesh and Chinchilli 1997):

$$[4] \quad p(\mathbf{y} | \mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\rho}, \mathbf{D}) = \prod_{i=1}^s \int p(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\lambda}, \boldsymbol{\rho}, \mathbf{b}_i) p(\mathbf{b}_i, \mathbf{D}) d\mathbf{b}_i$$

where s is the number of sampling units. Maximization of this expression requires the computation of the integral over the random effects. This is not an easy task, since $\mathbf{b}_i$'s enter the model in a nonlinear fashion. Several approaches have been developed to deal with this problem, including methods such as the adaptive Gaussian quadrature (Pinheiro and Bates 1995), which tries to solve the integral numerically. This method is available in the SAS NLMIXED procedure (SAS Institute Inc. 2000).

The most commonly used methods are those based on a linear approximation to the marginal likelihood function by expanding it with a first-order Taylor series to be linear on $\mathbf{b}_i$. This expansion can be either at a value of $\mathbf{b}_i = 0$ (Sheiner and Beal 1980) or about a value $\mathbf{b}_i^*$ close to $\mathbf{b}_i$ (Lindstrom and Bates 1990). The last option is available in the SAS macro NLINMIX (Wolfinger 1997a).

The NLINMIX macro, after linear expansion of marginal likelihood function over $\mathbf{b}_i^*$, generates a "pseudoresponse"

vector $\mathbf{w}_i$, linear both in the random parameters $\mathbf{b}_i$ as well as in the fixed parameters $\boldsymbol{\lambda}$, by expanding around $\boldsymbol{\lambda}^*$ close to $\boldsymbol{\lambda}$ (Gregoire and Schabenberger 1996). After this, algorithms for linear mixed-effects models (Searle 1971; Laird and Ware 1982) can then be used, thus simplifying the process.

The Taylor series expanded version of the $\mathbf{y}_i$ vector is given by the expression

$$[5] \quad \mathbf{y}_i \approx f(\Phi_i^*, \mathbf{x}_i) + \mathbf{Z}_i(\Phi_i^*, \mathbf{x}_i)(\mathbf{b}_i - \mathbf{b}_i^*) + \mathbf{X}_i(\Phi_i^*, \mathbf{x}_i)(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*) + \mathbf{e}_i$$

where

$$\begin{aligned} \Phi_i^* &= \mathbf{A}_i \boldsymbol{\lambda}^* + \mathbf{B}_i \mathbf{b}_i^* \\ \mathbf{Z}_i(\Phi_i^*, \mathbf{x}_i) &= \frac{\partial}{\partial \mathbf{b}_i} f(\Phi_i^*, \mathbf{x}_i) \\ \mathbf{X}_i(\Phi_i^*, \mathbf{x}_i) &= \frac{\partial}{\partial \boldsymbol{\lambda}} f(\Phi_i^*, \mathbf{x}_i) \end{aligned}$$

defining

$$\mathbf{w}_i = \mathbf{y}_i - f(\Phi_i^*, \mathbf{x}_i) + \mathbf{Z}_i(\Phi_i^*, \mathbf{x}_i)(\mathbf{b}_i^*) + \mathbf{X}_i(\Phi_i^*, \mathbf{x}_i)(\boldsymbol{\lambda}^*)$$

and rearranging terms in eq. 5:

$$[6] \quad \mathbf{w}_i = \mathbf{X}_i \boldsymbol{\lambda} + \mathbf{Z}_i \mathbf{b}_i + \mathbf{e}_i$$

In this case, the pseudoresponse vector $\mathbf{w}_i$ is linear both in $\mathbf{b}_i$ and $\boldsymbol{\lambda}$, being then identified with a linear mixed model (Harville 1977; Vonesh and Chinchilli 1997). Marginal moments for the pseudoresponse vector are determined by the following expressions:

$$\begin{aligned} [7] \quad E(\mathbf{w}_i | \mathbf{x}_i) &\approx \mathbf{X}_i(\Phi_i^*, \mathbf{x}_i) \boldsymbol{\lambda} \\ \mathbf{V}_i = \text{Var}(\mathbf{w}_i | \mathbf{x}_i) &\approx \mathbf{Z}_i(\Phi_i^*, \mathbf{x}_i) \mathbf{D} \mathbf{Z}_i^T(\Phi_i^*, \mathbf{x}_i) \\ &\quad + \mathbf{R}_i(\boldsymbol{\lambda}, \mathbf{b}_i^*, \boldsymbol{\rho}) \end{aligned}$$

In accordance with these considerations, the inference of $\boldsymbol{\lambda}$ can be carried out using linear theory (Searle 1971), obtaining generalized least squares estimates for $\boldsymbol{\lambda}$, identified as $\hat{\boldsymbol{\lambda}}$:

$$[8] \quad \hat{\boldsymbol{\lambda}} = \left(\sum_{i=1}^s \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^s \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{w}_i$$

Estimates for the variance components for the matrices $\mathbf{D}$ and $\mathbf{R}_i$, $\hat{\mathbf{D}}$ and $\hat{\mathbf{R}}_i$, respectively, are obtained in the SAS macro NLINMIX by solving an iterative Newton-Raphson algorithm, allowing solutions for both maximum likelihood and restricted maximum likelihood (Harville 1977) methods. From these matrices, the estimate for $\mathbf{V}_i$, $\hat{\mathbf{V}}_i$, is easily computed using eq. 7.

Finally, under linear and normal conditions, $\mathbf{b}_i$ is approximated by the value of the empirical best linear unbiased predictor (EBLUP) for the parameters of the random effects in plot i , $\hat{\mathbf{b}}_i$. EBLUP indicates that it is not the real best linear unbiased predictor (BLUP), since it has been estimated using estimates for the variance components (Gregoire et al. 1995; Fang and Bailey 2001). The equation for these EBLUPs is then given by (Searle 1971):

$$[9] \quad \hat{\mathbf{b}}_i = \hat{\mathbf{D}}\mathbf{Z}_i^T \hat{\mathbf{V}}_i^{-1}(\mathbf{w}_i - \mathbf{X}_i \hat{\boldsymbol{\lambda}})$$

Goodness-of-fit criteria

Models can be compared using both the Akaike information criterion (AIC) and the $-2 \times$ logarithm of likelihood function ($-2LL$). AIC has been widely used to compare the reliability of different models with alternative suites of fixed and random effects or to compare non-nested models. Models preferred are those where AIC is larger. A likelihood ratio test, taking into account the value of $-2LL$ between two models, was conducted only in those cases where a model was a special case of the other (nested models).

Model construction

According to Fang and Bailey (2001), during the construction of a mixed model it is necessary to determine the following:

- (1) Which of the parameters in the model are going to be considered as parameters of a mixed effect, composed of a fixed part (common to all data in the sample) and of a random part (specific for every sampling plot), or as parameters of a purely fixed effect.
- (2) The within-plot variance–covariance structure for explaining variability among trees in the same plot.
- (3) The structure of the among-plot variance–covariance matrix.
- (4) Which covariates can be used to explain among-plot variability.
- (5) Whether it is necessary to include regional effects to model interregional variability in the height–diameter relationship.

Steps 1–3 define the basic nonlinear mixed model, while steps 4 and 5 enable among-plot variability to be reduced, explaining it by the inclusion of both stand and regional variables in the model.

Parameter effects

Determination of which parameters are going to vary randomly among plots is a flexible decision subject to debate (Gregoire et al. 1995). If convergence is met, Pinheiro and Bates (1995) propose in a first attempt to consider all the parameters in the model as mixed, composed of a fixed and a random varying part.

If the convergence of the model is not met when considering all the parameters as mixed, parameters with high variability and less overlap in confidence intervals across the plots among those estimated in the previous individual fit to every plot can be proposed as random ones (Fang and Bailey 2001).

Within-plot variance–covariance structure

Commonly, for height–diameter relationships there is a pattern of increasing error variance as a function of the dependent variable within a plot (Larsen and Hann 1987; Cañadas 2000; Huang et al. 2000). We can also identify plot-level observations that are more intercorrelated than the average, as is the case when trees are arranged in clumps or we have repeated measurements from the same tree. In any case, heteroskedasticity in residual variance pattern or correlation among observations violate basic assumptions of statistical inference. To avoid this, a special structure for the within-plot variance–covariance matrix $\mathbf{R}_i(\boldsymbol{\lambda}, \mathbf{b}_i, \boldsymbol{\rho})$ can be

proposed, to include both correlation effects and weighting factors to balance error variance. In these cases general expression for the matrix is given by

$$[10] \quad \mathbf{R}_i(\boldsymbol{\lambda}, \mathbf{b}_i, \boldsymbol{\rho}) = \sigma^2 \mathbf{G}_i^{0.5} \boldsymbol{\Gamma}_i \mathbf{G}_i^{0.5}$$

where for a given plot i with n_i height–diameter measurements $\mathbf{G}_i$ is a $n_i \times n_i$ diagonal matrix describing non-homogenous variance, since its components are standard deviations of the residual errors; $\boldsymbol{\Gamma}_i$ is a $n_i \times n_i$ matrix showing the structure of the correlation among observations for plot i ; and σ^2 is a scaling factor for the error dispersion (Gregoire et al. 1995), given by the value of the residual variance of the model. In our case, as data did not show any pattern of correlation among observations from the same plot, $\boldsymbol{\Gamma}_i$ was the identity matrix $n_i \times n_i$.

Among-plot variance–covariance structure

The variance–covariance matrix for the random effects, $\mathbf{D}$, common for all the plots, defines variability existing among plots. Multiple structures, including spatial and temporal correlation, have been proposed (Gregoire 1987; Wolfinger 1997b). In our study case, as tested functions have two parameters, we first propose to fit a mixed model including two random parameters u and v (Pinheiro and Bates 1995). Then $\mathbf{D}$ is a 2×2 matrix. The variance–covariance matrix proposed is the following:

$$[11] \quad \mathbf{D} = \begin{bmatrix} \sigma_u^2 & \sigma_{uv}^2 \\ \sigma_{uv}^2 & \sigma_v^2 \end{bmatrix}$$

where σ_u^2 is the variance for the random effect u , σ_v^2 is the variance for the random effect v , and σ_{uv}^2 is the covariance among random effects.

Fitting basic nonlinear mixed model

Once we have selected which effects are going to be designated as mixed, as well as the structure for both within-plot and among-plot variance–covariance matrices, we can calculate the components of the variance–covariance matrix for the random effects, the residual variance of the model, the estimates for the mean value, and the confidence intervals for the parameters of the fixed effects, and we can predict the value of the random effects EBLUPs for every plot. The SAS macro NLINMIX (SAS Institute Inc. 2000) was used, and maximization of the marginal likelihood function was achieved using the EBLUP approximation (Lindstrom and Bates 1990), with a first-order Taylor series expansion about the estimates of the random parameter. The method used to estimate variance components was restricted maximum likelihood (Harville 1977).

Covariate modelling

To diminish variability detected among plots, variables characterizing the stand are included in the mixed model as fixed effects (Hokka 1997; Pinheiro and Bates 1998; Singer 1998). Stand variables tested were basal area, mean square diameter, dominant height, age, density, site index (Calama et al. 2003), and the value of the difference between the 90th and 10th percentiles from the plot diameter distribution, as proposed by Fang and Bailey (1998). In the analysis, logarithmic transformations of these variables were also tested.

Linear models including different stand variables were fit to explain variation in the random effects. Linear modelling of the random effects was achieved using the SAS REG procedure. The algorithm selected for including variables in the model was maximum increase in the determination coefficient R^2 , a method which selects the best model including one, two, and successive variables.

Interregional variability

The hierarchical structure of the data set, trees within plots, justifies mixed modelling as proposed. In the same way, plots are located in four different regions, creating a multilevel hierarchical structure (Singer 1998; Hall and Bailey 2001; Sheu and Suzuki 2001).

Interregional variability in height–diameter models has been sparsely treated in forest research. Huang et al. (2000) built separate height–diameter models for different ecological regions, testing homogeneity in parameters among regions using a nonlinear sum of squares F test, derived from the likelihood ratio test. Jayaraman and Lappi (2001) developed a multilevel height–diameter mixed model, which included random effects at the plot and forest territorial division level, and analysed the variability of the EBLUPs among territorial units.

In this study, interregional variability in height–diameter model was studied by including regional effects as dummy categorical variables. These variables take a value of 1 if the observation comes from the region and 0 if it comes from another region.

In a first step, a model including all the regional effects was fitted to the data. To evaluate interregional differences, models incorporating additional hypothesis of equal regional effects were also fitted. After testing the significance of the parameters, a log-likelihood ratio test was carried out to compare the constrained models with the model including all regional dummies. If the ratio test is nonsignificant, it means that the constrained model can be used instead of the full model, indicating that null hypothesis of equality among regions cannot be rejected.

Prediction

The main purpose of the model is to be used as a predictive tool in forest management. The introduction of random parameters into the model allows us to consider two different situations in a predictive role (Vonesh and Chinchilli 1997; Fang and Bailey 2001; Fang et al. 2001):

- (1) Fixed-effects response pattern: prediction of tree height in those stands where diameter and stand variables included in the model are measured.
- (2) Calibrated response pattern: prediction of tree height in stands where, together with the diameter of the trees and stand variables, heights are measured in a subsample of trees.

To evaluate the predictive ability of the model, data included in the validation data set are used. Evaluation is carried out using the following statistics:

$$\text{Modelling efficiency (EF)} = 1 - \frac{\sum_{i=1}^s \sum_{j=1}^{n_i} (h_{ij} - \hat{h}_{ij})^2}{\sum_{i=1}^s \sum_{j=1}^{n_i} (h_{ij} - \bar{h})^2}$$

$$\text{Error} = \frac{\sum_{i=1}^s \sum_{j=1}^{n_i} (h_{ij} - \hat{h}_{ij})}{n}$$

$$\text{Sum of squares error (SSE)} = \sum_{i=1}^s \sum_{j=1}^{n_i} (h_{ij} - \hat{h}_{ij})^2$$

where h_{ij} is the value of the observed height, $\hat{h}_{ij}$ is the value for the predicted height obtained when applying the model, $\bar{h}$ is the mean value for the measured heights, n is the total number of observations used in the validation, s is the number of plots in validation data set, and n_i is the number of trees in the i th plot within validation data set (s plots).

Case 1. Prediction over a particular stand with no prior height–diameter observations: fixed effects typical response

In this case, we only have diameter data as well as values for the stand and regional variables included in the model. As we do not have height measurements, we cannot make predictions about the value of the EBLUPs for the random effects $\mathbf{b}_i$. We then consider that the best linear unbiased predictor of $\mathbf{b}_i$ is the null $q \times 1$ vector (where q is the number of random effects in the model). The general expression for the nonlinear mixed model is the following:

$$\begin{aligned} \hat{h}_{ij} &= f(\hat{\Phi}_i, d_{ij}) + e_{ij} \\ [12] \quad \hat{\Phi}_i &= \begin{bmatrix} \hat{\alpha}_i \\ \hat{\beta}_i \end{bmatrix} = \mathbf{A}\hat{\lambda} + \mathbf{B}_i\hat{\mathbf{b}}_i \end{aligned}$$

where d_{ij} is the diameter at breast height of the j th tree in the i th plot. $\mathbf{A}_i$ and $\hat{\lambda}$ are, respectively, design matrix and estimated vector for the fixed effects, $\mathbf{B}_i$ is design matrix for random effects, and $\hat{\mathbf{b}}_i$ is a $q \times 1$ null vector. $\hat{h}_{ij}$ is the predicted height for the j th tree in the i th plot. This function represents the pattern of typical response and defines the mean behaviour of the pattern of variation in height versus diameter for a given stand and regional conditions.

Case 2. Prediction is carried out over stands with prior diameter–height observations: calibrated response

Prediction is required for a plot where diameter data and stand and regional variables are available, and in a subsample of k trees both diameter and height have been measured. In this situation, we can predict the value of the vector $\mathbf{b}_i$ of random effects, approaching the behaviour of the height–diameter relationship for a given stand. Prediction of $\mathbf{b}_i$ is carried out with the following expression (Vonesh and Chinchilli 1997):

$$[13] \quad \hat{\mathbf{b}}_i \approx \hat{\mathbf{D}}\hat{\mathbf{Z}}_i^T(\hat{\mathbf{R}}_i + \hat{\mathbf{Z}}_i\hat{\mathbf{D}}\hat{\mathbf{Z}}_i^T)^{-1}\hat{\mathbf{e}}_i$$

where $\hat{\mathbf{D}}$ is the $q \times q$ variance–covariance matrix (where q is the number of random effects included in the model) for the among-plot variability, common for all plots and estimated in the general fitting of the model; $\hat{\mathbf{R}}_i$ is the $k \times k$ variance–covariance matrix for within-plot variability; $\hat{\mathbf{e}}_i$ is the residual vector $k \times 1$, whose components $\hat{e}_{ij}$ are given by the difference between the observed height value for j th tree included in the subsample, h_{ij} , and the value predicted using the model including only fixed effects:

$$\hat{e}_{ij} = h_{ij} - f(\hat{\Phi}_i, d_{ij})$$

where $\hat{\Phi}_i$ in this case equals $\mathbf{A}_i\hat{\lambda}$, including only the fixed part of the estimated vector of parameters, and $\hat{\mathbf{Z}}_i$ is the $k \times q$ matrix evaluated at $\hat{\lambda}$:

$$[14] \quad \begin{bmatrix} \frac{\partial f(d_{i1}, \hat{\Phi}_i)}{\partial \lambda_1} & \dots & \frac{\partial f(d_{i1}, \hat{\Phi}_i)}{\partial \lambda_q} \\ \vdots & & \vdots \\ \frac{\partial f(d_{ik}, \hat{\Phi}_i)}{\partial \lambda_1} & \dots & \frac{\partial f(d_{ik}, \hat{\Phi}_i)}{\partial \lambda_q} \end{bmatrix}$$

where $\hat{\Phi}_i$ is as previously stated, $\lambda_1, \dots, \lambda_q$ are the fixed part of the mixed coefficients components of the vector for estimated fixed effects $\hat{\lambda}$, and d_{ij} is the diameter of the j th tree.

Once $\hat{b}_i$ is predicted, the value of the vector of heights, $\hat{h}_i$, for plot i , i.e., the calibrated response vector, is defined by the expression

$$[15] \quad \begin{aligned} \hat{h}_i &= f(\hat{\Phi}_i, d_i) + e_i \\ \hat{\Phi}_i &= \begin{bmatrix} \hat{\alpha}_i \\ \hat{\beta}_i \end{bmatrix} = \mathbf{A}_i\hat{\lambda} + \mathbf{B}_i\hat{b}_i \end{aligned}$$

Comparing different sample sizes for prediction

In this section, we determine and compare the predictive efficiency of the model over the data included in the validation data set. SSE, EF, and error statistics are used to evaluate different sampling design alternatives, with or without prediction for the random effects from the measurement of a subsample of tree heights in the plots. Different subsample sizes are also tested. Proposed alternatives are the following:

- (1) Only stand variables, together with the diameter of the trees, are included in the model.
- (2) Stand variables and the height of, respectively, 1, 2, 4, 6, 8 and 10 trees, randomly selected in each case, are measured.
- (3) Stand variables and the height of the four dominant trees, considered as the largest trees in the plot, are measured.

These alternatives are evaluated, and they are also compared with the value obtained in the individual fit, following ONLS techniques, of the selected function to all the data included in each plot. In those alternatives where the measurement of a subsample of randomly selected trees is required, results for the statistics SSE, mean error, and EF are estimated as the mean value after 500 realizations.

Results

Function selection

Results from the comparative analysis between studied functions, including values of the statistics E , V , MS , and MSE , are shown in Table 3. Although results are similar for every function, II.2 and II.5 show a slightly better predictive ability than the rest. When fitting function II.5 individually, meaningless parameter estimates were obtained for a few plots. Therefore, we decided to select function II.2, a general expression of the relationship between height and diameter widely used in forest research (Wykoff et al. 1982; Larsen and Hann 1987; Moore et al. 1996):

Table 3. Performance criteria for tested height–diameter models.

Function No.	E	V	MS	MSE
II.1	0.009 10	0.035 15	0.058 73	0.636 54
II.2	0.000 39	0.035 16	0.035 17	0.611 60
II.3	0.000 56	0.035 20	0.035 21	0.612 25
II.4	−0.000 99	0.035 35	0.035 40	0.614 17
II.5	−0.001 24	0.035 14	0.035 19	0.611 25

Note: E , mean residual; V , model precision; MS , mean square error; MSE , mean square error of prediction.

$$[16] \quad h_{ij} - 1.3 = \exp\left(\alpha_i + \frac{\beta_i}{d_{ij} + 1}\right) + e_{ij}$$

From this point, h_{ij} is the height (metres) of the j th tree in the i th plot, and d_{ij} is the diameter at breast height (centimetres) of the mentioned tree. This function was used in the development of the stand model PROGNOSIS (Wykoff et al. 1982), and it is based on the linear relation proposed by Curtis (1967), where a value of +1 is added to the diameter to avoid meaningless estimates when values of breast height diameter are near zero.

Model construction

Parameter effects

In the first step, we proposed that parameters α_i and β_i in eq. 16 be considered as mixed, i.e., composed of a fixed part and a random part, specific for each plot. The general expression for the mixed model so defined is given by

$$[17] \quad h_{ij} - 1.3 = \exp\left[(a + u_i) + \frac{b + v_i}{d_{ij} + 1}\right] + e_{ij}$$

Under the basic assumptions

$$\begin{aligned} \begin{bmatrix} u_i \\ v_i \end{bmatrix} &\sim N(0, \mathbf{D}) \\ e_i &\sim N[0, \mathbf{R}_i(\lambda, b, \rho)] \end{aligned}$$

where a and b are considered fixed parameters, common to every plot; u_i and v_i are random parameters, specific for plot i ; d_{ij} and h_{ij} are as defined before; and e_{ij} represents the estimation error for the j th observation in the i th plot.

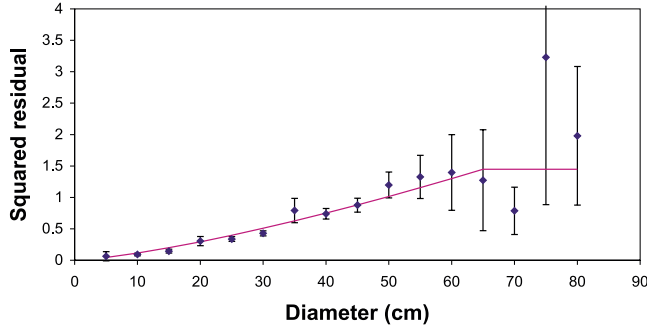
Within-plot variance–covariance structure

Graphical analysis of the mean squared residuals in 5-cm diameter classes obtained when fitting eq. 16 individually to the data of each plot indicated that the variance of the errors tended to increase as tree size increased (Fig. 1). The variance of the residual error can then be expressed as a function of the diameter, using the following variance power function:

$$\text{Var}(e_{ij}) = p[\min(d_{ij}, 60)^q]$$

where p and q are parameters, and d_{ij} is breast-height diameter (centimetres). Appropriate values for p and q were searched for by fitting the power function with the group

Fig. 1. Mean squared residuals and standard errors by tree diameter classes of eq. 16 fitted using ordinary nonlinear least squares in each plot. The solid line indicates estimated variance function.



means of squared residual in diameter classes. Values so obtained were 0.0051 and 1.353 for p and q , respectively.

In this case, the within-plot variance–covariance matrix is the diagonal matrix

$$[18] \quad \mathbf{R}_i(\boldsymbol{\lambda}, \mathbf{b}_i, \boldsymbol{\rho}) = \sigma^2 \mathbf{G}_i^{0.5} \mathbf{I}_{n_i} \mathbf{G}_i^{0.5}$$

where σ^2 is the residual variance of the model; $\mathbf{G}_i$ is the diagonal matrix of range the number of observations per plot, whose components are given by the variance power function; and $\mathbf{I}_{n_i}$ is the identity matrix with range the number of observations in the plot.

$$[19] \quad h_{ij} - 1.3 = \exp \left[(a + fH_{o_i} + gP_{9010_i} + u_i) + \frac{b + j \log(\text{dens}_i) + v_i}{d_{ij} + 1} \right] + e_{ij}$$

Results from fitting this model to the fitting data set under variance–covariance structures proposed previously are included in Table 4. When comparing these results with those from the fit of the basic model, we find a substantial increment in AIC, despite the inclusion of new parameters in the model. In addition to this, the values for the components of the variance associated with the random effects decrease significantly. Residual within-plot variance remains almost constant, showing that there is a pattern of variability not explained by among-plot differences.

Interregional variability

To determine and evaluate differences among regions, the following dummy variables are created: (1) $V_1 = 1$ if the observation comes from the Central Range, 0 in the rest of cases, (2) $V_2 = 1$ if the observation comes from Catalonia, 0 in the rest of cases, and (3) $V_3 = 1$ if the observation comes from West Andalusia, 0 in the rest of cases. Although the analysis is being carried out over four regions, the inclusion of three dummy variables is enough, since regional parameters associated with the Northern Plateau are identified as zeroes.

The nonlinear mixed model after including regional variables as fixed effects is then defined by the following expression:

$$[20] \quad h_{ij} - 1.3 = \exp \left[(a + fH_{o_i} + gP_{9010_i} + \phi_1 V_1 + \phi_2 V_2 + \phi_3 V_3 + u_i) + \frac{b + j \log(\text{dens}_i) + \psi_1 V_1 + \psi_2 V_2 + \psi_3 V_3 + v_i}{d_{ij} + 1} \right] + e_{ij}$$

Results from fitting eq. 20 are included in Table 4. The regional parameter estimates for the Central Range, ϕ_1 and ψ_1 , are not significant at an α level of 5%. Then we can consider a common regional effect for the Central Range and the Northern Plateau, since the regional effect in the latter region is considered null. Equation 20, defined as the full regional model, was compared with different constrained regional models, based on the assumption of equality among different regional effects. Comparisons were carried out on the basis of the log-likelihood ratio test. Results for the com-

Basic nonlinear mixed model

Results from fitting nonlinear mixed model defined by eq. 17 under the assumptions proposed in eqs. 11 and 18 to the data from the 7754 trees included in the fitting data set are shown in the first column of the Table 4.

Inclusion of stand covariates

EBLUPs of the random effects were explained using linear regression over the stand variables previously mentioned. To avoid overparameterization and collinearity in the model, only those variables largely correlated with the EBLUPs were included. The final proposition for explaining random parameters was the following:

$$u^* = a + fH_o + gP_{9010} + u$$

$$v^* = b + j \log(\text{dens}) + v$$

where f , g , and j are considered fixed parameters for dominant height (H_o) the difference between the 90th and 10th percentiles from the diameter distribution (P_{9010}) and the logarithm of the density ($\log(\text{dens})$), respectively; a and b are fixed parameters indicating the intercept of the model; u and v are expanded plot-specific random effects, and u^* and v^* are original plot-specific random effects.

After including stand covariates as fixed effects, the general expression for the nonlinear mixed model is as follows:

parison are shown in Table 5. These results lead us to propose a common regional model for the Central Range and the Northern Plateau and another model for Catalonia and West Andalusia. To this end, we propose the inclusion of a single regional variable in eq. 19: $K = 1$ if the observation comes from Catalonia or West Andalusia; 0 if it comes from the Central Range or the Northern Plateau.

The expression for the nonlinear mixed model that includes both stand covariates and a single regional variable is as follows:

Table 4. Fitting parameters and comparison of mixed model effects performance for different alternatives of fixed effects inclusion.

	Parameter	Basic nonlinear mixed model	Stand covariates inclusion	Regional effects and stand covariates	Definite model
Fixed parameters	a	2.4749	1.6657	1.7226	1.7306
	b	-13.3366	-22.1311	-26.2565	-25.2776
	f	—	0.0888	0.08874	0.0882
	g	—	-0.0067	-0.0055	-0.0062
	j	—	1.6878	1.8463	1.6999
Regional fixed effects	ϕ_1	—	—	-0.0212	-0.0936
	ϕ_2	—	—	-0.0776	—
	ϕ_3	—	—	-0.0649	—
	ψ_1	—	—	0.7379	4.7431
	ψ_2	—	—	4.5692	—
	ψ_3	—	—	1.0464	—
	σ_u^2	0.1518	0.0170	0.0115	0.0129
Variance components	σ_v^2	37.8072	20.7510	12.6687	13.3239
	σ_{uv}^2	-1.3771	-0.4963	-0.3089	-0.3273
	σ^2	1.4393	1.4776	1.4845	1.4847
	σ^2	1.4393	1.4776	1.4845	1.4847
Model performance	AIC	-9903.45	-9421.0	-9378.95	-9378.2
	-2LL	19 798.9	18 834.0	18 749.9	18 748.4

Note: AIC, Akaike information criterion; -2LL, $-2 \times$ logarithm of likelihood function.

$$[21] \quad h_{ij} - 1.3 = \exp \left[(a + fH_{o_i} + gP_{9010_i} + \phi_1 K + u_i) + \frac{b + j \log(\text{dens}_i) + \psi_1 K + v_i}{d_{ij} + 1} \right] + e_{ij}$$

Results after fitting previous function are included in Table 4. AIC values for both the full regional model (eq. 20) and the reduced regional model (eq. 21) are very similar. Since in the reduced model all the parameters are significantly different from zero, the number of parameters to estimate is smaller, and the log-likelihood test comparing both models is nonsignificant, we selected eq. 21 as the definite model.

After substituting in eq. 21 the value of the fixed parameters estimates, the general expression for the interregional height-diameter model for stone pine in Spain is the following:

$$[22] \quad h_{ij} - 1.3 = \exp \left[(1.7306 + 0.0882H_{o_i} - 0.0062P_{9010_i} - 0.0936K + u_i) + \frac{-25.2776 + 1.6999 \log(\text{dens}_i) + 4.743K + v_i}{d_{ij} + 1} \right] + e_{ij}$$

where

$$\mathbf{b}_i = \begin{bmatrix} u_i \\ v_i \end{bmatrix} \sim N \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.0129 & -0.3273 \\ -0.3273 & 13.3239 \end{bmatrix} \right\}; \quad \mathbf{e}_i \sim N(0, \mathbf{R}_i = 1.4847 \mathbf{G}_i^{0.5} \mathbf{I}_{n_i} \mathbf{G}_i^{0.5})$$

The units for the variables included in the formulae are metres for h and H_o , centimetres for d and P_{9010} , and stems per hectare for dens . K is a nondimensional dummy variable.

Prediction

Comparing sample size for prediction

Results from the analysis are included in Table 6. The largest values of SSE are obtained when applying typical fixed-effects response, without predicting random parameters. Furthermore, in this case error is significantly different from zero. When more measurements are included, successively, in the subsample, SSE and the mean value for the error decrease substantially, and the probability that the error

is nonsignificant is increased. The SSE value obtained when including in the subsample the four largest trees (20%) is larger than that obtained when calibrating with a random sample of four trees. A main reason for this could be that the sample of the largest trees is biased within the plot, unlike the random samples. Despite this, the prediction error in this case is nonsignificant.

Logically, the smallest value for SSE, together with largest probability of nonsignificance for error, is given by the individual fit of the model using ONLS.

An example of calibration

To show the predictive use of the mixed model, we finally

Table 5. Comparison of model performance for different alternatives of regional dummy variables inclusion.

Model	Separated regions	Combined regions	Regional parameters to estimate	-2LL	<i>p</i> LRT
1	WA, Cat, CR, NP	—	6*	18749.9	—
2	Cat, CR	WA-NP	4	18819.8	<0.0001
3	NP, CR	WA-Cat	4	18752.8	0.234
4	NP, Cat	WA-CR	4	18778.0	<0.0001
5	WA, NP	Cat-CR	4	18767.5	0.0001
6	WA, Cat	NP-CR	4	18745.5	1
7	WA, CR	NP-Cat	4	18794.0	<0.0001
8	CR	WA-Cat-NP	2	18826.7	<0.0001
9	Cat	WA-CR-NP	2	18825.2	<0.0001
10	NP	WA-Cat-CR	2	18778.9	<0.0001
11	WA	Cat-CR-NP	2	18791.5	<0.0001
12 [†]	—	WA-Cat and CR-NP	2	18748.4	1
13	—	WA-CR and Cat-NP	2	18819.5	<0.0001
14	—	WA-NP and CR-Cat	2	18834.0	<0.0001
15	—	WA-Cat-CR-NP	0	18834.0	<0.0001

Note: WA, West Andalusia; Cat, Catalonia; CR, Central Range; NP, Northern Plateau. Likelihoods ratios are calculated with respect to model 1. *p* LRT is the *p* value for the likelihood ratio test. Parameters for Northern Plateau are always 0.

*Not all the estimated parameters are significant.

[†]Model 12 is the selected model.

Table 6. Comparison among different subsample size alternatives.

Alternative*	SSE	EF (%)	Error (m)	<i>t</i>	Pr > <i>t</i>
1. Only stand variables are measured	854.86	92.69	0.1540	4.58	<0.0001
2. Stand variables + height measurement from 1 tree	743.39	93.64	0.1172	3.69	0.0002
3. Stand variables + height measurement from 2 trees	684.05	94.15	0.0939	3.08	0.0021
4. Stand variables + height measurement from 4 trees	617.95	94.72	0.0676	2.34	0.0196
5. Stand variables + height measurement from 6 trees	582.25	95.02	0.0474	1.68	0.0917
6. Stand variables + height measurement from 8 trees	562.23	95.19	0.0368	1.33	0.1825
7. Stand variables + height measurement from 10 trees	548.39	95.31	0.0299	1.09	0.2722
8. Stand variables+ height measurement from 4 (20%) largest trees	706.67	93.96	0.0594	1.93	0.0545
9. Individual fit to data from each plot (ONLS)	466.16	96.01	0.0003	0.01	0.9904

Note: SSE, sum of squares of the error; EF, modelling efficiency; error = $\frac{\sum(h_{ij} - \hat{h}_{ij})}{n}$; *t*, *t* statistic; Pr >|*t*|, probability for the absolute value of *t* to be distributed following a student *t* with *n* - 1 degrees of freedom.

*For alternatives 2–8, the means are obtained from 500 realizations.

develop a case study. Within the validation data set, we selected a plot from each region. Plots selected were 28105 (Central Range), 21057 (West Andalusia), 17043 (Catalonia), and 47060 (Northern Plateau). In these plots, stand variables as well as the diameter and the height of all the trees included in the plot were measured.

Case 1. Prediction over a particular stand with no prior height observations: fixed-effects typical response

Let us consider that in these plots, we are going to carry

out predictions using only stand variables as well as diameter data. This is defined as the fixed-effects typical response pattern. In this case, the expression used to define tree height would be

$$h_{ij} = \exp\left(\alpha_i + \frac{\beta_i}{d_{ij} + 1}\right) + e_{ij}$$

where

Table 7. Height–diameter model performance when applied to four plots from the validation data set.

Plot	Stand characterization			Case 1		Case 2		Case 3			
	H_o	P_{9010}	dens	SSE	Error	u	v	SSE	Error	SSE	Error
17043	13.1	17.5	245	24.933	0.752	0.0151	1.9233	13.755	−0.0757*	13.595	−0.0001*
21057	4.7	6.4	512	11.845	−0.685	0.0382	−2.7930	2.890	−0.1273*	1.903	0.0009*
28105	11.3	11.2	486	14.771	0.699	0.0154	1.1525	5.848	0.2188*	4.601	−0.0006*
47060	9.8	19.0	183	29.248	0.960	−0.0152	2.6033	14.989	0.6498	10.767	−0.0009*

Note: Case 1, typical fixed effects response (only stand covariates); case 2, calibrated response including empirical best linear unbiased predictors (EBLUPs) (four prior height measurements available); case 3, ONLS individual fit for the plot. H_o , dominant height (metres); P_{9010} , difference between the 90th and 10th percentiles from the diametric distribution (centimetres); dens, stand density (trees per hectare); u and v , EBLUPs for the random effects of the plot; SSE, sum of squares error; error = $\frac{\sum(h_{ij} - \hat{h}_{ij})}{n}$.

*Values are not significant with $p > 0.05$.

$$\begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} = \mathbf{A}_i \hat{\mathbf{X}} = \begin{bmatrix} 1 & H_{o_i} & P_{9010_i} & K_i & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \log(\text{dens}) & K_i \end{bmatrix} \begin{bmatrix} 1.7306 \\ 0.0882 \\ -0.0062 \\ -0.0936 \\ -25.2776 \\ 1.6999 \\ 4.7431 \end{bmatrix}$$

Results for the plots analysed, expressed in terms of mean residual error (mean values of the residual between predicted and observed values) and sum of squares error, are included in Table 7.

Case 2. Prediction is carried out over stands with prior height observations: calibrated response

In this case, let us suppose that in the plots we have available a measurement of the height of four trees randomly selected within the plot. As we have a subsample of trees with height measurements, we can predict the value of the EBLUPs, obtaining a calibrated function for each plot.

To predict the vector of random effects for the plot i , we use eq. 13:

$$\hat{\mathbf{b}}_i \approx \hat{\mathbf{D}}\hat{\mathbf{Z}}_i^T(\hat{\mathbf{R}}_i + \hat{\mathbf{Z}}_i\hat{\mathbf{D}}\hat{\mathbf{Z}}_i^T)^{-1}\hat{\mathbf{e}}_i$$

where matrices $\hat{\mathbf{D}}$ and $\hat{\mathbf{R}}_i$ are known, since we know the variance components for the model, estimated from eq. 22, and the structure of the matrices, $\hat{\mathbf{Z}}_i$ is developed using eq. 14, and $\hat{\mathbf{e}}_i$ is the residual vector, calculated as the difference between the real value of the trees measured in the subsample and the predicted value of the height for these trees, calculated as the typical response, including only fixed effects.

As we knew the value for the height of all the trees in these plots, we tested results obtained from both cases against real values, comparing the statistics sum of squares error, error, and level of significance of the error. Results from this comparison are included in Table 7. In Table 7, results for these statistics obtained in the individual fit for each plot using ONLS are also included. ONLS fit is carried out using all the trees from the plot. In this case, individual fit using ONLS is considered the best possible fit of the func-

tion to the available data from a plot and is then used for comparison. Figure 2 shows the graphical representation for results included in Table 7.

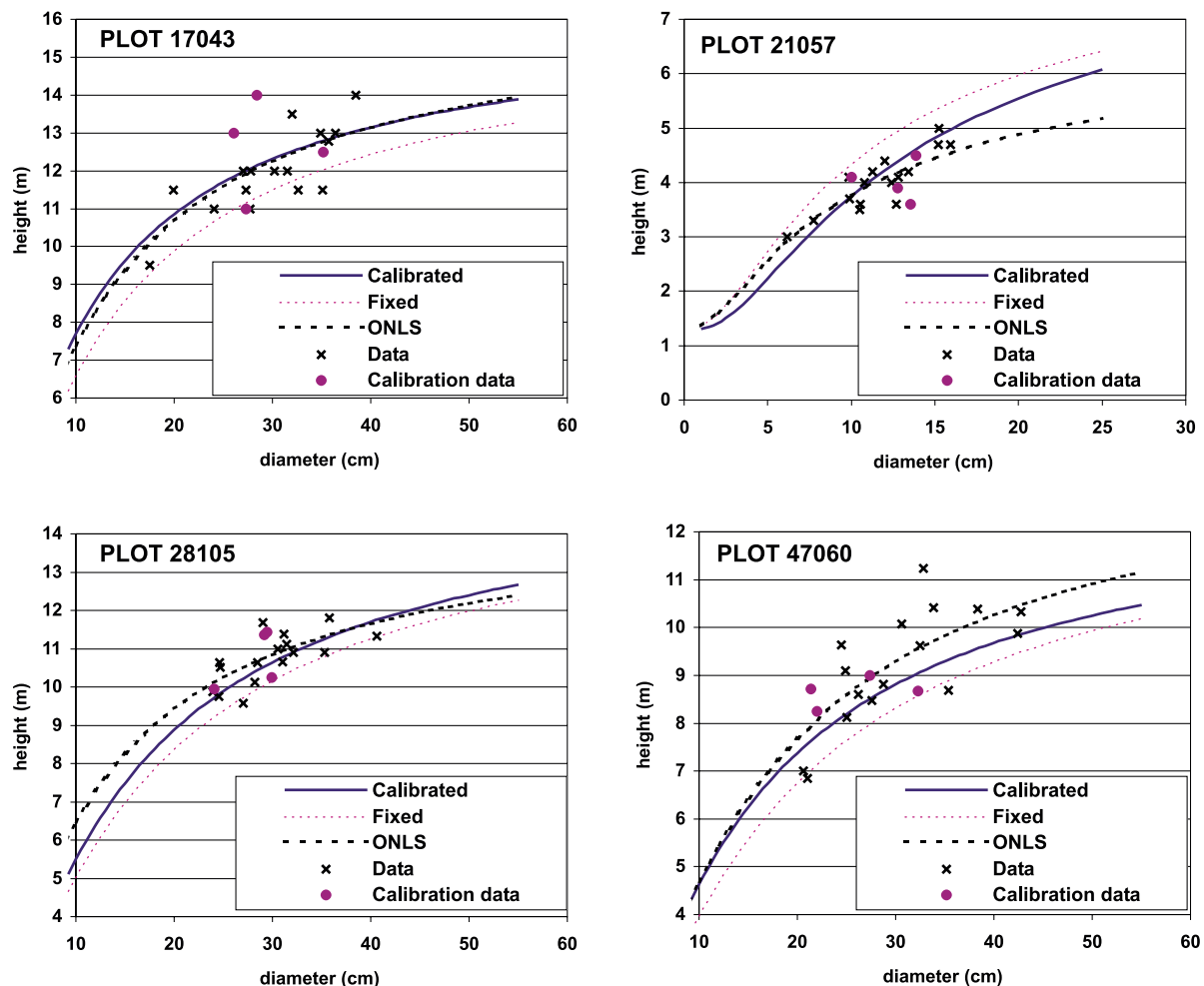
Discussion and conclusions

The height/diameter ratio is usually defined for a given stand or, even within a stand, for a given stratum (age, site index, density class), plot, or sampling unit. More general approaches define general expressions for the relationship to use them in different stand and regional scenarios. Among-plot variability is modelled in this work by including random effects specific to each plot. This allows us to deal with the lack of independence among observations derived from the special hierarchical structure of the data. Besides this, when using random component models, we can consider a common functional structure for the relationship, where a random part indicating plot deviation from the mean behaviour is added. The value of this deviation is explained including stand and regional variables.

In the model developed in this study, dominant height of the stand, density, and a measurement of the dispersion in the diameter distribution are included as covariates. The use of these covariates has already been proposed in height–diameter models by [Krumland and Wensel \(1978\)](#), [Larsen and Hann \(1987\)](#), and [Cañadas \(2000\)](#) in the case of dominant height; [Fang and Bailey \(1998\)](#) included an indicator of the diameter distribution of the stand in the height–diameter model for different forest species on Hainan Island. The use of density as a covariate in height–diameter models was proposed by [Nanos \(2001\)](#) and [Zhang and Burkhart \(1997\)](#).

Density and dominant height have been found to be positively correlated with height. Dominant height correlation

Fig. 2. Height–diameter performance of typical (only fixed effect) and calibrated (including random parameters) response, compared with individual ordinary nonlinear least squares (ONLS) fit in four plots from validation data set.



means that a relation exists between site index and stand height, as demonstrated by Eichorn's rule (Eichorn 1902). The relation between density and height is shown by the fact that for the same height, trees located in denser stands have smaller diameters than those located in less-populated stands, or, in other words, the height/diameter ratio is higher in denser stands.

We tested several other stand variables. Age was not included, as proposed by Zhang and Burkhardt (1997) and Jarayaman and Lappi (2001). We tested the inclusion of stand age within the model, but it did not improve it significantly. Nevertheless, if the model is used to predict the evolution of the stand in time, this requires the inclusion of a temporary variable. This can be achieved in an indirect way by developing temporal models for the exogenous variables, such as dominant height or diametric distribution.

From the results of the interregional variability analysis, we defined a model including only one regional variable, since the height–diameter relation has been shown to be constant for the Central Range and the Northern Plateau, and for Catalonia and West Andalusia. The establishment of these two regional groups can be explained by two circumstances. The Central Range and the Northern Plateau stands are in-

land, while Catalonia and West Andalusia are coastal regions. In addition to this, there are major differences in the silvicultural practices applied in these two areas, since in coastal regions management has been orientated towards timber production, therefore maintaining stands with high density, while in inland stands, management has considered both fruit (pine nuts) and timber production, therefore maintaining stands with lower densities.

The possibility of using data from a complementary measurement to calibrate the model for each specific occasion is a clear advantage of the mixed models when compared with other approaches to the problem of correlated observations. Table 7 and Fig. 2 show how prior knowledge of the height of a subsample of four trees in the plot increases significantly the predictive ability of the model and both prediction error and sum of squares error are reduced.

Together with the case study we have tested different alternatives to determine optimum subsample size in predicting EBLUPs. Calibration by inclusion of random parameters increases predictive ability even when EBLUPs are predicted from only a single tree randomly selected within the plot. This indicates that fixed-effects model can be calibrated using very few measurements, even less than the number of

random parameters to be predicted. Calibration using small subsamples of trees is one of the main utilities of mixed models in forest applications.

When subsample size is increased, prediction error decreases. The largest rate of decrease in SSE is obtained when comparing calibration from two random trees with calibration from four random trees (Table 6). Inclusion of a larger number of trees leads to a smaller rate of error decrease, and it involves a larger cost in inventory. As a consequence, prediction when height measurements are taken from four trees randomly selected can be considered a good option, although a utility model, considering both measurement costs and error, would be required to determine optimum subsample size.

Measuring the height of four trees by plot is a very common practice in forest inventories in Spain. The model developed enables these data to be used to calibrate a specific model for each plot, improving predictive ability for the height–diameter relationship and the estimation of productivity for the stand. Although measuring more heights in the plot improves results, it is not justified by the increased costs involved.

If the four trees where height is going to be measured are not randomly selected, but we select the four largest trees per plot, the reliability of height–diameter models is then increased if compared with the fixed-effects model, but predictive ability is lower than that attained by calibrating with four randomly selected trees. A main reason for this could be that a sample of the four largest trees is a biased sample within the plot.

Although the four largest trees in the plot constitute a biased sample, knowing their height allows us to obtain the dominant height for the plot, input variable for the model developed, as well as the site index or other related variables. When using these data in calibration, results are worse than those obtained with random samples, although we can obtain a more accurate calibration for the largest trees, those that account for more timber volume, and those that produce a more valuable timber.

This conclusion is similar to that obtained by [Krumland and Wensel \(1978\)](#), who proposed the measurement of the height of four dominant trees to calibrate height–diameter relationship for different species in California. [Houghton and Gregoire \(1993\)](#) compared several sampling designs and subsample size, finding the best results with a purposive sample (neither random nor systematic). In this case, optimum sample for a 10% accuracy level was also reached when including only four trees (the two largest and the two smallest from the plot).

Mixed models allowed us to define a height–diameter model for the four most important regions where stone pine occurs in Spain, and we found that separate models are necessary for coastal areas (Catalonia and West Andalusia) and inland areas (Central Range and Northern Plateau). A typical-response fixed-effects model can be determined for each plot if dominant height, density and diameter distribution are known for the stand. The predictive ability of the model is clearly improved if random parameters are predicted using data taken from a complementary observation in the same stand. When working with 20-tree plots, measuring the height of the four largest trees is considered a good

option, since these measurements are also required to calculate dominant height.

Acknowledgements

The authors express their gratitude to Dr. N. Cañadas (Central Range), Dr. C. García Güemes (Northern Plateau), Dr. M. Piqué (Catalonia), and J.A. Candela (West Andalusia) for their roles in establishing and maintaining the trial and coordinating data collection. We also thank Dr. Miren del Río for remarks and suggestions that improved significantly the manuscript. Finally, we also thank three anonymous reviewers for suggestions and comments that improved the contents and quality of the manuscript. The research was partially supported by a grant to R. Calama from the Consejería de Educación, Comunidad de Madrid. The research was carried out in the context of the Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria funded project SC-99-017.

References

- [Arabatzis, A.A., and Burkhardt, H.E. 1992. An evaluation of sampling methods and model forms for estimating height–diameter relationships in loblolly pine plantations. *For. Sci.* **38**: 192–198.](#)
- [Bates, D.M., and Watts, D.G. 1980. Relative curvature measures of nonlinearity. *J. R. Stat. Soc.* **42**: 1–16.](#)
- [Bégin, J., and Raulier, F. 1995. Comparaison de différentes approches, modèles et tailles d'échantillons pour l'établissement de relations hauteur–diamètre locales. *Can. J. For. Res.* **25**: 1303–1312.](#)
- [Burkhardt, H.E., and Strub, M.R. 1974. A model for simulation of planted loblolly pine stands. In *Growth models for tree and stand simulation*. Royal College of Forestry, Stockholm, Sweden. pp. 128–135.](#)
- [Calama, R., Cañadas, N., and Montero, G. 2003. Inter-regional variability in site index models for even-aged stands of stone pine \(*Pinus pinea* L.\) in Spain. *Ann. For. Sci.* **60**\(3\): 259–269.](#)
- [Cañadas, N. 2000. *Pinus pinea* L. en el Sistema Central \(Valles del Tiétar y del Alberche\): desarrollo de un modelo de crecimiento y producción de piña. Tesis doctoral, Universidad Politécnica de Madrid, Madrid, Spain.](#)
- [Curtis, R.O. 1967. Height–diameter and height–diameter–age equations for second-growth Douglas-fir. *For. Sci.* **13**: 365–375.](#)
- [Eerikäinen, K. 2001. Stem volume models with random coefficients for *Pinus kesiya* in Tanzania, Zambia, and Zimbabwe. *Can. J. For. Res.* **31**: 879–888.](#)
- [Eichorn, F. 1902. *Ertragstafeln* f.d. Weißtanne, Berlin.](#)
- [Fang, Z., and Bailey, R.L. 1998. Height–diameter models for tropical forests on Hainan Island in southern China. *For. Ecol. Manage.* **110**: 315–327.](#)
- [Fang, Z., and Bailey, R.L. 2001. Nonlinear mixed effects modeling for slash pine dominant height growth following intensive silvicultural treatments. *For. Sci.* **47**\(3\): 287–300.](#)
- [Fang, Z., Bailey, R.L., and Shiver, B.D. 2001. A multivariate simultaneous predictions system for stand growth and yield with fixed and random effects. *For. Sci.* **47**\(4\): 550–562.](#)
- [Ferguson, I.S., and Leech, J.W. 1978. Generalized least squares estimation of yield functions. *For. Sci.* **24**: 27–42.](#)
- [Fox, J.C., Ades, P.K., and Bi, H. 2001. Stochastic structure and individual-tree growth models. *For. Ecol. Manage.* **154**: 261–276.](#)

- Fulton, M.R. 1999. Patterns in height–diameter relationships for selected tree species and sites in eastern Texas. *Can. J. For. Res.* **29**: 1445–1448.
- García Güemes, C. 1999. Modelo de simulación selvícola para *Pinus pinea* L. en la provincia de Valladolid. Tesis doctoral, Universidad Politécnica de Madrid, Madrid, Spain.
- Gregoire, T.G. 1987. Generalized error structure for forestry yield models. *For. Sci.* **33**: 423–444.
- Gregoire, T.G., and Schabenberger, O. 1996. A nonlinear mixed-effects model to predict cumulative bole volume of standing trees. *J. Appl. Stat.* **23**: 257–271.
- Gregoire, T.G., Schabenberger, O., and Barrett, J.P. 1995. Linear modelling of irregularly spaced, unbalanced, longitudinal data from permanent plot measurements. *Can. J. For. Res.* **25**: 137–156.
- Hall, D.B., and Bailey, R.L. 2001. Modeling and prediction of forest growth variables based on multilevel nonlinear mixed models. *For. Sci.* **47**(3): 311–321.
- Harville, D.A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.* **72**: 320–340.
- Henderson, C.R., Kempthorne, O., Searle, S.R., and Krosigk, C.M. 1959. The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, **15**: 192–217.
- Hokka, H. 1997. Height–diameter curves with random intercepts and slopes for trees growing on drained peatlands. *For. Ecol. Manage.* **97**: 63–72.
- Houghton, D.R., and Gregoire, T.G. 1993. Minimum subsamples of tree heights for accurate estimation of loblolly pine plot volume. *South. J. Appl. For.* **17**: 124–129.
- Huang, S., Titus, S.J., and Wiens, D.P. 1992. Comparison of nonlinear height–diameter functions for major Alberta tree species. *Can. J. For. Res.* **22**: 1297–1304.
- Huang, S., Price, D., and Titus, S.J. 2000. Development of ecoregion-based height–diameter models for white spruce in boreal forests. *For. Ecol. Manage.* **129**: 125–141.
- Jayaraman, K., and Lappi, J. 2001. Estimation of height–diameter curves through multilevel models with special reference to even-aged teak stands. *For. Ecol. Manage.* **142**: 155–162.
- Krumland, B.E., and Wensel, L.C. 1978. A generalized height–diameter equation for coastal California species. *West. J. Appl. For.* **3**: 113–115.
- Laird, N.M., and Ware, J.H. 1982. Random-effects models for longitudinal data. *Biometrics*, **38**: 963–974.
- Lappi, J. 1986. Mixed linear models for analyzing and predicting stem form variation of scots pine. *Commun. Insti. For. Fenn.* **134**: 1–69.
- Lappi, J. 1991. Calibration of height and volume equations with random parameters. *For. Sci.* **37**(3): 781–801.
- Lappi, J. 1997. A longitudinal analysis of height/diameter curves. *For. Sci.* **43**(4): 555–570.
- Lappi, J., and Bailey, R.L. 1988. A height prediction model with random stands and tree parameters: an alternative to traditional site index methods. *For. Sci.* **34**(4): 907–927.
- Larsen, D.R., and Hann, D.W. 1987. Height–diameter equations for seventeen tree species in southwest Oregon. Oregon State University, Forest Research Laboratory, Corvallis, Ore.
- Lindstrom, M.J., and Bates, D.M. 1990. Nonlinear mixed effects for repeated measures data. *Biometrics*, **46**: 673–687.
- Meyer, H.A. 1940. A mathematical expression for height curves. *J. For.* **38**: 415–420.
- Moore, J.A., Zhang, L., and Stuck, D. 1996. Height–diameter equations for ten tree species in the Inland Northwest. *West. J. Appl. For.* **11**(4): 132–137.
- Nanos, N. 2001. Variabilidad y modelización geoestadística de producción de resina y madera de *Pinus pinaster* Ait. en los montes de Segovia. Tesis doctoral, Universidad Politécnica de Madrid, Madrid, Spain.
- Palmer, M.J., Phillips, B.F., and Smith, G.T. 1991. Application of nonlinear models with random coefficients to growth data. *Biometrics*, **47**: 623–635.
- Parresol, B.R. 1992. Baldcypress height–diameter equations and their prediction confidence interval. *Can. J. For. Res.* **22**: 1429–1434.
- Pinheiro, J.C., and Bates, D.M. 1995. Approximation to the log-likelihood function in the nonlinear mixed — effects model. *J. Comput. Graph. Stat.* **4**: 12–35.
- Pinheiro, J.C., and Bates, D.M. 1998. Model building for nonlinear mixed effects model. Department of Statistics, University of Wisconsin, Madison, Wis.
- SAS Institute Inc. 2000. SAS/STAT user's guide, version 8. SAS Institute Inc., Cary, N.C.
- Searle, S.R. 1971. Linear models. Wiley series in probability and mathematical statistics. John Wiley & Sons, New York.
- Sheiner, L., and Beal, S.L. 1980. Evaluation of methods for estimating population pharmacokinetic parameters. I. Michaelis–Mentem model: routine clinical pharmacokinetic data. *J. Pharmacokinet. Biopharm.* **8**: 553–571.
- Sheu, C.F., and Suzuki, S. 2001. Meta-analysis using linear mixed models. *Behav. Res. Methods Instrum. Comput.* **33**(2): 102–107.
- Singer, J.D. 1998. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *J. Educ. Behav. Stat.* **23**: 323–355.
- Stage, A.R. 1975. Prediction of height increment for models of forest growth. USDA For. Serv. Res. Pap. INT-164.
- Tukey, J.W. 1953. The problems of multiple comparisons. Princeton University Press, Princeton, N.J.
- Vonesh, E.F., and Chinchilli, V.M. 1997. Linear and nonlinear models for the analysis of repeated measurements. Marcel Dekker, Inc., New York.
- West, P.W., Ratkowsky, D.A., and Davis, A.W. 1984. Problems of hypothesis testing of regressions with multiple measurements from individual sampling units. *For. Ecol. Manage.* **7**: 207–224.
- Wolfinger, R.D. 1997a. Comment: experiences with the SAS macro NLINMIX. *Stat. Med.* **16**: 1258–1259.
- Wolfinger, R.D. 1997b. An example of using mixed models and PROC MIXED for longitudinal data. *J. Biopharm. Stat.* **7**: 481–500.
- Wykoff, W.R., Crookston, N.L., and Stage, A.R. 1982. User's guide to the Stand Prognosis Model. USDA For. Serv. Gen. Tech. Rep. INT-133.
- Zakrzewski, W.T., and Bella, E. 1988. Two new height models for volume estimation of lodgepole pine stands. *Can. J. For. Res.* **18**: 195–201.
- Zhang, S., and Burkhart, H.E. 1997. The influence of thinning on tree height and diameter relationships in loblolly pine plantations. *South. J. Appl. For.* **21**: 199–205.