Calibrating a Generalized Diameter Distribution Model with Mixed Effects

Mathieu Fortin, Chhun-Huor Ung, Louis Archambault, and Jean Bégin

Abstract: This article proposes a mixed model approach as an alternative to the traditional parameter prediction method in diameter distribution modeling. Unthinned and thinned plots established in mixed stands were used for calibrating a generalized diameter distribution model. The model was based on a two-parameter Weibull cumulative density function (cdf). This cdf was linearized through a complementary log–log link function. Plot random effects were included to account for autocorrelation and the dependent variable, i.e., cumulative stem frequency, was assumed to follow a binomial distribution. With respect to the parameter prediction method, this mixed model approach may generate a consistent estimator with consistent and unbiased variance for the vector of parameters when the variance-covariance matrix of the error terms is properly parameterized. Moreover, the approach enables a better assessment of the different variance components. For the whole group of plots, it provides a predicted average diameter distribution. At the plot level, the random effects can be considered as a departure from this average distribution. As long as the diameter distributions of the individual plots do not exhibit major departures from unimodality, the method proposed in this study should be used to calibrate generalized diameter distribution models. For. Sci. 52(6):650–658.

Key Words: Two-parameter Weibull function, random effects, link function, generalized linear mixed models.

The diameter distribution of trees is an important factor in forest management as it is often used to characterize forest structure (Smith et al. 1997, p. 14). Harvest options are evaluated by projecting the diameter distributions into the future and comparing the resulting forest structures (e.g., Bowling et al. 1989). To address the concern for ecosystem diversity and stability, new forest policies also require information on forest structure at different scales (Lähde et al. 1999). As a consequence, an effort is needed to develop diameter distribution models suitable for different levels of analysis.

Until now, three major trends have distinguished investigations of diameter distributions. The first approach has focused on developing and testing different, mostly nonlinear, statistical models to fit the diameter distributions of individual plots (e.g., Bailey and Dell 1973, Cao and Burkhart 1984, Borders et al. 1987, Zhang et al. 2001, 2003). The second approach consists of comparing different groups of plots through indices such as skewness, kurtosis, or the Shannon Index (cf. Shannon and Weaver 1949, p. 19). These indices are initially calculated for each individual plot. Then, an analysis of variance (ANOVA) is performed on the individual values to identify significant differences among groups of plots (e.g., Groot and Horton 1994, Kuuluvainen et al. 1996, Linder et al. 1997, Linder 1998, Lähde et al. 2001). The third approach attempts to link the previous two by generalizing a single population diameter distribution model to a group of plots (e.g., Magnussen 1986, Bowling et al. 1989, Maltamo et al. 1995, 2000, Maltamo 1997, Siipilehto 1999, 2000, Robinson 2004). These generalized diameter distribution models aim at providing information at two levels, namely the individual plot and the group of plots.

From a statistical standpoint, calibrating a generalized diameter distribution model from a sample of plots is often hindered by autocorrelated data. This problem is discussed extensively by García (1992). In most regular forest surveys, the plot is the sampling unit and many diameter measurements are taken within the plot. As a result, we can expect this kind of grouping to induce plot random effects in the model. Even though they are often

Mathieu Fortin, Department of Wood and Forest Sciences, Laval University, Quebec City, Quebec G1K 7P4, Canada—Phone: (418) 656-2131 ext. 8917; Fax: (418) 656-5262; mathieu.fortin.2@ulaval.ca; Current address: Direction de la recherche forestière, Ministère des Ressources naturelles et de la Faune du Québec, 2700 Einstein, Quebec City, Quebec G1P 3W8, Canada—Phone: (418) 648-5834; Chhun-Huor.Ung@nrcan-rncan.gc.ca. Louis Archambault, Natural Resources Canada, Canadian Forest Service, Laurentian Forestry Centre, 1055 du P.E.P.S., PO Box 3800, Quebec City, Quebec GIV 4C7, Canada—Phone: (418) 648-7230; Louis.Archambault@nrcan-rncan.gc.ca.

Acknowledgments: Funding for this work was provided by le Fonds québécois de la recherche sur la nature et les technologies (FQRNT), the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Forest Service, the Ministère des Ressources naturelles, de la Faune et des Parcs du Québec, and Abitibi-Consolidated, Inc. (Jean Girard). Special thanks are due to Michèle Bernier-Cardou (Canadian Forest Service), Louis-Paul Rivest (Laval University), and three anonymous referees for their reviews of an early version of this manuscript, to William F.J. Parsons (University of Sherbrooke) for editing the text, and to the people involved in field measurements. Thanks to le Parc national du Mont-Tremblant (Claire Ducharme) for support and access to the study area. This work would not have been possible without the exceptional dedication of R.J. Hatcher in establishing and conducting the experiment at Ouareau River Observation Area.

Manuscript received July 7, 2005, accepted May 30, 2006 Copyright © 2006 by the Society of American Foresters
considered as negligible, the effects of autocorrelated data on statistical inferences and predictions may be important. The variances estimated by traditional methods, such as ordinary least squares, are biased (Sullivan and Clutter 1972, Gregoire 1987, Gregoire et al. 1995). Hence, the selection of the best covariates is hampered (Gregoire et al. 1995) and the fit of the model cannot be correctly assessed. Consequently, the model predictions are inefficient, as their confidence intervals are biased.

Over the last two decades, the forestry literature has provided some examples of the flexibility of mixed-effects models for considering the departures from the assumption of independent errors (e.g., Gregoire 1987, Gregoire et al. 1995, Hall and Bailey 2001, Garber and Maguire 2003). Gregoire et al. (1995) pointed out that mixed models were still infrequently used in growth modeling due to their complexity and the previous lack of software suitable for estimation. The current situation in diameter distribution modeling is even worse. The stem frequencies that are inherent to diameter distributions follow a discrete distribution, whereas most statistical tools for mixed model regressions are designed to work with normally distributed variables.

A mixed model approach applied to a generalized diameter distribution model would make it possible to consider different levels of uncertainty due to the data structure. Moreover, this approach would enable valid statistical inferences since it accounts for autocorrelated data. The purpose of this study was to demonstrate the feasibility of a mixed model approach applied to diameter distribution modeling. Thinned and unthinned plots that were located in mixed stands of red spruce (*Picea rubens* Sarg.), balsam fir (*Abies balsamea* [L.] Mill.), and yellow birch (*Betula alleghaniensis* Britton.) were used to calibrate a two-parameter Weibull function with a mixed model approach. The discussion focuses on the advantages and limitations of the approach.

**Materials and Methods**

**The Study Site**

The Ouareau River Observation Area (OROA) (46°26′N, 74°10′W) was established in 1953 by the Department of Northern Affairs and National Resources of the Government of Canada. The study site, which is 13 km² in area, is located within the Laurentian Section (L.4a) of the Great Lakes-St. Lawrence Forest Region (cf. Rowe 1972, p. 96). The nearby municipality of Saint-Donat-de-Montcalm has an average daily temperature of 3.2°C (running average for 1971–2000, Environment Canada 2002). January is the coldest month whereas July is the warmest, with respective average temperatures of -13.6 and 18.1°C. Total annual precipitation averages 1,128 mm, with a mean annual snowfall of about 295 cm. The OROA has an average elevation of 515 m above sea level. It is characterized by a succession of low-altitude hills. Most soils are glacial tills. The red spruce-balsam fir-yellow birch association is the most abundant forest type (Hatcher 1954). It is usually located on well-drained flat lands and lower slopes (Heimburger 1941).

Between 1948 and 1956, a diameter-limit cutting was conducted over the entire OROA territory. All spruce and fir greater than 17.8 cm (7 in.) in diameter at stump height (30 cm) were logged for pulpwood. As early as 1953, permanent plots were established within the observation area for monitoring purposes. In these 404.7 m² (0.1 ac) plots, all trees with a dbh greater than 1.3 cm (0.5 in.) were tallied in 2.5 cm (1 in.) classes. At that time, about one-third of the OROA was still uncut and the plots established within that area could be considered as a representative sample of the conditions before thinning (Hatcher 1954). By 1956, the cutting had encompassed almost the entire study site, with the uncut patches being mostly inaccessible or uninteresting from a commercial standpoint (Hatcher 1954). During the period of 2002–2004, only the thinned plots located in the red spruce-balsam fir-yellow birch forest type were visited and remeasured.

The data set used in this case study included two groups of plots, the first of which was formed from the 1,953 unthinned plots located in the red spruce-balsam fir-yellow birch association (44 plots). The 2002–04 survey provided an additional 22 plot measurements, which represented mature forest conditions 50 years following thinning (Hatcher 1954). During the period of 2002–2004, only the thinned plots located in the red spruce-balsam fir-yellow birch forest type were visited and remeasured.

The data set used in this case study included two groups of plots, the first of which was formed from the 1,953 unthinned plots located in the red spruce-balsam fir-yellow birch association (44 plots). The 2002–04 survey provided an additional 22 plot measurements, which represented mature forest conditions 50 years following thinning (Hatcher 1954). During the period of 2002–2004, only the thinned plots located in the red spruce-balsam fir-yellow birch forest type were visited and remeasured.

The data set used in this case study included two groups of plots, the first of which was formed from the 1,953 unthinned plots located in the red spruce-balsam fir-yellow birch association (44 plots). The 2002–04 survey provided an additional 22 plot measurements, which represented mature forest conditions 50 years following thinning (Hatcher 1954). During the period of 2002–2004, only the thinned plots located in the red spruce-balsam fir-yellow birch forest type were visited and remeasured.

The data set used in this case study included two groups of plots, the first of which was formed from the 1,953 unthinned plots located in the red spruce-balsam fir-yellow birch association (44 plots). The 2002–04 survey provided an additional 22 plot measurements, which represented mature forest conditions 50 years following thinning (Hatcher 1954). During the period of 2002–2004, only the thinned plots located in the red spruce-balsam fir-yellow birch forest type were visited and remeasured.

**Statistical Developments**

Over the last several decades, the Weibull function has been one of the most, if not the most, popular models used in diameter distribution modeling (Johnson 2000, p. 260). An extensive description of the model can be found in Bailey and Dell (1973). The three-parameter version of this probability density function (PDF) is very flexible and suits a vast array of random variables such as dbh (Vanclay 1994, p. 23). As diameters are often measured in finite diameter classes, the cumulative density function (cdf) is used instead of the original PDF. Let i and j be, respectively, the plot and diameter class indices such that

<table>
<thead>
<tr>
<th>Unthinned group (44 plots)</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basal area (m²·ha⁻¹)</td>
<td>10.3</td>
<td>29.3</td>
<td>58.1</td>
</tr>
<tr>
<td>Density (stems·ha⁻¹)</td>
<td>445</td>
<td>2,182</td>
<td>8,970</td>
</tr>
<tr>
<td>Mean quadratic diameter (cm)</td>
<td>6.5</td>
<td>14.3</td>
<td>20.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Thinned group (22 plots)</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basal area (m²·ha⁻¹)</td>
<td>26.4</td>
<td>35.1</td>
<td>49.4</td>
</tr>
<tr>
<td>Density (stems·ha⁻¹)</td>
<td>766</td>
<td>3,007</td>
<td>8,525</td>
</tr>
<tr>
<td>Mean quadratic diameter (cm)</td>
<td>6.3</td>
<td>14.0</td>
<td>24.0</td>
</tr>
</tbody>
</table>
where $\pi_{ij}$ is the probability of having a dbh equal to or lower than $d_{ij}$, which is defined as the upper limit of the $i$th diameter class of the $j$th plot, and $\alpha_i$, $\lambda_i$, and $\chi_i$ are respectively the shape, the scale, and the location parameters of the cdf associated with the $i$th plot. When the lower bound of the diameter distribution is already known, let this limit be defined as $d_{\text{min},i}$. The limit value can be substituted, in turn, for the location parameter ($\chi_i$), thereby reducing Equation 1 to a two-parameter Weibull cdf (e.g., Magnussen 1986, Zhang et al. 2001).

Thus, for any tree of the $i$th plot, the probability of having $\text{dbh} \leq d_{ij}$ would be $\pi_{ij}$. Likewise, the probability of having $\text{dbh} > d_{ij}$ would be the complement of the probabilities, namely $1 - \pi_{ij}$. Consequently, given the number of trees with $\text{dbh} \leq d_{ij}$, let $s_{ij}$ be this variable, which is expected to follow a binomial distribution,

$$\Pr(s_{ij}; \pi_{ij}, ri) = \binom{ri}{s_{ij}} \pi_{ij}^{s_{ij}}(1 - \pi_{ij})^{ri - s_{ij}},$$

where $ri$ is the total number of trees in the $i$th plot. To refer to more classical terminology, $s_{ij}$ and $ri$ are the number of successes and trials, respectively.

Discrete variables are usually modeled through generalized linear models. An extensive description of these models is found in McCulloch and Searle (2001, §5). In addition to the specification of a distribution for the dependent variable (binomial in our case), these models require a transformation of the dependent variable to obtain a projection that can be modeled in a linear fashion. As the probability $\pi_{ij}$ is assumed to follow a Weibull cdf, we expect the transformation to provide a link between Equation 1 and its derived linear form. Such a linearization is obtained through a complementary log–log transformation (cf. Collett 2003, p. 57),

$$\ln(-\ln(1 - \pi_{ij})) = \beta_{0ij} + \beta_{1ij}(d_{ij} - d_{\text{min},i})$$

$$\beta_{0ij} = -\alpha_i \ln(\lambda_i)$$

$$\beta_{1ij} = \alpha_i,$$

where $d_{\text{min},i}$ has been substituted for the location parameter ($\chi_i$). Let the function $g(\pi_{ij})$ define the complementary log–log transformation of probability $\pi_{ij}$ as shown in Equation 3.

All the observations of the data set would be independent if each plot had a single $d_{ij}$. However, the data typically provide much more information. Actually, the upper limit of any diameter class can be used as $d_{ij}$. Besides, considering more than one $d_{ij}$ per plot implies serially correlated data. To take into account this possible correlation, plot random effects can be integrated into the model. The resulting generalized linear mixed model (GLMM) can be expressed as

$$g(\pi_{ij}) = X_i \beta + Z_i u_i,$$

$$\pi_{ij} = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{im_i} )^T u_i \sim N(0, G),$$

where $X_i$ is a matrix of explanatory variables, including an intercept and $\ln(d_{ij} - d_{\text{min},i})$, $\beta$ is a vector of unknown parameters, $Z_i$ is a design matrix indicating the variables with which the random effects are associated, $u_i$ is a vector of unobserved random effects due to the $i$th plot, and $T$ denotes a transposition. We expect $u_i$ to be normally distributed with mean 0 and variance-covariance $G$. Note that the random effects are assumed to follow a normal distribution, whereas the response is binomially distributed.

With linear mixed models, the final estimator of $\beta$ is provided by the maximization of a marginal likelihood function (Pinheiro and Bates 2000, p. 62). This marginal likelihood function is obtained by integrating the original likelihood function involving the random variables with respect to the distribution of these variables (Collett 2003, p. 272). Once the model has converged, empirical best linear unbiased predictors (eBLUP) of the random effects are calculated from the design matrix $Z_i$, the vector of residuals, and the variance-covariance matrix associated with this vector (McCulloch and Searle 2001, p. 255). Nevertheless, with a GLMM, the link function hinders the integration of the likelihood function analytically (McCulloch and Searle 2001, p. 231).

An alternative consists of using a quasi-likelihood function instead of a marginal likelihood. With simple generalized linear models, the quasi-likelihood approach has several attractive features, including model robustness and less restrictive assumptions (McCulloch and Searle 2001, p. 281). Furthermore, according to this approach, the vector of mean probabilities $\pi_{ij}$ is conditional on the random effects,

$$E[p_{ij} | u_i] = \pi_{ij},$$

$$p_{ij} = s_{ij}/ri.$$
we assume a binomial distribution, the variance \( \text{var}(\pi_i) \) depends on the value of \( \pi_i \) as

\[
\text{var}(\pi_i) = \pi_i(1 - \pi_i)/n_i.
\]  

(7)

Matrix \( A \) is usually set to \( dI \), with \( I \) being the identity matrix and \( \phi \) being a parameter equivalent to the residual variance. To refer to traditional mixed models, matrix \( A \) corresponds to the within-plot variance-covariance matrix of the error terms. Since the plots are assumed to be independent from each other, matrix \( A \) is diagonally blocked, with each block \( A_i \) being the variance-covariance matrix of plot \( i \). In cases of persistent correlations, a covariance structure may be generalized to all the blocks \( A_i \) (Vonesh and Carter 1992). Common covariance structures, such as the compound symmetry, the autoregressive, and the moving average, are extensively described in Pinheiro and Bates (2000, §5.3.3).

To check the validity of the assumptions of independence and binomially distributed errors with homogeneous variances, a vector of scaled Pearson residuals \( (e_i) \) can be derived from mixed model theory (cf. Pinheiro and Bates 2000, p. 239),

\[
e_i = (C^{-1})^T(p - \pi),
\]

(8)

where \( C \) is the upper triangle resulting from the Cholesky decomposition of matrix \( R \) as shown in (6). If a covariance structure is specified within matrix \( A \), the calculation of the scaled residuals through matrix \( C \) takes into account this structure. Consequently, it provides the marginal contribution of each observation to the error of the model. If the covariance structure is appropriate, the scaled residuals should not be correlated.

### Model Specifications

There is no limit to the number of diameter classes in a plot, since each one of them can be a potential \( d_{ij} \), which could be used for calibration. However, large numbers of diameter classes for which no trees were tallied may be cumbersome during the optimization of the quasi-likelihood function. In this case study, the 40.6-cm (16-in.) diameter class represented the 99th percentile of the observed average diameter distribution. Larger diameter classes were not considered. Consequently, 16 distinct \( d_{ij} \) were used per plot, each one corresponding to the upper limit of one of the first 16 diameter classes. During the model calibration, all trees were considered without distinguishing among species.

Model 3 was first simplified by setting \( d_{ij} = 1 \) for all plots. Preliminary calibrations were performed on each individual plot to obtain estimates of the intercept \( (\beta_{0ij}) \) and the slope \( (\beta_{1ij}) \) of the model. Then, these estimates were plotted against several covariates to identify the explanatory variables, which were the most likely to be significant.

After a few trials, the following explanatory variables and random effects were selected and tested:

\[
g(\pi_i) = \beta_{0ij} + \beta_{1ij}\ln(d_{ij} - 1.3)
\]

\[
\beta_{0ij} = b_0 + b_1\text{density}_i + b_2\text{group}_i + u_{0ij}
\]

\[
\beta_{1ij} = b_3 + b_4\text{basalarea}_i + b_5\text{group}_i + u_{1ij}
\]

\[
\begin{bmatrix}
  u_{0ij} \\
  u_{1ij}
\end{bmatrix} \sim N_2(0, G),
\]

where \( d_{ij} \) is a diameter-limit as defined above (cm), density\(_i\) is the density of plot \( i \) (stems/ha\(^{-1}\)), \( b_0, b_1, b_2, b_3, b_4, \) and \( b_5 \) are the model parameters, and \( u_{0ij} \) and \( u_{1ij} \) are the plot random effects, which follow a bivariate normal distribution with mean 0 and variance-covariance \( G \) as previously mentioned.

Likelihood ratio tests were performed to test the significance of the random effects. Scaled residuals were plotted to verify the assumptions of binomial distribution, homoskedasticity, and independent error terms. Also, the predictors of the random effects were plotted to check for normality. The analyses were performed with the GENMOD procedure and the GLIMMIX macro available in SAS (Littell et al. 1996, §11; SAS Institute, Inc. 2002).

### Results

Even when plot random effects were included in the model, within-plot residuals still exhibited persistent correlations. The two-band Toeplitz covariance structure (cf. SAS Institute, Inc. 2002) seemed to be the most appropriate structure. So, the structure of matrix \( A \) was modified according to the Toeplitz covariance structure as

\[
A = I_6 \otimes \phi
\]

where \( \rho \) is the Toeplitz parameter and \( \otimes \) denotes the Kronecker product.

The GLMM was recalibrated with the inclusion of this new feature. The residuals exhibited homogeneity of variances (Figure 1a), and there was no departure from the assumption of binomially distributed errors (Figure 1b). Moreover, none of the correlations was greater than 0.40 in terms of their absolute values (Figure 1c). Both random effects \( (u_{0ij} \) and \( u_{1ij} \)) did not exhibit major departures from the assumption of normality.

The estimates resulting from the final calibration of
model 8 are shown in Table 2. Except for the single effect of the group and the interaction \( \ln(d_{ij} - 1.3) \times \text{group} \), the parameter estimates related to the fixed effects all exhibited probabilities close to or less than 2%. Referring to Equation 3, a single effect is related to both the scale and the shape parameters (\( \lambda_i \) and \( \alpha_j \)) of the Weibull cdf as shown in (1). Interactions with \( \ln(d_{ij} - 1.3) \) indicate an effect on the shape \( \alpha_j \) of the cdf only. Consequently, the plot density had a significant effect on both the scale and the shape of the Weibull function. However, the basal area had a significant effect on the shape of the cdf only. With regard to random effects and the covariance structure, the probabilities associated with the Toeplitz parameter estimate and the standard deviations of both \( u_{0,i} \) and \( u_{1,i} \) were less than 0.01%. Therefore, including the random effects and the correlation structure in the model significantly improved its likelihood, justifying these additional features.

Each random effect can be interpreted as the between-plot variability around the population average response, which is estimated through the fixed effects. Actually, \( u_{0,i} \) and \( u_{1,i} \) express this variability around the estimates of \( b_0 \) and \( b_3 \), respectively. A 95% confidence interval around the fixed effect is easily computed by multiplying \( \pm 1.96 \) by the SD of the appropriate random effect. In this case study, the contribution of the random effects to the fit of the model was not negligible, since their 95% confidence intervals represented a variability of \( \pm 69\% \) and \( \pm 87\% \) around their respective fixed effects.

The fit of the model was also assessed by plotting the observed and predicted probabilities for each plot. Predictions based solely on fixed effects usually were not accurate. This result was expected since the variability at the plot level was important. Basically, three typical patterns were observed. These are illustrated in Figure 2. First, plot 51 illustrates one of the few examples of reliable predictions based on fixed effects only (Figure 2a). Second, in most cases, the inclusion of the empirical predictors of the random effects (eBLUP) resulted in more accurate predictions, with the fit of the model being nearly perfect (Figure 2b). The improvement was more important for the smallest \( d_{ij} \), for which the probabilities were the most variable. The third pattern is represented by a few plots for which the fit with the random effect predictors still exhibited non-negligible errors (Figure 2c). The observed cumulative probabilities of these plots had a very irregular shape that the Weibull cdf was unable to fit perfectly.

### Discussion

Ordinary least-squares (OLS) regressions are based on the assumptions of homogeneous, normally, and independently distributed errors (Steel et al. 1997, §7.10). In this context, diameter distribution modeling is often hindered by data that are measured on a discrete scale, such as diameter class. Even when tree diameters are precisely measured, the observation remains a frequency, which cannot be negative.
Consequently, the assumption of normally distributed errors about the mean probability is irrelevant, as it may result in confidence intervals including inconsistent values, namely probabilities outside of the range (0, 1).

Even if the OLS regressions produce in a model with good fit, the errors are no longer entirely independent of the predicted probabilities. For predicted probabilities close to 1, the distribution of the residuals tends to be left-skewed. Likewise, right-skewed residuals are associated with predicted probabilities close to 0. Consequently, regular statistical analyses are likely to provide biased inferences since the assumptions of normality and independent errors no longer hold.

Generalized linear models enable the analysis of a response variable whose distribution belongs to the exponential family of distributions, such as the binomial, the Poisson, and the normal distribution (McCulloch and Searle 2001, §5.2). This analysis requires the specification of an appropriate distribution and a link function. The link function makes it possible to model the transformed probabilities as a linear function of explanatory variables over an interval \((-\infty, +\infty)\) (Collett 2003, p. 56). Regular statistical tests are then suitable for such transformed probabilities. Although the calibration of GLMM is more complex than regular mixed models, statistical tools such as the GLIMMIX macro available in SAS (Littell et al. 1996, §11) perform relatively well with one level of grouping within the data structure. These tools facilitate the application of mixed models theory to diameter distribution modeling. This study is an example of this application.

The complementary log–log link function allows the Weibull cdf to be linearized. The linearized form results in an intercept and a slope related to the natural logarithm of the dbh. To calibrate the GLMM, the probability generated by the Weibull cdf is used as a probability of success in a binomial distribution for which the number of trees in the plot represents the number of trials. To our knowledge, no other study has addressed the matter of generalized diameter distribution models in this fashion.

Because there is more than one successes/trials ratio per plot, random effects have to be included in the model. To ensure that all appropriate random effects are present, Hall and Bailey (2001) recommended that all basic fixed-effect parameters should be allowed to vary across the different levels of grouping in the data set. The conversion of the linearized Weibull cdf into a mixed model form is relatively easy to perform, since the model only has required two random effects, i.e., one associated with the intercept \((\beta_0)\) and one related to the slope parameter \((\beta_1)\).

Even though the appropriate random effects were included, the within-plot residuals still exhibited significant correlations in our case study. This was essentially due to the nature of the response variable. As cumulative frequencies are used, a deficit in a particular diameter class is likely to induce a deficit in the following diameter classes. A persistent correlation between the residuals of a mixed model is not an entirely new phenomenon, as it has been already observed in growth and stem taper modeling (Gregoire et al. 1995; Garber and Maguire 2003). To overcome this problem, Gregoire et al. (1995) and Garber and Maguire (2003) used an autoregressive structure, which assumes a positive correlation decreasing as the distance increases between the observations (cf. Pinheiro and Bates 2000, p. 228). In this case study we attempted to specify this structure, without success. The negative correlations observed for some distances between \(d_j\) and \(d_{j+1}\) (Figure 1c) may explain why the model failed to converge. A two-band Toeplitz covariance structure accounts for the correlation between a given observation, namely the ratio of

![Figure 2. The three typical patterns related to the fit of the model.](https://academic.oup.com/forestscience/article/52/6/650/4617309)
successes/trials, and its nearest neighbor in terms of $d_i$. Adding this structure to the GLMM significantly improved the likelihood and decreased the correlations between the residuals, although a pattern of persistent correlations could still be observed (Figure 1c). The implementation of a covariance structure enabling a sinusoidal pattern might solve the problem of persistent correlations. Multinomial models with cumulative ordinal responses use covariance structures (cf. McCullagh and Nelder 1989, §5.3.2), which might be adapted to the context of diameter distribution modeling. However, the gain of reliability might be rather small since the persistent correlations resulting from the Toeplitz structure are already low.

The performance of the penalized quasi-likelihood (PQL) estimator has not been thoroughly evaluated. The PQL estimates are subject to serious bias for correlated binary data (Breslow and Lin 1995, Lin and Breslow 1996). For discrete data with moderate to large cell frequencies, the method has been demonstrated to work reasonably well (Breslow and Lin 1995). Usually, the properties of the PQL and maximum likelihood estimators are assumed to be similar. The reliability of the PQL estimator is beyond the scope of this study and it remains an open question.

In this case study, the scaled residuals of the model provided strong evidence that the assumptions of binomial distribution, homogeneous variances, and independent errors were globally respected (Figure 1). Consequently, the estimator for the vector of parameters could be considered as consistent and asymptotically normal (Pinheiro and Bates 2000, p. 81) with consistent and nearly unbiased variance. This feature is particularly useful because it enabled more reliable statistical comparisons among different groups of plots. In this case study, for instance, there were no significant differences between the unthinned and the thinned groups. Besides, two continuous explanatory variables, i.e., plot basal area and plot density, had significant effects on the slope and intercept of the model. These two variables accounted for most of the variability between the two groups.

Until now, the classical approach for the calibration of generalized diameter distribution models has been two-stage regression (cf. Steimer et al. 1984), an early method for integrating random effects into nonlinear models (Vonesh and Carter 1992). [Two-stage regression should not be confused with two-stage least squares. The first is an early approach in nonlinear mixed-effects modeling, whereas the second is a common statistical method in econometrics for solving simultaneous equations.] The model is first calibrated for each individual plot. Then, the resulting parameter estimates are regressed against several plot covariates. In the forestry literature, two-stage regression is referred to as the parameter prediction or parameter recovery approach (cf. Vanclay 1994, p. 23). Different fitting techniques for the parameter prediction approach are discussed by Robinson (2004).

Although it accounts for plot random effects, two-stage regression has two major flaws in addition to the assumption of normally distributed error terms. First, the regression performed during the second stage does not take into account the original data, which were used during the first-stage individual calibrations. Consequently, there is no guarantee of convergence on a maximum for the likelihood function. Second, it requires that the diameter distribution model converge for each individual plot during the first stage. As a consequence, it is possible that data from some subjects may be ignored, if the first stage regression fails to converge for certain plots (Vonesh and Carter 1992).

A GLMM offers a solution to these problems. Basically, it combines the aforementioned two stages into a single step. The variability between the individual plots is no longer assessed with a second regression, but rather through random effects. As the plot random effects and the fixed-effect parameters are simultaneously estimated, the GLMM ensures the convergence of the likelihood function on a local or a global maximum, whenever convergence is reached. Moreover, there is no possibility of omitting data from certain plots as no preliminary calibrations are required. In fact, the GLMM provides a predicted average diameter distribution for one or many groups of plots, depending on the model specifications. The random effects allow a departure from this average distribution at the plot level. Therefore, the model enables the analysis of diameter distributions at two different levels, namely the individual plot and the group of plots. This assessment of different error components is a direct response to the issue of variance estimation outlined by García (1992).

A major limitation of the method is the subjectivity associated with selecting the number of $d_i$ to be used for the model calibration. Actually, an infinite number of diameter limits is available. However, problems of convergence arise when the ratio of successes/trials approaches 1. To avoid such problems, we recommend using evenly spaced limits up to the 99th percentile of the observed average diameter distribution as we did in this case study. Using such a boundary does not limit the ability of the model to predict the stem frequency above the 99th percentile. Actually, the modeling approach considers the number of stems above this limit as the difference between the total number of trees and the number of trees below the 99th percentile. Although it provides a stem frequency for the trees above this threshold, the prediction of their diameters relies on the assumption that the diameter distribution of larger trees follows the trend observed in smaller diameters. If this assumption does not hold, the results are limited to the distribution below the selected boundary. Modelers should keep in mind that this recommendation about a threshold is a trade-off between convergence and model accuracy with regard to the diameter predictions of the largest stems. Even though the predicted diameter distributions of different groups of plots are similar, differences might be observed with respect to the distribution of the largest stems, as reported by Fortin et al. (2003).

In this case study, the predictions based on fixed effects only tended to be inaccurate at the plot level (Figure 2b,c). We did not investigate all possible relationships between parameters of the Weibull cdf and the explanatory variables,
nor did we include all possible explanatory variables in the model. To demonstrate the feasibility of the mixed model approach, we tried to keep the analysis as simple as possible. Indeed, nonlinear modeling and additional covariates might improve the accuracy of the predictions based on fixed effects only, but this remains to be investigated. The results of past studies relying on two-stage regression might prove a good framework for future developments for generalized diameter distribution modeling calibrated with mixed effects.

Conclusions

The application of mixed model theory to diameter distribution modeling is possible through a generalized linear mixed model (GLMM). If the random effects and/or the covariance structure provide the appropriate correction for correlation, a GLMM may generate an estimator for the vector of parameters that is consistent and asymptotically normally distributed with consistent and unbiased variance. This feature enables not only valid statistical tests on parameter estimates, but also coherent diameter distribution predictions compatible with stand variables. Thereby, it is possible to compare different groups of plots on a statistical basis by including dummy variables in the model. A GLMM also makes it possible to obtain an average diameter distribution for a group of plots. The random effects can be seen as departures from this average distribution at the plot level.

In this case study, the covariance structure was quite complex and none of the available predefined structures could fit the observed pattern perfectly. As a result, the within-plot residuals still exhibited low persistent correlations, even after specifying random effects and a covariance structure. Note that the residuals of GLMM are conditional on the random effects. Therefore, the correlations between the within-plot residuals do not consider the correlation absorbed by the random effects. For instance, if the random effects account for most of the within-plot variability, the correlations calculated with the conditional residuals may be negligible overall. In this case study, the random effects clearly absorbed a large part of variability (Figure 2b) and the specification of a covariance structure was a matter of fine fitting. Indeed, the implementation of a covariance structure that fits the observed sinusoidal pattern would provide a consistent estimator with truly unbiased variance for the vector of parameters. Multinomial models theory might prove useful for the specification of such a structure. Meanwhile, the variance of the estimator should be considered as “nearly” unbiased.

However, the mixed model approach is subject to convergence problems when many ratios of successes/trials approach 1, viz., when no tree has a diameter greater than the limit $d_{ij}$. As a consequence, a threshold must be set to avoid repeating such ratios unnecessarily. In our case study, we demonstrated that the 99th percentile of the empirical average diameter distribution may be an appropriate upper value. The method also is limited by the link function. This constraint may be important, especially if observed diameter distributions are irregular. Although a finite mixture of two Weibull functions may fit a bimodal distribution (Zhang et al. 2001), problems arise when this model has to be linearized. A complex PDF cannot be linearized with regular link functions. Developments in both irregular distribution modeling and complex link functions might prove a significant advance. Meanwhile, as long as the individual distributions do not exhibit major departures from unimodality, the proposed approach should be used as it ensures compatibility between plot scale and landscape scale.

Literature Cited


