
Predicting Parameters of a Weibull Function for Modeling Diameter Distribution

Quang V. Cao

ABSTRACT. The Weibull function has been widely used to characterize diameter distributions in a forest stand. Methods have been developed to predict the Weibull parameters, either directly or indirectly, from stand age, density, and dominant height. In this study, four existing methods and two new methods of obtaining the Weibull parameters were evaluated using data from a loblolly pine (*Pinus taeda* L.) plantation. While both new methods performed well, one of them produced consistently better goodness-of-fit statistics than those from the existing methods. FOR. SCI. 50(5):682–685.

Key Words: Weibull distribution, parameter prediction, parameter recovery, maximum likelihood, *Pinus taeda*.

THE ABILITY TO PREDICT the distribution of diameter in a stand is essential for foresters to make management decisions. Bailey and Dell (1973) introduced the use of the Weibull distribution in modeling diameter distribution because it is flexible and produces probabilities easily without the need for numerical integration. Parameters of the distribution have been predicted in a variety of ways, but no clear logic exists to justify any one approach over another.

The objective of this study was to develop and evaluate two new methods to predict the parameters of Weibull functions that modeled diameter distributions.

Data

Data from a loblolly pine (*Pinus taeda* L.) plantation at the Hill Farm Research Station, Homer, LA, were used in this study. Bareroot, 1–0 seedlings were planted at a 1.8 m × 1.8 m spacing. Twenty 0.62-ha measurement plots with buffer strips were established at age 5. Each of the five stepwise thinning treatments was applied to four plots starting at age 5. The treatments yielded the following residual densities by age 8: (1) 2,470 trees/ha; (2) 1,483 trees/ha; (3) 741 trees/ha; (4) 494 trees/ha; and (5) 247 trees/ha.

The plots were measured at ages 5, 6, 8, 9, 11, 13, 20, and 21. At age 21, average height of the dominants and codominants ranged from 17.7 to 21.1 m, and average basal

area ranged from 21.4 to 43.2 m²/ha. The data were randomly divided into a fit data set (50%) for estimating regression parameters and a validation data set (50%) for evaluating the methods.

Methods

The Weibull distribution was used in this study to characterize the diameter distribution. The probability density function (pdf) of the Weibull has the following form:

$$f(x) = \left(\frac{c}{b}\right) \left(\frac{x-a}{b}\right)^{c-1} \exp\left[-\left(\frac{x-a}{b}\right)^c\right] \quad (1)$$

where a , b , and c are the location, scale, and shape parameters of the Weibull distribution, respectively, and x is tree diameter at breast height (dbh).

The general form for regression equations used in this study was as follows:

$$y = \exp [b_1 + b_2 RS + b_3 \ln(N) + b_4 \ln(H) + b_5/A] + \epsilon \quad (2)$$

where:

y = a specific Weibull parameter, diameter percentile, or mean diameter;

Quang V. Cao, School of Renewable Natural Resources, Louisiana State University AgCenter, Baton Rouge, LA 70803—Phone: (225) 578-4218; Fax: (225) 578-4227; qcao@lsu.edu.

Manuscript received July 29, 2003, accepted March 4, 2004.

Copyright © 2004 by the Society of American Foresters

RS = relative spacing, $(10,000/N)^{0.5}/H$, which is the ratio of average distance between two adjacent trees (assuming square spacing) and dominant height;
 N = number of trees per ha;
 H = dominant height in meters, or average height of the dominants and codominants;
 A = stand age in years;
 $\ln(\cdot)$ = natural logarithm;
 b_i 's = regression parameters, and;
 ϵ = random error.

The methods used to obtain the Weibull parameters were evaluated and are discussed below.

Method 1: Parameter Prediction

Originally developed by Clutter and Bennett (1965), this method predicted the parameters of the diameter distribution directly from stand attributes. The system consisted of three regression equations, each in the form of Equation 2. The dependent variables of these equations were Weibull parameters a , b , and c , which had previously been estimated for each plot measurement using the method of maximum likelihood. Because cross-equation correlations existed among error components of these equations, they were treated as a system of nonlinear, seemingly unrelated equations (Borders 1989).

Method 2: Moment Estimation

The Weibull location parameter was computed from \hat{D}_0 , the predicted minimum diameter in the stand. Frazier (1981) found that $a = 0.5\hat{D}_0$ gave best results in terms of goodness-of-fit. The b and c parameters were recovered from the first two moments of the diameter distribution (Cao et al. 1982):

$$b = (\hat{D} - a)/\Gamma_1 \quad (3)$$

$$\hat{D}_q^2 + a^2 - 2a\hat{D} - b^2\Gamma_2 = 0 \quad (4)$$

where \hat{D} and \hat{D}_q are predicted arithmetic and quadratic mean diameters, respectively, and $\Gamma_i = \Gamma(1 + i/c)$. Borders' (1989) method was used to simultaneously estimate regression parameters of equations to predict D_0 , \bar{D} , and D_q .

Method 3: Percentiles

Bailey et al. (1989) used this method to compute the Weibull parameters from the predicted quadratic mean diameter (\hat{D}_q), minimum diameter (\hat{D}_0), 25th diameter percentile (\hat{D}_{25}), 50th percentile (\hat{D}_{50}), and 95th percentile (\hat{D}_{95}):

$$a = (n^{1/3}\hat{D}_0 - \hat{D}_{50})/(n^{1/3} - 1) \quad (5)$$

$$c = 2.343088/[\ln(\hat{D}_{95} - a) - \ln(\hat{D}_{25} - a)] \quad (6)$$

$$b = -a\Gamma_1/\Gamma_2 + [(a/\Gamma_2)^2(\Gamma_1^2 - \Gamma_2) + \hat{D}_q^2/\Gamma_2]^{0.5} \quad (7)$$

where n is the number of trees in the plot. This method was not strictly percentiles, as D_q was used. Regression parameters of equations to predict D_q , D_0 , D_{25} , D_{50} , and D_{95} were simultaneously estimated using Borders' (1989) method.

Method 4: Hybrid

As in Method 2, the Weibull location parameter was computed from $a = 0.5\hat{D}_0$. Baldwin and Feduccia (1987) developed this hybrid method in which the b and c parameters were recovered from two values: a moment (the quadratic mean diameter) and a percentile (the 93rd diameter percentile or \hat{D}_{93}):

$$b = (\hat{D}_{93} - a)/2.65926^{1/c} \quad (8)$$

$$a^2 - \hat{D}_q^2 + 2a(\hat{D}_{93} - a)\Gamma_1/2.65926^{1/c} + a(\hat{D}_{93} - a)^2\Gamma_2/2.65926^{2/c} = 0 \quad (9)$$

where $2.65926 = -\ln(1 - 0.93)$. The system of equations to predict D_q , D_0 , and D_{93} was fitted using Borders' (1989) method.

Method 5 (New): Maximum Likelihood Estimator (MLE) Regression

The Weibull location parameter was computed from $a = 0.5\hat{D}_0$. The following two equations were used for predicting the scale parameter (b) and shape parameter (c):

$$b = \exp[b_1 + b_2 RS + b_3 \ln(N) + b_4 \ln(H) + b_5/A] \quad (10)$$

$$c = \exp[c_1 + c_2 RS + c_3 \ln(N) + c_4 \ln(H) + c_5/A] \quad (11)$$

The coefficients b_i and c_i were iteratively searched (see Appendix) to maximize the sum of the log-likelihood values from all plots:

$$\sum_{i=1}^p \ln(L_i)/n_i \quad (12)$$

where:

p = number of plot-age combinations;
 n_i = number of trees in the i th plot-age combination (the term $1/n_i$ was included to give equal weights to plots containing different number of trees);

$\ln(L_i) =$

$$\sum_{j=1}^{n_i} \left[\ln(c) - \ln(b) + (c-1) \ln\left(\frac{x_{ij} - a}{b}\right) - \left(\frac{x_{ij} - a}{b}\right)^c \right] \quad (13)$$

or log-likelihood value for the i th plot-age combination, and;

x_{ij} = dbh of tree j in the i th plot-age combination.

This approach was similar to simultaneously fitting two regression equations for b and c , but the goal was not to minimize the sum of squares of error with respect to b and c but rather to maximize the total log-likelihood (hence the name MLE regression, where MLE stands for maximum likelihood estimator).

Method 6 (New): Cumulative Distribution Function (CDF) Regression

This method is similar to Method 5, except that the coefficients b_i and c_i in Equations 10 and 11 were iteratively searched (see Appendix) to minimize the following function:

$$\sum_{i=1}^p \sum_{j=1}^{n_i} (F_{ij} - \hat{F}_{ij})^2 / n_i \quad (14)$$

where:

F_{ij} = observed cumulative probability of tree j in the i th plot-age combination;

\hat{F}_{ij} = $1 - \exp\{ -[(x_{ij} - a)/b]^c \}$ or the value of the CDF of the Weibull distribution evaluated at x_{ij} ; and

x_{ij} = dbh of tree j in the i th plot-age combination.

Again, this approach was similar to simultaneously fitting two regression equations for b and c , but the objective was to minimize the sum of squares of error with respect to the CDF (hence the name CDF regression) rather than to b and c .

In the above methods, equations to predict D_q , D_0 , D_{25} , D_{50} , D_{93} , and D_{95} had the same form as Equation 2. The following equation was used to predict \bar{D} to ensure $\hat{\bar{D}} < \hat{D}_q$:

$$\bar{D} = \hat{D}_q - \exp[b_1 + b_2 \ln(N) + b_3/A] + \epsilon \quad (15)$$

Evaluation

Three goodness-of-fit statistics were computed for each method and for each plot-age combination of the fit and validation data as follows:

1. The Anderson-Darling (AD) statistic (Anderson and Darling 1954) for the i th plot-age combination:

$$AD_i = -n_i - \sum_{j=1}^{n_i} (2j-1) [\ln(u_j) + \ln(1 - u_{n-j+1})] / n_i \quad (16)$$

where $u_j = F(x_j) = 1 - \exp\{ -[(x_j - a)/b]^c \}$, n_i = number of trees in the i th plot-age combination, and

the x_j are dbh, sorted in ascending order for each plot-age combination ($x_1 \leq x_2 \leq \dots \leq x_{n_i}$).

2. The one-sample Kolmogorov-Smirnov (KS) statistic for the i th plot-age combination:

$$KS_i = \max\{\max_{1 \leq j \leq n_i} [(j/n_i) - u_j], \max_{1 \leq j \leq n_i} [u_j - (j-1)/n_i]\} \quad (17)$$

3. Reynolds et al.'s (1988) error index (EI) for the i th plot-age combination:

$$EI_i = \sum |n_{ik} - \hat{n}_{ik}| \quad (18)$$

where n_{ik} and \hat{n}_{ik} are the observed and predicted number of trees per hectare in diameter class k for the i th plot-age combination, and the sum includes all diameter classes in the i th plot-age combination.

Results and Discussion

The summaries of the AD and KS goodness-of-fit statistics and the error index for both data sets are displayed in Table 1. With a few exceptions, results from these statistics were very similar. Method 1 (parameter prediction) was the poorest performer, ranking last for both fit and validation data. The "parameter recovery" group, which involved recovery of the Weibull parameters from moments (Method 2), percentiles (Method 3), or a combination of both (Method 4), produced similar evaluation statistics.

The new methods (5 and 6) ranked best among the methods evaluated, with Method 6 (CDF regression) consistently producing the lowest goodness-of-fit statistics for both fit and validation data. Compared with Method 4 (hybrid), which was the best of the existing methods in this study, this new method reduced the AD statistic by 29%, the KS statistic by 9%, and the error index by 29% on the average.

Originally, Methods 5 and 6 involved finding appropriate values of the coefficients of three equations for the a , b , and c parameters of the Weibull. The problem with this approach was that the predicted values of a , the location parameter, were always close to zero, even for older plots with high \hat{D}_0 values. The methods were then modified by predicting D_0 separately, setting $a = 0.5 \hat{D}_0$ and reducing

Table 1. Means^a (and standard deviations) of goodness-of-fit statistics for the fit and validation data, by method.

Method ^b	Anderson-Darling statistic		Kolmogorov-Smirnov statistic		Error index	
	Fit data	Validation data	Fit data	Validation data	Fit data	Validation data
1	10.043 (12.528)	9.991 (12.531)	0.193 (0.099)	0.181 (0.083)	517 (386)	526 (374)
2	3.698 (4.162)	7.821 (7.516)	0.143 (0.070)	0.172 (0.084)	359 (264)	464 (330)
3	7.317 (16.126)	6.971 (7.203)	0.149 (0.069)	0.165 (0.082)	361 (298)	382 (273)
4	4.407 (4.864)	6.411 (7.454)	0.148 (0.063)	0.164 (0.082)	333 (272)	348 (241)
5	3.092 (3.061)	5.220 (5.221)	0.144 (0.068)	0.156 (0.072)	252 (167)	297 (189)
6	3.061 (3.028)	4.573 (4.435)	0.142 (0.068)	0.149 (0.066)	218 (145)	248 (158)

^a The smaller the statistic value, the better the fit.

^b The methods were: (1) parameter prediction, (2) moment estimation, (3) percentiles, (4) hybrid, (5) MLE regression, and (6) CDF regression.

the system to two equations for b and c . This approach worked well for both methods.

The poor performance of Method 1 might be due to low R^2 values (0.185–0.853) obtained in predicting Weibull parameters. On the other hand, Methods 2–4 produced distributions to match some plot-level attributes such as diameter moments and/or percentiles, whereas the new methods aimed at matching the observed diameter distribution itself. Even though the moments/percentiles were very well predicted in this study (high R^2 values ranging from 0.945–0.995), these methods were still inferior to the new methods in which the entire predicted diameter distribution contributed to the fitting criterion (either maximizing the total log-likelihood value or minimizing the CDF error sum of squares). It needs to be mentioned that the list of existing methods evaluated in this study was not exhaustive because variations on each method are possible with various combinations of moments and percentiles.

It would seem logical to expect Method 5 (MLE regression) to perform well, as it can be considered as the maximum likelihood predictor, which is the counterpart of the maximum likelihood estimator. On the other hand, because Method 6 (CDF regression) aimed at fitting the CDF of the diameter distribution, it seemed reasonable to expect this method to produce low KS statistics, which measure maximum deviation between observed and predicted CDFs. Furthermore, good performance of Method 6 in terms of lowest values of all goodness-of-fit statistics for both fit and validation data indicated that the new CDF regression method was really the superior method to predict diameter distribution from stand attributes, based on the data used in this study.

Appendix: SAS Program for Methods 5 and 6

```
data one;
input plot age n Hd TPH d F;
* age = stand age,
n = number of trees in plot,
Hd = dominant height,
TPH = number of trees per ha,
d = tree dbh,
F = observed cumulative proportion (count/n);
RS = ((10,000/TPH)**0.5)/Hd;
Tlog = log(TPH);
Hdlog = log(Hd);
pD0 = exp(8.4136 - 1.1626*RS - 0.8209*Tlog -
3.8647/age);
a = 0.5 * pD0;
Y = 0;
proc nlin method = newton;
parms b1 = 4 b2 = -1 b3 = -0.3 b5 = -2 c1 = 8 c2 =
-1 c3 = -0.5 c4 = -1 c5 = -5;
b = exp(b1 + b2*RS + b3*Tlog + b5/age);
c = exp(c1 + c2*RS + c3*Tlog + c4*Hdlog + c5/age);
```

```
Yhat = 0;
if d > a
then do;
x = (d - a)/b;
xlog = log(x);
Yhat = (-log(c/b) - (c - 1)*xlog + exp(c*xlog)) ** 0.5;
end;
else do;
xlog = -100;
Yhat = (-log(c/b) - (c - 1)*xlog + exp(c*xlog)) ** 0.5;
end;
_weight_ = 1/n;
model Y = Yhat;
title 'Method 5: MLE Regression to minimize [-sum
ln(L)/n]';
proc nlin;
parms b1 = 4 b2 = -1 b3 = -0.3 b4 = 0.1 b5 = -2 c1 =
8 c2 = -1 c3 = -0.5 c4 = -1 c5 = -5;
b = exp(b1 + b2*RS + b3*Tlog + b4*Hdlog + b5/age);
c = exp(c1 + c2*RS + c3*Tlog + c4*Hdlog + c5/age);
if d > a
then Fhat = 1 - exp(-(((d - a)/b)**c));
else Fhat = 0;
_weight_ = 1/n;
model F = Fhat;
title 'Method 6: CDF Regression to minimize [sum
(F-Fhat)**2]'
```

Literature Cited

- ANDERSON, T.W., AND D.A. DARLING. 1954. A test of goodness of fit. *J. Am. Stat. Assoc.* 49:765–769.
- BAILEY, R.L., AND T.R. DELL. 1973. Quantifying diameter distributions with the Weibull function. *For. Sci.* 19:97–104.
- BAILEY, R.L., T.M. BURGAN, AND E.J. JOKELA. 1989. Fertilized midrotation-aged slash pine plantations—Stand structure and yield prediction models. *South. J. Appl. For.* 13:76–80.
- BALDWIN, V.C. JR., AND D.P. FEDUCCIA. 1987. Loblolly pine growth and yield prediction for managed West Gulf plantations. *USDA For. Serv. Res. Pap. SO-236.* 27 p.
- BORDERS, B.E. 1989. Systems of equations in forest stand modeling. *For. Sci.* 35:548–556.
- CAO, Q.V., H.E. BURKHART, AND R.C. LEMIN. 1982. Diameter distributions and yields of thinned loblolly pine plantations. *VPI & SU Sch. For. Wildl. Res. Pub. FWS-1-82.* 62 p.
- CLUTTER J.L., AND F.A. BENNETT. 1965. Diameter distribution in old-field slash pine plantations. *Ga. For. Res. Counc. Rep.* 13. 9 p.
- FRAZIER, J.R. 1981. Compatible whole-stand and diameter distribution models for loblolly pine plantations. Ph.D. dissertation, Virginia Polytech. Inst. and State Univ., Blackburg, VA. 125 p.
- REYNOLDS, M.R., T.E. BURK, AND W.C. HUANG. 1988. Goodness-of-fit tests and model selection procedures for diameter distribution models. *For. Sci.* 34:373–399.