Forest Sci., Vol. 30, No. 1, 1984, pp. 129–137

A Segmented Distribution Approach for Modeling Diameter Frequency Data

QUANG V. CAO
HAROLD E. BURKHART

ABSTRACT. Different functions, each in the form of a modified Weibull cumulative distribution function (cdf), were joined together to form a segmented cdf to approximate diameter distributions in forest stands. Estimates of five percentile points (the 0th, 25th, 50th, 75th, and 100th percentiles) were used to determine a segmented cdf. The segmented cdf and the Weibull distribution were fit to diameter data from 252 plot measurements from both thinned and unthinned loblolly pine plantations. Based on the one-sample Kolmogorov-Smirnov goodness-of-fit statistic, the segmented cdf was found to perform no better than the three-parameter Weibull distribution for data from unthinned stands, but was superior in case of thinned stands. The segmented cdf is very flexible; it should be useful for modeling irregular data such as diameter distributions of thinned stands or mixed stands. FOREST SCI. 30:129–137.

ADDITIONAL KEY WORDS. *Pinus taeda,* cumulative distribution function, stand-level model, yield prediction, loblolly pine, Weibull distribution, thinning.

DIAMETER DISTRIBUTIONS in forest stands have been modeled by different probability density functions (pdf's) such as the log-normal (Bliss and Reinker 1964), gamma (Nelson 1964), beta (Clutter and Bennett 1965), Weibull (Bailey and Dell 1973), and Johnson's S_B distributions (Hafley and Schreuder 1977). These pdf's can describe unimodal distributions quite well but they may be inadequate for situations where highly irregular (e.g., multimodal) shapes are found.

When the empirical distribution is irregular, as it is in some cases in thinned stands and mixed stands, a more flexible approach is needed. In this study a method was developed to join different segments of cumulative distribution functions (cdf's) together to form a single smooth cdf which is flexible enough to model irregular diameter distributions.

THE SEGMENTED CUMULATIVE DISTRIBUTION FUNCTION

Suppose X is a continuous random variable which is defined over the interval $[x_{\min}, x_{\max}]$. Let us consider n points

$$x_1 < x_2 < \ldots < x_n,$$

where $x_1 = x_{\min}$ and $x_n = x_{\max}$. The x_j's will be referred to as either percentile points or join points. The cumulative probabilities (p_j's) corresponding to the x_j's are given by:

The authors are, respectively, Assistant Professor in the School of Forestry and Wildlife Management, Louisiana Agricultural Experiment Station, Louisiana State University Agricultural Center, Baton Rouge, LA 70803, and Thomas M. Brooks Professor of Forestry in the Department of Forestry, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061. This study was funded in part by the Loblolly Pine Growth and Yield Research Cooperative at Virginia Polytechnic Institute and State University. The authors thank the North Carolina State Forest Fertilization Cooperative for the data used in this study. Manuscript received 3 May 1982.

$$p_j = Pr(X \leq x_j), j = 1, 2, \ldots, n.$$

It follows that $p_1 = 0$ and $p_n = 1$. A segmented cdf, $F(x)$, is defined here as

$$F(x) = \begin{cases} 0, & x \leq x_1, \\ F_1(x), & x_1 \leq x < x_2, \\ F_2(x), & x_2 \leq x < x_3, \\ \vdots & \\ F_j(x), & x_j \leq x < x_{j+1}, \\ \vdots & \\ F_{n-1}(x), & x_{n-1} \leq x \leq x_n, \\ 1, & x_n \leq x. \end{cases}$$

$F_j(x)$ has to satisfy the following conditions:

(1) $F_j(x)$ must be monotonic nondecreasing.
(2) $F_j(x)$ must be continuous.
(3) $F(x)$ must be continuous at the join points, i.e.,

$$F_j(x_{j+1}) = F_{j+1}(x_{j+1}) = P_{j+1}, \quad j = 1, 2, \ldots, n - 1.$$

(4) $f_j(x)$, the derivative of $F_j(x)$ with respect to x, must be continuous at the join points, i.e.,

$$f_j(x_{j+1}) = f_{j+1}(x_{j+1}), \quad j = 1, 2, \ldots, n - 1.$$

Conditions (1, 2, and 3) ensure that $F(x)$ is a continuous cdf. Condition (4) requires that $F(x)$ be smooth and the corresponding pdf, $f(x)$, continuous at the join points. Graphical representation of a segmented cdf and its corresponding probability density function is shown in Figure 1.

In this study, a modified form of the Weibull cdf was used for each segment. Two coefficients (e_j and d_j) were added to the Weibull cdf resulting in the following segmented cdf:

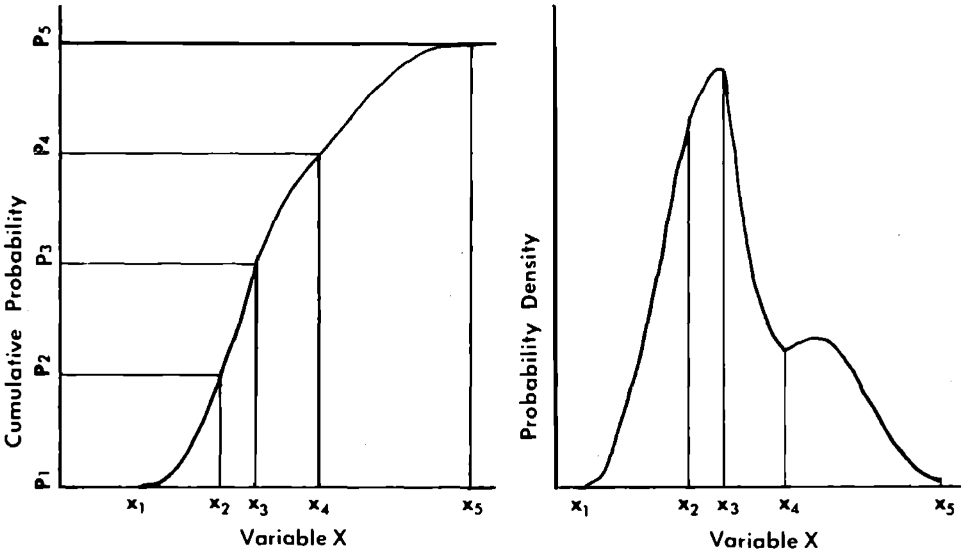

FIGURE 1. Graphical representation of a segmented distribution function and its corresponding probability density function.

$$F_j(x) = e_j\left\{1 - d_j \exp\left[-\left(\frac{x - a_j}{b_j}\right)^{c_j}\right]\right\}$$

$$\text{for } x_j \leq x \leq x_{j+1}, \quad j = 1, 2, \ldots, n - 1.$$

The roles of d_j and e_j will be discussed in the following section.

A Method for Computing the Coefficients of the Segmented CDF.—Computation of the coefficients (a_j, b_j, c_j, d_j, and e_j) of the segmented cdf from the x_j's and p_j's is carried out in the following steps:

1. One function is used for the first two segments ($j = 1, 2$), therefore $a_1 = a_2$, $b_1 = b_2$, $c_1 = c_2$, $d_1 = d_2$, and $e_1 = e_2$. Since d_1 and e_1 are set at 1, the function now has 3 coefficients (a_1, b_1, and c_1) and its curve must pass through 3 points (see the left diagram of Fig. 1): (x_1, p_1), (x_2, p_2), and (x_3, p_3). This method was employed due to its ease of computation.

First, one sets $d_1 = e_1 = 1$ and $a_1 = x_1$. Then from $F_2(x_2) = p_2$ and $F_2(x_3) = p_3$, the remaining coefficients can be derived as

$$b_2 = \frac{x_2 - x_1}{[-\ln(1 - p_2)]^{1/c_2}}$$

and

$$c_2 = \frac{\ln\left[\dfrac{\ln(1 - p_3)}{\ln(1 - p_2)}\right]}{\ln\left[\dfrac{x_3 - x_1}{x_2 - x_1}\right]}$$

where $\ln(x)$ denotes natural logarithm of x.

2. On the subsequent segments ($j = 3, 4, \ldots, n - 2$), d_j and e_j are again set at 1. Then for the interval $[x_j, x_{j+1}]$, $F_j(x)$ has to satisfy the following conditions

$$\begin{aligned} &(1) \quad F_j(x_j) &&= p_j, \\ &(2) \quad f_j(x_j) &&= f_{j-1}(x_j), \text{ and} \\ &(3) \quad F_j(x_{j+1}) &&= p_{j+1}, \end{aligned}$$

where $f_j(x) = dF_j(x)/dx$.

The above conditions require that $F_j(x)$ pass through the two points, (x_j, p_j) and (x_{j+1}, p_{j+1}), and that the derivative of the segmented cdf at x_j be continuous. These conditions result in the following relationship between the parameters:

$$b_j = \frac{x_j - a_j}{\left[-\ln\left(\dfrac{e_j - p_j}{d_j e_j}\right)\right]^{1/c_j}}$$

$$c_j = \frac{(x_j - a_j)\left[\dfrac{f_{j-1}(x_j)}{e_j - p_j}\right]}{\left[-\ln\left(\dfrac{e_j - p_j}{d_j e_j}\right)\right]}.$$

The location coefficient, a_j, is the solution of

$$g(a_j) = F_j(x_{j+1}) - p_{j+1} = 0.$$

The Newton-Raphson method (Hornbeck 1975) is used to search for a_j.

Thus: $a_j^{(2)} = a_j^{(1)} - g(a_j)/g'(a_j)$

where $a_j^{(2)}$ = improved estimate of a_j based on initial guess $a_j^{(1)}$,

$g'(a_j) = dg(a_j)/da_j \doteq [g(a_j + h) - g(a_j)]/h$,

h = a small positive value.

The search is terminated when $g(a_j)$ is very close to 0. Then the current a_j will be the solution. When convergence cannot be reached, i.e., the three-parameter Weibull segment is not flexible enough to satisfy the above three conditions, adjustment of d_j or e_j is necessary. If x increases from a_j to ∞, $F_j(x)$ increases from $e_j(1 - d_j)$ to e_j; or, the range of $F_j(x)$ is $(1 - d_j, 1)$ when $e_j = 1$ and $(0, e_j)$ when $d_j = 1$. Thus d_j and e_j are the "scale" parameters of $F_j(x)$. If convergence does not occur, d_j or e_j has to be adjusted depending on which of the following two cases is true.

(a) $g(a_j) < 0$.

In this case the above procedures are repeated with

$$d_j = [1 - p_j/2], [1 - 3p_j/4], [1 - 7p_j/8], \ldots$$

until convergence is obtained.

(b) $g(a_j) > 0$.

In this instance the coefficient e_j is adjusted:

$$e_j = [p_{j+1} + (1 - p_{j+1})/2], \quad [p_{j+1} + (1 - p_{j+1})/4], \quad [p_{j+1} + (1 - p_{j+1})/8], \ldots$$

until convergence is reached.

3. In the last segment ($j = n - 1$), procedures similar to Step 2 are applied to the last segment of the cdf, except that e_{n-1} is initially set at 1.001. If $e_{n-1} = 1$ then $F_{n-1}(x)$ approaches but never equals 1 as x approaches ∞. Setting $e_{n-1} > 1$ results in $F_{n-1}(x) = 1$ when $x = x_{\max}$. If convergence does not occur and $g(a_{n-1}) < 0$, then the method for adjusting d_j mentioned in Case 2a is used. If $g(a_{n-1}) > 0$ (Case 2b), then e_{n-1} is adjusted as follows:

$$e_{n-1} = [1 + 0.001/2], [1 + 0.001/4], [1 + 0.001/8], \ldots$$

Properties of the Segmented CDF.—When $d_j = e_j = 1$, F_j is a segment of a Weibull cdf. If $a_j = a_{j+1}$, $b_j = b_{j+1}$, $c_j = c_{j+1}$ ($j = 1, 2, \ldots, n - 2$), and all the d_j's and e_j's equal 1, then $F(x)$ is a Weibull cdf.

The coefficients, a_j, b_j, c_j, d_j, and e_j, can be solved algebraically from the x_j's and p_j's. Since the p_j's are specified, the x_j's define a segmented cdf. Thus the parameters of this cdf are actually the x_j's. As a result, a segmented cdf has n parameters which are the n percentile points.

Selection of Join Points.—In this study, the segmented cdf was determined from five arbitrarily chosen percentile points (the 0th, 25th, 50th, 75th, and 100th percentiles). The 0th and 100th percentiles are the minimum and maximum values for the random variable; whereas the 25th, 50th, and 75th percentiles dictated the shape of the segmented cdf. Preliminary investigations (e.g., plotting of diameter distribution histograms) indicated that five points should be satisfactory for describing the range of shapes commonly found in both thinned and unthinned

TABLE 1. Description of various plot attributes from the North Carolina State Forest Fertilization Cooperative study.

Attribute and stand treatment	Minimum	Mean	Maximum
Number of trees/ha			
Control	734	1,416	2,100
Thinned	361	739	1,194
Basal area (m^2/ha)			
Control	26.4	38.8	55.2
Thinned	17.1	25.2	38.2
Total outside-bark volume (m^3/ha)			
Control	136	285	433
Thinned	100	185	331
Age (years)			
Control	11	18	25
Thinned	11	18	25
Average plot dbh (cm)			
Control	14.7	18.6	23.4
Thinned	17.0	20.9	25.9
Minimum plot dbh (cm)			
Control	4.8	11.2	16.0
Thinned	11.7	15.6	21.8
Maximum plot dbh (cm)			
Control	21.1	26.3	35.6
Thinned	20.1	27.0	35.6
Number of trees/plot			
Control	19	40	85
Thinned	6	25	32

pine plantations. The 0th, 50th, and 100th percentile points have meaningful interpretations as the minimum, median, and maximum values of the variable, respectively. Points midway between the 0th and 50th, and between the 50th and 100th percentiles (i.e., the 25th and 75th percentiles) were chosen for use in this investigation. Although many other choices of percentile points could be made, equally spaced points (every 25 percent) were selected.

Evaluation of the Segmented CDF

The segmented cdf was evaluated against the 3-parameter Weibull distribution using data from the regionwide 5 series of the North Carolina State Forest Fertilization Cooperative (NCSFFC) study. These data consist of permanent plot records from eleven installations located in old-field loblolly pine plantations in the Piedmont and Coastal Plain regions of Maryland, Virginia, Georgia, and North and South Carolina. Two treatments were applied: control and thinning. There were between three and five replications for each treatment at each installation, totaling 42 plots for each treatment. The plots ranged from $\frac{1}{30}$ to $\frac{1}{5}$ acre (0.0135 to 0.0809 hectare) in size. Three measurements were taken at 2-year intervals, resulting in a total of 126 plot measurements for each treatment. Table 1 describes the various plot attributes.

TABLE 2. The Kolmogorov-Smirnov statistics, p-*levels, and sum of the ranks for the Weibull and segmented distributions, by stand treatment (control or thinned).*[a]

Stand treatment and distribution	K-S statistic			Sum of the ranks[b]	p-level			Number of rejections at 0.05 level
	Minimum	Mean	Maximum		Minimum	Mean	Maximum	
Control								
Segmented cdf	0.0547	0.1042	0.1865	61	0.0329	0.7734	0.9999	2
Weibull	.0571	.1042	.2057	65	.2918	.7766	.9995	0
				126				
Thinned								
Segmented cdf	.0684	.1273	.2341	81	.3064	.8236	.9997	0
Weibull	.0622	.1428	.2944	45	.0817	.7429	.9997	0
				126				

[a] Data were from the regionwide 5 series of the North Carolina State Forest Fertilization Cooperative study.

[b] For each plot measurement, a rank of 1 was given to the distribution with a smaller value of the K-S statistic (i.e., a better fit to the real data), and the other distribution received a rank of 0.

Diameters at breast height (dbh) were measured to the nearest 0.01 inch (0.025 cm). The diameter distributions from plot measurements were fitted with a 3-parameter Weibull function and a segmented cdf. The method of maximum likelihood was employed to estimate the Weibull parameters[1] whereas the coefficients of the segmented cdf were computed from the 5 sample percentiles using the method outlined above. It is important to note that small sample size (small number of trees per plot) results in biased estimates of population extremes (minimum and maximum stand diameters). This bias was not explicitly considered in the analysis reported here. Furthermore, irregular diameter distributions may result from inadequate sample size as well as thinning and other factors. The range and average of numbers of trees per fitted distribution for thinned and unthinned plots are presented in Table 1.

One-sample Kolmogorov-Smirnov (K-S) statistics were then computed for both the Weibull and segmented cdf as a means of comparison. For each plot measurement, a rank of 1 was given to that distribution with a smaller value of the K-S statistic, whereas the other distribution received a rank of 0. The sum of the ranks for a specified distribution denotes the number of plot measurements where that distribution achieved a smaller K-S value than the other. Table 2 presents the K-S statistics, p-levels, and the sum of the ranks for the Weibull and segmented distributions by stand treatment (control or thinned).

Past work with the Weibull function showed that, for unthinned stands, this pdf was suitable for quantifying diameter distributions which are often unimodal (Smalley and Bailey 1974, Lohrey and Bailey 1976, Clutter and Belcher 1978, Dell and others 1979, Feduccia and others 1979). Although the means of the K-S statistics for both distributions were equal, the Weibull distribution performed slightly better (by a margin of 65 to 61) than the segmented cdf on the data from the unthinned plots based on the one-sample K-S statistics. The Weibull function was also from 2 to 17 times more efficient of computer time in application.

[1] The authors thank B. R. Zutter and R. G. Oderwald for the use of their Weibull fitting program.

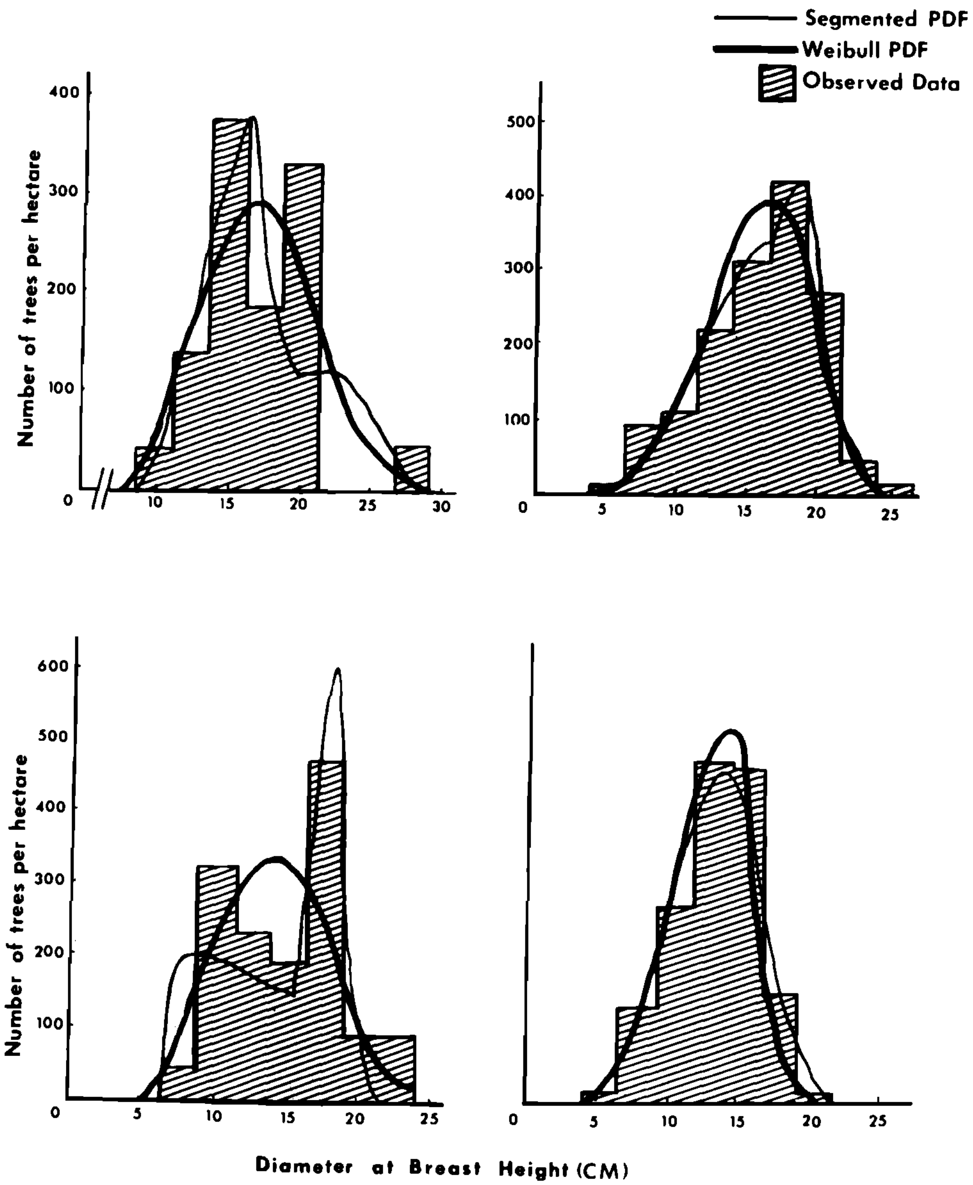

FIGURE 2. Plots of the Weibull and segmented distributions that approximate various observed diameter distributions in thinned stands.

The Weibull distribution has been recently employed for modeling diameter distributions in thinned plantations (Bailey and others 1981, Strub and others 1981, Cao and others 1982, Matney and Sullivan 1982). In this study, the segmented cdf was clearly superior to the Weibull cdf for the data from thinned plots (81 to 45). This is due to the greater flexibility of the segmented cdf which allows it to fit the irregularities of the diameter distributions in thinned stands. It should be noted, however, that the segmented cdf has 5 parameters, two more than the Weibull cdf, and that models with greater numbers of parameters can usually be expected to give closer fits to data.

Plots of the Weibull and segmented distributions that approximate different types of observed diameter distributions found in thinned stands are shown in Figure 2.

Application

The segmented cdf was used in a growth and yield model for thinned loblolly pine plantations to demonstrate the usefulness of this new technique (Cao 1981). The model consisted of two stages. In the first stage, stand-level attributes were predicted using regression techniques. The second stage involved determining the parameters of the segmented cdf so that the resulting diameter distribution would produce stand basal area and average dbh estimates compatible with those predicted from regression equations in the first stage. By linking these two stages, the size-class distribution information produced is conditioned to provide aggregate values that are consistent with the predicted overall stand attributes.

Stand-Level Model.—The stand-level model consisted of regression equations that predict (1) stand attributes (such as minimum, median, maximum, and average diameters) and (2) density (basal area and trees per unit area) of a stand in the future (age A_2) based on stand information at present (age A_1). Also needed was a mean height equation that predicts total height corresponding to a specified dbh for a given stand condition.

Deriving the Diameter Distribution from Stand Attributes.—The segmented cdf was employed to approximate diameter distributions. A total of five percentile points were used to determine the cdf:

$$D_{\min}, D_{0.25}, D_{\text{med}}, D_{0.75}, \text{ and } D_{\max},$$

where $D_{\min}$ = minimum dbh,
D_{med} = median dbh,
$D_{\max}$ = maximum dbh,
$D_{0.25}$ = the 25th percentile for dbh where $Pr[\text{dbh} \le D_{0.25}] = 0.25$,
$D_{0.75}$ = the 75th percentile for dbh.

$D_{\min}$, D_{med}, and $D_{\max}$ were predicted from stand variables, using regression techniques. Since five percentile points define a segmented cdf, the remaining two percentiles ($D_{0.25}$ and $D_{0.75}$) were determined in such a manner so as to ensure that the resulting segmented cdf would produce stand basal area and arithmetic mean dbh estimates identical to those predicted from regression equations. That is to say, $D_{0.25}$ and $D_{0.75}$ were given by following equations:

$$\hat{B} = KN \int_{D_{\min}}^{D_{\max}} x^2 f(x)\, dx$$

$$\hat{\bar{D}} = \int_{D_{\min}}^{D_{\max}} x f(x)\, dx$$

where $\hat{B}$ = stand basal area in m^2/ha, predicted from regression equation,
$\hat{\bar{D}}$ = predicted arithmetic mean dbh in cm,
K = $\pi/[4(100)^2] = \pi/40{,}000$,
N = number of surviving trees per hectare,
$f(x)$ = $dF(x)/dx$ = segmented diameter pdf,
$F(x)$ = segmented cdf defined by $D_{\min}$, $D_{0.25}$, D_{med}, $D_{0.75}$, and $D_{\max}$.

Possible Modifications and Refinements

In this study, a flexible method—the segmented cdf—was introduced to characterize diameter distributions in forest stands. The cdf adequately characterized diameter distributions in thinned stands. Due to its flexibility, the segmented cdf should be useful for describing data from other fields of biological science as well. Although satisfactory results were attained, there is still room for improvement in this approach. Three specific areas for further investigations are—

(1) Improvement in the flexibility of the segmented cdf approach by estimating more percentile points or join points. This is accompanied by a burden, however, of estimating these new parameters.

(2) A search is necessary for some method of meaningfully placing join points such that (a) good characterizations occur, (b) the join points can be easily predicted, so that (c) the d_j's and e_j's can be eliminated.

(3) Further investigations are also needed to identify an appropriate functional form (other than the modified form of the Weibull cdf used in this study) for the cdf segments that is both simple and flexible.

Literature Cited

Bailey, R. L., and T. R. Dell. 1973. Quantifying diameter distributions with the Weibull function. Forest Sci 19:97–104.

Bailey, R. L., N. C. Abernathy, and E. P. Jones, Jr. 1981. Diameter distributions models for repeatedly thinned slash pine plantations. *In* Proceedings of the first biennial southern silvicultural research conference (J. P. Barnett, ed), p 115–126. USDA Forest Serv Tech Rep SO-34, 375 p.

Bliss, C. I., and K. A. Reinker. 1964. A log-normal approach to diameter distributions in even-aged stands. Forest Sci 10:350–360.

Cao, Q. V. 1981. Empirical diameter distributions and predicted yields of thinned loblolly pine plantations. Ph D diss, Va Polytech Inst and State Univ. 100 p. (Diss Abstr Int 42:3890–B.)

Cao, Q. V., H. E. Burkhart, and R. C. Lemin, Jr. 1982. Diameter distributions and yields of thinned loblolly pine plantations. Sch For and Wildl Res, Va Polytech Inst and State Univ, FWS-1-82, 62 p.

Clutter, J. L., and D. M. Belcher. 1978. Yield of site-prepared slash pine plantations in the lower coastal plain of Georgia and Florida. *In* Growth models for long term forecasting of timber yields (J. Fries, H. E. Burkhart, and T. A. Max, eds), p 53–70. Sch For and Wildl Res, Va Polytech Inst and State Univ, FWS-1-78, 249 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.

Dell, T. R., D. P. Feduccia, T. E. Campbell, W. F. Mann, Jr., and B. H. Polmer. 1979. Yields of unthinned slash pine plantations on cutover sites in the West Gulf region. USDA Forest Serv Res Pap SO-147, 84 p.

Feduccia, D. P., T. R. Dell, W. F. Mann, Jr., T. E. Campbell, and B. H. Polmer. 1979. Yields of unthinned loblolly pine plantations on cutover sites in the West Gulf region. USDA Forest Serv Res Pap SO-148, 88 p.

Hafley, W. L., and H. T. Schreuder. 1977. Statistical distributions for fitting diameter and height data in even aged stands. Can J Forest Res 7:481–487.

Hornbeck, R. W. 1975. Numerical methods. Quantum Publishers, Inc, New York. 310 p.

Lohrey, R. E., and R. L. Bailey. 1976. Yield tables and stand structure for unthinned longleaf pine plantations in Louisiana and Texas. USDA Forest Serv Res Pap SO-133, 53 p.

Matney, T. G., and A. D. Sullivan. 1982. Compatible stand and stock tables for thinned and unthinned loblolly pine stands. Forest Sci 28:161–171.

Nelson, T. C. 1964. Diameter distribution and growth of loblolly pine. Forest Sci 10:105–115.

Smalley, G. W., and R. L. Bailey. 1974. Yield tables and stand structure for loblolly pine plantations in Tennessee, Alabama, and Georgia highlands. USDA Forest Serv Res Pap SO-96, 81 p.

Strub, M. R., D. P. Feduccia, and V. C. Baldwin, Jr. 1981. A diameter distribution method useful in compatible growth and yield modeling of thinned stands. *In* Proceedings of the first biennial southern silvicultural research conference (J. P. Barnett, ed), p 127–130. USDA Forest Serv Tech Rep SO-34, 375 p.