Precision and the Probability of Misclassification in Site Index Estimation

F. THOMAS LLOYD

WILLIAM L. HAFLEY

ABSTRACT. This work presents an estimate of variance for a site index predictor. The methodology was tested on height/age data for loblolly pine from the Coastal Plain of the southeastern United States. We quantified the effects of stand age, base age, and sample size on the variance of site index. The derived variance was used in the development of a measure of the probability of misclassification. We used this measure to show the effects of site index class interval width, stand age, base age, and sample size on classification performance of this site index predictor. The results show that changing the width of class intervals or the value of stand age has the greatest effect on the classification performance. The methods have practical value as guides to selecting sample sizes that produce uniformly reliable site index predictions. FOREST SCI. 23:493–499.

ADDITIONAL KEY WORDS. Cumulative height curves, sample size, stem analyses, polymorphic curves, variance, base-age invariance.

FORESTERS ROUTINELY ACCEPT SITE INDEX (SI) as a measure of the productive potential of forested sites. Since we have no way of estimating the variance of SI predictions, we do not know how many tree heights to measure—we just go out and measure a few. We assign silvicultural treatments to stands with certain SI classes (ranges in SI), but we do not know what the probability of misclassification from the sample measurements is. We suspect that SI predictions become increasingly unreliable as stand age decreases, but we have no measure of this change. This paper presents a method for addressing these problems.

Our approach differs from the only other definitive work on this problem (Heger 1971, 1973) in that our objective was to study the precision of a SI predictor directly, rather than indirectly via the estimation of approximate confidence intervals. Our reason for taking a different approach was predicated on the belief that the primary use of SI is in placing forested sites into SI classes. We assumed that classification success depends on the distribution of the true (but unknown) values of SI for sites in a given SI class interval and on the distribution of the estimator of SI for the particular site being indexed. Intuitively, such a measure should be a function of the variance of the SI predictor, since it—along with the true SI—is a random variable in the joint bivariate probability function used in defining the measure of misclassification.

We also developed a method that was base-age invariant, a property recommended by Bailey and Clutter (1974). This property is best described as one requiring that *only one* height-over-age curve be associated with any single height/age point. It says that once the point defined by the mean height of the sampled trees and the corresponding stand age is located, the same height growth curve should be estimated by the SI process no matter what base age is selected.

The authors are, respectively, Mathematical Statistician, Southeastern Forest Experiment Station, USDA Forest Service, Charleston, South Carolina 29403; and Professor of Forestry and Statistics, North Carolina State University, Raleigh, North Carolina 27607. This paper is based on the senior author's Ph.D. research at North Carolina State University. Manuscript received February 28, 1977.

The Test Data

Our results are demonstrated on a set of stem analysis data from naturally regenerated loblolly pine (*Pinus taeda* L.) trees in the southeastern Coastal Plain of the United States. The numerical results from the fit of our model are applicable only to the indicated species and region. The sample trees were cut and analyzed by Trousdell and others (1974) to develop polymorphic SI curves. Two sample trees were cut from each of 22 sites covering the full range in site quality and containing stands at least 50 years old. All trees were dominant or codominant, and examination of increment cores revealed no periods of suppressed growth.

Total age was estimated from an increment core taken 1 foot above the ground. After felling, each tree was cut into 6-, 8-, or 10-foot sections, the length depending on the total height of the tree. An increment core was taken from each section to determine the age of the nearest primary whorl below the cut. The sections were split when primary whorls could not be recognized on the surface. The process yielded 12 to 16 pairs of height/age measurements for each tree.

The height-over-age model we used required that height measurements be available at the same ages for all sites. Therefore, heights were estimated where necessary by linear interpolation for ages 10 to 55 by 5-year increments. The 5-year increment was chosen because it resulted in no more than one interpolated height between measured heights. Choosing 55 years as the maximum age resulted in dropping 3 sites containing stands less than that age, but the remaining sites covered the same range in SI as the entire data set. Heights at ten ages from each of two trees on 19 sites resulted in 380 height/age observations.

The Estimation Process

Close inspection of SI estimation reveals it to be a two-step process. The first step is the *selection* of an expected height-over-age relationship from a family of curves, where the selection is accomplished by choosing the curve that passes through the height/age point defined by the mean of the sample tree heights and stand age. The family model can be envisioned as a continuous version of the familiar set of SI curves. The second step is *to predict* a height from the selected curve at a specified age, called base (or index) age. We started our solution by hypothesizing an infinite-sized family of height-over-age curves where all member curves had a set of parameters in common (called recurrent parameters), but where each member curve was distinguished by the value of a single parameter (called the indexing parameter). The solution sequence was first to estimate the recurrent parameters from data on past growth. Next, we derived an estimator for the indexing parameter, which was used to select a member curve from the family model as the cumulative height growth curve for the site. Finally, the SI estimate was obtained via its definition by evaluating the selected curve at the chosen base age.

The Family Model

As long as only a single height/age point is observed for curve selection, the family model can have only one indexing parameter per member curve. Therefore, ingenuity is needed in constructing a model that has this property, yet which is at the same time flexible enough to adequately describe the observed height growth patterns. After some trial and error, we arrived at the model

$$h = \gamma t(1 - 0.01t) + \beta_1 t^2 + \beta_2 t^3 + \varepsilon \qquad (1)$$

which was derived by constraining the third degree polynomial

$$h = \gamma t + \beta_1{}^* t^2 + \beta_2 t^3 + \varepsilon \qquad (2)$$

with the relationship

$$\beta_1{}^* = \beta_1 + c\gamma \,, \tag{3}$$

where γ, the indexing parameter, takes values over some interval. The beta's (β_1 and β_2) are the recurrent parameters.

We chose to make the model linear in its parameters by taking $c = -0.01$. This value of c was obtained by fitting Equation (2) separately to the individual site stem analysis data and then fitting Equation (3) to the estimates of the pairs of parameters, $\beta_1{}^*$ and γ. The R-square for the fit of Equation (3) was high ($R^2 = 0.951$). It remains for future work to investigate whether the nonlinear model, which results when c is a parameter, will allow a tractable solution and, if it does, whether it is worth the additional effort.

Estimating the Recurrent Parameters

The recurrent parameters are estimated from past growth data, like the stem analysis data described above. The model used to fit these loblolly pine data consists of a subset of 19 member curves (for the 19 sites) from the family model given in Equation (1). The resulting cumulative height growth model is

$$h_{ijk} = \sum_{i=1}^{p} \gamma_i t_k (1 - 0.01 t_k) \theta_{i'} + \beta_1 t_k{}^2 + \beta_2 t_k{}^3 + \varepsilon_{ijk} \tag{4}$$

where h_{ijk} is the height at t_k ($k = 1, 2, \ldots, n$) of the j^{th} tree ($j = 1, 2, \ldots, s_i$) on the i^{th} site ($i = 1, 2, \ldots, p$), p (equal to 19) equals the number of sampled sites, s_i (equal to 2) equals the number of stem-analyzed trees on the i^{th} site, n (equal to 10) equals the number of ages at which heights were observed, and $\theta_{i'}$ ($i' = 1, 2, \ldots, p$) is a dummy variable that takes the values 1 when $i = i'$ and 0 when $i \neq i'$. It is helpful when trying to visualize Equation (4) to write it out in matrix form.

The error structure for the vector of ε_{ijk}'s from Equation (4) is not the familiar diagonal matrix $I\sigma^2$ (I being an identity matrix) from ordinary (unweighted) least squares (LS) theory. Specifically, nonzero covariances were assumed between the error terms ε_{ijk} and $\varepsilon_{ijk'}$ ($k \neq k'$) because they are associated with measurements on the same tree, while all other covariances between the ε_{ijk}'s are assumed equal to zero. Furthermore, these data supported the assumption that the variance of ε_{ijk} is homogeneous ($\mathrm{Var}(\varepsilon_{ijk}) = \sigma^2$) over the observed age range. This assumption might not hold for all species, in which case, solutions that use this approach would be more difficult to obtain. We will see later how the existence of a constant variance over age can be used to explain some unexpected results, namely, why variance of the SI predictor is not necessarily minimum at base age. Finally, we found that for these loblolly pine data the variance of the parameter estimates were not materially reduced by using weighted LS estimators, so we used the easier-to-work-with ordinary LS estimators. This choice should be reevaluated (see Rao 1967) when this methodology is applied to a different species.

Estimating the Indexing Parameter

Let

$$h_{j'}{}^* = \gamma^* t_\tau (1 - 0.01 t_\tau) + \beta_1 t_\tau{}^2 + \beta_2 t_\tau{}^3 + \varepsilon_{j'} \tag{5}$$

be the height-over-age model for the site being indexed where $h_{j'}{}^*$ ($j' = 1, 2, \ldots, r$) is the height of the j'^{th} sample tree (the * is used to distinguish the new sample tree data from the initial stem analysis data), r equals the number of tree heights measured, t_τ is the stand age, and the other terms are as previously defined. Since t_τ is the same for all measured heights, one's intuition suggests estimating γ^* with

$$\hat{\gamma}^* = (\bar{h}^* - \hat{\beta}_1 t_T^2 - \hat{\beta}_2 t_T^3)/[t_T(1 - 0.01 t_T)] \tag{6}$$

where $\bar{h}^*$ is the mean of the sample tree heights, $\hat{\beta}_1$ and $\hat{\beta}_2$ are the LS estimators from Equation (4), and t_T is the common stand age. Lloyd (1975) shows that the LS estimator of γ^* from a linear model like Equation (4) for the augmented vector of heights made up of both the stem analysis data and the sample tree data is the same as Equation (6) above. This enabled us to use ordinary LS theory in deriving the variance of the resulting SI predictor.

The SI Predictor and Its Variance

The SI predictor is obtained by applying the definition of SI to the selected curve obtained by estimating γ^*. The resulting SI predictor is

$$\begin{aligned} \hat{h}_{t_I}^* &= \hat{\gamma}^* t_I(1 - 0.01 t_I) + \hat{\beta}_1 t_I^2 + \hat{\beta}_2 t_I^3 \\ &= \bar{h}^* \phi + \hat{\beta}_1(t_I^2 - t_T^2 \phi) + \hat{\beta}_2(t_I^3 - t_T^3 \phi) \end{aligned} \tag{7}$$

where t_I is the selected base age and

$$\phi = t_I(1 - 0.01 t_I)/t_T(1 - 0.01 t_T)\ . \tag{8}$$

Traditional SI curves can be obtained from Equation (7) by solving for $\bar{h}^*$ in terms of base age (t_I), stand age (t_T), and SI $(\hat{h}_{t_I}^*)$, and then plotting $\bar{h}^*$ over t_T for selected values of base age and SI.

The variance of $\hat{h}_{t_I}^*$ is

$$\begin{aligned} \mathrm{Var}(\hat{h}_{t_I}^*) = {} & \phi^2 \frac{\sigma^2}{r} + (t_I^2 - t_T^2 \phi)^2 \,\mathrm{Var}(\hat{\beta}_1) \\ & + (t_I^3 - t_T^3 \phi)^2 \,\mathrm{Var}(\hat{\beta}_2) \\ & + 2(t_I^2 - t_T^2 \phi)(t_I^3 - t_T^3 \phi)\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2)\ , \end{aligned} \tag{9}$$

because $\mathrm{Cov}(\bar{h}^*, \hat{\beta}_1) = \mathrm{Cov}(\bar{h}^*, \hat{\beta}_2) = 0$. The estimators used for $\mathrm{Var}(\hat{\beta}_1)$, $Var(\hat{\beta}_2)$, and $\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2)$ are given by Lloyd (1975, Appendix D). The estimator for σ^2 is obtained by pooling the separate variance estimates from the 190 pairs of the tree heights at each age on each site.

Results from Fitting the Model

The statistics of fit for Equation (4) are $R^2 = 0.988$ and MSE $= 7.79$. The estimate of σ^2 is $\hat{\sigma}^2 = 8.71$. We tested for lack-of-fit (see Draper and Smith 1966, p. 57) and found no evidence of model bias (F $= 0.78$) when compared with the data used to fit the model. We believe the SI estimates from Equation (7) are equally unbiased for the intended population of sites because the SI curves from this model compared closely with the polymorphic curves from Trousdell and others (1974), which were tested for bias on an independent set of data. Also, the similarity of these curves to those from the polymorphic, nonlinear model used by Trousdell and others (1974) (over the common age range) attests to the flexibility of our model. However, this model would not be expected to work well over age ranges where height growth is nearly zero for an extended time. The unweighted LS estimates of the recurrent parameters are $\hat{\beta}_1 = 0.0017272$ and $\hat{\beta}_2 = 0.0001018$. Their estimated variances and covariance are $\hat{\mathrm{Var}}(\hat{\beta}_1) = 4.8292 \times 10^{-6}$, $\hat{\mathrm{Var}}(\hat{\beta}_2) = 9.9833 \times 10^{-10}$, and $\hat{\mathrm{Cov}}(\hat{\beta}_1, \hat{\beta}_2) = -6.8834 \times 10^{-8}$.

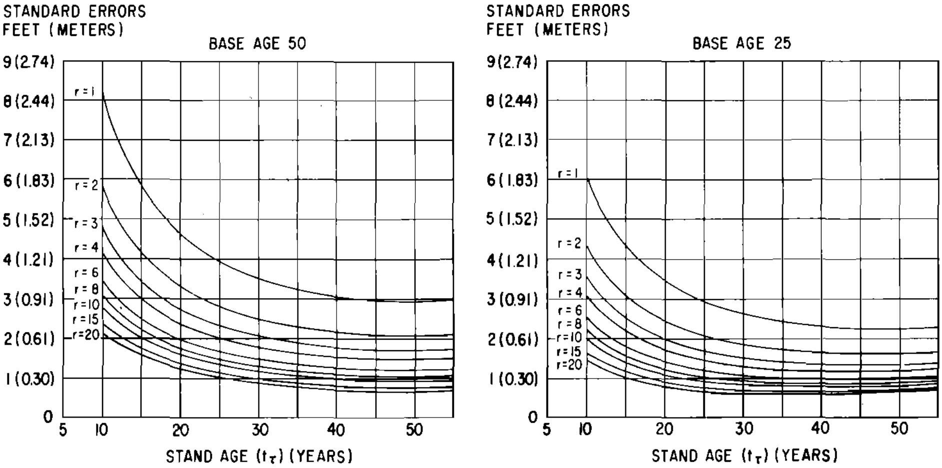

FIGURE 1. Estimated standard errors for the site index predictor plotted over stand age (for selected sample sizes (r)).

DISCUSSION

Standard errors obtained from an estimate of the variance given by Equation (9) are plotted in Figure 1 for selected sample sizes and for base ages 50 and 25, respectively. It is easy to see from an inspection of Equation (9) or from the graphs in Figure 1 that the variance of SI is the same for all base ages whenever stand age equals base age. It is not so easy, however, to see why variance is not always minimum at base age (see Figure 1, base age 25). It is also not easy to explain why this can happen. We start by asking, why should the variance of the SI predictor be minimum when stand age equals base age? If cumulative height growth curves are converging with decreasing stand age and the variance of $\bar{h}^*$ is constant (as assumed) over the observed age range, then projecting repeated samples of $\bar{h}^*$ from the same site backwards to a younger base age would produce projected heights (SI estimates) with less variance. Now superimpose on this the fact from regression theory that the variance of a predicted value increases as the independent variable (base age) takes values further from the mean value of the data used to fit the model. Consider also the reversed effect of stand age (beyond some age) caused when curves start to converge with *increasing* age—as was the case with these data (it is not uncommon for the poorer sites to start catching up in cumulative height growth in later years). We see then that the variance of a SI predictor is not a simple, symmetrical function of the distance of stand age from base age; but rather, is an interplay of the value of stand age, of whether the cumulative growth curves are diverging or converging with increasing age, the distance of base age from the mean age of the initial data, and the distance of stand age from base age. In summary, our results show the variance of SI increasing at an increasing rate as stand age becomes younger, so it is not impossible for a young base age to have a larger variance of SI when stand age equals that base age than for a stand age older than the selected base age.

The probabilities of misclassification plotted in Figure 2 were obtained by assuming a uniform distribution for the true SI ($h_{t_I}{}^*$) over the SI class interval (h_l,h_u) and a normal distribution for the ε_{ijk}'s from the stem analysis data. The probability estimates were calculated by numerically integrating the joint probability density over the region defined by true SI in the interval (h_l,h_u) and estimated SI outside

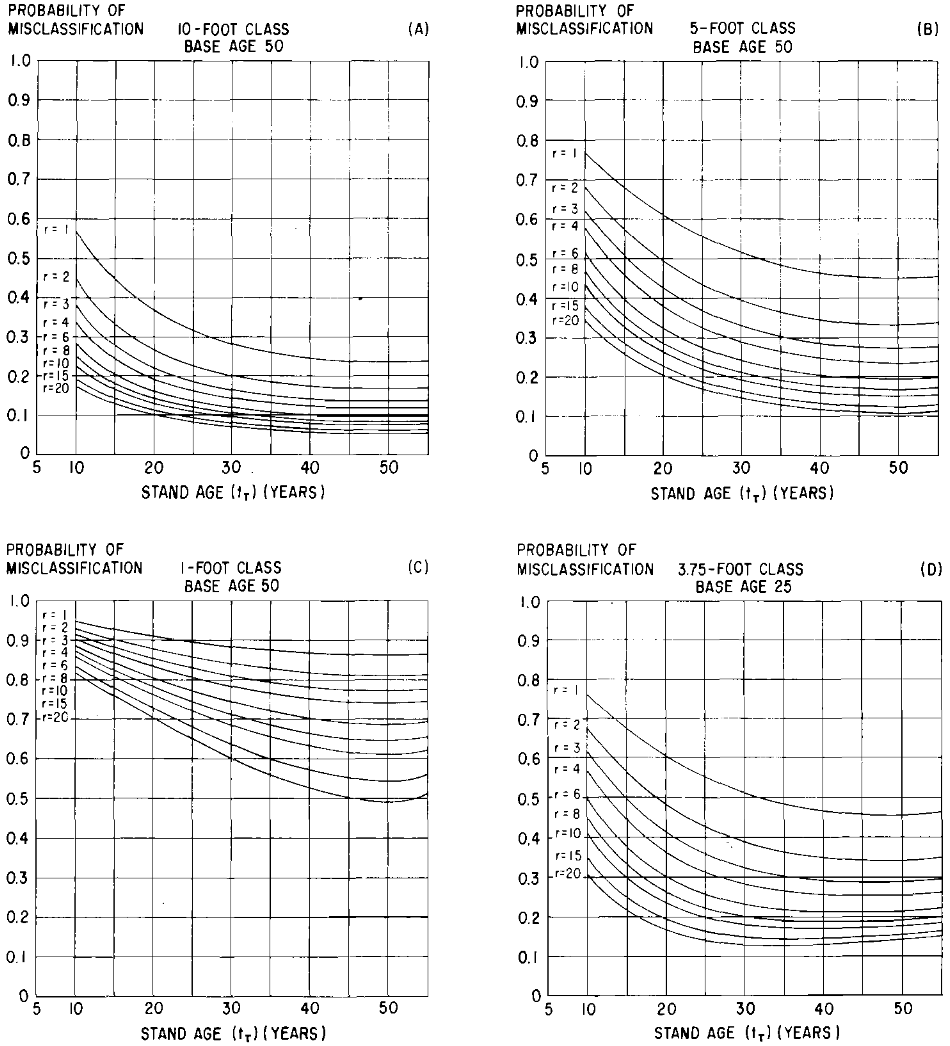

FIGURE 2. Estimates of the probability of misclassification plotted over stand age (for selected sample sizes (r)).

the interval (h_l, h_u). The resulting probability of misclassification is a function of the width of the SI class interval $(h_u - h_l)$, the sample size (r), the base age (t_I) and the stand age (t_T).

Figures 2A, 2B, and 2C illustrate the effect the SI class interval has on the probability of misclassification. Obviously, the choice is not arbitrary and there is a practical limit to how small they can be chosen. Even the often used 5-foot class (Figure 2B) requires measuring approximately 6 tree heights at age 10 to have a chance of being right half of the time (1-out-of-2 odds). It takes approximately 20 tree heights at age 10 to decrease the odds of misclassification to 1-out-of-3.

Figures 2B and 2D give an example of the effect of changing base age. In this model, curves that are 5 feet apart at age 50 are 3.75 feet apart at age 25, so the SI classes were changed accordingly to facilitate the comparison. The example shows that lowering base age for younger stands does not necessarily offer great gains in classification success. It also demonstrates again the dominant effect of stand age.

Application

This information is useful in choosing the number of sample trees relative to stand age, given the desired odds of misclassification, the SI class, and base age. For example, figure 2B shows it would be necessary to sample from stands aged 50, 40, 30, 20, and 10 years approximately 2, 2, 3, 6, and 20 tree heights, respectively, to maintain 1-out-of-3 odds for misclassification. This type of information could be tabulated to aid data collection teams in obtaining uniformly reliable SI classifications.

This type of analysis could help the forester use SI more wisely. In particular, it offers an objective method of determining whether particular demands on the SI tool are realistic. For example, Figure 2 shows that requiring the 1-out-of-20 odds used as an error rate in most statistical inference work in research would require prohibitively large sample sizes, even for the widely used 10-foot SI class. The forester might find that he has to go to 20- or 30-foot SI classes in order to have classification success commensurate with his willingness to measure sample tree heights.

Literature Cited

Bailey, R. L., and J. L. Clutter. 1974. Base-age invariant polymorphic site curves. Forest Sci 20(2):155–159.

Draper, N. R., and H. Smith. 1967. Applied regression analysis. John Wiley and Sons, New York. 407 p.

Heger, L. 1971. Confidence intervals for site index using curves based on stem analyses. Can J For Res 1:241–245.

Heger, L. 1973. Effect of index age on the precision of site index. Can J. For Res 3(1): 1–6.

Lloyd, F. T. 1975. An estimator of variance for a site index predictor using a family of cumulative total height growth curves as the model. Ph.D. Thesis. N. C. State Univ. 73 p. University Microfilms, Ann Arbor, Mich. (Diss. Abstr. 75-17, 103).

Rao, C. R. 1967. Least squares theory using an estimated dispersion matrix and its application to measurement of signals. Fifth Berk Symp on Math Stat and Prob 1:355–372.

Trousdell, K. B., D. E. Beck, and F. T. Lloyd. 1974. Site index for loblolly pine in the Atlantic Coastal Plain of the Carolinas and Virginia. USDA Forest Serv Res Pap SE-115, 11 p. Southeast Forest Exp Stn, Asheville, N. C.