Comparing and Combining Stem Analysis and Permanent Sample Plot Data in Site Index Models

Oscar García

ABSTRACT. Most site index models are built either from repeated measurements on permanent sample plots or by using data obtained by stem analysis. Sometimes both types of data are available. It is then of interest to test them for consistent height-age trends, and to estimate model parameters using all the data. Efficient procedures should take into account the difference in error patterns between the two data sources. A stochastic differential equation model is proposed in which trend and error parameters can take either different or common values for various data groups. All parameters are simultaneously estimated by maximum-likelihood, and computed likelihoods and likelihood-ratio tests can be used for comparing model variants and for hypothesis testing. The methods are applied to Pinus pinaster plantations in northern Spain.


Key Words: Height growth, Pinus pinaster, differential equations, Richards, Spain.

SITE INDEX MODELS relate height, age, and site quality in even-aged stands (Jones 1969, Carmean 1975, Hägglund 1981). They are commonly obtained from one of three kinds of data: temporary sample plots, permanent sample plots, or stem analysis. Temporary sample plots provide one-time measurements, their use requiring strong assumptions that are often violated in practice (Clutter et al. 1983, Walters et al. 1989). When available, sequences of consecutive stand height measurements on permanent sample plots (PSP) are widely considered the best basis for site index model development (Spurr 1952, Clutter et al. 1983, Hägglund 1981). Stem analysis (SA) reconstructs past height growth from growth ring observations on dissected sample trees (Dahms 1963, Curtis 1964). With SA, long-term data collection is not needed, and measurements tend to be more precise than those for heights on standing trees in PSPs, but there is a potential for serious biases due to changes in dominance, tree selection, and other reasons (Dahms 1963).

Sometimes, both PSP and SA data are available. I present methods for combining these data sources in site index modeling, and for testing consistency of the PSP and SA measurements. The modeling takes into account the different stochastic structure of the data. The techniques are demonstrated in a model for Pinus pinaster Ait. in Galicia, northwestern Spain, for which regional differences are also tested.

In what follows, first a site modeling framework from García (1983) is explained. It consists of a stochastic differential equation (SDE) that generates Richards’s height-age curve, including random environmental perturbations and measurement error. Parameters are estimated by maximum-likelihood, and likelihood ratio tests can be used for inference about various hypotheses. The original approach is extended to handle differences between data sources. Then, the data and results obtained for Pinus pinaster are described. A Summary and Conclusions section closes the article.

Methods

Site-index curves are one-parameter families of curves [1] of the form

$$H = f(t, q),$$

(1)

where $H$ is top height, $t$ is age, and $q$ is a site-dependent parameter defining different curves in the family (Bailey and Clutter 1974, Hägglund 1981). A stand in a certain site is supposed to follow one of these curves. The most commonly used site-dependent parameter is the site index, the height reached at some given reference age (index age or base age). Site index is related to any other $q$ by substituting in Equation 1 the site index for $H$ and the base age for $t$.

In reality, observed heights are subject to measurement and sampling errors, and growth fluctuates over time according to weather and other factors. Equation 1 can be thought of as representing the expected or most likely result for stands growing in a given site. Rational estimation needs to consider the deviations from the site-index curve trend.

A general model would describe height growth rate as

$$\frac{dH}{dt} = g(H, t, q, u),$$

(2)

with an initial height $H(t_0) = H_0$ at some age $t_0$ (often both taken as zero); $u(t)$ is a stochastic process representing environmental fluctuations. In addition, observed heights

University of Northern British Columbia, 3333 University Way, Prince George, British Columbia V2N 4Z9, Canada—Phone: (250) 960-5004; Fax: 960-5539; garcia@unbc.ca.

Acknowledgments: Much of this work was done at the Centro de Investigaciones Forestales de Lourizán, Pontevedra, Spain, while the author was employed by the Xunta de Galicia and the Universidad de Santiago de Compostela. The assistance of Pablo Vila Lameiro made this project possible. I am grateful to Lourizán and its staff for the use of data, and for their collaboration and hospitality.

Manuscript received January 9, 2004, accepted October 28, 2004

Copyright © 2005 by the Society of American Foresters

Forest Science 51(4) 2005 277
are subject to measurement/sampling errors represented by random variables $\varepsilon$:

$$h_i = H(t_i) + \varepsilon_i.$$  

(3)

Although the definition of site index is straightforward and unambiguous in a deterministic world, once we introduce randomness it needs to be made more precise. Some researchers have implicitly defined site index as the actual height reached by a specific stand at the base age. That is, site index would be a property of the stand. However, more in the spirit of the traditional concept is to view site curves and site index as some average over hypothetical stands that might grow in the site. Site index is then a property of the site. Confusion can arise from failing to distinguish between these two different definitions, which we might dub “stand site-index” and “site site-index” (e.g., Heger 1973, Hägglund 1981, Bailey and Cieszewski 2000, and references therein). I take the site site-index view, with site index being the most likely top height at a base age. Therefore, the base age is here essentially arbitrary, a device for labeling members from the family of curves; results should be base-age invariant, in the sense of Bailey and Clutter (1974).

**Models**

A flexible and convenient model for the SDE (Equation 2) is

$$\frac{dH}{dt} = b(a' - H) + \sqrt{\sigma} w(t),$$

(4)

with $w(t)$ being “white noise,” the formal derivative of a Wiener stochastic process (García 1983). The deterministic part integrates to the Richards growth curve (von Bertalanffy 1949, Richards 1959).

$$H = a[1 - (1 - H_0/a')e^{-b(1-t_o)}]^{1/c},$$

(5)

one of the most commonly used height-age models[2]. The serially independent white noise perturbations cause a particular stand to deviate from the most likely trajectory given by the Richards. Their cumulative effect produces autocorrelated errors with a variance increasing over time, similar to theoretical expectations and empirical observation.

Measurement/sampling errors are modeled by

$$h_i^* = H(t_i) + \sigma_m \varepsilon_i,$$

(6)

for independently distributed standard normal $\varepsilon_i$.

Any of the parameters $a$, $b$, or $c$, or some specified function of $a$, $b$, and $c$, can be taken as the site-dependent parameter $q$, varying across sample plots (in what follows called “local,” for short), while the others are common to all plots (called “global”). The variance parameters $\sigma$ and $\sigma_m$ may be local or global, although in previous work I have generally made them global (García 1999). There are three other optional parameters associated to the initial conditions (if not used they are fixed at zero): $\sigma_0$ represents additional variability in early growth, before the first plot measurement, and $t_0$ and $H_0'$ are mutually exclusive parameters that allow for varying the curve origin.

**Estimation**

The data consist of a sequence of top height observations $h_i$ at ages $t_i$ for each sample plot (either a PSP or an SA plot). Ages and number of measurements usually vary across plots, and measurements need not be evenly spaced in time. There is a large number of parameters to be estimated: a few global parameters, plus one or more local parameters that are specific to each plot. All these are estimated simultaneously by maximum-likelihood (ML). In fact, the local parameter estimates are not used in the final site index model, although they can provide plot site indices for other purposes. From the point of view of Equation 1, only the global parameters (excluding the variances) are needed; the rest are what are called nuisance parameters.

Model 4 can be analytically integrated and combined with Equation 6 to give the probability of observing a certain sequence of height-age pairs in a plot (García 1983, Seber and Wild 2003). Neglecting possible correlations between measurements in different plots, the product of all these plot probabilities is the probability of observing such a data set, given the model and the parameter values. Considered as a function of the parameters, this is the likelihood function. The ML estimates are the parameter values that maximize this function.

Optimizing a nonlinear function over hundreds or thousands of variables (parameters) is not trivial. Computational efficiency and reliability have been achieved with a customized procedure that exploits the special structure of the problem (García 1983, 1999). In combining PSP and SA data sets we may want to have the same values for the global parameters in Equation 5, but to allow the variance parameters to differ between data sets, reflecting different error characteristics (see below). For this purpose, the original programs were modified to allow some of the global parameters to take different values in different groups of plots, with other parameters forced to be the same. The groups can be indicated by position in the data file or by flags in the plot identifiers. Software and technical details are available from www.unbc.ca/forestry/forestgrowth/sde.

A useful property of ML estimators is their transformation invariance: the ML estimate for any function of the parameters is the same as the function of the parameter ML estimates (Zacks 1971). Intuitively, redefining what is to be estimated does not change the position of the maximum. Therefore, quantities such as site indices and heights calculated from the estimated parameter values are automatically also ML estimates. In particular, this does away with questions of methods for predicting height development not being the best for classifying plots according to site index, and vice versa (e.g., Strand 1964, Curtis et al. 1974, Hägglund 1981).
Model Selection and Hypothesis Testing

Many models based on Equations 4 and 6 can be formulated, differing in which parameters are global, local, or fixed at given values. Furthermore, one may want to examine joint and separate modeling alternatives for groups of data, such as geographical regions or data sources (e.g., SA versus PSP). Models with the same number of parameters can be compared and chosen according to the value of the computed likelihood, or of its logarithm, the log-likelihood. In comparing models with different number of parameters, extra flexibility needs to be penalized. Various theories suggest adding penalties of one-half to three log-likelihood units per parameter (Edwards 1972, Akaike 1973, Stone 1977, Atkinson 1980). Akaike’s information criterion, perhaps the most popular, prescribes one unit. Note that the likelihood value by itself is essentially meaningless; it is the relative values, or differences in the logarithms, that are important.

Hypothesis tests can be based on the computed likelihoods for the models that correspond to the null hypothesis and alternatives. One approach considers as “significant” a log-likelihood difference of about two units, after adjusting for parameter numbers with a half-unit penalty (Edwards 1972). More orthodox is a likelihood-ratio test. It is based on the fact that, under the null hypothesis, twice the log-likelihood difference is asymptotically distributed as a $\chi^2$ with degrees of freedom equal to the difference in number of parameters (Wilks 1962, Seber and Wild 2003).

García (1999) demonstrated the use of these and other statistical inference criteria in evaluating a set of regional site index models. Here I apply these techniques to the comparison of SA and PSP data.

Combining Data Sets

Regardless of comparison results, it may still be necessary to obtain a model using both SA and PSP data. Even if a test indicates inconsistency, a pair of separate SA and PSP models is of little use for future estimation or prediction. The alternatives are to develop a model that somehow combines the information from both sources of data, or to ignore one or the other data set.

One possibility is to simply pool all the data to fit a joint model. I show how this works for geographical regions, and with SA and PSP for comparison purposes. This is too restrictive, however; SA and PSP observations might follow the same overall height-age trends, but their variability would be expected to be different. Specifically, SA is less subject to uncorrelated measurement and sampling errors, so that $\sigma_m$ should be smaller. However, issues such as changes in dominance might contribute variation correlated over time, increasing the value of $\sigma$. Therefore, I propose a combined SA + PSP model where the global parameters in the deterministic component (Equation 5) are the same for both data sources, but $\sigma$ and $\sigma_m$ are allowed to differ.

Data

*Pinus pinaster* is an important plantation species in Galicia, Portugal, and southern France. In Galicia, Echeverría (1948) obtained normal yield tables for Pontevedra province, including site index curves based on mean height, using PSP data and graphical methods. Bara Temes and Toval Hernández (1983) collected SA information from the whole of Galicia, and used it to fit separate height-age curves by site quality classes, and to relate site quality to environmental variables. Later, Rodríguez Soalleiro et al. (1993) (also in Alvarez González et al. 1999) used the same SA data to develop site index models for a coastal and an interior region, through a “parameter prediction” method (Clutter et al. 1983). I use these and additional PSP data to investigate differences between regions and between data sources.

For convenience in the development and use of the models, I consider the provinces of Pontevedra and A Coruña as a coastal region (to be called “Coast”), and Orense and Lugo as interior regions (to be called “Interior”; coastal Lugo was not significantly represented in the data). The boundaries are not far from those between the coastal and interior regions used by Rodríguez Soalleiro et al. (1993), and provincial boundaries would be easier to apply in practice.

The SA data come from the 96 plots of Bará Temes and Toval Hernández (1983), 50 in the Coast and 46 in the Interior. Two “mean dominant” trees were selected in 1974–1975 from plots of approximately 625, 1,000, 1,400, or 1,600 m², plot size being chosen to include at least 60 trees. Rings were counted on cross-sections at every 2 m up the tree, with a final section at a diameter of 7 cm. Total height and age from planting were also recorded. I graphed the data from both trees in each plot to check for consistency and indications of past suppression. Only a few doubtful measurements, and no trees, were rejected. Then both heights and the pairs of ring counts at each cross-section were averaged, discarding the 7-cm counts, which could not be reliably combined. The averaged count from each cross-section was used to calculate the respective age and height, applying Carmean’s correction for height (Dyer and Bailey 1987) [3].

PSP data came from three sources: from Echeverría (1948), from plots established by G. Toval at the Loutrizán Forest Research Center (1975–1978, unpublished), and from a study described by Pita Carpenter and Moreno Pinedo (1973). A height-diameter regression $H = b_0 + b_1 \log D$ was fitted to sample trees to estimate tree heights. Top height was then calculated as the mean height of the trees corresponding to the largest 100 per hectare. All data were plotted, and obviously erroneous measurements were eliminated.

Table 1 shows the numbers of plots and total measurements after data screening. The data are displayed graphically in Figure 1. Clearly, the age coverage of the PSPs is far from ideal, but they can still contribute valuable information.
It is conceivable that stem analysis heights could be made more compatible with a standard top height definition by using the plot diameter distribution information in some way. In addition, it has been shown that top heights calculated by the usual “100 largest per hectare” method vary somewhat with plot size (García 1998, García and Batho 2004); I have not attempted to adjust for this effect here. Top heights might also have been slightly affected by uncertain numbers of trees in some of the plots. Although representative of common practices, this data set is presented more to illustrate the methodology than as an example of the best results that could be achieved with a heavier investment into data checking and more sophisticated data processing.

Results

Models were estimated separately for all the region and data type (SA, PSP) combinations, and also with data pooled across regions and types. Various models were tried: with \( a \) or \( b \) as the site-dependent (local) parameter, and with or without free \( t_0 \) and \( H_0 \) parameters. For all data groupings the anamorphic models, with the asymptote \( a \) as local, had a lower likelihood value than those with \( b \) local.

Within the models with site-dependent \( b \), including a \( t_0 \) or \( H_0 \) parameter produced negligible changes in the log-likelihood for the PSP data. For the SA there were more substantial increases in the log-likelihood: e.g., for the pooled Coast + Interior, 1.8 units with \( t_0 \) and 16.8 units with \( H_0 \). The differences in the height-age curves, however, are practically insignificant, and the fact that the estimated \( H_0 \) were negative would be slightly inconvenient. Therefore, in what follows we restrict attention to models with the parameter \( b \) as site-dependent, and with curves forced through the origin (\( t_0 = H_0 = 0 \)).

The log-likelihood results are given in Table 2. Separate refers to the fitting of separate models, where the log-

![Figure 1. Data. Consecutive measurements in the same plot are joined by lines.](https://academic.oup.com/forestscience/article/51/4/277/4617597)
likelihood equals the sum of the data group values (e.g., 826.9 = 503.3 + 323.6). Pooled is using a common model. Examining first the effect of the Coast/Interior classification, when fitting separate models we find differences of 826.9 - 821.2 = 5.7 and 822.9 - 815.2 = 7.6 log-likelihood units for SA and PSP, respectively, for an additional four parameters (a second set of a, c, σ, and σ_m). The differences of 5.7/4 = 1.4 and 7.6/4 = 1.9 units per parameter would favor either separate or joint models, depending on the various likelihood criteria mentioned above.

A likelihood-ratio test for the null hypothesis of no difference between Coast and Interior would use a χ^2 value of -2 times the logarithm of the likelihood ratio, with 4 degrees of freedom. That is, twice the log-likelihood differences, 11.4 and 15.2. This results in P-values of 0.02 and 0.004 for SA and PSP, respectively (probability of observing values like these or greater, under the null hypothesis). The P-values are low, but not extremely so. Considering the practical advantages of using a single model for the whole of Galicia, I believe pooling the data from the two regions is justified. Of course, we are talking about “shape” of the site index curves; on average, site indices in the Coast are clearly higher than in the Interior.

Turning now to the data types, looking in the Pooled column, it is clear that separate models for SA and PSP fit each data set much better than a common model. There is a log-likelihood difference of (1,636.4 - 1,525.9)/4 = 28 units per additional parameter, or a χ^2 P-value of 8 x 10^{-13}. However, as explained above, because of the different error structure this SA + PSP pooling is not realistic, and it seems more reasonable to consider a combined SA + PSP model where the parameters a and c are the same for SA and PSP, but the stochastic component parameters σ and σ_m may differ.

The estimates for the different models, pooling Coast and Interior, are given in Table 3. For the combined model, the SA and PSP variance parameters are given separated by slashes, in that order. Zero estimates for σ have been observed and discussed before (García 1983, 1999). In addition to the fact that ML estimators are in general not unbiased, both sources of variation are heavily confounded and difficult to separate in most data sets. Nevertheless, the differences in σ and σ_m between SA and PSP are in the expected direction (see above).

The combined model resulted in a log-likelihood of 1,606.0, with two parameters less than the separate models. The difference is 17 units per parameter, somewhat better than the pooled model. The P-value is still about the same, 9 x 10^{-13}.

Therefore, there is compelling statistical support for a difference in the height-age patterns of the SA and PSP data sets. From a practical point of view, however, it is not obvious which data set is the “correct” one, or the best. SA is prone to biases due to interchanges of dominance, to sample trees not representing the stand top height, and possibly to measurement and calculation procedures. In contrast, our PSP data cover much narrower ranges of age, measurement error is higher, and there may be some bias associated to plot size. Both data sets provide useful information. Until more knowledge becomes available, the best option seems to be to use the combined model.

The equation is (cf. Equation 5)

\[ H = 28.26(1 - e^{-b\tau}/0.7677). \]  

The site-dependent parameter b can be obtained from site index S and base age τ (or estimated from any other height-age pair) as

\[ b = -\ln[1 - (S/28.26)^{0.7677}]\tau. \]  

Alternatively, Equation 8 may be substituted into Equation 7,

\[ H = 28.26[1 - (1 - (S/28.26)^{0.7677})^{\tau/0.7677}]^{10.7677}. \]  

Site index curves (base age 25) and the full data are shown in Figure 2.

Some may feel uneasy about my apparent dismissal of P-value verdicts. However, apart from the SA versus PSP issues discussed in the Combining Data Sets section, hypothesis testing is different from decision making, despite the use of language like accept and reject. Although the merits of various approaches are controversial (Barnett 1999), it is fair to say that classical statistics attempts to summarize one part of the information that is considered as “objective.” Its conclusions are conditional on a model, which unavoidably depends much on the researcher’s prejudices. In addition, to make a decision those results must be combined with other (prior) information, and with perceived costs and benefits, either formally, in decision theory (Ferguson 1967), or informally in the mind of the user. I

### Table 3. Global parameter estimates, pooled Coast and Interior

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>c</th>
<th>σ</th>
<th>σ_m</th>
<th>Log-likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>26.39</td>
<td>0.7731</td>
<td>2.063</td>
<td>0.05192</td>
<td>821.2</td>
</tr>
<tr>
<td>PSP</td>
<td>35.19</td>
<td>0.7579</td>
<td>0</td>
<td>0.1350</td>
<td>815.2</td>
</tr>
<tr>
<td>Pooled SA + PSP</td>
<td>27.24</td>
<td>0.7806</td>
<td>0</td>
<td>0.2266</td>
<td>1,525.9</td>
</tr>
<tr>
<td>Combined SA + PSP</td>
<td>28.26</td>
<td>0.7677</td>
<td>2.2020</td>
<td>0.02739/0.1471</td>
<td>1,606.0</td>
</tr>
</tbody>
</table>
take the hypothesis tests as a guide and warning flag, indicating uncertainty, but the single combined model still seems to be the best alternative.

Summary and Conclusions

Enhanced understanding and statistical inference in height growth and site index models require an adequate representation of the error structure. Stochastic differential equations constitute a natural framework for this purpose.

A special case, where the deterministic part coincides with the widely used Richards growth model, has been successfully used in a number of applications. All the parameters can be estimated simultaneously by maximum likelihood, and likelihood-based inference may be used to evaluate model versions and to test various hypotheses.

The approach was extended to deal with common parameters in nested groupings of plots. In particular, the method is useful for evaluating the consistency of stem analysis and permanent sample plot data sources, and for combining the data into one model while allowing for different error properties.

The techniques were applied to Pinus pinaster in Galicia. Statistical tests point to differences between the height-age trends observed in stem analysis and in permanent sample plots. Nevertheless, a combined model seems appropriate for forest management use, pending future investigations. Although data checking and data processing could be improved, the model obtained should compare favorably to the alternatives currently available.

Endnotes

[1] An infinite set of curves whose equations can be obtained from a given equation by varying a constant, here $q$ (James 1992, Kreyszig 1993). Although the family is often described through a differential equation obtained by eliminating $q$ between the original equation and its derivative, that differential equation should not be confused with the one for height growth used below.

[2] Similar differential equation linearizations exist for most growth models, although substantial changes to the current software would be needed. As pointed out by a reviewer, it is of course possible that a model other than Richards’s could fit the data better.

[3] A simpler alternative to the procedures discussed by Dyer and Bailey (1987) has been suggested by Jim Goudie (Nigh 1995). It consists of subtracting 0.5 years from the cross-section age, instead of extrapolating the growing tip location.

Literature Cited


