ELSEVIER

Ecological Modelling 98 (1997) 1–12

ECOLOGICAL MODELLING

Evaluating forest growth models

J.K. Vanclay [a], J.P. Skovsgaard [b,*]

[a] *Center for International Forestry Research, P.O. Box 6596 JKPWB, Jakarta 10065, Indonesia*
[b] *Danish Forest and Landscape Research Institute, Hørsholm Kongevej 11, DK-2970 Hørsholm, Denmark*

Accepted 12 April 1996

Abstract

Effective model evaluation is not a single, simple procedure, but comprises several interrelated steps that cannot be separated from each other or from the purpose and process of model construction. We draw attention to several statistical and graphical procedures that may assist in model calibration and evaluation, with special emphasis on those useful in forest growth modelling. We propose a five-step framework to examine logic and bio-logic, statistical properties, characteristics of errors, residuals, and sensitivity analyses. Empirical evaluations may be made with data used in fitting the model, and with additional data not previously used. We emphasize that the validity of conclusions drawn from all these assessments depends on the validity of assumptions underlying both the model and the evaluation. These principles should be kept in mind throughout model construction and evaluation. © 1997 Elsevier Science B.V.

Keywords: Forest growth modelling; Model evaluation; Validation; Verification; Testing

1. Introduction

Model evaluation is an important part of model building, and some examination of the model should be made at every stage of model design, fitting and implementation. A thorough evaluation of a model involves several steps, including two which are often called verification and validation.

In forest growth modelling, verification and validation usually denote qualitative and quantitative tests of the model, respectively. However, there are some objections to these terms (e.g. Oreskes et al., 1994):

1. They are value-loaded, and it is preferable to use neutral language to assess model performance.
2. The same terms are used in other branches of mathematics and logic to denote other meanings: a model is valid if the logic is correct, and verified if it is 'true'.

* Corresponding author.

PII S0304-3800(96)01932-1

3. Verity implies truth, but it is impossible to prove a model 'true' (except in the special case of a closed system). The only truth that can be established in a growth model is (e.g. in the context of Goulding (1979)) that the model is a faithful representation of what the modeller intended. Similarly, the only sense of validity that can be demonstrated for an empirical model is the 'reasonableness' of the statistical assumptions.

Thus it is appropriate to avoid the terms verification and validation, and to use alternatives. We use the term model evaluation to encompass both these aspects. Thorough model evaluation comprises several steps, each of which may involve qualitative and quantitative aspects. Some steps involve examination of the structure and properties of a model, with or without supplementary data, to confirm that it has no internal inconsistencies and is biologically realistic. Others require comparisons with additional data to quantify the performance of the model, and have become known in some forestry literature as benchmarking (cf. surveyor's reference mark). Ideally, benchmark tests should involve data which are in some sense unlike the data used to fit the model, but useful insights can also be obtained with the calibration data.

These tests cannot prove a model to be 'correct', but may be used in attempts to falsify inferences made from the model. The quality of a model can only be evaluated in relative terms, and its predictive ability always remains open to question. However, the failure of several attempts to falsify a model should increase its credibility and build user confidence. This is the role of model evaluation. Thus, model evaluation should be an on-going procedure which commences during model design and continues throughout model construction and for as long as the model remains in use.

Soares et al. (1995) and Vanclay (1994) recently reviewed ways to evaluate forest growth models. Here, we give a brief overview of the framework they suggest, and offer some new insights. We stress that model evaluation should not be a mere mechanical procedure to examine a model's technical credentials, but should also involve philosophical considerations by modellers and model users.

2. Procedures for evaluating growth models

Model evaluation should try to reveal any errors and deficiencies in the model, in part, by establishing (Vanclay, 1994):

- whether the equations used adequately represent the processes involved;
- if the equations have been combined correctly in the model;
- whether the numerical constants obtained in fitting the model are the 'best' estimates;
- whether the model provides realistic predictions throughout the likely range of application;
- if the model satisfies specified accuracy requirements;
- how sensitive model predictions are to errors in estimated coefficients and input variables.

An evaluation requires more than a decision regarding the acceptability of a model for a defined use. It should provide as much information as possible about the model's behaviour and predictive ability, to allow users to decide if it is adequate for their intended uses. It should also reveal where future data collection and model revision efforts may be most useful.

Evaluation should not be a mere afterthought to model construction, but should be considered at every stage of model design and construction, when component functions are formulated and fitted to data, and when these components are assembled to provide the completed model. Here we deal primarily with regression techniques, but recognise that other approaches may also be used in modelling. Model evaluation includes both theoretical and empirical issues, and is dealt with in standard texts on applied regression analysis (e.g. Gilchrist, 1984; Ratkowsky, 1990). Key aspects may be grouped under several interrelated headings (with some selected examples):

1. Examine the model and its components in terms of logic structure and from theoretical and biological views (e.g. Hamilton, 1990; Oderwald and Hans, 1993; Sievänen and Burk, 1993; Zhang et al., 1993) to see if they are:

- parsimonious;
- biologically realistic;
- consistent with existing theories of forest growth;
- predict sensible responses to management actions.

2. Ascertain the statistical properties of the model in relation to data (e.g. Bates and Watts, 1988; Ratkowsky, 1983; Seber and Wild, 1989), including:
   - nature of the error term (i.e. additive or multiplicative, independence, etc.);
   - estimation properties of parameters in model functions.
3. Characterize errors (e.g. Power, 1993; Reynolds, 1984; Reynolds and Chung, 1986) in terms of:
   - accuracy;
   - nature of residuals (distribution, dependencies on initial stand conditions and length of projection);
   - confidence intervals and critical errors;
   - contributions by each model component to total error.
4. Test, using statistical approaches (e.g. D'Agostino and Stephens, 1986; Gregoire and Reynolds, 1988; Mayer and Butler, 1993; Power, 1993; Reynolds et al., 1988) for:
   - bias and precision of the model and its components;
   - goodness-of-fit of predicted size distributions;
   - patterns in, and distribution of residuals;
   - correlations over time and between components.
5. Conduct sensitivity analyses to determine (e.g. Botkin, 1993; Gertner, 1987; Jørgensen, 1986; Mowrer, 1991; Van Henten and Van Straten, 1991):
   - how model components influence predictions;
   - how inputs to the model influence predictions;
   - how errors propagate through the model.

These analyses need not be sequential, but all relevant aspects should be examined in each model component and in the assembled model. Each of these steps could involve both graphical analyses as well as statistical indices.

2.1. Logical and biological consistency

Each model component and the model as a whole should be logically consistent and biologically realistic. Many model properties can be examined for consistency, e.g. (after Oderwald and Hans, 1993):

1. Do variables included in, and omitted from the model agree with expectations?
2. Do the sign and magnitude of coefficients agree with expectations?
3. Are extrapolations outside the range of the development data reasonable?
4. Are transformations of model predictions reasonable (e.g. do model forecasts of future diameters also provide reasonable estimates of diameter increments, future volumes, mean increment curves, etc.)?
5. Are any contradictions present within the model?
6. Do derivatives, limits, maxima, minima, inflections, etc. agree with expectations?

Although these questions seem appropriate, in some respects they avoid the real issue, namely, what constitutes a 'reasonable expectation'. Clearly, these questions introduce a subjective element, and reflect our previous observation that model quality can only be evaluated in relative terms. Fortunately, these decisions are not without precedent, and most models can be contrasted with empirical studies.

Matrix plots of simulated stand development trajectories showing a range of property–time and property–property relationships (Leary, 1988, 1997) may offer useful insights into model behaviour, and may provide an efficient way to reveal discrepancies in model predictions. Care is required in resolving an apparent discrepancy between model predictions and expectation: it may be the expectations, and not the model, that is wrong!

Parameter estimates and model forecasts should agree with both empirical data and current understanding of growth processes. Experienced foresters and other experts may indicate areas where model predictions are deficient. Several researchers have advocated formalizing this procedure as a Turing test in which experts are asked to

discriminate between simulated and real world data, but this does not provide a good basis for comparison. If the real and simulated data are sufficiently alike to offer a realistic test, they should be amenable to statistical testing which avoids potential difficulties with personal bias. Conversely, if the data are unsuited to statistical testing, it is likely that they will contain certain identifiable features which may make the distinction easy.

Simulations at extremes of stand condition may be particularly revealing. Such simulations may encompass not only the upper and lower limits of site quality and stand density represented in the data, but also alternative stand structures (e.g. even- vs. uneven-aged, pure vs. mixed, thinned vs. unthinned, pruned vs. unpruned, etc.).

Optimization studies may provide a discriminating test of a model, since optimizers seem remarkably efficient at exploiting seemingly minor quirks in models to arrive at unrealistic solutions (e.g. Monserud, 1989). Thus, optimization studies coupled with expert insights may provide a good basis for model evaluation. However, a model should not be rejected simply because it behaves in a counter-intuitive fashion; it may be our preconceptions that are wrong. Thus, discrepancies should cause a critical reappraisal of the model, the data, and of preconceptions.

2.2. Statistical properties

With linear regression models, $\boldsymbol{Y} = \boldsymbol{Xb} + \boldsymbol{e}$, it is usually assumed that the random errors e_i are additive, independent and identically normally distributed with zero mean and constant, but unknown variance ($e_i \sim N(0, \sigma^2)$). Departures from these assumptions may result in parameter estimates with undesirable statistical properties. Several transformations and weighting techniques may be used where data do not satisfy these assumptions, but some problems may remain (e.g. multiplicative errors in models with additive terms that preclude logarithmic transformations).

In forestry applications, several measurements are often taken from each sampling unit (e.g. measurements on a single tree, trees on a plot, or re-measures of a plot). These repeated measurements are not statistically independent, and ordinary least squares techniques may underestimate the variance of parameters, leading to the acceptance of more complex models than would otherwise be indicated. Several suggestions have been given to deal with this problem of longitudinal data (see for example West (1995) and Gregoire et al. (1995)).

Parameter estimates of non-linear growth models may not possess the same desirable statistical properties as their linear counterparts (i.e. unbiased, normally-distributed, minimum variance estimators). However, non-linear models which are 'close-to-linear' approach these properties asymptotically, and many models may be reparameterized so that they behave in a close-to-linear fashion (Ratkowsky, 1983, 1990).

In the standard regression model, the explanatory variables are assumed to be free of error. This assumption is rarely tenable in forest growth models, where there is joint variation in the variates, and this means that derived relationships could be grossly in error. It is possible to correct for this (e.g. Seber and Wild, 1989; Weisberg, 1985), but the procedures may be tedious. Failure to account for the nature of the response variable will lead to inflated estimates of variance, but the effect can be minimized by ensuring a large range of each explanatory variable relative to its error.

Most forest growth models are constructed from several equations independently fitted to data. Simultaneous estimation of all model components minimizes overall model errors and provides a variance–covariance matrix for the model as a whole (e.g. Gallant, 1987; Seber and Wild, 1989), but few forest growth models have been constructed in this way (e.g. Furnival and Wilson, 1971; García, 1984; Leary, 1970).

The standard regression assumptions are ideals that real situations (models and data in conjunction) may approach without ever exactly attaining. Fortunately, least-squares techniques tend to be relatively robust in practice (at least for parameter estimation, if not for assessing precision). Irrespective of this, evaluation of a model, before and after fitting to data, should include the appraisal of the statistical properties of the model and the data.

2.3. Characterizing model error

One of the most efficient ways to examine model performance is to plot residuals or standardized residuals for all possible combinations of tree and stand variables to detect possible autocorrelation and other dependencies. Such plots may be interpreted visually, but formal tests are also available (e.g. Draper and Smith, 1981; Weisberg, 1985).

Two simple criteria, in conjunction, provide a summary of the overall model performance: average model bias ($\Sigma(y_i - \hat{y}_i)/N$) and mean absolute difference ($\Sigma|y_i - \hat{y}_i|/N$) (e.g. Burk, 1986). Average model bias measures the expected error when several observations are to be combined by totalling or averaging, and mean absolute difference measures the average error associated with a single prediction. Error dependencies on projection length or initial forest condition can be shown graphically. Regression analysis and principal component analysis may help to detect possible dependencies. These techniques apply equally when checking the model against data used for model calibration, and when testing the model with additional data.

The error structure and the contribution of each model component to total error may be more revealing than a mere evaluation of total model performance. Thus, a map of variance components of the model may help to identify weaknesses and define priorities for future research (e.g. Hann, 1980; Gertner et al., 1995).

2.4. Statistical tests

Many statistical tests of model performance have been suggested, but no single criterion can incorporate all aspects of model evaluation, and it is desirable to use several simple tests to examine different facets of model behaviour.

One simple but efficient technique is based on linear regression of observed vs. predicted data. Some useful insights into the quality of predictions may be given by R^2 and the slope and intercept of the fitted line, and a good test for bias is the simultaneous F-test for slope = 1 and intercept = 0 (e.g. Dent and Blackie, 1979; Mayer and Butler, 1993; Mayer et al., 1994).

Another useful technique is to compare predictions directly with observed data using a statistic analogous to R^2, and sometimes called modelling efficiency:

$$\mathrm{EF} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

This statistic provides a simple index of performance on a relative scale, where 1 indicates a 'perfect' fit, 0 reveals that the model is no better than a simple average, and negative values indicate a poor model indeed.

In addition to overall appraisals, it is desirable to partition data (e.g., by age, site index or stand density), and examine model performance in each of several strata (e.g. Mayer and Butler, 1993). The most revealing insights may be obtained by devising strata based on a knowledge of the biological system, as well as model and data characteristics. However, the absence of inadequacies in any particular stratification does not imply that weaknesses will not be found in an alternative stratification. Nor does consistency with empirical data confirm the quality of a model, since several alternative model formulations may have equal merit (e.g. Kincaid, 1996).

2.5. Sensitivity analyses

A sensitivity analysis should reveal how model predictions depend upon inputs, parameters, relationships and submodels. Commonly, sensitivity analyses focus on parameters which, when perturbed, cause the greatest fluctuations in model predictions. These studies may reveal model components with low and high sensitivity, both of which are of interest. Insensitive components may contribute little towards model predictions and could be targets for omission from the model during model revisions. Conversely, it is useful to know about model components with high sensitivity, because these may have the greatest impact on model predictions. All model parameters and inputs should be estimated accurately, but particular care is required with the most sensitive variables.

In theory, the sensitivity of model parameters can be examined analytically (e.g. by taking derivatives), but in practice this may be complicated by the interaction of various model components and feedback loops. Thus sensitivity analyses are often carried out as simulation studies in which the parameters or components are changed to observe corresponding effect on predicted outputs. In practice, meaningful sensitivity studies are difficult, as the estimate of sensitivity depends both on the values of the inputs and the model parameters, so that many simulations may be necessary to complete the picture. This may be a tedious undertaking, especially where there are many parameters. Results of sensitivity tests may reveal parameters critical to model predictions, and parameters which may be redundant. Knowledge of sensitive parameters may guide applications (especially extrapolations) and the planning of model enhancements.

The implications of functional relationships and submodels should also be examined. The decision to use a particular relationship (e.g., in process models, the decision to use the Michaelis–Menten rather than the exponential equation) may have an influence on model outcomes, and the implications of such decisions for model predictions should be examined.

Similarly, it is important that users have a knowledge of the model's sensitivity to inputs. Studies of error propagation (Gertner, 1987; Mowrer, 1991) may reveal model limitations, and are particularly useful in offering insights into the interaction of errors in the input data and in the simulation, but can only be used when the model under consideration is completely defined by a set of empirical equations.

One application of stochastic simulation studies is to investigate the 'quality' of predictions. Variance approximation provides an efficient alternative to such studies, and enables the variance of predictions to be estimated deterministically. It also enables the variance of the input data to be incorporated into the analysis. Mowrer and Frayer (1986) and Gertner (1987) used a simple first-order Taylor series to estimate the errors propagated through growth and yield projections.

3. Empirical data for model evaluation

Several aspects of model evaluation relate to comparisons with empirical data, and these comparisons may be more rigorous when made with data not used in fitting the model (benchmark tests). Thus it has become customary in the evaluation of forest growth models to reserve some data to provide an 'independent' benchmark test of the model (e.g. Snee, 1977; West, 1981; Shifley, 1987).

This raises several questions about the merits of setting data aside for such tests, about the nature and amount of data used for such comparisons, and about the nature of the population of interest. In effect, this involves a compromise between the best possible parameter estimates (using all the data for calibration) and the best possible overall impression of precision (reserving some data for testing). Two options seem to offer the best of both worlds:

1. Fit the model using some data, test it against the remainder, and then recalibrate using the full data set.
2. Use re-sampling techniques such as cross-validation.

3.1. Partitioning data

The most rigorous test of a model requires independent data, ideally from controlled and replicated trials measured over a long period. Unfortunately, such data may not be available, and the only 'independent' data readily available may relate to other regions or species. These, however, may not reflect the population of interest, and it is not always clear how to interpret the results of tests with such data. Thus, growth modellers often have to decide whether it is worthwhile splitting data into two subsets, one for development, and the other for testing the model. This is not a trivial decision, especially when data are scarce. Setting some data aside may provide for a better test of the model, but may result in inferior parameter estimates.

The role of an independent or benchmark test cannot be divorced from the nature of the model. If the model fitting exercise is intended to reveal

possible causal parameters, then the costs of benchmarking may be greater than the benefits (e.g. in medical epidemiology, Hirsch, 1991). Partitioning data to allow benchmarking may help to reduce type I errors (i.e. falsely rejecting the null hypothesis, and thus e.g. including irrelevant variables in the model), but fewer data for calibration mean a reduction in the precision of parameter estimates, and an increase in type II errors (i.e. falsely accepting the null hypothesis, and thus, e.g. incorrectly concluding that a variable contributes little and should be omitted from the model). However, if empirical data are used to calibrate a model deliberately formulated to represent biological processes, then the goal is a different one: namely to accurately estimate parameters rather than to identify possible explanatory variables. In this latter case, benchmark data may serve a more useful role in illustrating the robustness of the model. Clearly, an assessment of the utility of independent benchmark data cannot be divorced from the purpose of the model.

If a decision is made to partition a data set, the modeller must avoid the temptation to weaken the tests, for example, by reducing the number of data available for testing, despite a desire to find the model acceptable. The outcome of benchmark tests can be influenced by the selection of data: 'like' data will provide a more optimistic result than comparisons with 'unlike' data from another population. Thus, the most convincing demonstration of model quality can be made only if the test data are in some sense unlike the development data. A single sample split into two parts is no substitute for test data from controlled, replicated trials. Vanclay (1994) discussed the dangers of constructing a growth model from passive monitoring data in which stand density and site productivity are confounded. Splitting such data into calibration and benchmark sets may not reveal the fallacy of a positive correlation between stand density and tree growth; this can only be refuted (empirically) using data from thinning and spacing trials (depending somewhat on how the data are divided).

Unfortunately, the ideal, a series of properly replicated trials, is rarely available. However, data which are independent spatially (e.g. different location or site), silviculturally (e.g. different management regime), temporally (e.g. more recent), or logistically (e.g. collected by a different agency) may provide a convincing test if they can be reserved without compromising the range of site and stand conditions represented in the model. Plots established for long periods with regular remeasurement, particularly those remaining undisturbed (i.e. no thinning), may prove useful as a discriminating test. Objective procedures (e.g. Snee, 1977) may be used to select benchmark data to minimize the dangers of bias. Following testing, the benchmark and calibration data may be pooled and the model recalibrated to obtain the best parameter estimates.

One possible frustration with benchmarking may arise when the initial calibration of the model seems inadequate in benchmark trials, since there is no way to test if recalibration using the pooled data will result in an improvement. The change in parameter estimates may serve as a guide (and may even serve as a good benchmark criterion, see for example Sievänen and Burk, 1993), but do not reveal if the recalibration is 'adequate'. However, if the model is the best that can be obtained with existing resources, it must be considered acceptable, even if inadequate in some sense, since there is no alternative other than to invest more resources and wait for new data and techniques. Perhaps the real test of a model (in a practical sense, if not in an epistemological sense) is if forest managers have sufficient confidence in it to use it as the basis for management decisions.

3.2. Resampling procedures

An efficient alternative to independent data is to mimic these tests with resampling techniques such as cross-validation, boot-strapping and jackknifing (e.g. Efron and Gong, 1983; Weisberg, 1985). Cross-validation is the logical generalization of partitioning the data for model calibration and benchmarking (e.g. Burk, 1990). Rather than omitting some data, each datum is deleted in turn and the model is fitted to the remaining $n-1$ data. Benchmark tests are averaged from the individual deleted data. If the test statistic is squared

error and the model is linear, the cross-validation estimate of true error is n times the PRESS statistic computed by many regression packages. A variation on these single-observation resampling procedures is to omit groups of data, for example, according to geographic location, management strategy, or other criteria (e.g. Tarp-Johansen et al., 1997).

One shortcoming of any resampling procedure lies in its dependence on the data. The sample should adequately represent the variability and other characteristics of the population of interest, or the resampling procedure will not provide an adequate test of the model. Unfortunately, these are the very circumstances under which the model itself should come under heaviest criticism (Burk, 1990).

Despite the efficiency of re-sampling procedures, it seems impossible to avoid the use of some benchmark data, since resampling to test a complete model involving many relationships and assumptions seems impractical.

4. Other considerations

A technical appraisal of a model does not constitute a complete evaluation. There are several other important qualitative aspects which should also be considered. Many of these have already been considered, at least in part, but it remains important to re-examine several aspects:

1. Does the model satisfy the needs of clients?
2. Are underlying concepts sound, and visible to users (in the model or documentation)?
3. Have concepts been implemented faithfully, unconstrained by resources or technology (e.g. has the IF ... THEN ... ELSE ... ENDIF structure of the computer language led to the use of on-off behaviour rather than a gradual phasing in and out, even though the latter may be more appropriate)?
4. Is the model parsimonious, satisfying the general principle of science that *entities should not be multiplied beyond necessity*? (The principle is known as Ockham's razor; see Keuzenkamp and McAleer, 1995, for a recent review).

Some of these aspects have been explored more thoroughly in the social sciences where it is more difficult to obtain quantitative benchmark data than in the natural sciences. Thus, it is interesting to explore some experiences of that discipline. In a review of several models for social policy analysis, Meadows and Robinson (1985) observed that "tests tend to be weak, marginal, unsymmetrical and very biased. In part this is due to oversized models whose complete testing would be impossibly expensive and tedious. It is also due to a general lack of imagination, motivation, training, client pressure and agreed-upon methods for testing". Although this criticism was levelled specifically at the social sciences, it also applies to some extent, in forest growth modelling.

Meadows and Robinson (1985) collated specific advice to overcome these limitations, including (and followed by our responses):

1. Modellers should think more and wield tools less (Majone, 1977)—the 'new toy' syndrome is a hazard that is also prevalent in forest growth modelling.
2. Models should be given to an independent evaluation agency for testing (Quade and Boucher, 1968)—several independent evaluations have been published in refereed journals (e.g. Reynolds, 1984; Oderwald and Hans, 1993; Soares et al., 1995).
3. Modellers should test each part of their model, not just the summary output (Biggs and Cawthorns, 1962)—this may be tedious and time-consuming, but is important to gain a good insight into the model (e.g. Hann, 1980; Soares et al., 1995).
4. Modellers should test their results against the real world, rather than against a set of artificial rules or formulas (Brewer, 1973)—this seems to be one thing that is done well on the rare occasions that forest growth models are thoroughly benchmarked (e.g. Hann, 1980; Reynolds, 1984; Soares et al., 1995; West, 1981).

It is disconcerting to reflect that this advice remains as necessary, and as rarely applied today, as when it was first offered. Meadows and Robinson (1985) concluded with a warning that "modelling efforts often succumb to a slow ...

drift ... away from what is important to what is ... tractable, away from unconventional viewpoints and toward established wisdom. At each little decision point ... the guiding question should be 'would it help solve the problem?'. ...[T]he criterion for decision should always be what will most help real-world decisions, not what the modeller will find easy or fun, or what the client will find ... uncontroversial." A decade later, this warning remains timely and pertinent.

Some readers may find our stance too idealistic, but while we accept that modelling may be constrained by knowledge, data and resources, we echo the sentiments of Ziman (1978): "[one] learns how easy it is to persuade oneself of the validity of a model which later turns out to be false, and comes to realize that even in very strongly mathematical and well-defined scientific issues, it may take a long time, much criticism and the death of many promising conjectures before a reliable theory is [established]".

5. Practical relevance

It is appropriate to conclude by exploring the practical relevance of model evaluation with some case studies. We tried to find documented instances of well-tested models that nonetheless went wrong and of untested models that revealed useful insights, but found this difficult, probably because modellers rarely admit model weaknesses or failures, and usually publish material relating to model development rather than model applications. However, some insights can be offered.

The JABOWA model (Botkin et al., 1972; Botkin, 1993) has been remarkably successful in many respects. It laid the foundations for the 'gap-phase' modelling approach; has been used extensively in academia, research and teaching; has been re-calibrated for many different forests ranging from the tropics to the boreal zone; provided the basis for several models derived more-or-less directly from JABOWA (e.g. CLIMACS, FIRESUM, FORET, FORENA, LINKAGES, SILVA, ZELIG; e.g. Liu and Ashton, 1995); and has been cited in the formal literature more than 160 times (Botkin, 1993). Despite this prominence, the model and its derivatives have apparently not been used in operational forest management or planning. In his attempt to adapt the JABOWA model for British woodlands, Spilsbury (1991) noted that JABOWA had several serious deficiencies involving growth patterns, calibration procedures, and performance in empirical tests, especially relating to the diameter frequency distribution.

Another relatively prominent model, the STEMS model and its derivatives (Leary, 1979; Belcher et al., 1982), has both been formally evaluated (e.g. Holdaway and Brand, 1983, 1986; Brand and Holdaway, 1989) and used operationally by forest managers. However, the operational use of this model may be as much a result of packaging, marketing and institutional affiliation, as of performance testing, since several deficiencies are apparent in the model. By studying predicted trajectories in a Bakuzis' matrix, Leary (1997) found several shortcomings in STEMS projections of the growth of red pine in pure even-aged stands. Specifically, there was little dependence of any stand property on site, and mortality rates, height growth and stand basal area development appeared questionable. And following an empirical benchmark study, Brand and Holdaway (1989) recommended 'cautious use of STEMS85 and TWIGS for:

- stands with high basal area;
- lowland hardwood and northern hardwood stands in western lower Michigan;
- stands with many trees slightly smaller than sawtimber size.'

Management acceptance of a model and its predictions may rest on many factors other than formal evaluations of model performance. The management response to a series of yield forecasts for north Queensland rainforests (Vanclay, 1996) depended not on formal test results (there were no such tests), but on subjective evaluations, personalities, the difficulty of implementation, and on politics and economics. This may reflect on the immaturity of modelling, and the poor linkage between modelling and management. Most managers have little experience of modelling, don't know what kind of model evaluation to expect, and have no real basis for appraising models (see

for example Brand and Holdaway, 1983). Thus, modellers need to be more proactive in discussing their work with forest managers and other model users. This lucid communication may be especially important if the model is being used to make inferences about the sustainability of forest practices (e.g. Moir and Mowrer, 1995).

6. Synthesis of evaluation procedures

These few simple suggestions are not intended as a comprehensive review of model evaluation procedures, but merely highlight some important and sometimes overlooked aspects. We stress that evaluation is not one simple procedure, but consists of a number of interrelated steps that cannot be separated from each other or from model construction. Our five-point checklist urges modellers to examine:

1. logic and bio-logic;
2. statistical properties;
3. characteristics of errors;
4. residuals;
5. sensitivity analyses.

Several statistical tests, as well as graphical procedures, may be useful, both with data used for model calibration and with data used for 'independent' evaluation of the model. However, the validity of conclusions depends on the validity of assumptions and the application in question. These principles should be kept in mind throughout model construction and evaluation.

Acknowledgements

Tom Burk, Oscar García, Jerry Leech, Naomi Oreskes and Stanley Wood provided helpful suggestions and thought-provoking comments on the draft manuscript.

References

Bates, D.M. and Watts, D.G., 1988. Nonlinear Regression Analysis and its Applications. Wiley, New York, xiv + 365 pp.

Belcher, D.W., Holdaway, M.R. and Brand, G.J., 1982. A description of STEMS: the stand and tree evaluation and modelling system. USDA For. Serv. Gen. Tech. Rep., NC-79, 18 pp.

Biggs, A.G. and Cawthorns, A.R., 1962, quoted in: P.W. House and J. McLeod (Editors), Large-Scale Models for Policy Evaluation. Wiley, New York, p. 73.

Botkin, D.B., 1993. Forest Dynamics: An Ecological Model. Oxford University Press, Oxford, xv + 309 pp.

Botkin, D.B., Janak, J.F. and Wallis, J., 1972. Some ecological consequences of a computer model of forest growth. J. Ecology, 60: 849–872.

Brand, G.J. and Holdaway, M.R., 1983. Users need performance information to evaluate models. J. For., 81: 235–237, 254.

Brand, G.J. and Holdaway, M.R., 1989. Assessing the accuracy of TWIGS and STEMS85 volume predictions: a new approach. Northern J. Appl. For., 6: 109–114.

Brewer, G.D., 1973. Politicians, Bureaucrats and the Consultant. Basic Books, New York.

Burk, T.E., 1986. Growth and yield model validation: Have you ever met one that you liked? In: A. Allen and T.C. Cooney (Editors), Data Management Issues in Forestry. Forests Resources Systems Institute, Florence, AL, pp. 35–39.

Burk, T.E., 1990. Prediction error evaluation: preliminary results. In: L.C. Wensel and G.S. Biging (Editors), Forest Simulation Systems: Proc. IUFRO Conf., 2–5 Nov. 1988. University of California, Division of Agriculture and National Research, Bulletin, 1927, pp. 81–88.

D'Agostino, R.B. and Stephens, M.A. (Editors), 1986. Goodness-of-fit Techniques. Marcel Dekker, New York, xviii + 560 pp.

Dent, J.B. and Blackie, M.J., 1979. Systems Simulation in Agriculture. Applied Science, London.

Draper, N.R. and Smith, H., 1981. Applied Regression Analysis. Wiley, New York, 709 pp.

Efron, B. and Gong, G., 1983. A leisurely look at the bootstrap, the jackknife and cross-validation. Am. Stat., 37: 36–48.

Furnival, G.M. and Wilson, R.W., 1971. Systems of equations for predicting forest growth and yield. In: G.P. Patil, E.C. Pielou and W.E. Walters (Editors), Statistical Ecology, Vol. 3. Pennsylvania State University Press, pp. 43–57.

Gallant, R.A., 1987. Nonlinear Statistical Models. Wiley, New York, xii + 610 pp.

García, O., 1984. New class of growth models for even-aged stands: Pinus radiata in Golden Downs Forest. N.Z. J. For. Sci., 14: 65–88.

Gertner, G., 1987. Approximating precision in simulation projections: an efficient alternative to Monte Carlo methods. For. Sci., 33: 230–239.

Gertner, G.Z., Cao, X. and Zhu, H. 1995. A quality assessment of a Weibull based growth projection system. For. Ecol. Manage., 71: 235–250.

Gilchrist, W., 1984. Statistical Modelling. Wiley, Chichester, xv + 339 pp.

Goulding, C.J., 1979. Validation of growth models used in forest management. N.Z. J. For., 24: 108–124.

Gregoire, T.G. and Reynolds, M.R., 1988. Accuracy testing and estimation alternatives. For. Sci., 34: 302–320.

Gregoire, T.G., Schabenberger, O. and Barrett, J.P., 1995. Linear modelling of irregularly spaced, unbalanced, longitudinal data from permanent-plot measurements. Can. J. For. Res., 25: 137–156.

Hamilton, D.A., 1990. Extending the range of applicability of an individual tree model. Can. J. For. Res., 20: 1212–1218.

Hann, D.W., 1980. Development and evaluation of an even- and uneven-aged ponderosa pine/Arizone fescue stand simulator. USDA For. Serv. Res. Pap., INT-267, 95 pp.

Hirsch, R.P., 1991. Validation samples. Biometrics, 47: 1193–1194.

Holdaway, M.R. and Brand, G.J., 1983. An evaluation of the STEMS tree growth projection system. USDA For. Serv. Res. Pap., NC-234, 20 pp.

Holdaway, M.R. and Brand, G.J., 1986. An evaluation of Lake States STEMS85. USDA For. Serv. Res. Pap., NC-269, 10 pp.

Jørgensen, S.E., 1986. Fundamentals of Ecological Modelling. Elsevier, Amsterdam, 389 pp.

Keuzenkamp, H.A. and McAleer, M., 1995. Simplicity, scientific inference and econometric modelling. Econ. J., 105: 1–21.

Kincaid, H., 1996. Philosophical Foundations of the Social Sciences: Analyzing Controversies in Social Research. Cambridge, New York.

Leary, R.A., 1970. Systems identification principles in studies of forest dynamics. USDA For. Serv. Res. Pap., NC-45, 38 pp.

Leary, R.A., 1979. Design. In: A generalized forest growth projection system applied to the lake states region. USDA For. Serv. Gen. Tech. Rep., NC-49: 5–15.

Leary, R.A., 1988. Some factors that will affect the next generation of forest growth models. In: A.R. Ek, S.R. Shifley and T.E. Burk (Editors), Forest Growth Modeling and Prediction. Proc. IUFRO Conf., 24–28 Aug. 1987, Minneapolis, MN. USDA For. Serv., Gen. Tech. Rep., NC-120: 22–32.

Leary, R.A., 1997. Testing models of red pine plantation dynamics using a modified Bakuzis matrix of stand properties. Ecol. Model., 98: 35–46.

Liu, J. and Ashton, P.S., 1995. Individual-based simulation models for forest succession and management. For. Ecol. Manage., 73: 157–175.

Majone, G., 1977. Pitfalls of analysis and analysis of pitfalls. Urban Ana., 4: 235.

Mayer, D.G. and Butler, D.G., 1993. Statistical validation. Ecol. Model., 68: 21–32.

Mayer, D.G., Stuart, M.A. and Swain, A.J., 1994. Regression of real world data on model output: an appropriate overall test of validity. Agric. Syst., 45: 93–104.

Meadows, D.H. and Robinson, J.M., 1985. The Electronic Oracle: Computer Models and Social Decisions. Wiley, Chichester, xv + 445 pp.

Moir, W.H. and Mowrer, H.T., 1995. Unsustainability. For. Ecol. Manage., 73: 239–248.

Monserud, R.A., 1989. Optimizing single-tree simulators. In: H.E. Burkhart, M. Rauscher and K. Johann (Editors), Artificial intelligence and growth models for forest management decisions. Proceedings IUFRO meeting, Vienna, 18–22 Sept 1989. VPI and SU, Blacksburg VA, FWS-1-89, pp. 308–321.

Mowrer, H.T., 1991. Estimating components of propagated variance in growth simulation model projections. Can. J. For. Res., 21: 379–386.

Mowrer, H.T. and Frayer, W.E., 1986. Variance propagation in growth and yield projections. Can. J. For. Res., 16: 1196–1200.

Oderwald, R.G. and Hans, R.P., 1993. Corroborating models with model properties. For. Ecol. Manage., 62: 271–283.

Oreskes, N., Shrader-Frechette, K. and Belitz, K., 1994. Verification, validation and confirmation of numerical models in the earth sciences. Science, 263: 641–645.

Power, M., 1993. The predictive validation of ecological and environmental models. Ecol. Model., 68: 33–50.

Quade, E.S. and Boucher, W.I. (Editors), 1968. Systems Analysis and Policy Planning. Elsevier, New York.

Ratkowsky, D.A., 1983. Nonlinear Regression Modeling. Marcel Dekker, New York, viii + 276 pp.

Ratkowsky, D.A., 1990. Handbook of Nonlinear Regression Models. Marcel Dekker, New York, ix + 241 pp.

Reynolds, M.R., 1984. Estimating the error in model predictions. For. Sci., 30: 454–469.

Reynolds, M.R. and Chung, J., 1986. Regression methodology for estimating model prediction error. Can. J. For. Res., 16: 931–938.

Reynolds, M.R., Burk, T.E. and Huang, W., 1988. Goodness-of-fit tests and model selection procedures for diameter distribution models. For. Sci., 34: 373–399.

Seber, G.A.F. and Wild, C.J., 1989. Nonlinear Regression. Wiley, New York, xx + 768 pp.

Shifley, S.R., 1987. A generalized system of models forecasting Central States growth. USDA For. Serv. Res. Pap., NC-279, 10 pp.

Sievänen, R. and Burk, T.E., 1993. Adjusting a process-based growth model for varying site conditions through parameter estimation. Can. J. For. Res., 23: 1837–1851.

Snee, R.D., 1977. Validation of regression models: methods and examples. Technometrics, 19: 415–428.

Soares, P., Tomé, M., Skovsgaard, J.P. and Vanclay, J.K., 1995. Validating growth models for forest management using continuous forest inventory data. For. Ecol. Manage., 71: 251–266.

Spilsbury, M.J., 1991. Computer modelling of mixed age, polyspecific broadleaf woodland in the United Kingdom. D. Phil. Thesis, University of Oxford, 215 pp.

Tarp-Johansen, M.J., Skovsgaard, J.P., Madsen, S.F., Johannsen, V.K. and Skovgaard, I., 1997. Compatible stem taper and stem volume functions for oak in Denmark. Ann. Sci. For., 54: in press.

Vanclay, J.K., 1994. Modelling Forest Growth and Yield: Applications to Mixed Tropical Forests. CAB International, Wallingford, UK, xvii + 312 pp.

Vanclay, J.K., 1996. Lessons from the Queensland rainforests: steps towards sustainability. J. Sustainable For., 3(2/3): 1–27.

Van Henten, E.J. and Van Straten, G., 1991. Sensitivity analysis of a dynamic growth model of lettuce. J. Agric. Eng. Res., 59: 19–31.

Weisberg, S., 1985. Applied Linear Regression, 2nd edn. Wiley, New York, xiv + 324 pp.

West, P.W., 1981. Simulation of diameter growth and mortality in regrowth eucalypt forest of southern Tasmania. For. Sci., 27: 603–616.

West, P.W., 1995. Application of regression analysis to inventory data with measurements on successive occasions. For. Ecol. Manage., 71: 227–234.

Ziman, J., 1978. Reliable Knowledge: An Exploration of the Grounds for Belief in Science. Cambridge University Press, Cambridge, 197 pp.

Zhang, L., Moore, J.A. and Newberr, J.D., 1993. Disaggregating stand volume growth to individual trees. For. Sci., 39: 295–308.