MODEL EVALUATION: FROM MODEL COMPONENTS TO SUSTAINABLE FOREST MANAGEMENT INDICATORS

Validación de Modelos: de los componentes del modelo a los indicadores de gestión forestal sostenible

Paula Soares & Margarida Tomé

Centro Estudos Florestais. Instituto Superior de Agronomia. Universidade Técnica Lisboa Tapada da Ajuda. 1349 017-LISBOA (Portugal). Correo electrónico: paulasoares@isa.utl.pt; magatome@isa.utl.pt

Abstract

Model evaluation is a basic procedure during model development or when selecting the "best available" model for a specific application. The evaluation process should involve several procedures, including qualitative as well as quantitative examinations of the model. Despite the fact that there are no standardized rules for model evaluation, a methodology is presented. The evaluation of different types of conceptual models should consider the particularities of these models. In this work, evaluation of empirical and process-based models is discussed. The use of such models at the landscape level implies their implementation in a Decision Support System (DSS). The evaluation of the output scenarios from the DSS is also discussed.

Keywords: *Model evaluation, Qualitative and quantitative procedures, Empirical models, Process-based models, Implementation of models at landscape level*

Resumen

La validación y evaluación de un modelo es un procedimiento básico tanto durante la fase de construcción del mismo como en el momento de seleccionar el mejor modelo disponible para una aplicación específica. La validación de un modelo comprende distintos procesos, incluyendo análisis cuantitativos y cualitativos del modelo. A pesar de que no existen reglas estandarizadas para la validación del modelo, se proponen diferentes metodologías que deben considerarse y seleccionarse de acuerdo a la tipología y particularidades del modelo a validar. En el presente trabajo se discute la validación de modelos empíricos y de modelos de proceso, así como la implementación de los mismos en herramientas de apoyo a la decisión (DSS) y la validación de los escenarios de salida derivados de estas herramientas.

Palabras clave: *Validación, Procedimientos cuantitativos, Procedimientos cualitativos, Modelo empírico, Modelo de procesos, Implementación del modelo a escala paisaje*

ISSN: 1575-2410

INTRODUCTION

Defining models as abstractions of reality, normally represented by systems of mathematical equations implemented in a computer program, makes it obvious that model evaluation is essential to a model's increased credibility and to ensure that model predictions reflect the most likely outcome of the reality. However, PRISLEY & MORTINER (2004) found some literature that minimizes the importance of the evaluation step, defending that evaluation "entirely depends on the purpose for which the model is intended" (RYKIEL, 1996). To us, model evaluation is, as much as an aid in choosing the "best available" model for one specific application, a crucial part of the process of model development itself. Of course, models can and should be evaluated in relation to specific applications but there are a series of evaluation procedures that do not depend on any specific application.

TERMINOLOGY AND CONCEPTS

The work of PRISLEY & MORTINER (2004) presents an exhaustive list of published definitions of model validation, verification and calibration. From that list we distinguish a definition by VANCLAY & SKOVSGAARD (1997): "In forest growth modelling, verification and validation usually denote qualitative and quantitative tests of the model, respectively". Validation is identified as the "test of a model by comparing model results with observations not used to develop the model" and verification as the "process to demonstrate that the modelling formalism is correct" (HELMS, 1998 and RYKIEL, 1996, respectively). As other authors, we defend the use of the term "evaluation" to designate the process that includes qualitative as well as quantitative examinations of the model. VANCLAY & SKOVSGAARD (1997) emphasize that "model evaluation should be an ongoing procedure which commences during model design and continues throughout model construction and for as long as the model remains in use".

METHODOLOGIES TO EVALUATE MODELS

As a model is a set of submodels, fitted separately or simultaneously, model evaluation should focus on the analyses of the model components as well as on the overall model. Aspects such as spatial and temporal scales, complexity –as defined by the number of variables and processes included– and also applicability of the model must be considered.

There are no standardized rules for model evaluation. Here we present a list of several procedures that we consider essential for such evaluation.

Qualitative evaluation

Qualitative evaluation of a model does not use real data and is aimed to verify the logical consistency and biological realism of the model. LOEHLE (1997) wrote: "It is not sufficient that a model fits field data if it does so by employing biologically unreasonable behaviours or processes". The qualitative evaluation involves (ODERWARLD & HANS, 1993) the analysis of:

1. Consistency of the model with up-to-date knowledge of forestry growth theories
2. Consistency of the relationships between submodels of a model
3. Variables included in and omitted from the submodels
4. Signs and values of coefficients in the submodels, namely asymptotes
5. Location of inflexion points (values of the asymptotes can be obtained in literature and location of inflexion points can be verified if measurements of young stands are available)
6. Agreement of the outputs of the model with results from designed experiments, for instance where the modeler should examine the location of maximum mean annual increment in volume for different site indices and spacings
7. Extrapolation of the model outside the range of fitting data
8. Invariance for projection length

LEARY (1988) suggests the use of a matrix (after BAKUZIS, 1969) to evaluate stand property-time and property-property relations, by displaying all possible combinations of model variables.

Quantitative evaluation

Quantitative evaluation of a model requires real data and compares similarity between model results to observations. Data used to evaluate a model should be independent of the modelling data but, due to the lack of such data, the modelling dataset is commonly split into two data subsets of different percentages (e.g. 50-50%, 75-25%) (e.g. SNEE, 1977). One is used to fit the model and the other to evaluate it; the final model is recalibrated using the total dataset. However, two difficulties in this methodology can be pointed out (SNEE, 1977; VANCLAY & SKOVSGAARD, 1997):

1. How to split data into two subsets when criteria are not always clear? In most of the applications, data splitting is randomly based, resulting in data subsets with similar characteristics. As a consequence, a model is good for both fitting and evaluation datasets.
2. How to define the compromise between data splitting and the loss of quality in parameter estimates, especially when data are scarce?

KOZAK & KOZAK (2003) showed that evaluation based on data splitting provides little, if any, additional information because data subsets are not independent and they present the same statistical structure. An alternative is to use resampling techniques such as cross-validation (e.g. EFRON & GONG, 1983; JONES & CARBERRY, 1994). Cross-validation is the logical generalization of partitioning the data for model calibration and benchmarking (e.g. VANCLAY & SKOVSGAARD, 1997). 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. Jackknifing and bootstrapping are common techniques. A particular case is the statistic based on the PRESS residuals (e.g. MYERS 1986). This entails omitting, in turn, each observation (y_i) from the data, fitting the model to the remaining observations, predicting the response for the omitted observation ($\hat{y}_{i,-i}$) and comparing the prediction with the observed value ($e_{i,-i}$):

$$y_i - \hat{y}_{i,-i} = e_{i,-i} (i=1,2\ldots n)$$

The PRESS residuals are true prediction errors with $y_{i,-i}$ being independent of y_i. Each (sub)model has *n* PRESS residuals associated with it, and the PRESS (PREdiction Sum of Squares) statistic is defined as:

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2 = \sum_{i=1}^{n} (e_{i,-i})^2$$

When the dataset contains several data from the same individual (e.g. trees or plots) the prediction residuals can be computed by fitting the model as many times as the number of individuals, omitting one individual at a time. This method has been used by SÁNCHEZ-GONZÁLEZ et *al*. (2005). In this case, bias and precision are computed for each individual and then combined in order to obtain "overall" measures of bias and precision (e. g. SÁNCHEZ-GONZÁLEZ et *al*., 2005; BRAVO-OVIEDO et *al*., 2007).

The comparison of model outputs with observed data can be based on:

1. Statistical tests
2. Graphical and visual analyses
3. Model efficiency computation
4. Bias and precision statistics analysis

Statistical tests

Literature analyzing the usefulness of statistical tests in model evaluation is extensive (e.g. SOARES et *al*., 1995; HUANG et *al*., 2003; YANG et *al*., 2004; PINJUV et *al*., 2006). Both parametric and non-parametric tests have commonly been analyzed: the paired *t* test, the χ^2 test, the Theil's inequality test, the simultaneous *F* test, the Kolmogorov-Smirnov test, the sign test and the Wilcoxon signed-rank test. Most authors agree that the usefulness of statistical tests in model evaluation is very limited. The use of different tests to evaluate models, with the same dataset, can lead to antagonistic conclusions. This is why YANG et *al*. (2004) alert to the important need "to reduce and remove any potential personal bias in selecting a favourite test". PINJUV et *al*. (2006) refuse the use of statistical tests when repeated measurements have been taken from the same plots. A list of consequences is presented:

1. Estimators of the regression coefficients may no longer have minimum variance but will still be unbiased and consistent
2. Standard errors of coefficients in the regression will be underestimated

3. Any significance tests or confidence limits constructed using *t* or *F* distributions are likely to be incorrect, since assumed independence of errors is violated

Model efficiency

Model efficiency (ME) is a measure of model performance and is described by:

$$ME = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value and $\bar{y}$ is the average observed value. The statistic is usually computed with an independent dataset or with "true" prediction residuals, as explained before. 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 very poor model.

Graphical and visual analyses

Graphical and visual analyses are among the most efficient tools in the model evaluation process. Several graphs can be plotted:

1. Observed values *versus* predicted values: in a good model, data should lie in a diagonal pattern
2. Observed or predicted values *versus* age, so as to analyze whether the prediction variance is changing with time or if the model is behaving well across the entire range of the age factor
3. Residuals (or standardized residuals) versus explanatory variables or predicted values, to detect possible autocorrelation and other dependencies or systematic patterns

The major disadvantage of the graphical and visual analyses is that they can be subjective. This is why it is necessary to validate a model by combining graphical approaches with other methods.

Bias and precision analyses

Bias can be assessed through histograms of the residuals and computation of the mean of the residuals. The interquantile range of the residuals, the residual mean square and the mean of the absolute residuals can be computed as measures of precision. Average model bias measures the error when several observations are combined by totaling or averaging, and mean absolute difference measures the average error associated with a single prediction (VANCLAY & SKOVSGAARD, 1997). In a good model, the mean of the residuals should not differ significantly from zero and the precision of prediction should not exceed certain limits.

Both measures of bias and precision may also be expressed as percentages, which may be particularly useful when the y values are not of the same magnitude. More information can be obtained by partitioning data - e.g. by age, site index or stand density - and examining model performance in each of the strata as well as by examining bias and precision along a gradient of these variables.

Sensitivity analysis

The sensitivity analysis (SA), sometimes called "what-if" analysis, examines how model predictions depend upon inputs, parameters, relationships and submodels. SA provides insight into the influence of different model parameters, which can promote a general understanding of model robustness. Results of sensitivity tests may reveal parameters critical to model predictions and parameters which may be redundant. Knowledge of sensitive parameters may guide applications and the planning of model enhancements.

In practice the SA is carried out by changing the parameter or component and observing the corresponding effect on predicted outputs. However, meaningful SAs are difficult, as the estimate of sensitivity depends both on the values of the inputs and the model parameters, so that many model runs may be necessary to complete the scenarios. This may be a tedious undertaking, especially when there are many parameters (VANCLAY & SKOVSGAARD, 1997). HUANG et *al*. (2003) refer two simple and practical methods of conducting SA in growth and yield modelling: (a) computation of a sensitivity index (PANNELL, 1997 in HUANG et *al*. 2003) and (b) graphical techniques of sensitivity analysis (FREY & PATIL, 2002 in HUANG et *al*., 2003).

The SA is especially important when models are applied outside the range of conditions presented in the dataset used to build them.

EVALUATION OF DIFFERENT TYPES OF MODELS

Forest growth and yield models have evolved with the evolution of forest management objectives: forest models were originally empirical growth and yield models, but today there is a large spectrum of models which range from state-space stand-level models through complex process-based eco-physiological models (RENNOLLS et *al*., 2007). BATTAGLIA & SANDS (1998) categorized models according to their dimension of resolution (spatial and temporal scale), level of complexity (number of environmental variables and processes included) and generality (situations to which the model can be applied). Different models require specific procedures during model evaluation. In this work, the evaluation of empirical and process-based models is discussed. The use of such models at landscape level requires implementation in a Decision Support System (DSS). The evaluation of the output scenarios resulting from the DSS has specific problems.

Evaluation of empirical and process based models

Empirical growth and yield models are based on large amounts of field data and describe growth rate as a regression function of several stand/tree variables, using site index as the main driving variable. They are appropriate for predicting growth for a range of silvicultural practices and site conditions but are generally site-specific and cannot simulate the results of changing conditions, thus restricting their applicability.

The applicability of the previously presented methodology requires a profound knowledge of the system to be modeled as well as of the model itself and the relationships between submodels. In empirical models the *type* of data used for model building is determinate of the good performance of the model during the evaluation process. The dataset should include long-term series data and should sample the full range of site and stand conditions, including extremes of stand conditions (even if they will never be applied); these are fundamental to the proper definition of the response surface for growth models (VANCLAY, 1994).

Process-based models aim to simulate the growth pattern of stands in terms of the physiological processes that determine growth. As a consequence, they are useful for long-term predictions, especially in changing conditions of management and climate. All the submodels of a process-based model are representations of processes at the same conceptual level of hierarchy and can be calibrated independently, based on measurements designed for that purpose (SHARPE & RYKIEL, 1991 in MÄKELA et *al*., 2000). However, the modelling of light interception, photosynthesis, stomata conductivity, water relations and nutrition includes many uncertainties and requires the use of many poorly known parameters (MOHREN & BURKHART, 1994).

The outstanding difference between evaluations of empirical and process-based models lies at the level of qualitative evaluation, due to the great conceptual difference between these two types of models. Qualitative evaluation of empirical models deals mainly with the simple analysis of asymptotes, signs and values of parameters and of the agreement of long-term simulations with the theories of forest growth. In contrast, this type of evaluation in process-based-models focuses on the simulation of physiological processes and on the effect on the model outputs of the simplifications that are usually assumed by each model.

In order to characterize model error, the methodology described for quantitative evaluation can be applied to both types of models. The main difference lies in the interpretation of results, namely the analysis of possible causes of the model's detected failures. The error or inadequacies detected in empirical models are usually related to one of the following problems: (1) lack of quality of the dataset; (2) wrong specification of one or more of the growth functions used as submodels. The first problem implies poor estimations of the parameters included in the growth functions, even if those are correct.

The problem is different with process-based models. MÄKELÄ et *al*. (2000) presented, a reflection on the use of process-based models for forest ecosystem management and in this document they resumed the particularities of evaluation of this type of models: "When evaluating a process based model of forest growth it is often difficult to determine whether deviations from predicted performance are caused by variation in the system, by inadequacies of the model, or by incorrect values for submodel parameters." Strong emphasis should be put, in the case of process-based models, on the analysis of the correctness of the parameters included in the submodels. The evaluation of these models can be complemented by the comparison of their outputs with those from empirical models available for specific species and regions. One advantage of process-based models is the fact that they only need a small dataset for calibration (e.g. FONTES et *al*., 2006, who used data from 12 plots, 4 from one trial with irrigation and fertilization and 8 from a spacing trial for calibration of the 3PG model for eucalyptus plantations in Portugal). This leaves a large independent dataset for the quantitative evaluation. The problem of data splitting is not an issue for process-based models.

For the evaluation of both types of forest growth models, data from long-term series, spacing trials, physiological field measurements and laboratory analyses are fundamental.

Application of models at landscape level

Empirical and process based models, as well as hybrid models (models with both process based and empirical components), can be used in forest management at the landscape level. This implies implementation of the models in a decision support system (DSS) that encompasses the simulation of stand-specific management alternatives, the evaluation of total production and the respective net present value for each combination of prescriptions. DSS are computer-based systems that integrate database management systems with analytical and operational research models, graphic display, tabular reporting capabilities and the expert knowledge of decision makers to assist in solving specific problems (FISCHER et *al*., 1996). The general architecture of a conventional DSS typically includes: the decision support system generator, a database management system (DBMS), a management system for the method base and the model base (MDMS) and a graphical user interface (GUI) (REYNOLDS et *al*., 2005). Data is organized and made available by the DBMS to models and methods in the MBMS that process and convert it into information and recommendations to the decision-maker. The DSS-generator allows modelling of the sequence of algorithms required to generate and evaluate decision alternatives (decision model) which should be adaptable to new decision problems within a particular decision-making domain. The GUI supports communication between the system and the decision-maker by the use of a help system and report management system.

Evaluation of DSS outputs is a difficult task. How can the user know if the optimal solution obtained for a specific region has something similar to reality? And what is the effect on the reality of the output of the errors existing in all the components of the DSS? This is a challenging subject for future research. At present we have several DSS for forest management available but, to our knowledge, few have been evaluated in the real world.

CONSIDERATIONS

Model evaluation is not a simple process and should require quantitative as well as qualitative procedures. To evaluate models, a diversified dataset is necessary: data from long-term series, spacing trials, physiological trials, detailed field measurements, soil and climatic information and laboratory analyses. The evaluation of empirical and process-based models should take into account the specific characteristics of these two types of models. While qualitative evaluation of empirical models deals mainly with the analysis of asymptotes, signs, and values of parameters, this type of evaluation in process-based models focuses on the simulation of the physiological processes and the effect on the model outputs of the simplifications that are usually assumed by each model. The quantitative evaluation procedures can be applied to both

types of models in order to characterize model error. However, the causes of the errors or inadequacies detected are different because conceptually, these models are distinct. Forest management at the landscape level implies the use of a decision support system (DSS). To evaluate the accuracy of the outputs of the DSS, all possible sources of errors should be analyzed and this is a very arduous task, including data, databases and interfaces, model programming, GIS implementation, optimization algorithm, graphical and tabular interfaces… The notion that model evaluation should focus on analyses of the model's components as well as on the whole model is of great importance with all types of models.

Models can only be evaluated in relative terms, and their predictive value is always open to questioning. Model evaluation is an ongoing process.

LITERATURE

BAKUZIS, E.V.; 1969. Forestry viewed in an ecosystems perspective. *In*: G.M. Van Dyne (eds.), *The Ecosystem Concept in Natural Resources Management:* 189-258. Academic Press. New York.

BATTAGLIA, M. & SANDS, P.J.; 1998. Process-based forest productivity models and their application in forest management. *Forest Ecol. Manage.* 102: 13-32.

BRAVO-OVIEDO, A.; RÍO, M. & MONTERO, G.; 2007. Geographic variation and parameter assessment in generalized algebraic difference site index modelling. *Forest Ecol. Manage*. 247: 107-119.

EFRON, B. & GONG, G.; 1983. A leisurely look at the bootstrap, the jackknife and cross-validation. *The American Statistician* 37(1): 36-48.

FISCHER, M.M.; SCHOLTEN, H.J. & UNWIN, D.; 1996. Geographic information systems, spatial data analysis and spatial modelling. *In*: M. M. Fischer, H.J. Scholten & D. Unwin (eds.), *Spatial Analytical Perspectives on GIS.* GISDATA Series 4: 3-19. Taylor and Francis.

FONTES, L.; LANDSBERG, J.; TOMÉ, J.A.; TOMÉ, M.; PACHECO, C.A.; SOARES, P. & ARAÚJO, C.; 2006. Calibration and testing of a generalized process-based model for use in Portuguese eucalyptus plantations. *Can. J. For. Res.* 36: 3209-3221.

FREY, H.C. & PATIL, S.R.; 2002. Identification and review of sensitivity analysis methods. *Risk Analysis* 22: 553-578.

HELMS, J.A.; 1998. *The dictionary of forestry.* Society of American Foresters. Bethesda.

HUANG, S.; YANG, Y. & WANG, Y.; 2003. A critical look at procedures for validating growth and yield models. *In*: A. Amaro, D. Reed & P. Soares (eds.), *Modelling Forest Systems*: 97-110. CAB International. Wallingford.

JONES, P.N. & CARBERRY, P.S.; 1994. A technique to develop and validate simulation models. *Agric. Systems* 46: 427-442.

KOZAK, A. & KOZAK, R.; 2003. Does cross validation provide additional information in the evaluation of regression models? *Can. J. For. Res.* 33: 976-987.

LANDSBERG, J.; 2003. Modelling forest ecosystems: state of the art, challenges, and future directions. *Can. J. For. Res.* 33: 385-397.

LEARY, R.A.; 1988. Some factors that will affect the next generation of forest growth models. *In*: A.R. Ek, S.R. Shifley & T.E. Burk (eds.), *Forest Growth Modelling and Prediction, Proc. IUFRO Conf.* 2: 1058-1065. General Technical Report NC-120. Forest Service, North Central Forest Experimental Station.

LOEHLE, C.; 1997. A hypothesis testing framework for evaluating ecosystem model performance. *Ecol. Model.* 97: 153-165.

MÄKELÄ, A.; LANDSBERG, J.; EK, A.R.; BURK, T. E.; TER-MIKAELIAN, M.; ÅGREN, G.I.; OLIVER, C.D. & PUTTONEN, P.; 2003. Process-base models for forest ecosystem management: current state of the art and challenges for practical implementation. *Tree Phys.* 20: 289-298.

MOHREN, G.M.J. & BURKHART, H.E.; 1994. Contrasts between biologically-based process models and management-oriented growth and yield models. *Forest Ecol. Manage.* 69: 1-5.

MYERS, R.; 1986. *Classical and modern regression with applications.* Duxbury Press. Boston.

ODERWARLD, R.G. & HANS, R.P.; 1993. Corroborating models with model properties. *Forest Ecol. Manage.* 62: 271-283.

PANNELL, D.J.; 1997. Sensitivity analysis of normative economic models: theoretical framework and practical strategies. *Agric. Economics* 16: 139-152.

PINJUV, G.; MASON, E.G. & WATT, M.; 2006. Quantitative validation and comparison of a range of forest growth model types. *Forest Ecol. Manage.* 236: 37-46.

PRISLEY, S.P. & MORTIMER, M.J.; 2004. A synthesis of literature on evaluation of models for policy applications, with implications for forest carbon accounting. *Forest Ecol. Manage.* 198: 89-103.

RENNOLLS, K.; TOMÉ, M.; MCROBERTS, R.E.; VANCLAY, J.K.; LEMAY, V.; GUAN, B.T. & GERTNER, G.Z.; 2007. Potential contributions of statistics and modelling to sustainable forest management: review and synthesis. *In*: K.M. Reynolds, A. Thomson, M. Shannon, M. Köhl, R. Duncan & K. Rennols (eds.), *Sustainable forestry: from monitoring and modelling to knowledge management and policy science*: 314-341. CAB International. Wallingford.

REYNOLDS, K.M.; BORGES, J.G.; VACIK, H. & LEXER, M.J.; 2005. Information and communication technology in forest management and conservation. *In*: L. Hetemaki & S. Nilsson (eds.), *Information Technology and the Forest Sector*: 150-171. IUFRO World Series vol. 18.

RYKIEL JR., E.J.; 1996. Testing ecological models: the meaning of validation. *Ecol. Model.* 90: 229-244.

SÁNCHEZ-GONZÁLEZ, M.; TOMÉ, M. & MONTERO, G.; 2005. Modelling height and diameter growth of dominant cork oak trees in Spain. *Ann. For. Sci.* 62: 633-643.

SHARPE, P.J.H. & RYKIEL, E.J. JR.; 1991. Modelling integrated response of plants to multiple stresses. *In*: H.A. Mooney, W.E. Winner & E.J. Pell (eds.), *Response of plants to multiple stresses*: 205-224. Academic Press. New York.

SNEE, R.D.; 1977. Validation of regression models: methods and examples. *Technometrics* 19: 415-428.

SOARES, P.; TOMÉ, M.; SKOVSGAARD, J.P. & VANCLAY, J.K.; 1995. Evaluating a growth model for forest management using continuous forest inventory data. *Forest Ecol. Manage.* 71: 251-265.

VANCLAY, J. & SKOVSGAARD, J.P; 1997. Evaluating forest growth models. *Ecol. Model.* 98: 1-12.

YANG, Y.; MONSERUD, R.A. & HUANG, S.; 2004. An evaluation of diagnostic tests and their roles in validating forest biometric models. *Can. J. For. Res.* 34: 619-629.