An evaluation of diagnostic tests and their roles in validating forest biometric models

Yuqing Yang, Robert A. Monserud, and Shongming Huang

Abstract: Model validation is an important part of model development. It is performed to increase the credibility and gain sufficient confidence about a model. This paper evaluated the usefulness of 10 statistical tests, five parametric and five nonparametric, in validating forest biometric models. The five parametric tests are the paired t test, the χ^2 test, the separate t test, the simultaneous F test, and the novel test. The five nonparametric tests are the Brown–Mood test, the Kolmogorov–Smirnov test, the modified Kolmogorov–Smirnov test, the sign test, and the Wilcoxon signed-rank test. Nine benchmark data sets were selected to evaluate the behavior of these tests in model validation; three were collected from Alberta and six were published elsewhere. It was shown that the usefulness of statistical tests in model validation is very limited. None of the tests seems to be generic enough to work well across a wide range of models and data. Each model passed one or more tests, but not all of them. Because of this, caution should be exercised when choosing a statistical test or several tests together to try to validate a model. It is important to reduce and remove any potential personal bias in selecting a favorite test, which can influence the outcome of the results.

Résumé : La validation est un aspect important dans le développement des modèles. Elle est nécessaire pour augmenter la crédibilité d'un modèle et lui faire suffisamment confiance. Cet article évalue l'utilité de 10 tests statistiques, cinq paramétriques et cinq non paramétriques, pour valider les modèles biométriques utilisés en foresterie. Les cinq tests paramétriques sont le test t jumelé, le test du χ^2, le test t séparé, le test F simultané et le test nouveau. Les cinq tests non paramétriques sont le test de Brown–Mood, le test de Kolmorogov–Smirnov, le test de Kolmorogov–Smirnov modifié, le test des signes et le test de rang de Wilcoxon. Neuf séries de données repères ont été sélectionnées pour évaluer le comportement de ces tests lors de la validation des modèles. Trois séries de données ont été collectées en Alberta et six autres proviennent de publications. Il a été démontré que l'utilité des tests statistiques est très limitée pour la validation des modèles. Aucun de ces tests ne semble assez générique pour bien performer avec une vaste gamme de modèles et de données. Chaque modèle a résussiréussi un test ou plus, mais pas tous les tests. C'est pourquoi le choix d'un ou de plusieurs tests statistiques doit être fait avec prudence pour essayer de valider un modèle. Il est important de réduire ou d'éliminer tout biais personnel dans le choix d'un test pour ne pas influencer le résultat.

[Traduit par la Rédaction]

Introduction

Statistical models have been widely used in many disciplines for various purposes. Decision makers often use the results from these models to assist and support the decision-making process. The developers and users of these models, and the people affected by the decisions based on these models, are all concerned with whether or not these models are acceptable representations of the real world (Morehead 1996). This concern is normally addressed through model validation (Sargent 1999). Many statistical models are developed for prediction purposes based on empirical data. Once a prediction model is constructed, an assessment of its valid-

ity using an independent data set is often needed to ensure that model predictions represent the most likely outcome of the real world.

Shugart (1984) defined model validation as "procedures, in which a model is tested on its agreement with a set of observations that are independent of those observations used to structure the model and estimate its parameters". Note that data-splitting methods (Snee 1977) that randomly split one data set into a model estimation and a model testing portion are not validation, for the testing portion of the data set is not independent of the estimation portion; it will have the same statistical structure. This is clearly shown by Kozak and Kozak (2003), who demonstrated that validation by data splitting provides little, if any, additional information in the process of evaluating models. Model validation is sometimes a prickly issue for both model builders and model users (Gass and Thompson 1980; Monserud 2003). There are many types of validation methods available; some are qualitative and others are quantitative (Holmes 1983; Sargent 1999). It is important to recognize that model validation is not used to prove that a model is correct (Popper 1963), but rather to show that model predictions are close enough to independent data and that decisions made based on the model are defensible. If a model provides acceptable predictions with small prediction errors and low variances, it is consid-

Y. Yang[1] and S. Huang. Forest Management Branch, Land and Forest Division, Alberta Sustainable Resource Development, 8th Floor, 9920-108 Street, Edmonton, AB T5K 2M4, Canada.
R.A. Monserud. USDA Forest Service, Pacific Northwest Research Station, 620 SW Main Street, Suite 400, Portland, OR 97205, U.S.A.

[1]Corresponding author (e-mail: yuqing.yang@gov.ab.ca).

doi: 10.1139/X03-230

ered appropriate. Otherwise, the model needs to be adjusted or even reestimated. The ultimate goal of model validation is to increase the credibility and gain sufficient confidence about a model.

Model validation is important in any empirical analysis. It is an integral part of model building as well as quality assurance and quality control. Unfortunately, it is often ignored, a rich literature notwithstanding (Balci and Sargent 1984; Morehead 1996; Kleijnen 1999; Sargent 1999). Perhaps modelers (and employers) are reluctant to subject their newly discovered model to the fires of independent testing after surviving the rigors and expense of development and publication. Nevertheless, the attempt to reject hypotheses, especially favorite hypotheses, is the modus operandi of science. This is especially true for validating forest biometric models.

Forest biometric models often consist of several submodels or functions, each independently or simultaneously fitted using different techniques (e.g., Hasenauer et al. 1998; Huang and Titus 1999). Because the behavior of individual components within a system plays an important role in determining the overall outcome, these individual components need to be validated separately. The validity of each individual component, however, does not guarantee the validity of the overall outcome, which is usually considered more important in practice. Therefore, the overall system outcome should be validated as well.

Validation (testing with independent data) is a subset of the general problem of model evaluation. As a standard for comparison, an ideal model would be able to forecast important traits of interest with a high degree of accuracy and precision for any time horizon and any given stand condition (Curtis 1972; Ritchie and Hann 1997). The first step in model evaluation is the clear specification of the criteria for the evaluation (Brand and Holdaway 1983; Robinson and Monserud 2003). For Robinson and Monserud (2003), their criterion was adaptability of the model for extension into new populations and applications. Buchman and Shifley (1983) offer a detailed checklist of questions helpful for evaluating a growth projection system and address features such as documentation and system support, data requirements, accuracy and precision of component models and overall system prediction, flexibility for modification for new situations or resources, and the logic and biological realism of predictions. Goelz and Burk (1992) list several desirable criteria for evaluating site index equations, including logical behavior and a theoretical basis. Vanclay and Skovsgaard (1997) also list biological and theoretical criteria. They ask whether the structure is parsimonious, biologically realistic, consistent with existing theories of forest growth, and whether it predicts sensible responses to management activities.

By their nature, models are designed to predict the average expected behavior of the systems that they represent. Recall that Sir Francis Galton's 1885 paper first discussing regression was actually "regression to the mean" (Draper and Smith 1981). Models will fail to predict for extreme situations because they are designed to be correct in the long run, over many different sets of circumstances. Forest growth models arguably provide their greatest advantage when used in a large-scale planning operation because the most important characteristic is to describe the average response of stands over a wide range of conditions (A. Stage, Moscow, Idaho, personal communication, 1999). Model predictions tell us about what behavior is to be expected on average; the residuals tell us about behavior that is deviant. While statistical assessments of a model may conclude that there is insignificant bias and high precision, this describes the mean behavior of the model and does not guarantee its accuracy in any particular application (A. Robinson, University of Idaho, personal communication, 2001).

A large number of validation procedures can be used to quantitatively assess the predictive ability of a fitted model. Among them, many statistical tests have been routinely employed. These tests are often interpreted as more objective ways for judging the predictive ability of a model (Harrison 1990; Mayer et al. 1994). They assess the size, direction, and dispersion of prediction errors in various ways. In forestry, modelers have adopted and developed various procedures for validating models (Ek and Monserud 1979; Marshall and LeMay 1990; Huang et al. 1999; Waller et al. 2003). Some of the statistical tests on means, variances, autocorrelations, and overall distributions have been applied (Freese 1960; Stephens 1974; Reynolds et al. 1981; Harrison 1990). Vanclay and Skovsgaard (1997) list several statistical properties useful in model evaluation: properties of the model parameters (e.g., unbiased), error characterization (evaluate the accuracy, residual patterns, confidence intervals, and contribution of individual model components to the total error), statistical tests (bias and precision of the model and components, goodness-of-fit of predicted distributions, patterns and distribution of residuals, and correlations over time), and sensitivity analyses (determine how model components influence predictions, the sensitivity of the model to its inputs, and how errors propagate through the model). Further examples can be found in Desanker et al. (1994), Robinson (1998), and Yaussy (2000).

The use of statistical tests in model validation has attracted much debate since the work of Wright (1972). With varying criteria for the "value" of models and the methods of determining it, validation has proven to be a complex issue (Mayer et al. 1994; Morehead 1996). Each model is unique and no validation technique or method enjoys widespread application. It is also relatively easy to choose a favorite procedure and then comparatively rank one model above or below another one (Huang et al. 2003).

The main objective of the study was to evaluate the usefulness of various statistical tests on validating individual forest biometric models. Ten tests, five parametric and five nonparametric, were selected for this purpose. All tests, whether they are parametric or nonparametric, require some specific assumptions (Marshall et al. 1995; Conover 1999). The assumption of normality is especially common in this regard (Gregoire and Reynolds 1988). Parametric tests generally assume, among others, that prediction errors are normally distributed. When this assumption is upheld, parametric tests are preferred because they are more powerful in revealing significant differences between model predictions and actual observations. When the normality assumption is not satisfied, nonparametric tests should be used, unless a normalizing transformation can be found. Because of the large variety of data involved in forest biometric studies,

Table 1. Summary of observed and predicted values for the variable of interest (y) as well as prediction errors obtained from nine benchmark data sets.

Data set	n	Observed, y				Predicted, $\hat{y}$				Prediction error, $e_i = (y_i - \hat{y}_i)$			
		Mean	Min.	Max.	SD	Mean	Min.	Max.	SD	Mean	Min.	Max	SD
I	46	21.6	5.5	31.1	5.0	21.0	8.4	27.5	4.2	0.63	–4.9	5.62	2.79
II	81	278.4	22.3	502.3	138.9	273.8	21.6	510.2	137.7	4.56	–58.65	41.88	16.53
III	247	14.0	5.4	22.4	3.6	14.0	7.0	22.4	3.6	0.03	–7.99	4.37	1.84
IV	15	25.8	7.0	64.8	15.1	25.8	6.0	76.1	17.8	0.03	–11.32	5.87	4.87
V	54	2.2	1.3	2.8	0.3	2.1	1.3	2.0	0.3	0.07	–0.02	0.17	0.05
VI	10	4712	2380	8320	1863	4423	2320	6780	1456	289	–110	1540	505.50
VII	21	56.0	37.0	75.0	10.9	56.2	39.0	73.0	10.3	–0.19	–8.00	7.00	4.06
VIII	16	99.2	19.8	167.1	38.4	90.9	12.8	138.2	35.3	8.32	–3.00	28.90	7.37
IX	63	179.5	32.3	511.8	125.8	170.8	44.2	373.5	92.9	8.75	–113.50	198.90	55.61

Note: n is number of observations; mean, min., max., and SD are the mean, minimum, maximum, and standard deviation, respectively. Details about the data sets are provided in the Data section.

nine benchmark data sets of different types were selected to demonstrate the roles of the 10 statistical tests in model validation.

Data

A total of nine data sets (labeled I, II, III, …, IX) were used to assess the behavior of the 10 diagnostic tests. Each data set is independent of the data used to fit the model to be validated. The first three were collected in Alberta. The rest were published data sets from elsewhere: IV (Montgomery and Peck 1992), V (Neter et al. 1989), VI (Rawlings 1988), VII (West 1981), VIII (Ek and Monserud 1979), and IX (Reynolds and Chung 1986). We purposely used these published data in addition to our own data because they have been used in several validation studies, and users can easily access them. For each data set, the observed values for the variable of interest and their respective predictions from the model developed on an independent data set are obtained. A summary of the observed and predicted values, as well as the prediction errors (the difference between observed and predicted values) associated with each data set, is shown in Table 1. A brief description of the nine data sets follows:

(1) Height–diameter data (data set I) — Forty-six lodgepole pine (*Pinus contorta* var. *latifolia* Engelm.) trees were felled in the Lower Foothills natural subregion of Alberta (Alberta Environment Protection 1994). Accurate measurements for tree height (HT, m) and tree diameter at breast height (DBH, cm) were taken from these trees. A predicted tree height (HT_P, m) was obtained for each tree based on the height–diameter model developed by Huang (1999):

[1] $\quad HT_P = 1.3 + 29.4214$
$$\times \, [1 - \exp(-0.0457DBH)]^{1.1381}$$

The predicted heights and the observed height–diameter data are shown in Fig. 1.

(2) Plot volume data (data set II) — Eighty-one sample plots of 200 m^2 each were established in pure lodgepole pine stands in the Lower Foothills natural subregion of Alberta. The diameters of all trees taller than 1.3 m were tallied, and the total basal area of each plot (BA, m^2/ha) was obtained. The heights of the two largest diameter trees in each plot were measured to obtain

top height (TH, m). Individual tree volumes within each plot were calculated following the procedures described by Huang (1994) and summarized to give plot volume (VOL, m^3/ha). The following plot volume prediction equation used in Alberta was applied to obtain a predicted plot volume (VOL_P, m^3/ha) for each plot:

[2] $\quad VOL_P = 0.2125TH^{1.2856}BA^{1.0638}$
$$\times \exp(-0.00155BA - 0.0166TH)$$

The observed volume, basal area, and top height, and the predicted volume from eq. 2 are shown in Fig. 2.

(3) Site index data (data set III) — Two hundred and forty-seven lodgepole pine trees, covering three natural subregions (Sub-Alpine, Upper Foothills, and Lower Foothills, see Alberta Environment Protection 1994), were felled, and the stems were analyzed. Each tree was sectioned at stump height (0.3 m), breast height (1.3 m), 2.3, 3.3, and 4.3 m above ground, and equal lengths of 2.5 m thereafter to the top of the tree. The ring count at the top of each section was measured, and the height–age trajectories were obtained (Fig. 3). The observed site index of each tree was defined as the tree height at breast-height age (Bhage) 50 years. A corresponding predicted site index (SI_P, m) was solved from the following site index model for lodgepole pine (Huang et al. 1997):
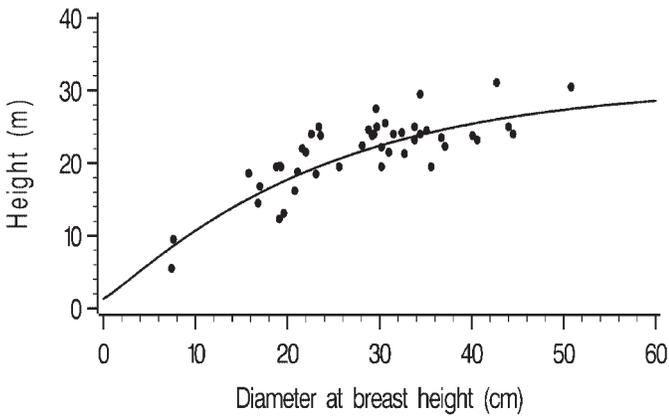
[3] $\quad HT = 1.3 + (SI_P - 1.3)$
$$\times \left(\frac{1 + \exp[b_1 + b_2 \ln(50 + b_3) - \ln(SI_P - 1.3)]}{1 + \exp[b_1 + b_2 \ln(Bhage + b_3) - \ln(SI_P - 1.3)]} \right)$$

where the coefficients b_1, b_2, and b_3 for different natural subregions are given in Huang et al. (1997).

(4) Data set IV (Montgomery and Peck 1992) — A model was developed to predict the amount of time required by a route driver to service the vending machines in an outlet. Fifteen new observations were collected independently of the modeling data to validate the developed model. The observed and predicted values of service time were obtained by Montgomery and Peck (1992).

(5) Data set V (Neter et al. 1989) — Fifty-four new observations were obtained to validate a survival model fitted

Fig. 1. Observed height–diameter data from 46 felled lodgepole pine trees (data set I), overlaid with the height–diameter curve generated from eq. 1.

Fig. 2. Plot volume data (data set II) from 81 plots. (*a*) Observed volume (m³/ha) versus basal area (m²/ha) data. (*b*) Observed volume versus top height data. (*c*) Observed versus predicted volumes. The predicted volumes are obtained using eq. 2.

on a different data set. The dependent variable of the survival model is survival time. Four independent variables are used to predict the survival time.

(6) Data set VI (Rawlings 1988) — Data from 10 new watersheds were used to evaluate a prediction model of the maximum rates of runoff (ft³/s) from watersheds following rainstorms. The observed and predicted runoff rates for these 10 new watersheds were obtained, and their differences were calculated.

(7) Data set VII (West 1981) — A simulation model for stand basal area growth was developed. Data from 21 remeasurement plots were collected in a southern Tasmania eucalypt forest to evaluate the performance of the simulation model. The observed and simulated stand basal areas (m²/ha) for each of these stands given by West (1981) were used.

(8) Data set VIII (Ek and Monserud 1979) — Sixteen northern hardwood plots from the Wisconsin Timber Harvest Forests were used to evaluate the performance of two stand growth models, FOREST and SHAF. Only the basal area data and the FOREST model were considered in this study. The observed and predicted basal areas (ft²/acre) at the end of growth period were used.

(9) Data set IX (Reynolds and Chung 1986) — New data from 63 plots located in loblolly pine (*Pinus taeda* L.) plantations in Virginia were used to validate the growth and yield simulator PTAEDA. Stand volume to a 10-cm top diameter was the variable of interest. The observed and predicted stand volumes (m³/ha) were shown in Reynolds and Chung (1986).
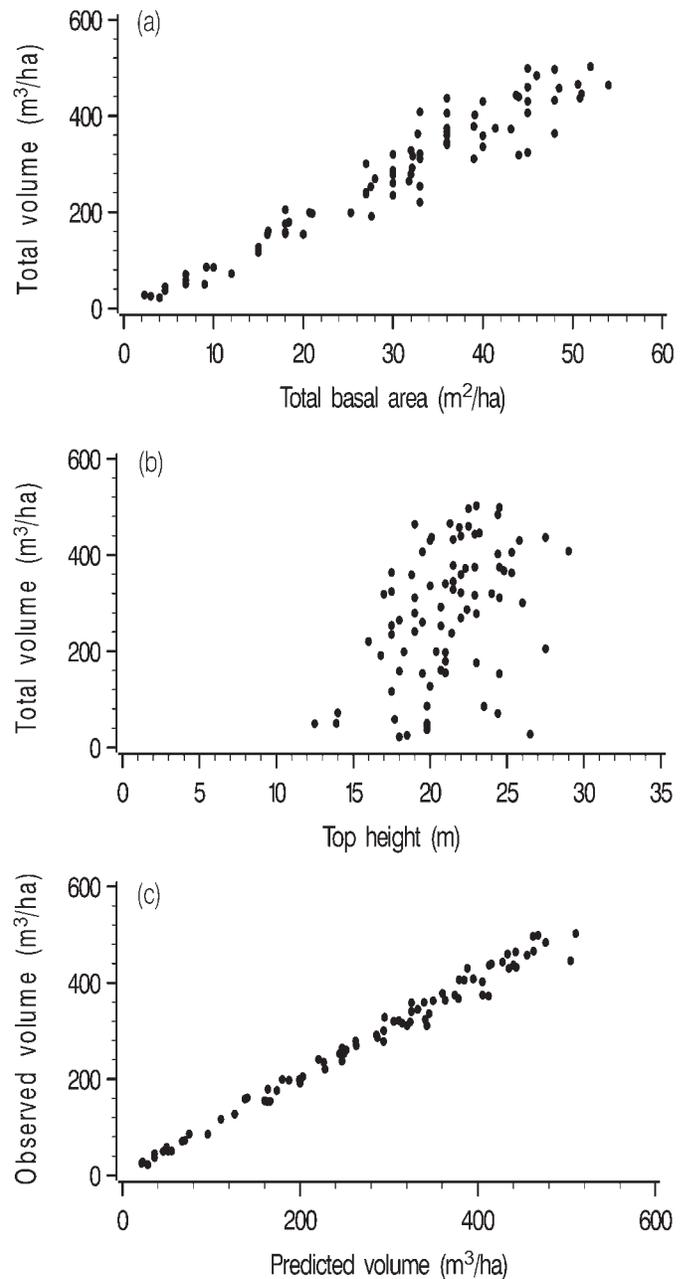
Methods

Ten commonly used statistical tests in model validation were used to assess the predictive ability of fitted models. Since the normality of the error distribution of each data set determines whether a parametric or a nonparametric test is more appropriate to use, tests for normality were conducted first.

Normality tests

Four normality tests were conducted to evaluate the normality of the prediction errors associated with each of the nine data sets. All four are commonly used in testing normality. A brief description of the four tests follows, where n refers to the sample size of a data set, and i is the ith observation in the data set.

The Shapiro–Wilk test

The Shapiro–Wilk test evaluates whether a random sample comes from a normal distribution. It is basically the square of Pearson's correlation coefficient calculated between the order statistics of the sample data and some constants that represent what the order statistics should look like

Fig. 3. Observed tree sectioning data from 247 felled lodgepole pine trees (data set III), overlaid with the site index (SI) curves generated from eq. 3. Site index values at 50 years are given for each site index curve.

if the population were normal. It is calculated as (Conover 1999):

$$[4] \qquad W = \left[\sum_{i=1}^{k} a_i (X^{(n-i+1)} - X^{(i)}) \right]^2 / \sum_{i=1}^{n} (X_i - \overline{X})^2$$

where X_i is the ith observation of the sample data, $\overline{X}$ is the sample mean, $X^{(i)}$ is the ith order statistic, and $a_1, a_2, ..., a_k$ are the constants generated from the means, variances, and covariances of the order statistics of a sample of size n from a normal distribution, with k equals to $n/2$ (Conover 1999). The sample data is considered normal if the W value in eq. 4 is close to 1.0. Small values of W imply departure from normality.

The Kolmogorov–Smirnov test

The Kolmogorov–Smirnov (KS) test assesses how well the cumulative distribution of the sample data conforms to that of a hypothesized theoretical distribution (in this case, normal distribution). The test focuses on the maximum vertical difference between the two cumulative distribution functions, which is intuitively appealing. The test takes the following form (Scheaffer and McClave 1995):

$$[5] \qquad D = \max |F(x) - F_n(x)|$$

where $F(x)$ and $F_n(x)$ are the theoretical and sample cumulative distribution functions, respectively. The KS test is distribution free in the sense that the critical values do not depend on the specific distribution being tested. To test for normality of the sample data, $F(x)$ is defined as a cumulative normal distribution.

The Cramér – von Mises test

The Cramér – von Mises test is calculated by (Stephens 1974; D'Agostino and Stephens 1986):

$$[6] \qquad W^2 = \sum_{i=1}^{n} [F(X^{(i)}) - (2i-1)/(2n)]^2 + 1/(12n)$$

where F is the cumulative distribution function of the specified distribution, and $X^{(i)}$'s are the ordered data. This is in the KS family of tests. However, unlike the Kolmogorov

test, the Cramér – von Mises test not only considers the largest difference, but also considers n differences between the two curves (the hypothesized and the empirical cumulative distribution functions). Intuitively it appears that the Cramér – von Mises test statistic makes more complete use of the data and therefore should be more effective than the Kolmogorov test statistic, but the facts fail either to prove or disprove such intuition (Conover 1971).

The Anderson–Darling test

This test evaluates whether the sample data come from a population with a specific distribution. It is a modification of the KS test and gives more weight to the tails. Unlike the KS test, the Anderson–Darling test uses the specific distribution in calculating critical values. This allows for a more sensitive test. But the disadvantage is that critical values must be calculated for each particular distribution (Stephens 1974). The test statistic of the Anderson–Darling test, A^2, is computed as follows:

$$[7] \qquad A^2 = - \frac{\sum_{i=1}^{n} (2i-1)\{\ln(F(X^{(i)}) + \ln[1 - F(X^{(n+1-i)})]\}}{n} - n$$

where F is the cumulative distribution function of the specified distribution, and $X^{(i)}$'s are the ordered data.

Parametric and nonparametric validation tests

Ideally, for a good model, the mean prediction error should not differ significantly from zero, so that the prediction is unbiased, and the precision of prediction (i.e., the variance) should not exceed some limits, so that the variation of the prediction is within some tolerance. Almost all statistical tests are designed to evaluate (1) whether the mean prediction error is significantly different from zero, and (2) whether the variance of the prediction is larger than some critical value. The 10 statistical tests used in this study were assessed for their usefulness in validating forest biometric models, based on the nine benchmark data sets. These tests are described below, the first five being parametric and the remaining five being nonparametric.

The paired t test (Snedecor and Cochran 1980)

The null hypothesis of the paired t test is that the mean prediction error is zero. It is written as follows:

$$[8] \qquad t = \overline{e} / \sqrt{\sum (e_i - \overline{e})^2 / [n(n-1)]}$$

where e_i is the ith prediction error, and $\overline{e}$ is the mean prediction error. This test statistic has $n - 1$ degrees of freedom (df). Large t values indicate large mean prediction errors.

The chi-square (χ^2) tests

The χ^2 test can be written in various forms. The following three formulations presented by Freese (1960) were used in this study. Equation 9a is the standard χ^2 test with df = n; eq. 9b is a bias-free χ^2 test with df = $n - 1$, assuming the bias being constant; and eq. 9c is another bias-free χ^2 test with df = $n - 2$, assuming the bias being a linear function of y_i:

$$[9a] \qquad \chi_n^2 = \sum (y_i - \hat{y}_i)^2 / \lambda^2$$

where $\lambda^2 = E^2/\tau^2$

$$[9b] \quad \chi^2_{n-1} = \left[\sum (y_i - \hat{y}_i)^2 - nB^2\right]\Big/\lambda^2$$

where $B = \sum (y_i - \hat{y}_i)/n$

$$[9c] \quad \chi^2_{n-2} = SSE_{y \, on \, \hat{y}}/\lambda^2$$

where y_i and $\hat{y}_i$ are the observed and predicted values of the ith observation, respectively, λ^2 is an acceptable accuracy specified in the form of a hypothesized variance or a critical error, E is the allowable error in units of the observed y, τ is a standard normal deviate at the specified probability level (e.g., $\tau = 1.960$ when $\alpha = 0.05$), B is the bias, and SSE is the error sum of squares of a simple linear regression of $y_i = b_0 + b_1\hat{y}_i$. For consistency, an allowable error of $\pm7.16\%$ within the true (observed) mean (i.e., within ±30 of the observed mean of 419) as specified by Freese (1960) was used in this study for all data sets.

Separate t tests

A widely used method of validating a model is to evaluate the simple linear regression between observed and predicted y values, $y_i = b_0 + b_1\hat{y}_i$. For an acceptable model, the regression will be a 45° line through the origin. This is illustrated in Fig. 4. The adequacy of the model can be evaluated by testing separately whether $b_0 = 0$ and $b_1 = 1$ through separate t tests (Montgomery and Peck 1992). To test $b_0 = 0$, the t value can be read directly from the fit of $y_i = b_0 + b_1\hat{y}_i$. To test $b_1 = 1$, the test statistic is calculated by

$$[10] \quad t_{b_1} = (b_1 - 1)\Big/\sqrt{\left[\sum (y_i - \tilde{y}_i)^2/(n-2)\right]\Big/\sum (\hat{y}_i - \bar{\hat{y}})^2}$$

where $\tilde{y}_i$ is the predicted value from $y_i = b_0 + b_1\hat{y}_i$, and $\bar{\hat{y}}_i$ is the mean of $\hat{y}_i$ values. This test statistic has df = $n - 2$.

Simultaneous F test

Such as the separate t test, the simultaneous F test is designed to evaluate the regression $y_i = b_0 + b_1\hat{y}_i$ by testing whether $b_0 = 0$ and $b_1 = 1$ simultaneously (Montgomery and Peck 1992):

$$[11] \quad F = \frac{n(b_0 - 0)^2 + 2\sum \hat{y}_i(b_0 - 0)(b_1 - 1) + \sum \hat{y}_i^2(b_1 - 1)^2}{2\sum (y_i - \tilde{y}_i)^2/(n-2)}$$

where the variables are as previously defined in the separate t tests. The degrees of freedom are 2 and $n - 2$ for the numerator and denominator, respectively.
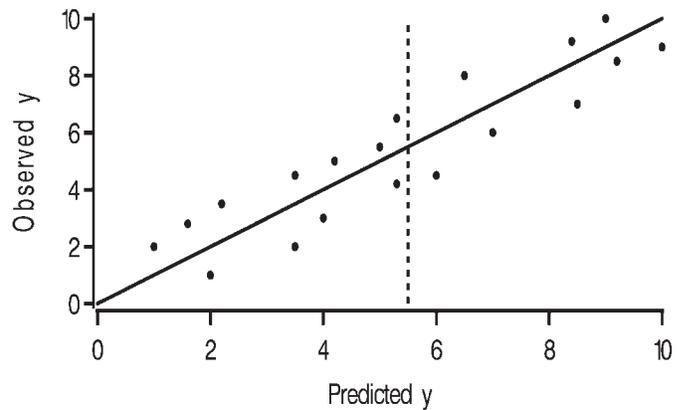
The novel test

Kleijnen et al. (1998) showed that it is inadequate to validate a model by testing whether the regression line $y_i = b_0 + b_1\hat{y}_i$ passes through the origin with a 45° slope. They proposed a new test termed "the novel test". This test regresses the prediction errors on the sum of observed and predicted values via a simple linear regression: $e_i = b_0 + b_1(y_i + \hat{y}_i)$. The validity of the model being tested is determined by testing $b_0 = b_1 = 0$ jointly based on the extra sum of squares method (Kleijnen et al. 1998):

$$[12] \quad F = [(n - 2)(SSE_R - SSE_F)]/(2SSE_F)$$

where $SSE_R = \sum e_i^2$ and $SSE_F = \sum(e_i - \hat{e}_i)^2$ are the reduced and the full sums of squared errors, respectively. The de-

Fig. 4. An illustration of the simple linear regression between observed (y) and predicted ($\hat{y}$) values, where the regression line is defined by $y_i = b_0 + b_1\hat{y}_i$. The vertical broken line refers to the median of $\hat{y}$.

grees of freedom are 2 and $n - 2$ for the numerator and denominator, respectively.

The Brown–Mood test

An alternative to the simultaneous F test and the novel test is the nonparametric Brown–Mood test (Daniel 1990). It can be used to jointly test whether the intercept and the slope of the regression line are equal to some hypothesized values. As with the simultaneous F test and the novel test, the regression line $y_i = b_0 + b_1\hat{y}_i$ is evaluated by testing whether $b_0 = 0$ and $b_1 = 1$. The validation data are plotted first as a scatter diagram, and the hypothesized diagonal line $y_i = \hat{y}_i$ and a vertical line through the median of $\hat{y}_i$ are drawn on the diagram (see Fig. 4). Let n_1 be the number of data points above the hypothesized line $y_i = \hat{y}_i$ and to the left of the vertical line, and let n_2 be the number of data points above the hypothesized line $y_i = \hat{y}_i$ and to the right of the vertical line. The test statistic is then calculated as

$$[13] \quad X^2 = (8/n)[(n_1 - n/4)^2 + (n_2 - n/4)^2]$$

The test statistic defined in eq. 13 follows approximately a χ^2 distribution with two degrees of freedom.

The Kolmogorov–Smirnov test

The nonparametric Kolmogorov–Smirnov (KS) test used for testing normality (see eq. 5) is also commonly used to evaluate whether the prediction errors are normally distributed with mean of zero, so that the model is unbiased and the prediction errors are symmetric around zero.

The modified KS test

The modified KS test (D_1) was developed by Stephens (1974, p. 732) and is written as

$$[14] \quad D_1 = D\left(\sqrt{n} - 0.01 + 0.85/\sqrt{n}\right)$$

where D is the KS statistic calculated in eq. 5. Stephen's modification assumes an underlying normal distribution, but allows for the true mean and variance to be unknown. Critical values of this test are given in Stephens (1974).

Table 2. Results of the four normality tests conducted on nine data sets described in Table 1.

	Shapiro–Wilk		Kolmogorov–Smirnov		Cramér – von Mises		Anderson–Darling	
Data set	W	$p < W$	D	$p > D$	W^2	$p > W^2$	A^2	$p > A^2$
I	0.9764	0.4654	0.0756	>0.1500	0.0355	>0.2500	0.2519	>0.2500
II	0.9575	0.0089*	0.0865	0.1383	0.0823	0.1979	0.6591	0.0858
III	0.9643	<0.0001*	0.0511	0.1159	0.132	0.0428*	1.0721	0.0084*
IV	0.9216	0.2036	0.1708	>0.1500	0.0674	>0.2500	0.4255	>0.2500
V	0.9816	0.5713	0.0729	>0.1500	0.0430	>0.2500	0.2703	>0.2500
VI	0.7416	0.0028*	0.3229	<0.0100*	0.1984	<0.0050*	1.0909	<0.0050*
VII	0.9211	0.0911	0.1868	0.0540	0.1303	0.0415	0.7350	0.0473
VIII	0.9096	0.1147	0.1557	>0.1500	0.0690	>0.2500	0.4848	0.2041
IX	0.8602	<0.0001*	0.1759	<0.0100*	0.5834	<0.0050*	3.2123	<0.0050*

Note: An asterisk indicates significant difference between observed and theoretical distributions ($p < 0.05$) at 95% probability level ($\alpha = 0.05$). Details about the data sets are provided in the Data section.

The sign test

The sign test is used to assess whether the median of prediction errors is zero. The test statistic is the smaller number between the number of positive prediction errors ($T+$) and the number of negative prediction errors ($T–$). Tied pairs (zero prediction errors) are excluded from the calculation. If the null hypothesis is true, $T+$ and $T–$ should be approximately the same, and the test statistic follows a binomial distribution (Daniel 1990).

The Wilcoxon signed-rank test

The Wilcoxon signed-rank test uses both the magnitudes and the signs of the differences and is more powerful than the sign test (Conover 1999). As with the sign test, ties (zero prediction errors) are excluded when computing the test statistic. The remaining n_1 prediction errors are ranked without considering the signs. The sign of each prediction error is then attached to the rank and the sums of positive ($T+$) and negative ranks ($T–$) are calculated. The smaller of the two becomes the test statistic. If the sample size is greater than 50, normal approximation is typically used:

$$[15] \quad T = \sum_{i=1}^{n_1} R_i \Big/ \sqrt{\sum_{i=1}^{n_1} R_i^2}$$

where R_i is the rank of the ith observation with the corresponding sign (Conover 1999).

Results and discussion

Table 2 lists the results of the four normality tests conducted on the nine data sets. The calculated test statistics, the p values of the critical values smaller than the test statistics for the Shapiro–Wilk test, and the p values of the critical values greater than the test statistics for the remaining three tests are provided. Smaller p values lead to the rejection of normality. In practice, the calculated p values are often compared with 0.05 (5% significance level or $\alpha = 0.05$, which is used throughout this study). If a calculated p value is smaller than 0.05, the normality hypothesis is rejected.

Results in Table 2 showed that all four tests failed to reject the hypothesis of normally distributed prediction errors for data sets I, IV, V, VII, and VIII and did reject the hypothesis of normally distributed prediction errors for data sets VI and IX. Mixed results occurred for data sets II and III. For

Fig. 5. Normal probability plot of prediction errors for the site index data (data set III). The three data points on the left lead to the rejection of normality from the Shapiro–Wilk test, the Cramér – von Mises test, and the Anderson–Darling test.
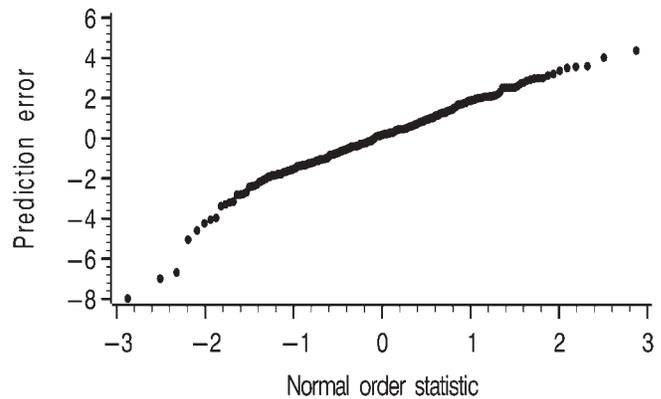
data set II, the Shapiro–Wilk test rejected the normality hypothesis, while all three other tests failed to reject it. It was found that of 81 observations, one single observation was responsible for the rejection from the Shapiro–Wilk test. Therefore, the prediction errors associated with data set II were considered normally distributed in this study. For data set III, only the KS test failed to reject the normality hypothesis, while the other three tests rejected it. A close examination of the normal probability plot shown in Fig. 5 indicates that three observations on the left led to the rejection of the normality assumption.

The KS test has been said to be extremely sensitive to slight deviations from normality and is very likely to reject the normality hypothesis even when data appear to be approximately normal (Pett 1997). However, our results suggest that the Shapiro–Wilk test is the most sensitive test. It tends to reject normality more often than the other three tests (as on data set II). On the contrary, the KS test is not that sensitive and tends to fail to reject normality when it is rejected by the other tests (as on data set III).

The tests of normality suggested that data sets I, II, IV, V, VII, and VIII have normally distributed prediction errors, whereas data sets III, VI, and IX do not. Hence, parametric tests were applied to data sets I, II, IV, V, VII, and VIII, and nonparametric tests were applied to data sets III, VI, and IX. Table 3 shows the results of these tests. The results are both

Table 3. Results of parametric and nonparametric tests conducted on data sets with normal and non-normal prediction errors.

Test	Equation	Data set								
		I	II	III	IV	V	VI	VII	VIII	IX
Paired t test	8	1.52	2.48*		0.022	11.08*		–0.22	4.51*	
χ^2 tests										
Test 1	9a	590.50*	227.76*		373.75*	62.71		78.86*	146.32*	
Test 2	9b	561.59*	211.46*		373.74*	18.91		78.67*	62.04*	
Test 3	9c	561.59*	211.44*		220.91*	17.9		78.53*	55.17*	
Separate t tests										
$b_0 = 0$		0.3	1.02		2.51*	–0.2		0.14	0.39	
$b_1 = 1$	10	–0.014	0.09		–3.00*	1.72		–0.18	1.32	
Simultaneous F test	11	1.1327	3.0496		4.497*	65.158*		0.0391	11.566*	
Novel test	12	3.6917*	3.2538*		2.777	69.228*		0.252	13.057*	
Brown–Mood test	13			19.891*			2.000			8.968*
Kolmogorov–Smirnov test	5			0.0564			0.4139*			0.1831*
Modified Kolmogorov–Smirnov test	14			0.8889			1.4160*			1.4711*
Sign test				132			7			28
Wilcoxon signed-rank test				0.9626			9.5			–0.1712

Note: An asterisk indicates significant differences between observed and predicted y values ($\alpha = 0.05$), which leads to model rejection. Prediction errors are normally distributed for data sets I, II, IV, V, VII, and VIII, and not normally distributed for data sets III, VI, and IX. Details about the data sets are provided in the Data section.

interesting and puzzling. Each model passed one or more tests, but none passed all tests. In applications, this could mean that one could apply different tests, choose the one that satisfies his or her objective, and accept or reject a model at one's disposal. This type of flexibility and the lack of a consistent outcome are very undesirable when validating a model.

In model validation, the inappropriateness of the paired t test has been discussed by a number of researchers. Freese (1960) pointed out that the paired t test "uses one form of accuracy (precision) to test for the other form (freedom from bias), frequently with anomalous results". It was used in this study because it is still applied in practice for model validation. It is important to know that while the paired t test can still be used to evaluate whether the mean prediction error is different from zero, the test itself overlooks several important aspects and cannot be relied on for determining the validity of a model (Freese 1960; Ek and Monserud 1979; Reynolds et al. 1981; Huang et al. 2003).

Freese (1960) suggested the χ^2 test as an alternative to the paired t test for testing the accuracy and determining the applicability of forestry models (see also Reynolds 1984). Of the six models tested, only one model associated with data set V was not rejected (Table 3). A more striking phenomenon occurred on data set VII. Studies done elsewhere suggested that the model fitted the data almost perfectly (West 1981; Huang et al. 2003). This was supported by all other tests except for the χ^2 test, which still indicated that the predictions were unacceptable. All these imply that the χ^2 test tend to reject valid models too often. Another limitation of the χ^2 tests is that they require a statement of an acceptable accuracy, which is often set subjectively.

Recall the optimism principle of model selection: "A model chosen by some selection process provides a much more optimistic explanation of the data used in its derivation

than it does of other data that will arise in a similar fashion" (Picard and Cook 1984). This is largely due to overfitting the idiosyncrasies of those particular data (Mosteller and Tukey 1977). It follows that a validation test result is almost certain to indicate that the model is "poorer" on an equally likely validation data set, which may lead to wrongly rejecting the model (Reynolds et al. 1981; Picard and Cook 1984).

The separate t tests and the simultaneous F test of $b_0 = 0$ and $b_1 = 1$ from the fit of $y_i = b_0 + b_1 \hat{y}_i$ also produced mixed signals when they are compared with each other and with other tests (see Table 3). For instance, both tests failed to reject models associated with data sets I, II, and VII and rejected the model associated with data set IV. However, conflicting results occurred for data sets V and VIII.

The simultaneous F test has been widely used in model validation. It is based on the rationale that if a model is a good one, the regression line between observed and predicted values should be a 45° line through the origin (see Fig. 4). This rationale, however, has been challenged by some researchers (Harrison 1990; Kleijnen et al. 1998; Kleijnen 1999), who showed that even if a model gives good predictions, it could still lead to the rejection of $b_0 = 0$ and $b_1 = 1$ for the regression line $y_i = b_0 + b_1 \hat{y}_i$. In fact, Kleijnen et al. (1998) called the simultaneous F test a "naïve test" and listed it as an "example of wrong validation". Instead, they proposed a new test, called the novel test, to overcome the problems. Mayer et al. (1994), on the other hand, used a Monte-Carlo experiment and proved that simultaneous F test is a very powerful test for model validation and is able to correctly reject invalid models and accept valid ones when prediction errors are independent of each other. The rejection rate (of a valid model) becomes much higher if prediction errors are correlated. Mayer et al. (1994) also pointed out that because the simultaneous F test evaluates the degree of relationship between the observed and predicted values by

a model, it is a lot more powerful than many other tests, which only assess the overall bias.

Our results (Table 3) suggested that the simultaneous F test and the novel test reached the same conclusions for data sets V, VII, and VIII. For data sets I, II, and IV, conflicting results occurred. The novel test rejected two models associated with data sets I and II, while the simultaneous F test failed to reject them. The model associated with data set IV was not rejected by the novel test but was rejected by the simultaneous F test. Data sets I and II are from Alberta, and the prediction equations have been carefully evaluated and proved to be sufficiently accurate for predictions in many cases. Therefore, we typically consider them as valid models. The test results shown in Table 3 appear to support the conjecture that the novel test rejects a valid model more often than the simultaneous F test. This is opposite to the observations made by Kleijnen et al. (1998). The novel test is relatively new and, as of now, not widely used in dealing with actual data. Additional research is needed to evaluate this test on other independent data sets.

The tests discussed so far are parametric tests. Parametric tests in theory can only be used for data sets that follow the assumed distribution, which is usually normal. When the normality assumption is violated, two choices are available: apply a transformation satisfying normality, or use nonparametric tests. Some parametric tests, such as the simultaneous F test, are reasonably robust under low to moderate departure from normality (Rawlings 1988). Other researchers have shown that, for example, the power of parametric tests such as the χ^2 tests proposed by Freese (1960) and Reynolds (1984) can be seriously distorted by the non-normality of the data (Gregoire and Reynolds 1988). It is therefore very important to verify the underlying assumptions before choosing a particular test for model validation. This may help to determine the "best" choices of tests and eliminate some of the less pertinent tests.

Results of the nonparametric tests on all three data sets (III, VI, and IX) with non-normal prediction errors are also shown in Table 3. It is clearly evident that mixed outputs were produced and no consistent conclusion could be reached for any one of the three models tested. Each model passed one or more tests, but not all five tests.

For data set III, only the Brown–Mood test rejected the site index model (eq. 3); the other four nonparametric tests failed to reject it. This site index model has been thoroughly evaluated and proven to be very accurate for site index predictions in Alberta. It is considered a valid model in most cases. This is evident from the results of the KS test, the modified KS test, the sign test, and the Wilcoxon signed-rank test (Table 3). The Brown–Mood test has been found to be fairly sensitive to the number of influential observations and the total number of observations in the samples (Daniel 1990). It failed to reject the model on data set VI with 10 observations, but rejected the models on data sets III and IX with 247 and 63 observations, respectively.

For data set VI, the results of the nonparametric tests in Table 3 showed that the KS test and the modified KS test led to the rejection of the model predictions, whereas the Brown–Mood test, the sign test, and the Wilcoxon signed-rank test failed to reject the model predictions. For data set IX, the Brown–Mood test, the KS test, and the modified KS

test indicated that the model was unacceptable, whereas the sign test and the Wilcoxon signed-rank test failed to reject the model. No consistent acceptance or rejection of a model was obtained from the nonparametric tests for either one of the two data sets.

The Kolmogorov–Smirnov test is historically a popular nonparametric test. In this study, both the KS test and its modified form presented by Stephens (1974) were evaluated. Both tests led to the same conclusion for all three data sets with non-normal prediction errors (Table 3). This suggests that these two tests have the same functionality in model validation and result in the same outcome. It appears unnecessary to conduct both tests at the same time, if a KS test is to be conducted at all.

The sign test and the Wilcoxon signed-rank test also produced compatible results for all three data sets (III, VI, IX): none of the three models were rejected. The Wilcoxon signed-rank test statistics for data sets III and IX were calculated using the normal approximation (eq. 15), as shown in Daniel (1990) and Conover (1999). Sprent and Smeeton (2001) pointed out that the sign test is particularly useful for skewed distributions or for small sample sizes. Although in many cases the Wilcoxon signed-rank test is more powerful than the sign test, the sign test can be more powerful under certain circumstances (Conover 1999).

A comparative look at several parametric or nonparametric tests on any specific data or model often produces mixed outcomes. This is clear in Table 3, where no consistent conclusion could be reached for any one of the six models from the parametric tests or three models from the nonparametric tests. In fact, for any model on any one of the six data sets with normal prediction errors, or for any model on any one of the three data sets with non-normal prediction errors, one could apply different parametric or nonparametric tests, accept or reject a model, by "accidentally" or "purposely" choosing one or several tests that satisfy a researcher's objectives. Obviously, this type of intended or unintended subjectiveness and arbitrariness must be removed from model validation.

Whether or not model validation needs to involve testing a specific validity hypothesis may still be open for discussion, and an agreed-upon solution may not exist (Wright 1972; Gass and Thompson 1980; Sargent 1999). Reynolds (1984) and Burk (1986) questioned the need to include specific statistical tests in validating growth and yield models and suggested using other methods instead. A. Robinson (University of Idaho, personal communication, 2001) contends that most statistical assessments of models using a sample of field data are necessarily ambiguous. A failure to accurately predict field observations could be due to flaws in the underlying structure of the model, to the difference between the population from which the model was parameterized and the population against which it is being compared, or simply to sampling error. This objection can also be made for normality tests. Problems will also arise if these data are from a limited geographical or ecological portion of the target population, as is commonly the case with long-term permanent plots (Sterba and Monserud 1997), or if the time span of the data is too short to ensure thorough assessment of long-term experimental or model behavior (Monserud 2002). Furthermore, Goodall (1972) draws the distinction between predic-

tion errors that are statistically significant yet practically unimportant.

Our results (Table 3) obtained from different data sets appeared to support Reynolds (1984) and Burk (1986). They found that the usefulness of statistical tests in model validation is very limited. Each model could either be accepted or rejected depending on the tests chosen. This lack of consistency in the outcome of the tests can be precarious in application. It could lead to the acceptance of a poor model and the rejection of a good model by accidentally or purposely choosing a test. In fact, statistical tests applied in model validation can be easily misused, misinterpreted, and misleading. The array of available tests that could potentially be used in model validation and the varied outcome from these tests made the distinction between valid and invalid models extremely difficult. It is important to exercise caution and understand the limitations of statistical tests in validating forest biometric models.

Conclusions

This study demonstrated that the usefulness of statistical tests in model validation is very limited. Over the years many statistical tests have been used in model validation, and new and "improved" tests are still being proposed. Unfortunately, not one of them seems to be generic enough to work well across a diverse range of models. It appears that testing is a relative concept, relative to a unique situation, a particular set of data, assumptions, or constraints. For this reason, some researchers have suggested using a number of tests to look at different facets of model behavior. But the results of this study showed that using various tests is hardly a solution. In fact, it may be the source of confusion.

A recommended general strategy concerning model validation should be to look at how well a model fits new, independent data, rather than use a statistical test to determine whether it is good enough, which may be different depending on the strength of the inherent relationship, the data, model types, study objectives, the "comfort level" of the individual(s) involved, and most importantly, the type of statistical tests selected to be used. Statistically testing the null hypothesis that a model is invalid will likely lead to a simple conclusion of yes or no. This may not be appropriate because model validation should be far more comprehensive than a single pass–fail test. Recall that model validation is simply an attempt to judge whether or not a model is an acceptable representation of reality. Model validation is a composite process that may involve some form of tests, but statistical test results should not be used as the sole criterion for deciding the validity of a model (Morehead 1996). Perhaps the exception would be a model that is unanimously rejected by a wide range of tests, a situation that was not found in this study. Instead, we found conflicting results, with the same model being accepted or rejected seemingly haphazardly as the test changed. The availability of an array of tests and subsequent potential choices of test results made the use of statistical tests and their outcome extremely subjective, arbitrary, and unreliable. Statistical tests should be used with caution and in combination with other validation methods.

References

Alberta Environment Protection. 1994. Natural regions and subregions of Alberta. Alberta Environment Protection, Edmonton, Alta. Publ. I/531.

Balci, O., and Sargent, R.G. 1984. A bibliography on the credibility assessment and validation of simulation and mathematical models. Simuletter, **15**(3): 15–27.

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

Buchman, R.G., and Shifley, S.R. 1983. Guide to evaluating forest growth projection systems. J. For. **81**: 232–234, 254.

Burk, T.E. 1986. Growth and yield model validation: have you ever met one you liked? *In* Data Management Issues in Forestry. Proceedings of a Computer Conference and 3rd Annual Meeting of the Forest Resources Systems Institute, 7–9 April 1986, Atlanta, Ga. *Edited by* S. Allen and T.M. Cooney. Forest Resource Systems Institute, Florence, Ala. pp. 35–39.

Conover, W.J. 1971. Practical nonparametric statistics. John Wiley & Sons, New York.

Conover, W.J. 1999. Practical nonparametric statistics. 3rd ed. John Wiley & Sons, New York.

Curtis, R.O. 1972. Yield tables past and present. J. For. **70**(1): 28–32.

D'Agostino, R.B., and Stephens, M.A. 1986. Goodness-of-fit techniques. Marcel Dekker Inc., New York.

Daniel, W.W. 1990. Applied nonparametric statistics. 2nd ed. PWS-KENT Publishing, Boston, Mass.

Desanker, P.V., Reed, D.D., and Jones, E.A. 1994. Evaluating forest stress factors using various forest growth modeling approaches. For. Ecol. Manage. **69**: 269–282.

Draper, N.R., and Smith, H. 1981. Applied regression analysis. 2nd ed. Wiley, New York.

Ek, A., and Monserud, R.A. 1979. Performance and comparison of stand growth models based on individual tree and diameter-class growth. Can. J. For. Res. **9**: 231–244.

Freese, F. 1960. Testing accuracy. For. Sci. **6**(2): 139–145.

Gass, S.I., and Thompson, B.W. 1980. Guidelines for model evaluation. Oper. Res. **28**(2): 431–479.

Goelz, J.C.G., and Burk, T.E. 1992. Development of a well behaved site index equation: jack pine in north central Ontario. Can. J. For. Res. **22**: 776–784.

Goodall, D.W. 1972. Building and testing ecosystem models. *In* Mathematical models in ecology. *Edited by* J.N.J. Jeffers. Blackwell, Oxford, U.K. pp. 173–194.

Gregoire, T.G., and Reynolds, M.R. 1988. Accuracy testing and estimation alternatives. For. Sci. **34**(2): 302–320.

Harrison, S.R. 1990. Regression of a model on real-system output: an invalid test of model validity. Agric. Syst. **34**: 183–190.

Hasenauer, H., Monserud, R.A., and Gregoire, T.G. 1998. Using simultaneous regression techniques with individual tree growth models. For. Sci. **44**(1): 87–95.

Holmes, W.M. 1983. Confidence building in simulation models as a practical process. *In* Proceedings of the 1983 Summer Computer Simulation Conference, 11–13 July 1983, Vancouver, B.C. The Society for Modeling and Simulation International, San Diego, Calif. pp. 195–199.

Huang, S. 1994. Ecologically based individual tree volume estimation for major Alberta tree species. Alberta Land and Forest Service, Edmonton, Alta. Tech. Rep. Publ. T/288.

Huang, S. 1999. Ecoregion-based individual tree height–diameter models for lodgepole pine in Alberta. West. J. Appl. For. **14**(4): 186–193.

Huang, S., and Titus, S.J. 1999. Estimating a system of nonlinear simultaneous individual tree models for white spruce in boreal mixed-species stands. Can. J. For. Res. **29**: 1805–1811.

Huang, S., Titus, S.J., and Klappstein, K. 1997. Subregion-based compatible height and site index models for young and mature stands in Alberta: revisions and summaries (Part I). Alberta Land and Forest Service, Edmonton, Alta. For. Manage. Res. Note Publ. T/389.

Huang, S., Titus, S.J., Price, D., and Morgan, D.J. 1999. Validation of ecoregion-based taper equations for white spruce in Alberta. For. Chron. **75**(2): 281–292.

Huang, S., Yang, Y., and Wang, Y. 2003. A critical look at procedures for validating growth and yield models. *In* modelling forest systems. *Edited by* A. Amaro, D.D. Reed, and P. Soares. CABI Publishing, Wallingford, U.K. pp. 271–294.

Kleijnen, J.P.C. 1999. Validation of models: statistical techniques and availability. *In* Proceedings of the 1999 Winter Simulation Conference, 5–8 December 1999, Phoenix, Ariz. *Edited by* P.P. Farrington, H.B. Nembhard, D.T. Sturrock, and G.W. Evans. Institute of Electrical and Electronics Engineers, New York. pp. 647–654.

Kleijnen, J.P.C., Bettonvil, B., and Groenendaal, W.V. 1998. Validation of trace-driven simulation models: a novel regression test. Manage. Sci. **44**(6): 812–819.

Kozak, A., and Kozak, R. 2003. Does cross validation provide additional information in the evaluation of regression models? Can. J. For. Res. **33**: 976–987.

Marshall, P., and LeMay, V. 1990. Testing prediction equations for application to other populations. School of Forestry, Wildlife Resources, Virginia Tech, Blacksburg, Va. Publ. FWS-3-90. pp. 166–173.

Marshall, P., Szikszai, T., LeMay, V., and Kozak, A. 1995. Testing the distributional assumptions of least squares linear regression. For. Chron. **71**(2): 213–218.

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.

Monserud, R.A. 2002. Large-scale management experiments in the moist maritime forests of the Pacific Northwest. Landsc. Urban Plann. **59**(3): 159–180.

Monserud, R.A. 2003. Evaluating forest models in a sustainable forest management context. For. Biom. Modell. Inf. Sci. **1**: 35–47.

Montgomery, D.C., and Peck, E.A. 1992. Introduction to linear regression analysis. John Wiley & Sons, New York.

Morehead, L.A. 1996. Determining the factors influential in the validation of computer-based problem solving systems. Ph.D. dissertation, Portland State University, Portland, Ore. [University Microfilms UMI No. 9628864.]

Mosteller, F., and Tukey, J.W. 1977. Data analysis and regression. Addison-Wesley, Reading, Mass.

Neter, J., Wasserman, W., and Kutner, M.H. 1989. Applied linear regression models. 2nd ed. Irwin, Chicago, Ill.

Pett, M.A. 1997. Nonparametric statistics for health care research. Sage Publications, Thousand Oaks, Calif.

Picard, R.R., and Cook, R.D. 1984. Cross-validation of regression models. J. Am. Stat. Assoc. **79**(387): 575–583.

Popper, K.R. 1963. Conjectures and refutations. Routledge and Kegan Paul, London.

Rawlings, J.O. 1988. Applied regression analysis: a research tool. Wadsworth, Pacific Grove, Calif.

Reynolds, M.R. 1984. Estimating the error in model predictions. For. Sci. **30**(2): 454–468.

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

Reynolds, M.R., Jr., Burkhart, H.E., and Daniels, R.F. 1981. Procedures for statistical validation of stochastic simulation models. For. Sci. **27**: 349–364.

Ritchie, M.W., and Hann, D.W. 1997. Implications of disaggregation in forest growth and yield modeling. For. Sci. **43**(2): 223–233.

Robinson, A.P. 1998. Forest ecosystem dynamics: a systematic approach to modelling in a model-rich environment. Ph.D. dissertation, University of Minnesota, St. Paul, Minn.

Robinson, A.P., and Monserud, R.A. 2003. Criteria for comparing the adaptability of forest growth models. For. Ecol. Manage. **172**: 53–67.

Sargent, R.G. 1999. Validation and verification of simulation models. *In* Proceedings of the 1999 Winter Simulation Conference, 5–8 December 1999, Phoenix, Ariz. *Edited by* P.A. Farrington, H.B. Nembhard, D.T. Sturrock, and G.W. Evans. Institute of Electrical and Electronics Engineers, New York. pp. 39–48.

Scheaffer, R.L., and McClave, J.T. 1995. Probability and statistics for engineers. Duxbury Press, Belmont, Calif.

Shugart, H.H. 1984. A theory of forest dynamics. Springer-Verlag, New York.

Snedecor, G.W., and Cochran, W.G. 1980. Statistical methods. 7th ed. Iowa State University Press, Ames, Iowa.

Snee, R.D. 1977. Validation of regression models: methods and examples. Technometrics, **19**(4): 415–428.

Sprent, T., and Smeeton, N.C. 2001. Applied nonparametric statistical methods. 3rd ed. Chapman & Hall, New York.

Stephens, M.A. 1974. EDF statistics for goodness-of-fit and some comparisons. J. Am. Stat. Assoc. **69**: 730–737.

Sterba, H., and Monserud, R.A. 1997. Applicability of the forest stand growth simulator PROGNAUS for the Austrian part of the Bohemian Massif. Ecol. Modell. **98**: 23–34.

Vanclay, J.K., and Skovsgaard, J.P. 1997. Evaluating forest growth models. Ecol. Modell. **98**: 1–12.

Waller, L.A., Smith, D., Childs, J.E., and Real, L.A. 2003. Monte Carlo assessments of goodness-of-fit for ecological simulation models. Ecol. Modell. **164**: 49–63.

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

Wright, R.D. 1972. Validating dynamic models: an evaluation of tests of predictive powers. *In* Proceedings of 1972 Summer Computer Simulation Conference, 13–16 June 1972, San Diego, Calif. Simulation Councils, La Jolla, Calif. pp. 1286–1296.

Yaussy, D.A. 2000. Comparison of an empirical forest growth and yield simulator and a forest gap simulator using actual 30-year growth from two even-aged forests in Kentucky. For. Ecol. Manage. **126**: 385–398.