Does cross validation provide additional information in the evaluation of regression models?

Antal Kozak and Robert Kozak

Abstract: A detailed study using seven data sets, two standing tree volume estimating models, and a height–diameter model showed that fit statistics and lack of fit statistics calculated directly from a regression model can be well estimated using simulations of cross validation or double cross validation. These results suggest that cross validation by data splitting and double cross validation provide little, if any, additional information in the process of evaluating regression models.

Résumé : Une étude détaillée basée sur sept séries de données, deux modèles d'estimation du volume sur pied et un modèle de diamètre–hauteur montre que les statistiques d'ajustement ou d'absence d'ajustement calculées directement à partir d'un modèle de régression peuvent être adéquatement estimées à l'aide des simulations de validation croisée simple ou double. Ces résultats suggèrent que la validation croisée simple en scindant les données et la validation croisée double fournissent peu sinon aucune information additionnelle dans le processus d'évaluation des modèles de régression.

[Traduit par la Rédaction]

Introduction

One of the main objectives of regression analysis is to select a model that "best" predicts a dependent variable. The last and perhaps most important step in regression analysis is to carry out a thorough evaluation of the selected model. Two procedures are normally used in this process, both of which are based on an examination of prediction errors or fit statistics calculated from the ordinary residuals $y_i - \hat{y}_i$, where y_i and $\hat{y}_i$ are the observed and predicted values of the dependent variable, respectively. In the first procedure the calculation of the predicted errors is based on all of the n observations that were used in constructing the model. In the second procedure — cross validation — the available observations in the data set, n, are split into two groups such that $n = n_c + n_v$, where n_c is the number of observations used for model construction, and n_v is the number of observations used for model validation or calculation of prediction errors.

Many statisticians claim that the first procedure does not provide an acceptable indication of the predictive ability of a given model because prediction errors, calculated in this manner, are not independent of the data used to fit the model (e.g., Stone 1974; Snee 1977; Berk 1984; Ronchetti et al. 1997; Shao 1993). Although cross validation, by its very nature, addresses the issue of independence of the prediction errors, some statisticians have questioned the theoretical and methodological foundations of data splitting (e.g., Picard and Cook 1984). In addition, two distinct disadvantages of cross validation have been noted by Picard and Cook (1984). The first involves the accompanying loss of information in model development, that is, the model being validated is based on fewer observations than if all of the data were used in model development. Second, there are concerns about the stability of the validation estimates, mainly because the prediction errors estimated by cross validation are naturally based on fewer observations than if all data were used.

In this paper, we further question the practical utility of data splitting or cross validation in regression model development. Specifically, we will attempt to show that prediction errors and any other fit statistics based on these procedures provide little, if any, incremental information compared with calculating prediction errors and other fit statistics from an entire data set. Given that validation techniques have been widely used in forestry model development for over four decades, we recognize that these objectives may be viewed as somewhat contentious. To that end, we have performed Monte Carlo simulations using two well-known tree volume estimation equations and a height–diameter equation to substantiate our assertions. The two main objectives of this project were (*i*) to describe the relationships between prediction errors calculated from the entire data set used for model building and the estimates of the prediction errors calculated from two cross-validation procedures, and (*ii*) to investigate the effects of varying data set sizes (both for model building and validation) on the aforementioned relationships.

The reader should be cautioned that the scope of this paper is limited to models for which statistics (both fit statistics and lack of fit statistics) can be calculated analytically from the entire data set.

Received 8 April 2002. Accepted 7 January 2003. Published on the NRC Research Press Web site at http://cjfr.nrc.ca on 28 April 2003.

A. Kozak.[1] Department of Forest Resources Management, Faculty of Forestry, The University of British Columbia, 2045-2424 Main Mall, Vancouver, BC V6T 1Z4, Canada.
R. Kozak. Department of Wood Science, Faculty of Forestry, The University of British Columbia, 4041-2424 Main Mall, Vancouver, BC V6T 1Z4, Canada.

[1]Corresponding author (e-mail: kozak@interchange.ubc.ca).

 doi: 10.1139/X03-022

Background

Cross validation

According to Snee (1977), there are four procedures available for the validation of regression models: (*i*) a comparison of predictions and coefficients with physical theory; (*ii*) a comparison of results with those obtained by theory and simulation; (*iii*) the use of new data; and (*iv*) the use of data splitting or cross validation. Of these, the last two provide independent prediction residuals and are, therefore, preferred by most practitioners. Since the use of new data for model validation is frequently neither practical nor feasible, data splitting is regarded as an acceptable alternative, provided that the data set is large enough. Because of its popularity, a large number of publications addressing various procedures for cross validation have appeared in the literature in the past four decades (e.g., Stone 1974; Snee 1977; Berk 1984; Burk 1990; Ronchetti; Field and Blanchard 1997; Shao 1993).

In cross validation, the number of available observations, n, is split into two groups such that $n = n_c + n_v$; n_c is the number of observations used for model construction, and n_v is the number of observations used for model validation or calculation of prediction residuals (prediction error). Various recommendations are made in the literature both on how to split data into two groups and on what the size of the validation data set relative to the construction data set should be. Most frequently, the validation data set is set aside randomly, although formal algorithms are available to select the validation data set in such a way that it matches, as closely as possible, the distribution of the construction data set (Snee 1977). Usually, between 10 and 50% of the available observations are used for validation purposes.

Double cross validation is generally regarded as an improved version of cross validation, in which the total number of observations, n, is divided into k equal subsets, and $(k-1)/k$ portion of the data is used for model construction, while $1/k$ portion of the data is used for validation. This is repeated k times so that each subset is used for validation as well as for model fitting. Note that the condition of $(k-1)/k \geq 0.5$ must be met. An extreme case occurs when $k = n$, so that each iteration of the model is derived from $n-1$ observations, and the validation data set contains only one observation. This special case of double cross validation, widely used in model selection, results in what is known as the prediction sum of squares (PRESS) statistic (Shao 1993; Stone 1974; Picard and Cook 1984).

Fit statistics, prediction errors, and lack of fit statistics

To evaluate the ability of a model to predict a dependent variable (y), it is customary to calculate one or more statistics, based on the residuals in various forms, such as

(1) Residual sum of squares (SS_{RES}):

$$[1] \qquad SS_{RES} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where n is the number of observations, y_i is an observation of the dependent variable, and $\hat{y}_i$ is the predicted value of a given observation of the dependent variable. The residual sum of squares is frequently used in the form of the residual mean square (MS_{RES}) or residual variance:

$$[2] \qquad MS_{RES} = \frac{SS_{RES}}{n-m-1}$$

where m is the number of independent variables in the model. The residual sum of squares can also be used in the form of square root residual mean square (standard error of estimate). In this paper, we will use the expression of standard error of estimate (SEE), which is

$$[3] \qquad SEE = \sqrt{MS_{RES}}$$

(2) Coefficient of determination (R^2):

$$[4] \qquad R^2 = \frac{SS_{REG}}{SS_y} = 1 - \frac{SS_{RES}}{SS_y}$$

where SS_{REG} is the regression sum of squares, and SS_y is the corrected sum of squares of the dependent variable, calculated as

$$SS_y = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

where $\bar{y}$ is the mean of the dependent variable.

(3) Average absolute bias (AB):

$$[5] \qquad AB = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

(4) Average bias (B):

$$[6] \qquad B = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)}{n}$$

If calculated for an entire data set, the above statistics are usually referred to as fit statistics. If calculated for an independent data set (e.g., validation data), they are referred to as prediction statistics or prediction errors. It should be noted that when the model is derived by the ordinary least squares (OLS) procedure, by definition the average bias must be zero when computed for the entire data set. It is, therefore not a suitable measure of fit in this case. Also, if the coefficient of determination is not calculated from residuals based on OLS estimates, it is usually symbolized by I^2 (correlation index squared) instead of R^2.

Although statistics like average bias, average absolute bias, residual mean square, standard error of estimate, or coefficient of determination are good indicators of the effectiveness of a model, they do not necessarily single out the "best" model for prediction of the dependent variable. For selecting the most suitable regression model, it is generally advisable to use some measure of lack of fit in combination with one or more of the above statistics, e.g., plotting the residuals. Better yet, calculating biases and standard errors of estimate (or residual mean squares) for various subgroups of the independent variable(s), usually referred to as lack of fit statistics, is preferred. For multiple regression models with several independent variables, it is feasible and advisable to present the lack of fit statistics as a function of the dependent variable by constructing classes for the predicted dependent

variable. Unfortunately, in both practice and in the literature, it is not uncommon to see only measures of fit statistics being used to evaluated a model and to compare several models (Burk 1990). An example illustrating the importance of using lack of fit statistics will be presented in this paper.

Methods

Two models for total tree volume (V) estimation of standing trees from diameter at breast height (D or DBH) and total tree height (H) were used in this study. These are Schumacher's logarithmic volume equation (Schumacher and Hall 1933) and the combined variable volume equation (Spurr 1952), respectively,

Model 1:

$$\hat{V} = aD^bH^c$$

Model 2:

$$\hat{V} = b_0 + b_1D^2H$$

To compare the two procedures, model 2 was fitted by OLS, while model 1 was fitted using logarithmic transformations as recommended by Schumacher and Hall (1933). As a result, the fit statistics provided for model 1 are not based on least square estimates of the residual sum of squares. The two volume models contrasted here were selected only because they are well known to practicing foresters and forest scientists. It should be noted that the purpose of this paper was not to uncover or select a better volume equation, but rather to demonstrate the use of cross validation in evaluating these models. The points made in this paper can be applied to any simple, multiple, linear, or nonlinear regression model.

In addition, a height–diameter equation, which had a somewhat lower R^2 value (approximately 0.8) than the above two volume estimating models (greater than 0.95 for both), was studied. The purpose of including this model was simply to investigate the effect of coefficient of determination on cross validation. Because of its common use in practice, the selected model was

Model 3:

$$\hat{H} = b_0 + b_1D + b_2D^2$$

The data used in this study were provided by Resources Inventory Branch of the Ministry of Forests of British Columbia (presently: Vegetation Resources Inventory, Terrestrial Information Branch, Ministry of Sustainable Resource Management), who manage a data bank containing information on thousands of trees that is used in the development of volume and taper equations. For this study, 1000 coastal Douglas-fir (*Pseudotsuga menziesii* (Mirb.) Franco) trees were randomly selected from the data bank. The selected trees were characterized by DBHs ranging from 5.4 to 216.4 cm with a mean of 57.8 cm, heights ranging from 6.28 to 76.72 m with a mean of 34.49 m, and volumes ranging from 0.0074 to 77.3126 m^3 with a mean of 4.9404 m^3. To study the effects of data set size on cross validation, seven sets of size 30, 50, 70, 100, 200, 500, and 1000 were created such that trees were randomly selected from the largest (1000) set for the six smaller data sets and placed into separate files.

Initially, the two volume models were fit for each of the seven data sets, and the following fit statistics were calculated for all of the resulting 14 models: average bias, average absolute bias, residual mean square, standard error of estimate, and coefficient of determination. In addition, two lack of fit statistics were calculated for the 14 models by 10 cm DBH classes for trees below 95 cm, and by 20 cm DBH classes for trees above 95 cm: average bias and standard error of estimate.[2] Since a limited number of trees were present in some of the extreme DBH classes, the residual sum of squares term in the standard errors of estimate for the DBH classes and for all trees were calculated by dividing by n, instead of $n - m - 1$, as given in eq. 2.

Next, Monte Carlo simulation was used to estimate the same fit statistics and lack of fit statistics listed above for the two volume models using cross validation. In this simulation study, each of the seven data sets was subdivided such that 20% (commonly 10 to 50% is used in practice) of the set was randomly put aside for validation, and the remaining 80% of the set was used for model fitting. This process of subdivision was repeated 1000 times for each data set. For each of the 1000 simulations, a separate model was fitted (from 80% of the data set), and fit statistics and lack of fit statistics for each model were computed using the remaining 20% of the data. Next, averages of the 1000 simulated fit statistics and lack of fit statistics were computed. By invoking the central limit theorem, these averages might be considered "estimates" of the fit statistics and lack of fit statistics obtained for models derived from the entire data set.

In the second phase of the study, the effect of the validation data set size was investigated. For each of the seven data sets, 10, 20, 30, 40, and 50% of the data points were randomly put aside for validation, with the respective remainders being used for model fitting. In each case, the same fit statistics and lack of fit statistics were estimated from 1000 simulations, as described above.

In the third phase of the study, double cross validation was used to estimate the fit statistics and lack of fit statistics, using 10, 20, 33.3, and 50% of each of the seven data sets for validation, with the remainder being used for model fitting. As a special case of the double cross validation, the PRESS procedure was also used to estimate the fit statistics and the lack of fit statistics. For each case, averages of the same fit statistics and lack of fit statistics were estimated from 1000 simulations, as described above.

Lastly, the process of model fitting and simulations outlined above for the two volume models were repeated for model 3, the height–diameter equation.

Results

In the interest of simplifying the tables, lack of fit statistics in this study are presented only for DBH classes and not

[2] Since the number of trees above 95 cm DBH in the data set was limited, wider DBH classes were used for trees with a DBH greater than 95 cm.

Table 1. Average biases and standard errors of estimate (SEE) (both in cubic metres and in percent) by DBH class for two models frequently used in volume estimation.

		Model 1				Model 2			
DBH class (cm)	Frequency	Bias (m^3)	Bias (%)	SEE (m^3)	SEE (%)	Bias (m^3)	Bias (%)	SEE (m^3)	SEE (%)
0.1–15.0	54	–0.001	–2.3	0.007	10.8	–0.282	–447.1	0.282	447.5
15.1–25.0	120	0.006	2.2	0.032	12.1	–0.221	–83.4	0.226	83.4
25.1–35.0	101	–0.003	–0.5	0.073	10.8	–0.136	–20.2	0.170	25.1
35.1–45.0	133	0.021	1.5	0.146	10.9	–0.010	–0.8	0.186	14.0
45.1–55.0	115	0.023	1.1	0.255	12.0	0.061	2.9	0.326	15.4
55.1–65.0	139	0.016	0.5	0.375	11.6	0.149	4.6	0.481	14.8
65.1–75.0	98	0.049	1.1	0.466	10.2	0.232	5.1	0.564	12.3
75.1–85.0	59	–0.115	–1.9	0.764	12.6	0.082	1.4	0.789	13.0
85.1–95.0	50	–0.100	–1.2	0.985	11.6	0.170	2.0	1.109	13.0
95.1–105.0	35	0.133	1.2	0.978	8.6	0.346	3.0	1.059	9.3
105.1–125.0	44	–0.046	–0.3	1.861	11.8	0.106	0.7	1.985	12.6
125.1–145.0	27	0.045	0.2	3.445	15.0	–0.677	–3.0	3.383	14.8
145.1–165.0	13	1.172	3.6	4.262	12.9	–0.403	–1.2	4.289	13.0
<165.1	12	3.779	7.6	5.829	11.7	–0.021	–0.1	3.604	7.2
All trees	1000	0.066	1.3	1.134	22.9	0.000	0.0	1.053	21.3

Note: Model 1, $\hat{V} = 0.000\,0435\ D^{1.696}H^{1.187}$; $I^2 = 0.9799$; $MS_{RES} = 1.2898$; average absolute bias = 0.46. Model 2, $\hat{V} = 0.3045 + 0.000\,0213\ D^2H$; $R^2 = 0.9826$; $MS_{RES} = 1.1109$; average absolute bias = 0.53.

for total tree height (H) classes, the other independent variable in both models 1 and 2.[3]

Table 1 provides the regression coefficients for models 1 and 2 based on fitting all 1000 trees, as well as some of the more important fit statistics (R^2 or I^2 values, residual mean squares, standard errors of estimate, and average absolute biases) and lack of fit statistics (average and percent bias and average and percent standard error of estimate by DBH class). This example aptly demonstrates that the sole use of overall fit statistics (values below the table and in the "all trees" row) is oftentimes misleading. The higher R^2 and lower MS_{RES} (or SSE) values for model 2 would, for many, lead to the conclusion that it is superior to model 1. However, a closer examination of the biases and standard errors of estimate for model 2 reveals that there is a significant overestimation of the volumes for trees with a DBH of between 0 and 35 cm (close to one-third of the data set). There is also some, albeit small, overestimation of the volumes for the trees with a DBH greater than 125 cm, while trees in the middle of the data set (DBH-wise) are generally underestimated. It is true that model 1 underestimates very large trees, but this only affects 12 to 25 trees at most. Furthermore, while model 2 demonstrates a clear lack of fit and a tendency to bias in one direction, the positive and negative biases randomly alternate for model 1. Most forest mensurationists should select model 1 to be the superior predictive model, both because of its acceptable fit statistics and its superior lack of fit statistics.

Therefore, although all the analyses were completed for all three models, the results presented largely focus on model 1. Table 2 shows average biases and standard errors of estimate by DBH class and by data set size, calculated from data sets of various sizes in their entirety, and estimated from 1000 simulated cross validation data sets using 20% of the total data set observations. The results clearly indicate that estimates of bias and standard error of estimate from the cross validation data sets well approximate the values obtained using the entire data set. In addition, as the sample size of the data sets increase, the estimates of these lack of fit statistics improve. For the purposes of comparison, Tables 3 and 4 summarize the corresponding results for models 2 and 3, respectively. However, only four sample sizes are given, since the results by data set size were notably very similar for all three models.

Table 5 illustrates the effect of cross validation data set size on the estimation of average biases and average standard errors of estimate by DBH class based on 1000 simulations. Again, because of similarities among the results of the three models and the seven data set sizes, results are presented only for model 1 (the superior predictive model) and for data set sizes of 100 and 1000. These results indicate that changing the validation data set size from 10 to 50% of the total observations in the data set has little or no affect on the estimation of the various lack of fit statistics, especially for the larger data set size of 1000. Results for data set size 100 are somewhat problematic in so much as the two highest DBH classes contain only two trees and one tree, respectively. Thus, the estimates in these two classes are not very reliable or stable.

Table 6 summarizes the results of the double cross validation study, including the PRESS procedure. Results are presented for model 1 and for data set sizes of 100 and 1000 only. Like the results for cross validation seen in Tables 2, 3, and 4, it can be observed that the estimations of biases and standard errors of estimate using double cross validation are reliable for data set size 1000 and acceptable for data set size 100. Again, some of the problems in estimation for the data set size 100 stem from the fact that the two highest DBH classes contain very few trees. It is interesting to note that only negligible gains are made by using an adequate

[3] Results pertaining to lack of fit statistics in this study are very similar for both DBH and total tree height classes.

Table 2. Average biases (m^3) and standard errors of estimates (SEE, m^3) for model 1 calculated from data sets of various sizes (n) and

	n = 30				n = 50				n = 70				n = 100	
	Bias		SEE		Bias		SEE		Bias		SEE		Bias	
DBH class (cm)	E	S	E	S	E	S	E	S	E	S	E	S	E	S
0.1–15.0	0.002	0.002	0.006	0.009	0.000	0.001	0.005	0.007	0.003	0.003	0.004	0.005	–0.002	–0.002
15.1–25.0	–0.011	–0.016	0.025	0.033	0.001	0.002	0.025	0.028	0.005	0.005	0.018	0.019	0.010	0.009
25.1–35.0	–0.015	–0.016	0.015	0.018	–0.019	–0.020	0.026	0.028	–0.038	–0.041	0.051	0.055	0.007	0.007
35.1–45.0	–0.014	–0.014	0.059	0.066	–0.025	–0.025	0.059	0.061	0.056	0.047	0.160	0.160	–0.053	–0.055
45.1–55.0	–0.141	–0.151	0.156	0.169	–0.072	–0.067	0.119	0.120	–0.010	–0.018	0.181	0.190	–0.029	–0.031
55.1–65.0	0.183	0.200	0.401	0.442	0.117	0.132	0.363	0.390	0.005	–0.011	0.208	0.219	0.168	0.169
65.1–75.0	0.025	0.015	0.505	0.546	0.080	0.088	0.452	0.484	–0.030	–0.051	0.346	0.365	–0.035	–0.028
75.1–85.0	–0.215	–0.317	0.583	0.676	0.083	0.087	0.609	0.654	0.487	0.493	1.005	1.018	–0.000	0.008
85.1–95.0	–0.324	–0.370	0.378	0.454	–0.367	–0.380	0.422	0.447	0.029	–0.044	0.840	0.884	–0.402	–0.422
95.1–105.0	–0.490	–0.560	0.490	0.583	–0.356	–0.361	0.356	0.376	–0.447	0.465	1.063	1.078	0.260	0.199
105.1–125.0	0.000	0.000	0.000	0.000	0.729	0.879	1.778	1.946	0.039	0.020	0.872	0.914	0.447	0.460
125.1–145.0	3.184	3.651	3.184	3.659	0.820	0.648	2.750	2.963	0.532	0.557	0.659	0.734	0.043	–0.065
145.1–165.0									0.225	0.225	0.225	0.384	4.884	5.158
>165.1									3.266	3.523	3.266	3.552		
All trees	0.077	0.072	0.679	0.769	0.051	0.054	0.723	0.769	0.051	0.046	0.648	0.682	0.068	0.066
Difference*		0.005		0.090		0.003		0.046		0.005		0.034		0.002

Note: E, entire data set. Simulation (S) results are based on 1000 validation data sets of size 20% of the total data sets.
*Difference between data set and simulation values for all trees.

proportion (between 10 and 50%) of the data set for validation compared with using single observations for validation $n - 1$ times (PRESS), especially for the larger data set size of 1000.

Tables 7, 8, and 9 differ from previous results in that they provide comparisons of three overall fit statistics — absolute bias, coefficient of determination, and residual mean square — for all three models under various conditions. Table 7 shows fit statistics calculated from the total data sets of varying sizes and estimated from 1000 simulations using 20% of the data set for cross validation. Similar to previous findings (Tables 2, 3, and 4), the estimation of each statistic improves with increasing data set size. This phenomenon is more pronounced in model 3, which has a lower R^2 value than the other two models. Table 8 illustrates the effect of cross validation data set size on the estimation of the three fit statistics using data set of sizes of 100 and 1000. While the estimations seem acceptable for both data set sizes, they do improve with increasing cross validation data set size as well as increasing data set size. Again, these improvements are more pronounced for model 3. Table 9 summarizes the effect of double cross validation data size (including PRESS) on the estimation of the three fit statistics, again using data set sizes of 100 and 1000. Previous findings (Table 6) are verified, as the averages of the fit statistics calculated from double cross validations estimate the fit statistics calculated from the total data set very well. This is true for both data set sizes, although the estimations are somewhat better for the larger data set size of 1000.

This analysis raises a very interesting question. How often would fit statistics (average bias, average absolute bias, standard error of estimate, and coefficient of determination) suggest the use of one model, while lack of fit statistics (average biases and standard error of estimates by DBH classes) suggest the use of another? Table 10 summarizes the results of 100 simulations that show the proportion of times that model 1 and model 2 would be selected by calculating fit statistics and lack of fit statistics using cross validation. These simulations were carried out for data set sizes of 70, 100, and 1000 using 20% of the data set for validation. Simulations for data set sizes less than 70 were not attempted because lack of fit statistics for 10 or less observations (validation data set size) are very difficult, if not impossible, to evaluate. The results clearly indicate that the superiority of a model is highly dependent on whether fit statistics or lack of fit statistics are utilized in model evaluation. In this case, lack of fit statistics clearly indicate that model 1 is superior to model 2. However, the results are somewhat less evident with the sole use of fit statistics. Depending on the data set size and the specific fit statistic employed, there is an approximately equal probability of selecting model 1 and model 2.

The standard deviations (which could be called standard errors) of the estimated average biases and standard errors of estimate were also calculated from simulations of the three models used in this analysis. Because these statistics do not play an important role in our interpretation of the results, Table 11 summarizes the results of each DBH class for data set sizes of 30, 70, 100, and 1000 only. As expected, these statistics decrease with increasing data set sizes (and validation data set sizes), with the exception of the spread of the standard errors of estimate for the data set size of 100. This anomalous result can be explained by an increased proportion of large trees in the data set size of 100.

Lastly, in an attempt to substantiate the findings summarized above for the three models, a similar study was conducted on Kozak's (1988) variable-exponent taper model, which is a much more complicated model than the ones reported here, but also widely used in forestry applications. Taper models estimate diameters inside bark at given heights from ground, as well as log volumes, total tree volumes, and merchantable heights. Although the results obtained by applying the more complicated taper model are not shown in this paper, they support well the findings of this study. Simulations using cross validation and double cross validation es-

from 1000 randomly selected validation data sets of size 0.2*n*.

		n = 200				*n* = 500				*n* = 1000			
SEE		Bias		SEE		Bias		SEE		Bias		SEE	
E	S	E	S	E	S	E	S	E	S	E	S	E	S
0.006	0.007	–0.001	–0.001	0.007	0.008	–0.001	–0.001	0.006	0.006	–0.001	–0.001	0.007	0.007
0.022	0.023	0.002	0.003	0.028	0.029	0.006	0.007	0.031	0.031	0.006	0.006	0.032	0.032
0.050	0.050	0.021	0.020	0.066	0.067	–0.015	–0.015	0.082	0.081	–0.003	–0.003	0.073	0.073
0.098	0.101	0.043	0.045	0.103	0.105	0.040	0.039	0.154	0.155	0.021	0.022	0.146	0.146
0.147	0.151	–0.011	–0.016	0.208	0.211	0.029	0.030	0.300	0.302	0.023	0.023	0.255	0.255
0.432	0.444	0.008	0.003	0.391	0.393	–0.037	–0.037	0.310	0.310	0.016	0.018	0.375	0.375
0.390	0.401	0.024	0.026	0.514	0.524	0.061	0.063	0.473	0.475	0.049	0.044	0.466	0.470
0.469	0.475	–0.425	–0.424	0.664	0.674	0.063	0.052	0.841	0.840	–0.115	–0.144	0.764	0.769
0.812	0.836	0.168	0.164	0.993	1.001	–0.163	–0.176	0.900	0.917	–0.100	–0.100	0.985	0.987
0.986	0.979	–0.010	0.060	1.013	1.017	0.043	0.064	0.914	0.922	0.133	0.123	0.978	0.985
1.372	1.425	0.715	0.731	1.678	1.720	–0.379	–0.384	1.939	1.928	–0.046	–0.068	1.861	1.882
2.532	2.670	0.441	0.315	4.113	4.112	0.039	–0.031	3.253	3.336	0.045	0.063	3.445	3.443
4.884	5.179	0.576	0.765	3.917	4.139	2.030	2.068	5.102	5.166	1.172	1.165	4.262	4.278
		0.436	0.384	3.608	3.694	5.142	5.347	6.434	6.629	3.799	3.810	5.829	5.873
0.771	0.801	0.055	0.055	1.146	1.176	0.089	0.092	1.234	1.263	0.066	0.066	1.134	1.142
	0.030		0.000		0.030		0.003		0.029		0.000		0.008

timated the model fit statistics and lack of fit statistics well for all four dependent variables.

Discussion

The main objective of this study was to show that fit statistics and lack of fit statistics obtained from cross validation or double cross validation provide little, if any, additional information in comparison with those obtained from a model containing all of the available data points. Since exact distributional results of the statistics calculated from validation processes are not readily available in the literature and would be extremely difficult to obtain, Monte Carlo simulations were used to study this problem. Specifically, since the sole use of overall fit statistics can oftentimes lead to spurious conclusions (Table 1), simulations were used to produce both overall fit statistics as well as lack of fit statistics for two well-known tree volume equations and one height equation. The overall fit statistics studied included average bias, average absolute bias, coefficient of determination, residual mean square, and standard error of estimate. The lack of fit statistics studied included average bias and standard error of estimate categorized by subgroups of DBH, one of the independent variables. These two lack of fit statistics were selected because they are both meaningful and diagnostically useful in determining the predictive power of regression models — average bias indicates trends in lack of fit, and standard error of estimate indicates the extent of the spread of the residuals. Although these lack of fit statistics are presented as a function of the independent variable (DBH) in this study, they can be shown as a function of the predicted or actual (measured) dependent variable for multiple regression models containing several independent variables.

The results of this study decisively show that for all models studied, the averages of the various fit statistics and lack of fit statistics calculated from 1000 cross validations (Tables 2, 3, 4, and 7) approximate the values that can readily be calculated from the original model, especially for data sizes 70 and over, with improving approximations as the sizes of data sets increase. This result clearly points to the fact that the practical utility of cross validation is questionable, in spite of its ability to calculate prediction errors independently from the data used for model development.

Without question, OLS has the properties to produce regression models that will result in residual mean squares smaller than the validation residual mean squares. It follows, then, that the validation coefficients of determination (eq. 4) should always be less than (or equal to) the model coefficients of determination. Consistent with this, our results indicate that the model fit statistics and lack of fit statistics were generally superior to the prediction fit statistics and prediction lack of fit statistics. This means that the biases and absolute biases for the models were closer to zero than for the validations and that their standard errors of estimate were smaller. Interestingly, these findings were also true for model 1, which was not fit by OLS.

Our findings support the concerns regarding a loss of information raised by Picard and Cook (1984). Specifically, since cross validation models are based on subsets of entire data sets, they will not be as reliable as the overall models that they are validating. Myers (1986) maintains that problems related to this loss of information may be remedied, at least in part, by fitting the final model using the entire data set once the selection of the appropriate model by cross validation is completed. However, according to Picard and Cook (1984), data splitting results in fit statistics and lack of fit statistics calculated from the validation portion of the data that are not as stable as if they were calculated from entire data set. Since n_v is a subset of n, this statement is obvious.

Table 10 clearly shows that if cross validation is used to choose between two models, the decision will not be as reliable as the decision based on the overall model. That said, the decision to use cross validation improves with an increasing data set size (or validation data size). The statistics

Table 3. Average biases (m^3) and standard errors of estimates (SEE, m^3) for model 2 calculated from data sets of various sizes (n) and from 1000 randomly selected validation data sets of size 0.2n.

	n = 30				n = 70				n = 100				n = 1000			
	Bias		SEE		Bias		SEE		Bias		SEE		Bias		SEE	
DBH class	E	S	E	S	E	S	E	S	E	S	E	S	E	S	E	S
0.1–15.0	–0.098	–0.124	0.098	0.152	–0.347	–0.354	0.347	0.356	–0.222	–0.232	0.223	0.238	–0.282	–0.283	0.282	0.284
15.1–25.0	–0.074	–0.084	0.074	0.114	–0.273	–0.275	0.276	0.280	–0.180	–0.184	0.181	0.191	–0.221	–0.221	0.226	0.227
25.1–35.0	–0.042	–0.063	0.042	0.101	–0.259	–0.265	0.269	0.276	–0.089	–0.094	0.101	0.114	–0.136	–0.136	0.170	0.171
35.1–45.0	0.035	0.024	0.084	0.105	0.014	0.009	0.185	0.192	–0.054	–0.056	0.101	0.108	–0.010	–0.010	0.186	0.187
45.1–55.0	–0.089	–0.101	0.098	0.117	0.073	0.084	0.275	0.287	–0.022	0.021	0.136	0.141	0.061	0.060	0.326	0.324
55.1–65.0	0.302	0.314	0.539	0.558	0.012	0.024	0.353	0.359	0.297	0.296	0.535	0.543	0.149	0.150	0.481	0.482
65.1–75.0	0.145	0.194	0.618	0.665	0.310	0.323	0.445	0.459	0.104	0.108	0.432	0.443	0.232	0.228	0.561	0.565
75.1–85.0	–0.235	–0.266	0.506	0.550	0.817	0.881	1.305	1.348	0.178	0.187	0.489	0.500	0.082	0.083	0.789	0.797
85.1–95.0	–0.701	–0.745	0.716	0.794	0.307	0.309	0.641	0.658	–0.364	–0.375	0.817	0.820	0.170	0.168	1.109	1.108
95.1–105.0	–1.175	–1.304	1.175	1.324	–0.091	–0.069	0.862	0.873	0.142	0.122	0.983	1.011	0.346	0.338	1.059	1.065
105.1–125.0	0.000	0.000	0.000	0.000	0.106	0.022	1.143	1.179	–0.020	–0.029	1.170	1.262	0.106	0.085	1.985	2.001
125.1–145.0	1.071	2.072	1.071	2.090	0.445	0.484	1.160	1.287	–1.147	–1.227	3.025	3.442	–0.677	–0.658	3.383	3.391
145.1–165.0					–1.470	–1.792	1.470	1.814	1.327	2.021	1.327	2.198	–0.403	–0.421	4.289	4.348
>165.1					–0.083	–0.230	0.083	0.662					–0.021	–0.037	3.604	3.728
All trees	0.000	0.025	0.512	0.642	0.000	–0.005	0.607	0.645	0.000	0.004	0.654	0.727	0.000	–0.002	1.053	1.065
Difference*		0.025		0.130		–0.005		0.038		0.004		0.073		–0.002		0.012

Note: E, entire data set. Simulation (S) results are based on 1000 validation data sets of size 20% of the total data sets.
*Difference between data set and simulation values for all trees.

Table 4. Average biases (m) and standard errors of estimates (SEE, m) for model 3 calculated from data sets of various sizes (n) and from 1000 randomly selected validation data sets of size 0.2n.

	n = 30				n = 70				n = 100				n = 1000			
	Bias		SEE		Bias		SEE		Bias		SEE		Bias		SEE	
DBH class (cm)	E	S	E	S	E	S	E	S	E	S	E	S	E	S	E	S
0.1–15.0	0.137	–0.167	2.867	4.057	–3.314	–3.594	3.725	4.107	–0.995	–1.162	2.476	2.731	–2.624	–2.631	3.773	3.790
15.1–25.0	1.518	1.986	5.384	6.482	0.922	0.991	4.523	4.788	–2.250	–2.546	4.653	4.956	–0.357	–0.347	4.577	4.605
25.1–35.0	–1.642	–1.842	1.642	1.975	–1.440	–1.735	4.833	5.133	1.730	1.732	4.025	4.082	1.468	1.472	4.780	4.801
35.1–45.0	–0.024	–0.037	4.494	4.899	2.232	2.200	5.309	5.460	2.020	2.038	4.627	4.707	1.647	1.623	5.281	5.291
45.1–55.0	–3.431	–3.976	7.873	8.294	2.548	2.687	7.167	7.435	–4.063	–4.130	6.328	6.411	–0.197	–0.201	6.259	6.208
55.1–65.0	0.779	0.837	5.985	6.343	–3.825	–3.835	7.204	7.348	1.265	1.471	6.272	6.425	–0.136	–0.134	6.572	6.596
65.1–75.0	2.381	2.848	6.200	6.782	1.981	2.091	7.146	7.327	–0.382	–0.501	5.740	5.769	–0.920	–0.857	6.984	7.005
75.1–85.0	0.859	1.208	9.895	10.612	0.123	0.290	4.493	4.716	3.251	3.559	7.878	8.076	–1.577	–1.559	5.532	5.529
85.1–95.0	–0.310	–0.340	3.504	4.171	–2.163	–2.194	5.928	6.260	–2.222	–2.035	5.324	5.499	–0.161	–0.211	7.155	7.172
95.1–105.0	–2.844	–3.548	2.844	3.923	–0.441	–0.509	5.431	5.664	–2.758	–2.961	4.390	4.560	–0.145	–0.103	6.837	6.875
105.1–125.0	0.000	0.000	0.000	0.000	–2.819	–3.125	4.468	4.826	1.453	1.767	3.692	3.995	1.757	1.811	7.489	7.520
125.1–145.0	–0.356	–1.866	0.356	4.963	2.803	3.378	5.240	5.910	–2.242	–1.704	7.474	8.266	–1.566	–1.513	6.804	6.814
145.1–165.0					–3.539	–3.878	3.539	4.367	0.979	2.109	0.979	3.540	0.244	0.365	6.289	6.328
>165.1					3.961	9.051	3.961	9.352					1.410	1.540	4.959	5.163
All trees	0.000	–0.074	6.094	6.690	0.000	0.060	5.603	5.935	0.000	0.048	5.434	5.581	0.000	0.015	5.926	5.936
Difference*		–0.074		0.596		0.060		0.332		0.048		0.147		0.015		0.010

Note: E, entire data set. Simulation (S) results are based on 1000 validation data sets of size 20% of the total data sets.
*Difference between data set and simulation values for all trees.

Table 5. The effect of validation data set size on average biases (m^3) and standard errors of estimate (m^3) using model 1.

	Bias*						Standard error of estimate*					
DBH class (cm)	E	10%	20%	30%	40%	50%	E	10%	20%	30%	40%	50%
Data set size = 100												
0.1–15.0	–0.002	–0.002	–0.002	–0.002	–0.002	–0.002	0.006	0.007	0.007	0.007	0.007	0.007
15.1–25.0	0.010	0.008	0.009	0.009	0.010	0.009	0.022	0.023	0.023	0.023	0.024	0.024
25.1–35.0	0.007	0.007	0.007	0.007	0.007	0.007	0.050	0.050	0.050	0.050	0.051	0.052
35.1–45.0	–0.053	–0.054	–0.055	–0.054	–0.053	–0.054	0.098	0.100	0.101	0.100	0.101	0.102
45.1–55.0	–0.029	–0.033	–0.031	–0.034	–0.034	–0.032	0.147	0.149	0.151	0.151	0.152	0.153
55.1–65.0	0.168	0.184	0.169	0.168	0.170	0.170	0.432	0.453	0.444	0.442	0.444	0.441
65.1–75.0	–0.035	–0.022	–0.028	–0.029	–0.027	–0.030	0.390	0.400	0.401	0.404	0.406	0.406
75.1–85.0	–0.000	0.031	0.008	–0.002	–0.005	–0.001	0.496	0.478	0.475	0.483	0.492	0.501
85.1–95.0	–0.402	–0.398	–0.422	–0.396	–0.399	–0.393	0.812	0.830	0.836	0.842	0.845	0.845
95.1–105.0	0.260	0.213	0.199	0.233	0.228	0.264	0.986	0.977	0.979	0.993	0.991	1.018
105.1–125.0	0.447	0.479	0.460	0.477	0.466	0.442	1.372	1.425	1.425	1.451	1.451	1.453
125.1–145.0	0.043	0.088	–0.065	–0.047	–0.022	–0.026	2.532	2.643	2.670	2.679	2.693	2.713
145.1–165.0	4.884	5.155	5.158	5.186	5.172	5.169	4.884	5.164	5.179	5.225	5.226	5.257
>165.1												
All trees	0.068	0.075	0.066	0.068	0.068	0.069	0.771	0.796	0.801	0.803	0.802	0.812
Difference†		0.007	0.002	0.000	0.000	0.001		0.025	0.030	0.032	0.031	0.041
Data set size = 1000												
0.1–15.0	–0.001	–0.001	–0.001	–0.001	–0.001	–0.001	0.007	0.007	0.007	0.007	0.007	0.007
15.1–25.0	0.006	0.006	0.006	0.006	0.006	0.006	0.032	0.032	0.032	0.032	0.032	0.032
25.1–35.0	–0.003	–0.003	–0.003	–0.003	–0.003	–0.003	0.073	0.073	0.073	0.073	0.073	0.073
35.1–45.0	0.021	0.021	0.022	0.022	0.021	0.021	0.146	0.145	0.146	0.145	0.146	0.146
45.1–55.0	0.023	0.022	0.023	0.024	0.024	0.024	0.255	0.256	0.255	0.256	0.255	0.255
55.1–65.0	0.016	0.020	0.018	0.018	0.018	0.018	0.375	0.374	0.375	0.377	0.377	0.377
65.1–75.0	0.049	0.046	0.044	0.046	0.046	0.047	0.466	0.473	0.470	0.471	0.470	0.469
75.1–85.0	–0.115	–0.115	–0.114	–0.117	–0.119	–0.112	0.764	0.768	0.769	0.768	0.768	0.768
85.1–95.0	–0.100	–0.100	–0.100	–0.108	–0.103	–0.104	0.985	0.985	0.987	0.983	0.988	0.985
95.1–105.0	0.133	0.116	0.123	0.126	0.127	0.125	0.978	0.981	0.985	0.986	0.983	0.982
105.1–125.0	–0.046	–0.100	–0.068	–0.048	–0.037	–0.041	1.861	1.892	1.882	1.879	1.878	1.869
125.1–145.0	0.045	0.124	0.063	0.060	0.048	0.068	3.445	3.460	3.443	3.454	3.469	3.475
145.1–165.0	1.172	1.211	1.165	1.175	1.173	1.185	4.262	4.348	4.278	4.291	4.283	4.280
>165.1	3.799	3.908	3.810	3.830	3.820	3.848	5.829	5.963	5.873	5.879	5.899	5.891
All trees	0.066	0.069	0.066	0.066	0.066	0.067	1.134	1.158	1.142	1.142	1.142	1.141
Difference†		0.003	0.000	0.000	0.000	0.001		0.024	0.008	0.008	0.008	0.007

Note: E, entire data set.
*The percentage values indicate the percentage of the data set used as validation data.
†Difference between data set and validation values for all trees.

Table 6. Comparison of average biases (m^3) and standard errors of estimate (m^3) for various levels of double cross validations (including the prediction sum of squares (PRESS) statistic), using model 1.

		Bias						Standard error of estimate					
DBH class (cm)	Frequency	E	PRESS	10%	20%	33%	50%	E	PRESS	10%	20%	33%	50%
Data set size = 100													
0.1–15.0	5	–0.002	–0.002	–0.003	–0.003	–0.001	–0.001	0.006	0.007	0.007	0.007	0.007	0.006
15.1–25.0	5	0.010	0.011	0.011	0.012	0.011	0.010	0.022	0.023	0.024	0.024	0.024	0.023
25.1–35.0	11	0.007	0.007	0.005	0.004	0.007	0.010	0.050	0.051	0.048	0.049	0.050	0.051
35.1–45.0	16	–0.053	–0.054	–0.053	–0.053	–0.055	–0.050	0.098	0.100	0.099	0.099	0.097	0.105
45.1–55.0	13	–0.029	–0.030	–0.031	–0.031	–0.022	–0.037	0.147	0.150	0.149	0.153	0.149	0.156
55.1–65.0	16	0.168	0.174	0.176	0.176	0.187	0.180	0.432	0.445	0.438	0.443	0.462	0.434
65.1–75.0	11	–0.035	–0.035	–0.045	–0.048	–0.044	–0.034	0.390	0.399	0.405	0.408	0.430	0.437
75.1–85.0	6	–0.000	–0.003	0.009	0.057	–0.057	–0.088	0.469	0.482	0.455	0.438	0.516	0.428
85.1–95.0	6	–0.402	–0.416	–0.400	–0.417	–0.365	–0.472	0.812	0.833	0.809	0.790	0.808	0.943
95.1–105.0	3	0.260	0.266	0.231	0.321	0.314	–0.017	0.986	1.013	1.001	0.907	0.953	0.921
105.1–125.0	5	0.447	0.463	0.453	0.420	0.557	0.346	1.372	1.419	1.429	1.478	1.419	1.279
125.1–145.0	2	0.043	0.029	0.018	–0.014	–0.195	–0.198	2.532	2.649	2.916	2.905	2.964	3.119
145.1–165.0	1	4.884	5.152	4.401	4.377	4.676	5.887	4.884	5.152	4.401	4.377	4.676	5.887
>165.1													
All trees	100	0.068	0.071	0.062	0.064	0.069	0.052	0.771	0.804	0.775	0.773	0.798	0.877
Difference*			0.003	0.006	0.004	0.001	0.016		0.033	0.004	0.002	0.027	0.106
Data set size = 1000													
0.1–15.0	54	–0.001	–0.001	–0.001	–0.001	–0.002	–0.001	0.007	0.007	0.007	0.007	0.007	0.007
15.1–25.0	120	0.006	0.006	0.006	0.006	0.006	0.006	0.032	0.032	0.032	0.032	0.032	0.032
25.1–35.0	101	–0.003	–0.003	–0.003	–0.003	–0.003	–0.003	0.070	0.073	0.073	0.073	0.074	0.072
35.1–45.0	133	0.021	0.021	0.021	0.021	0.020	0.021	0.146	0.146	0.145	0.146	0.145	0.146
45.1–55.0	115	0.023	0.023	0.023	0.023	0.023	0.023	0.255	0.255	0.255	0.255	0.255	0.256
55.1–65.0	139	0.016	0.016	0.016	0.016	0.017	0.018	0.375	0.376	0.375	0.376	0.374	0.376
65.1–75.0	98	0.049	0.050	0.050	0.049	0.051	0.050	0.466	0.467	0.466	0.466	0.465	0.469
75.1–85.0	59	–0.115	–0.115	–0.115	–0.116	–0.119	–0.114	0.764	0.766	0.764	0.763	0.765	0.765
85.1–95.0	50	–0.100	–0.099	–0.101	–0.101	–0.106	–0.101	0.985	0.987	0.988	0.988	0.989	0.997
95.1–105.0	35	0.133	0.134	0.129	0.127	0.131	0.120	0.978	0.981	0.977	0.984	0.998	0.967
105.1–125.0	44	–0.046	–0.046	–0.045	–0.042	–0.020	–0.066	1.861	1.868	1.876	1.881	1.853	1.871
125.1–145.0	27	0.045	0.046	0.044	0.039	0.070	0.084	3.445	3.459	3.451	3.456	3.479	3.469
145.1–165.0	13	1.172	1.179	1.193	1.168	1.193	1.204	4.262	4.280	4.280	4.284	4.353	4.328
>165.1	12	3.799	3.823	3.770	3.655	3.971	3.867	5.829	5.863	5.841	5.785	6.073	5.893
All trees	1000	0.066	0.066	0.065	0.064	0.069	0.067	1.134	1.139	1.137	1.135	1.156	1.144
Difference*			0.000	0.001	0.002	0.003	0.001		0.005	0.003	0.001	0.022	0.010

Note: E, entire data set. For the PRESS statistic, $n_v = 1$; otherwise $n_v = 10$, 20, 33, or 50% of the data set used as validation data set.
*Difference between data set and validation values for all trees.

Table 7. Comparison of some fit statistics (absolute bias, coefficient of determination, and residual mean square) calculated from the data sets and from 1000 randomly selected validation data from each data set.

	Absolute bias		Coefficient of determination		Residual mean square	
Data set size	E	S	E	S	E	S
Model 1						
30	0.35	0.39	0.9694	0.9607	0.5123	0.6571
50	0.36	0.38	0.9783	0.9740	0.5561	0.6291
70	0.34	0.36	0.9930	0.9873	0.4387	0.4860
100	0.38	0.39	0.9837	0.9814	0.6128	0.6614
200	0.48	0.50	0.9792	0.9782	1.3333	1.4040
500	0.47	0.48	0.9793	0.9785	1.5319	1.6048
1000	0.46	0.46	0.9799	0.9796	1.2898	1.3081
Model 2						
30	0.37	0.43	0.9826	0.9525	0.2809	0.4416
50	0.44	0.50	0.9786	0.9634	0.5370	0.8971
70	0.45	0.47	0.9939	0.9859	0.3793	0.4283
100	0.38	0.40	0.9882	0.9806	0.4364	0.5393
200	0.59	0.61	0.9771	0.9741	1.4594	1.6318
500	0.52	0.53	0.9851	0.9821	1.1027	1.1452
1000	0.53	0.53	0.9826	0.9814	1.1110	1.1365
Model 3						
30	4.85	5.41	0.6970	0.4274	41.263	49.729
50	4.64	5.01	0.7394	0.6302	36.034	42.156
70	4.67	4.96	0.8331	0.7855	32.799	36.801
100	4.38	4.50	0.7904	0.7553	30.442	32.111
200	4.54	4.60	0.8224	0.8072	33.741	34.803
500	4.96	4.99	0.7988	0.7918	38.299	38.734
1000	4.73	4.74	0.8026	0.8000	35.223	35.342

Note: E, entire data set. Simulation (S) results are based on 1000 validation data sets of size 20% of the total data sets.

Table 8. The effect of validation data set size on the estimation of some fit statistics (absolute bias, coefficient of determination, and residual mean square) using cross validation.

	Absolute bias		Coefficient of determination		Residual mean square	
Size of validation data	100*	1000*	100*	1000*	100*	1000*
Model 1						
10%	0.39	0.47	0.9789	0.9791	0.6532	1.3450
20%	0.39	0.46	0.9814	0.9796	0.6614	1.3081
30%	0.39	0.46	0.9817	0.9795	0.6648	1.3081
40%	0.39	0.46	0.9818	0.9795	0.6631	1.3081
50%	0.39	0.46	0.9816	0.9796	0.6797	1.3058
All data	0.38	0.46	0.9837	0.9799	0.6128	1.2898
Model 2						
10%	0.39	0.53	0.9764	0.9802	0.5173	1.1364
20%	0.38	0.53	0.9806	0.9814	0.5393	1.1365
30%	0.40	0.53	0.9822	0.9815	0.5498	1.1429
40%	0.41	0.53	0.9816	0.9818	0.5755	1.1386
50%	0.41	0.53	0.9826	0.9818	0.5847	1.1451
All data	0.38	0.53	0.9882	0.9826	0.4364	1.1110
Model 3						
10%	4.46	4.73	0.7030	0.7975	31.356	35.247
20%	4.50	4.74	0.7553	0.8000	32.111	35.342
30%	4.57	4.73	0.7560	0.8005	32.852	35.306
40%	4.56	4.75	0.7641	0.8004	32.805	35.473
50%	4.62	4.74	0.7632	0.8004	33.577	35.497
All data	4.38	4.73	0.7904	0.8026	30.442	35.223

*Data set size.

Table 9. The effect of validation data set size on some fit statistics (absolute bias, coefficient of determination, and residual mean square) using double cross validation.

	Absolute bias		Coefficient of determination		Residual mean square	
Size of validation data	100*	1000*	100*	1000*	100*	1000*
Model 1						
PRESS†	0.39	0.46	0.9822	0.9797	0.6664	1.3012
10%	0.39	0.46	0.9835	0.9797	0.6192	1.2967
20%	0.39	0.46	0.9836	0.9798	0.6160	1.2921
33%	0.40	0.46	0.9825	0.9790	0.6565	1.3404
50%	0.41	0.46	0.9789	0.9795	0.7929	1.3127
All data	0.38	0.46	0.9837	0.9799	0.6128	1.2898
Model 2						
PRESS†	0.40	0.53	0.9854	0.9823	0.5408	1.1344
10%	0.40	0.53	0.9856	0.9822	0.5349	1.1386
20%	0.40	0.53	0.9856	0.9819	0.5334	1.1579
33%	0.39	0.54	0.9860	0.9809	0.5202	1.2213
50%	0.43	0.53	0.9829	0.9823	0.6336	1.1344
All data	0.38	0.53	0.9882	0.9826	0.4364	1.1110
Model 3						
PRESS†	4.50	4.74	0.7800	0.8016	31.950	35.414
10%	4.50	4.74	0.7786	0.8018	32.157	35.378
20%	4.57	4.74	0.7743	0.8015	32.782	35.426
33%	4.67	4.74	0.7705	0.8019	33.331	35.366
50%	4.43	4.75	0.7814	0.8018	31.755	35.378
All data	4.38	4.73	0.7904	0.8026	30.442	35.223

*Data set size.
†Prediction sum of squares statistic.

Table 10. Percentage of two models selected (model 1 and model 2) based on various fit and lack of fit statistics from 100 randomly selected cross validation data sets of size 0.2*n*.

Data set size	Model selected	Average bias	Average absolute bias	Standard error of estimate	Coefficient of determination	Lack of fit
70	1	41%	56%	44%	44%	94%
	2	59%	44%	56%	56%	6%
100	1	55%	62%	51%	51%	96%
	2	45%	38%	49%	49%	4%
1000	1	44%	98%	32%	32%	100%
	2	56%	2%	68%	68%	0%

Table 11. Standard errors of the biases (S_B) and standard errors of the standard error of estimates (S_S) calculated from 1000 randomly selected cross validation data sets of size 0.2*n*.

	n = 30		n = 70		n = 100		n = 1000	
	S_B	S_S	S_B	S_S	S_B	S_S	S_B	S_S
Model 1	0.325	0.472	0.200	0.278	0.185	0.330	0.081	0.225
Model 2	0.281	0.254	0.192	0.159	0.180	0.271	0.084	0.169
Model 3	3.007	1.769	1.831	0.818	1.408	0.741	0.488	0.259

presented in Table 1 show that a comparison of standard errors of estimate (residual mean square) and coefficients of determination would indicate the choice of model 2, while the average absolute bias and the lack of fit statistics would indicate the choice of model 1. The results of our simulation mimic the choice indicated by lack of fit statistics very well for all data set sizes, as well as the choice indicated by average absolute bias for a data set size of 1000. However, all other choices were not at all well modeled by the simulations. These results illustrate two key points. First, they demonstrate the importance of using lack of fit statistics, as well as fit statistics, for a fair model evaluation and compari-

son of two or more models. Second, they indicate that the practitioner is better served by using fit statistics and lack of fit statistics that are calculated from the entire model, rather than from a so-called "independent data set".

Varying the cross validation data sizes from 10 to 50% of the observations in the model data set had little effect on the estimation of fit statistics and lack of fit statistics calculated from 1000 validations (Tables 5 and 8). While cross validation sizes of 10% and, to a lesser extent, 50% resulted in some slight discrepancies, the differences are negligible for all practical purposes, especially for a data set size of 1000.

In double cross validation, every one of the observations is used to calculate the prediction errors and prediction lack of fit statistics independently of the model data set. As a result, many practitioners prefer its use over cross validation. This argument is valid if the cross validation is based on a single validation data set, as is usually the case, mainly because there is a chance (because of the random selection of the portion of the data for validation) that a single validation set will provide a poor estimation of the model fit statistics and lack of fit statistics. On the other hand, double cross validation is always based on at least two (if the data is split into two equal sets) or more validations, up to a maximum of $n - 1$ (the extreme case of PRESS). The simulations in this study showed that cross validation and double cross validation both estimate model fit statistics and lack of fit statistics equally well (compare Tables 6 and 9 with Tables 5 and 8). Furthermore, it can be concluded that double cross validation, using either observations one at a time (PRESS) or 10 to 20% splits, estimates the model fit statistics and lack of fit statistics somewhat better than by using 33 or 50% splits (Tables 6 and 9). This difference is more pronounced for a data set size of 100.

Recommendations

Both cross validation and double cross validation produce estimates of model fit statistics and lack of fit statistics that are similar to those derived from using the entire data set. This result, in and of itself, can only lead to the conclusion that validation techniques do not provide any additional information compared with the respective statistics obtained directly from models built from entire data sets. In fact, the fit statistics and lack of fit statistics calculated directly from the data used to fit the model provide a better and more reliable description of the prediction errors than any of the validation processes studied. Therefore, it is recommended that the model fit statistics and lack of fit statistics be used in the process of model evaluation and (or) model selection, and that many cross validation procedures may be unnecessary.

As stated earlier, the use of new data is the most preferred method of validation. It is hoped that the recommendations provided here will not hinder researchers from using new data for validation purposes, because if nothing else, their models will be improved in the end by incorporating new, added observations. That said, it must be pointed out that if the new data are taken from the same population as the model construction data, they will likely behave in exactly the same manner as a cross validation data set. If, on the other hand, new data is taken from a separate population, they will likely not provide much in the way of additional information for model evaluation or selection. Nor will they improve the model since these data ought not be added to the original data set for model improvement.

References

Berk, K.N. 1984. Validating regression procedures with new data. Technometrics, **26**: 331–338.

Burk, T.E. 1990. Prediction error evaluation: preliminary results. *In* Proceedings of the IUFRO Forest Simulation Systems Conference. University of California, Division of Agriculture and Natural Resources, Davis, Calif. Bull. 1927. pp. 81–88.

Kozak, A. 1988. A variable-exponent taper equation. Can. J. For. Res. **18**: 1363–1368.

Myers, H.M. 1986. Classical and modern regression with application. Duxbury Press, Boston, Mass.

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

Ronchetti, E., Field, C., and Blanchard, W. 1997. Robust linear model selection by cross-validation. J. Am. Stat. Assoc. **92**: 1017–1023.

Schumacher, F.X., and Hall, S.H. 1933. Logarithmic expression of timber tree volume. J. Agr. Res. **47**: 719–724.

Shao, J. 1993. Linear model selection by cross-validation. J. Am. Stat. Assoc. **88**: 486–494.

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

Spurr, S.H. 1952. Forest inventory. Ronald Press. Co., New York.

Stone, M. 1974. Cross-validation choice and assessment of statistical predictions. J. R. Stat. Soc. Series B, **36**: 111–147.