Forestry An International Journal of Forest Research

Institute of Chartered Foresters

Forestry 2014; **87**, 654–662, doi:10.1093/forestry/cpu025
Advance Access publication 2 July 2014

Suitability of five cross validation methods for performance evaluation of nonlinear mixed-effects forest models – a case study

Yuqing Yang* and Shongming Huang

Biometrics Unit, Forest Management Branch, Alberta Environment and Sustainable Resource Development, 8th Floor, 9920-108 Street, Edmonton, AB, Canada T5K 2M4

*Corresponding Author. Tel: +1 7804225250; Fax: +1 7804270085; E-mail: yuqing.yang@gov.ab.ca

Received 19 September 2013

Five cross validation methods, the *k*-fold, leave one plot out (LOP), leave one tree per plot out (LOT), 0.632 and 0.632+ bootstrap methods, were examined in this study for their suitability for performance evaluation of seven nonlinear mixed models based on a height–diameter relationship. The *k*-fold, LOP, 0.632 and 0.632+ methods used plot as the basic unit for data resampling, and applies to situations where predictions are needed for all trees in a new plot not used for model development. All four methods were suitable for evaluating the predictive performance of the selected model(s), and the 0.632 and 0.632+ methods were better than the *k*-fold and LOP methods. The LOT method used tree as the basic unit for data resampling, and applies to situations where predictions are needed for a portion of trees in a plot not used for model development, while the remaining trees of the plot are used for model development. The LOT method was not suitable for performance evaluation of the selected model(s).

Introduction

Most forest growth and yield models are developed for making predictions on datasets not used in model development. To develop a prediction model, different candidate models are often compared and the 'best' model is selected based on some model selection criteria. Once the 'best' prediction model is determined, its predictive performance on new data needs to be assessed. It is well known that model fitting statistics may not be a good indication on how well a model will predict. It is easy to over-fit the data by including too many covariates to subsequently inflate model fitting statistics. The best way to measure the predictive ability of a model is to test it on an independent dataset not used in parameter estimation. But an independent dataset is often not available or difficult and expensive to collect (Snee 1977).

One way to address the problem is through cross validation (CV). CV is a data resampling method by partitioning a dataset into two: a training dataset and a testing dataset. The training dataset is used to fit a model, and the testing dataset is used to evaluate the predictive performance of the fitted model through prediction errors. This process is repeated many times and the CV estimate of error is the average prediction error over the testing datasets. The idea of CV originated in the 1930s (Larson 1931) and was further developed by Mosteller and Turkey (1968) and others (e.g. Stone 1974; Gelfand *et al*. 1992; Shao 1993). A clear statement of CV first appeared in Mosteller and Turkey (1968).

In general, complex models achieve better fits as they exploit some local features of the dataset that may not be present globally. A complex model may not predict well in datasets not used for model fitting if these datasets do not have the features of the model fitting data. In CV, local features of a training set could be very different from local features of a testing set, resulting in poor predictive performance. Therefore, CV protects against over-fitting by selecting a model that captures the global patterns of a dataset and by avoiding models that exploit local features of a dataset.

Forest growth and yield models have traditionally been developed by least squares regression. In recent years, however, mixed-effects models have gained popularity in growth and yield modelling to achieve better local predictions and to handle residual autocorrelation from repeatedly measured data (Lappi 1991; Trincado and Burkhart 2006; Temesgen *et al*. 2008; Yang *et al*. 2009; Huang *et al*. 2009; Yang and Huang 2011). This modelling approach is very effective in dealing with hierarchical data structures (Greenland 2000). In addition to accounting for covariate or treatment effects through fixed parameters as in least squares regression, mixed-effects models can account for various sources of heterogeneity and randomness in the data caused by known and unknown factors through random parameters (Vonesh and Chinchilli 1997).

Although there exists an extensive and growing literature on CV for models fitted by least squares (e.g. Zhang 1997), the literature on CV for mixed-effects models is scarce. In fact, many studies (e.g. Budhathoki *et al*. 2008) only presented model fitting results when developing mixed-effects models and model prediction capability was not examined. Afshartous and de Leeuw (2004) cross validated four candidate linear mixed models to develop model selection criteria that assess the predictive ability of these candidate models. To take into account the hierarchical data structure, data resampling was based on the subject (school) where one student per school was randomly selected as testing data and the remaining students from each school were used as training data. Estimated parameters based on training data were then applied to testing data for making predictions. Mean square error (MSE) was calculated based on the observed and predicted values for the response variable. The procedure was repeated *m* times and a mean MSE can be computed over the *m* repetitions. Robinson and Wykoff (2004) compared the predictive capabilities of both nonlinear least squares models and mixed-effects models through CV. Their data resampling approach was to randomly split the available data into model fitting and model testing datasets, regardless of species and tree diameters. The criterion was that each sampling point contained four trees with diameters >7.62 cm in model fitting data. To our knowledge, other data resampling approaches have not been examined in the forestry literature for cross validating mixed-effects models while considering hierarchical data structures.

The objective of this study was to assess the suitability of various CV methods for evaluating the predictive performance of nonlinear mixed-effects forest models. To achieve this objective, we purposely selected a basic, yet common height–diameter relationship as an example. Seven nonlinear mixed-effects height–diameter models were developed, and several CV methods that considered the hierarchical data structure were examined for their abilities in providing reliable predictions on new data not used for model development.

Methods

Data

Two datasets were used for this study, a model fitting dataset and a model validation dataset. The datasets were collected by the Alberta government as a part of the provincial forest inventory database. Various plot sizes were used for data collection, from 200 to 2000 m^2 with 1000 m^2 being the most common. Within each plot, diameters at breast height of 1.3 m above ground were measured for trees taller than 1.3 m, and heights were only measured for a subsample of trees (at least 10–20 per cent) in a systematic or random fashion to cover the range of tree heights. A detailed description of the data collection procedures is provided in ASRD (2005). In total, there were 5942 trembling aspen (*Populus tremuloides* Michx.) trees from 600 plots in the model fitting data (Figure 1a) and 1388 trembling aspen trees from 99 plots in the model validation data (Figure 1b).

Model development and prediction

For this study, the response variable was individual tree height (HT, m) and the covariate was tree diameter at breast height (DBH, cm). First, a base model was selected based on nonlinear least squares fits. Three commonly used model forms, a power function, an exponential function and a Chapman–Richards function, were evaluated for the height–diameter relationship. The Chapman–Richards function had the best fit with the lowest root mean square error and the highest coefficient of determination (Table 1) and was selected as the base model:

$$\mathrm{HT} = 1.3 + \beta_1[1 - \exp(-\beta_2 \mathrm{DBH})]^{\beta_3}, \qquad (1)$$

where β_1, β_2 and β_3 are fixed parameters to be estimated.

Nonlinear mixed height–diameter models were subsequently developed by adding random parameters to the fixed parameters of the base model. All possible combinations of random parameter inclusion were examined, including adding one random parameter to each of the three fixed parameters, adding two random parameters to any two of the three fixed parameters, and adding three random parameters to all three fixed parameters. There were a total of seven models, three with one random parameter, three with two random parameters, and one with three parameters. These random parameters were plot specific, i.e. each plot had a unique set of random parameter(s). All nonlinear mixed models can be expressed in a general formula:

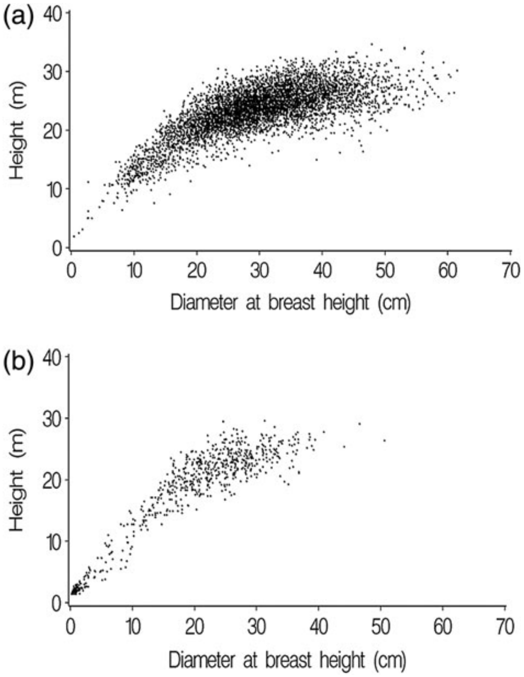

Figure 1 Model fitting (a) and model validation data (b) for aspen height–diameter relationship.

Table 1 Root mean square error (RMSE) and R^2 values for candidate base functions

Function	RMSE	R^2
Power	2.5587	0.5838
Exponential	2.4203	0.6277
Chapman–Richard	2.4100	0.6309

$$\mathbf{HT}_i = f\,(\mathbf{DBH}_i, \boldsymbol{\beta}, \mathbf{u}_i) + \boldsymbol{\varepsilon}_i, \qquad (2)$$

where $\mathbf{HT}_i$ is a vector of tree height for plot $i = 1, 2, \ldots, m$; m is the number of plots; $f(\cdot)$ is a nonlinear function of the covariate matrix $\mathbf{DBH}_i$, $\boldsymbol{\beta}$ is a fixed parameter vector common to all plots, $\mathbf{u}_i$ is a random parameter vector unique for plot i; and $\boldsymbol{\varepsilon}_i$ is a matrix of unknown within-plot errors. The $\mathbf{u}_i$ and $\boldsymbol{\varepsilon}_i$ are often assumed to be uncorrelated and normally distributed with mean zero and variance–covariance matrices $\mathbf{D}$ and $\mathbf{R}_i$, respectively, i.e. $\mathbf{u}_i \sim N\,(\mathbf{0}, \mathbf{D})$ and $\boldsymbol{\varepsilon}_i \sim N\,(\mathbf{0}, \mathbf{R}_i)$.

The ZERO expansion method under the SAS macro NLINMIX (Littell *et al.* 2006) was used for parameter estimation. This method uses a first-order Taylor series expansion of equation (2) around an estimator $\boldsymbol{\beta}^*$ close to $\boldsymbol{\beta}$ and an estimator $\mathbf{u}_i^*$ close to $\mathbf{u}_i$, with negligible terms (quadratics and cross-products) dropped. Maximum likelihood method was used to estimate model parameters.

A major distinction between mixed-effects models and least squares models is that mixed-effects models can produce localized (subject-specific) predictions. For this study, the subject was plot. To make plot-specific tree height predictions under the ZERO expansion method, random parameters were predicted first by equation (3), and plot-specific height predictions were predicted by equation (4):

$$\hat{\mathbf{u}}_i = \hat{\mathbf{D}}\mathbf{Z}_i'(\mathbf{Z}_i\hat{\mathbf{D}}\mathbf{Z}_i' + \hat{\mathbf{R}}_i)^{-1}[\mathbf{HT}_i - f(\mathbf{DBH}_i, \hat{\boldsymbol{\beta}}, \mathbf{0})], \qquad (3)$$

$$\overset{\wedge}{\mathbf{HT}}_i = f(\mathbf{DBH}_i, \hat{\boldsymbol{\beta}}, \mathbf{0}) + \mathbf{Z}_i\hat{\mathbf{u}}_i, \qquad (4)$$

where $\hat{\mathbf{D}}$ is an estimate of $\mathbf{D}$, $\hat{\mathbf{R}}_i$ is an estimate of $\mathbf{R}_i$, and $\mathbf{Z}_i$ is the partial derivative of $\mathbf{HT}_i$ with respect to $\mathbf{u}_i$ and $\mathbf{Z}_i = \partial f(\mathbf{DBH}_i, \boldsymbol{\beta}, \mathbf{u}_i)/\partial\mathbf{u}_i|_{\hat{\boldsymbol{\beta}},\mathbf{0}}$. With plot-specific height predictions ready, the seven candidate models were compared for their predictive capabilities through CV.

Cross validation

There are different types of CV methods that adopt different data resampling techniques. The simplest method is the holdout method, in which a dataset is divided into a training set and a testing set. Validation based on a single split is sometimes called simple validation (Arlot 2010). This method has been widely used for evaluating the predictive performance of a fitted model. However, this kind of evaluation depends heavily on which data points end up in the training set and which in the testing set. Results from such an evaluation can be substantially different depending on how the partition is made.

K-fold CV, the basic form of CV, is one way to improve over the holdout method. A dataset is randomly partitioned into k mutually exclusive equal (or nearly equal) subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as a testing set and the remaining $k-1$ subsets are put together to form a training set. Then the average prediction error on the testing data across all k trials is computed. For the k-fold method, data resampling is done without replacement, and it matters less how the data get divided. Every data point gets to be in the testing sets exactly once, and gets to be in the training sets $k-1$ times.

For the k-fold method, typical choice of k value is 5 or 10 (Hastie *et al.* 2009). For larger datasets, smaller k values are sufficient. For smaller datasets, however, larger k values are needed. To cross validate mixed-effects models, hierarchical data structures need to be considered (Afshartous and de Leeuw 2004). This study involved a two-level data structure with trees nested within plots, and plot was used as the basic unit for data resampling. Six different k values were examined: 5, 10, 20, 30, 40 and 50. For example, for a 5-fold CV, the 600 plots in the model fitting dataset were divided into five mutually exclusive subsets with each having 120 plots. Each time 480 plots were used as training data for parameter estimation and 120 plots were used as testing data for making predictions. Once model parameters were estimated from each training dataset, all trees in each plot from the corresponding testing dataset were used to predict random parameters by equation (3) and their heights were predicted by equation (4). Prediction errors were then calculated for these trees and subsequently used for calculating different validation statistics.The procedure was repeated five times so that each subset was used once as testing data. Mean validation statistics were then computed across the repetitions. The goal was to find the proper k value for reliable assessment of the predictive capability of each nonlinear mixed model.

Leave-one-out (LOO) method is a k-fold method taken to its logical extreme, with k equal to the number of data points in a dataset. The LOO CV is computationally expensive because it requires many repetitions of model fitting and model prediction, especially for large datasets. An accuracy estimate obtained using the LOO CV is known to be almost unbiased but it has high variance (Efron 1983). To take into account the hierarchical data structure used for fitting the nonlinear mixed height–diameter models, two variations of the LOO method were examined: leave one plot out (LOP) method and leave one tree per plot out (LOT) method.

Same as the k-fold method, plot was used as the basic unit for data resampling for the LOP method. Each of the 600 plots was reserved as testing data and the remaining 599 plots were used as training data. The number of repetitions was 600 so that each plot was reserved once as testing data. For each repetition, parameters estimated from the training data were used to make plot-specific height predictions for trees in the testing data. All trees in a plot of the testing data were used for predicting random parameters by equation (3) and their heights were predicted by equation (4). Prediction errors were then calculated for these trees and subsequently used for calculating different validation statistics. Mean validation statistics were calculated across the 600 repetitions.

For the LOT method, a tree was used as the basic unit for data resampling. One tree was randomly selected from each of the 600 model fitting plots, and the selected 600 trees were reserved as testing data. The remaining 5342 trees from the 600 plots were used as training data for model fitting. Estimated parameters were then used to predict random parameters using equation (3) for each tree in the testing data, and plot-specific height prediction was derived by equation (4) for that tree. To echo the LOP method, this process was repeated 600 times, and the mean validation statistics were calculated across the 600 repetitions.

The k-fold, LOP and LOT methods are based on data resampling without replacement. An alternative to these methods is the bootstrap method based on data resampling with replacement. The bootstrap family was formally introduced by Efron (1979) and fully described in Efron and Tibshirani (1997). The basic idea of a bootstrap method is to create a bootstrap sample by randomly selecting a sample of size n with replacement, with n being the sample size of the original data; estimate the prediction rule using the bootstrap sample; test the prediction rule on the original data; and estimate the prediction error. A large number of such bootstrap samples are drawn, and some measures of prediction error are then averaged over the number of repetitions. Such estimates of prediction error tend to be biased downward since some of the individuals in the testing data are also in the training data.

Efron (1983) proposed the 0.632 bootstrap CV method to improve the ordinary bootstrap method. Instead of using the full set of the original data, prediction errors are only calculated using the subset of the original data not included in the bootstrap sample. To draw a bootstrap sample of size n with replacement from a dataset of size n, the probability of any given instance not being chosen after n samples is $(1 - 1/n)^n \approx e^{-1} \approx 0.368$. The expected number of distinct instances from the original dataset appearing in the bootstrap sample is therefore $0.632 \times n$. The 0.632 bootstrap estimate of prediction errors ($E_{0.632}$) can be calculated as:

$$E_{0.632} = 0.368E_{\mathrm{app}} + 0.632E_{\mathrm{boot}} \qquad (5)$$

where E_{boot} is the bootstrap estimate of prediction errors of the data not occurring in the bootstrap sample, and E_{app} is the apparent error estimate based on the available data for both estimating the prediction rule and testing its performance. In the context of this study, the apparent error estimate was based on the model fitting data where candidate models were fitted and tested on the same data.

Efron and Tibshirani (1997) argued that highly over-fitting rules do not benefit from the 0.632 estimator. An alternative estimator that corrects for over-fitting, the 0.632+ estimator, was proposed:

$$E_{0.632+} = (1-\omega)E_{\mathrm{app}} + \omega E_{\mathrm{boot}}, \qquad (6)$$

where ω is the weight and is calculated as $\omega = 0.632/(1-0.368R)$, with $\omega = 0.632$ if $R = 0$ and $\omega = 1$ if $R = 1$; R is the relative over-fitting rate and is calculated as $R = (E_{\mathrm{boot}} - E_{\mathrm{app}})/(\gamma - E_{\mathrm{app}})$, R is in the range of [0,1], with $R = 0$ indicating no over-fitting and $R = 1$ indicating highly over-fitting; γ is the no-information error rate, the expected prediction error when the prediction rule is tested on data where the responses and predictors are independent, and can be estimated by permuting the response and predictor (Efron and Tibshirani 1997).

Both 0.632 and 0.632+ bootstrap CV methods were examined in this study. Similar to the *k*-fold and LOP methods, plot was used as the basic unit for data resampling for both methods. In fact, Ren *et al.* (2010) concluded that for multilevel hierarchical data, bootstrapping on the highest level was better than that on lower levels. The main reason may be that the bootstrap sampling at the highest level can accurately reflect original sample information. Again, 600 repetitions were used. All seven candidate models were fitted on each bootstrap sample, and prediction errors were computed on data not included in the bootstrap sample. Bootstrap estimates of prediction errors were then averaged across the 600 repetitions. The 0.632 and the 0.632+ estimates of prediction errors were subsequently computed. To calculate the no-information error rate γ, permutation of the response and predictor was also conducted at the plot level.

For this study, the five CV methods were grouped into two categories: the *k*-fold, LOP, 0.632 and 0.632+ methods in the first category and the LOT method in the second category by itself. Plot was used as the basic unit for data resampling for the four methods in the first category. All trees in a plot belonged to either a training dataset or a testing dataset. These CV methods can be used for situations where predictions are applied to all trees in a new plot not used for model fitting. For the LOT method, data resampling was based on trees, not plots. One tree of each plot belonged to a testing dataset while the remaining trees of the same plot belonged to a training dataset. Therefore, the resulting training and testing datasets were not independent. This method can be applied to situations where some trees from a plot are used for model fitting and the fitted model is used for predictions for the remaining trees of the same plot.

Table 2 summarizes the number of repetitions, data resampling method, number of plots and number of trees per sample for the five CV methods.

Validation statistics

For all five CV methods, the following statistics were calculated first on each testing dataset using all seven mixed height–diameter models, and averaged over all repetitions:

$$\mathrm{ME} = \tfrac{1}{N}\sum\nolimits_{i=1}^{p}\sum\nolimits_{j=1}^{q_i}(\mathrm{HT}_{ij} - \hat{\mathrm{HT}}_{ij}), \tag{7}$$

$$\mathrm{MAE} = \tfrac{1}{N}\sum\nolimits_{i=1}^{p}\sum\nolimits_{j=1}^{q_i}|\mathrm{HT}_{ij} - \hat{\mathrm{HT}}_{ij}|, \tag{8}$$

$$\mathrm{MAPE} = \tfrac{1}{N}\sum\nolimits_{i=1}^{p}\sum\nolimits_{j=1}^{q_i}(\mathrm{HT}_{ij} - \hat{\mathrm{HT}}_{ij})/\mathrm{HT}_{ij}, \tag{9}$$

$$\mathrm{MSE} = \tfrac{1}{N}\sum\nolimits_{i=1}^{p}\sum\nolimits_{j=1}^{q_i}(\mathrm{HT}_{ij} - \hat{\mathrm{HT}}_{ij})^2, \tag{10}$$

where HT_{ij} and $\hat{\mathrm{HT}}_{ij}$ are the *j*th observed and predicted tree heights in the *i*th plot, *p* is the number of plots in a testing dataset; q_i is the number of trees in the *i*th plot; $N = \sum_{i=1}^{p} q_i$ is the total number of trees in all plots in the testing dataset; ME is mean error; MAE is mean absolute error; MAPE is mean absolute percent error and MSE is mean square error.

The testing datasets used for deriving the four validation statistics from the five CV methods were considered as random samples of the entire model fitting data. Therefore, the calculated validation statistics were random variables with associated variation. To evaluate the fidelity of the four validation statistics, relative standard error, calculated as the percentage of the standard error for each validation statistics to the estimated mean value of the corresponding statistics, was also provided for all seven candidate models based on each of the five CV methods.

In this study, data partition for the five CV methods was based on model fitting data. To evaluate the suitability of these methods for performance evaluation of the candidate nonlinear mixed-effects forest models, the statistics in equations (7–10) were also calculated for the entire model fitting data (Figure 1a) and model validation data (Figure 1b). Parameters for each candidate model estimated from the entire model fitting data were used to make plot-specific tree height predictions for both model fitting and model

Table 2 Summary of the *k*-fold, LOP, LOT, 0.632 and 0.632+ CV methods

CV method	Data summary	Testing dataset	Training dataset
k-fold	*k* values	5, 10, 20, 30, 40, 50	5, 10, 20, 30, 40, 50
	Sampling unit	Plot	Plot
	Sampling method	One group of plots	the remaining ($k-1$) groups of plots
	No. of plots	600/*k*	600–600/*k*
	No. of trees	Variable	Variable
LOP	No. of repetitions	600	600
	Sampling unit	Plot	Plot
	Sampling method	One plot each time	The remaining plots
	No. of plots	1	599
	No. of trees	Variable	Variable
LOT	No. of repetitions	600	600
	Sampling unit	Tree	Tree
	Sampling method	Randomly select one tree per plot	The remaining trees of each plot
	No. of plots	600	600
	No. of trees	600	5342
0.632/0.632+	No. of repetitions	600	600
	Sampling unit	Plot	Plot
	Sampling method	The remaining plots	Randomly select 600 plots with replacement
	No. of plots	Variable	600
	No. of trees	Variable	Variable

No. of repetitions means number of training and testing datasets generated. For the 0.632 and 0.632+ bootstrap methods, some of the 600 plots in each training dataset were not unique due to the resampling method used.

validation data. The calculated statistics on model validation data were used as the basis for evaluating the five CV methods.

The calculated validation statistics on both model fitting and model validation data, as well as the validation statistics derived from the five CV methods on model testing data, were also used to rank the candidate models, with ranking 1 being the best model with the smallest absolute value of each statistics and ranking 7 being the worst model with the largest absolute value of each statistics.

Results

Table 3 shows the model numbers and AIC statistics for all seven candidate models. According to the AIC statistics, the best nonlinear mixed models with one, two and three random parameters fitted on the model fitting data were M1, M4 and M7. When all seven models were ranked together, the sequence from the best model to the worst model was: M7 – M4 – M5 – M1 – M6 – M2 – M3.

Table 3 Model number, model form and AIC statistics for seven candidate nonlinear mixed models

Model	Model form	AIC
M1	$\mathrm{HT}_{ij} = 1.3 + (\beta_1 + u_{i1})\{1 - \exp[-\beta_2 \mathrm{DBH}_{ij}]\}^{\beta_3}$	24484.1
M2	$\mathrm{HT}_{ij} = 1.3 + \beta_1\{1 - \exp[-(\beta_2 + u_{i1})\mathrm{DBH}_{ij}]\}^{\beta_3}$	24961.4
M3	$\mathrm{HT}_{ij} = 1.3 + \beta_1\{1 - \exp[-\beta_2 \mathrm{DBH}_{ij}]\}^{\beta_3 + u_{i1}}$	25439.9
M4	$\mathrm{HT}_{ij} = 1.3 + (\beta_1 + u_{i1})\{1 - \exp[-(\beta_2 + u_{i2})\mathrm{DBH}_{ij}]\}^{\beta_3}$	24398.4
M5	$\mathrm{HT}_{ij} = 1.3 + (\beta_1 + u_{i1})\{1 - \exp[-\beta_2 \mathrm{DBH}_{ij}]\}^{\beta_3 + u_{i2}}$	24420.8
M6	$\mathrm{HT}_{ij} = 1.3 + \beta_1\{1 - \exp[-(\beta_2 + u_{i1})\mathrm{DBH}_{ij}]\}^{\beta_3 + u_{i2}}$	24734.8
M7	$\mathrm{HT}_{ij} = 1.3 + (\beta_1 + u_{i1})\{1 - \exp[-(\beta_2 + u_{i2})\mathrm{DBH}_{ij}]\}^{\beta_3 + u_{i3}}$	24366.0

HT_{ij} and DBH_{ij} are the *j*th observed tree height and diameter in the *i*th plot of the model fitting data, β_1, β_2 and β_3 are fixed parameters and u_{i1}, u_{i2} and u_{i3} are random parameters for plot *i*, and AIC is Akaike's information criterion.

Figure 2 shows the mean errors, mean absolute errors, mean absolute percent errors and mean square errors for all seven candidate models derived from the *k*-fold method for different *k* values. Several trends were obvious. All mean errors were very small, with the largest one < 0.01 m in absolute value (Figure 2a). For each of the seven models, the six *k* values produced nearly identical results, indicating that a smaller *k* value was sufficient. To compare the *k*-fold method with other CV methods, results for $k = 5$ were used.

Table 4 shows the four validation statistics (equations 7 – 10) for all seven candidate models calculated on model fitting and model validation data, and the four validation statistics plus the associated relative standard errors from the 5-fold, LOP, 0.632+ and LOT methods calculated on model testing datasets that were generated from model fitting data. The rankings of the seven models based on the calculated validation statistics are also provided in the table. Note that 0.632 and 0.632+ bootstrap methods produced almost identical values for all four statistics, and the results were only presented for the 0.632+ method. The relative over-fitting rates (*R* in equation 6) were close to zero for all seven models, indicating no over-fitting for these models. As a result, ω was close to 0.632, and equations (5) and (6) were equivalent.

Of the four statistics, ME produced quite different model rankings on model fitting and validation data. The rankings based on *ME* were also quite different for the 5-fold, LOP and 0.632+

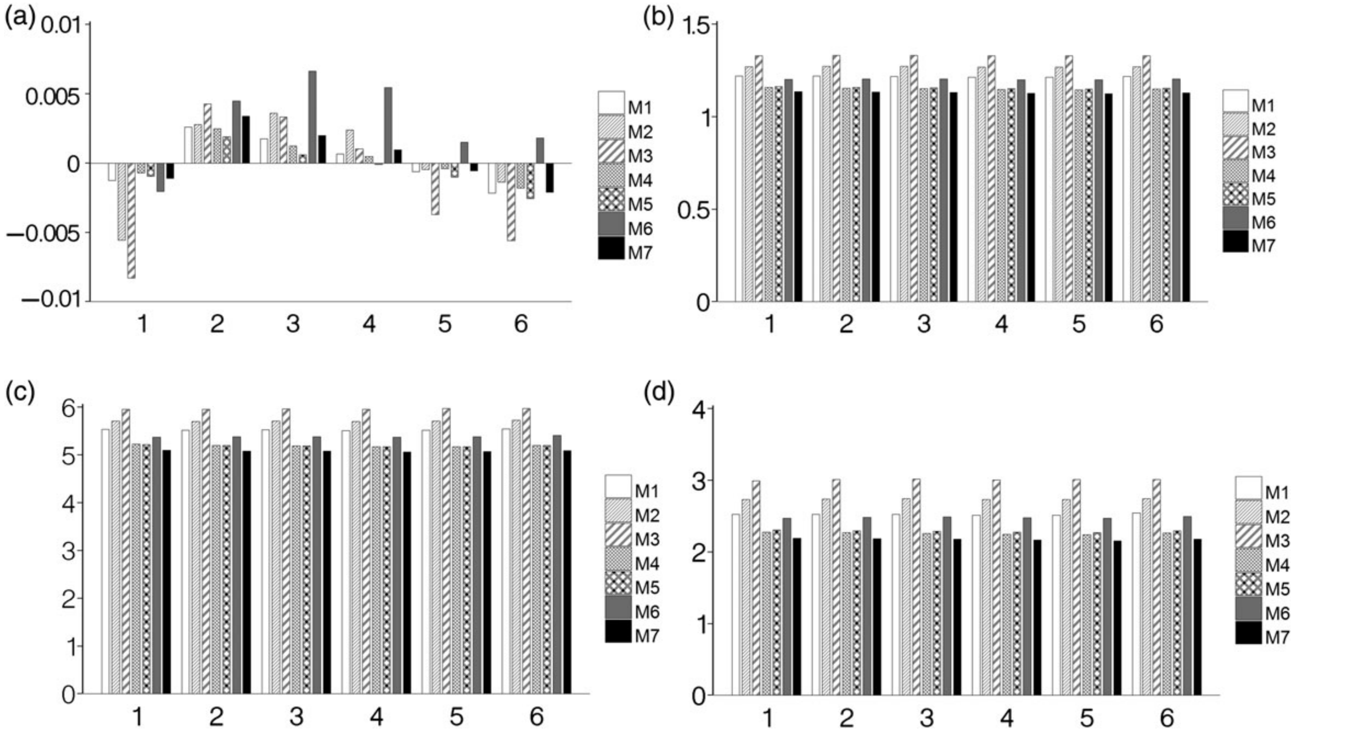

Figure 2 Mean errors (a), mean absolute errors (b), mean absolute percent errors (c) and mean square errors (d) for candidate models M1 – M7 by the *k*-fold method, where 1 – 6 on the *x*-axis are for $k = 5$, 10, 20, 30, 40 and 50.

Table 4 Validation statistics for the seven candidate models (M1–M7) calculated on model fitting and model validation data, and the four validation statistics plus their relative standard errors from the 5-fold, LOP, 0.632+ and LOT methods calculated on model testing data that were generated from model fitting data

Data/CV method	Model	ME	MAE	MAPE	MSE
Fitting data	M1	0.0009 (5)	1.232 (5)	5.577 (5)	2.574 (5)
	M2	−0.0002 (3)	1.286 (6)	5.772 (6)	2.792 (6)
	M3	−0.00011 (1)	1.340 (7)	6.001 (7)	3.045 (7)
	M4	0.0012 (7)	1.167 (2)	5.250 (3)	2.315 (2)
	M5	0.0010 (6)	1.172 (3)	5.248 (2)	2.343 (3)
	M6	−0.00012 (2)	1.218 (4)	5.445 (4)	2.532 (4)
	M7	0.0004 (4)	1.147 (1)	5.143 (1)	2.236 (1)
Validation data	M1	0.0522 (4)	1.391 (5)	6.762 (5)	3.397 (5)
	M2	0.1020 (6)	1.469 (6)	7.081 (6)	3.729 (6)
	M3	0.1968 (7)	1.519 (7)	7.238 (7)	3.998 (7)
	M4	0.0369 (1)	1.311 (2)	6.377 (3)	3.049 (2)
	M5	0.0424 (3)	1.313 (3)	6.340 (2)	3.052 (3)
	M6	0.0570 (5)	1.354 (4)	6.438 (4)	3.202 (4)
	M7	0.0386 (2)	1.271 (1)	6.132 (1)	2.900 (1)
5-Fold	M1	−0.0012 (4, −184)	1.222 (5, 1.55)	5.538 (5, 1.86)	2.530 (5, 3.31)
	M2	−0.0056 (6, −209)	1.272 (6, 1.54)	5.713 (6, 1.84)	2.732 (6, 3.22)
	M3	−0.0083 (7, −194)	1.329 (7, 1.64)	5.960 (7, 1.99)	2.994 (7, 3.43)
	M4	−0.0007 (1, −187)	1.159 (2, 1.57)	5.226 (3, 1.82)	2.281 (2, 3.31)
	M5	−0.0009 (2, −180)	1.165 (3, 1.57)	5.225 (2, 1.77)	2.309 (3, 3.31)
	M6	−0.0020 (5, −179)	1.204 (4, 1.56)	5.380 (4, 1.78)	2.470 (4, 3.27)
	M7	−0.0011 (3, −217)	1.136 (1, 1.56)	5.102 (1, 1.78)	2.193 (1, 3.30)
LOP	M1	0.0009 (5, −168)	1.233 (5, 1.55)	5.583 (5, 1.87)	2.578 (5, 3.32)
	M2	0.0002 (1, −240)	1.287 (6, 1.55)	5.779 (6, 1.84)	2.798 (6, 3.23)
	M3	0.0003 (2, −165)	1.342 (7, 1.64)	6.010 (7, 1.99)	3.054 (7, 3.44)
	M4	0.0012 (7, −169)	1.168 (2, 1.57)	5.256 (3, 1.81)	2.319 (2, 3.33)
	M5	0.0010 (6, −145)	1.174 (3, 1.58)	5.255 (2, 1.77)	2.348 (3, 3.33)
	M6	0.0005 (4, −206)	1.220 (4, 1.57)	5.454 (4, 1.80)	2.540 (4, 3.29)
	M7	0.0004 (3, −174)	1.148 (1, 1.57)	5.152 (1, 1.79)	2.242 (1, 3.33)
0.632+	M1	0.0012 (6, 25)	1.233 (5, 0.05)	5.586 (5, 0.06)	2.581 (5, 0.11)
	M2	0.0003 (1, −224)	1.288 (6, 0.06)	5.783 (6, 0.06)	2.803 (6, 0.12)
	M3	0.0004 (2, −281)	1.342 (7, 0.06)	6.015 (7, 0.07)	3.059 (7, 0.12)
	M4	0.0013 (7, 14)	1.169 (2, 0.05)	5.260 (3, 0.06)	2.323 (2, 0.11)
	M5	0.0010 (5, 64)	1.174 (3, 0.05)	5.259 (2, 0.06)	2.351 (3, 0.11)
	M6	0.0007 (4, −47)	1.221 (4, 0.06)	5.461 (4, 0.06)	2.546 (4, 0.12)
	M7	0.0006 (3, 14)	1.149 (1, 0.05)	5.158 (1, 0.06)	2.247 (1, 0.11)
LOT	M1	−0.0015 (1, −98)	0.906 (4, 0.12)	4.221 (5, 0.15)	1.364 (4, 0.24)
	M2	−0.0037 (2, −44)	0.978 (6, 0.13)	4.269 (6, 0.14)	1.773 (6, 0.28)
	M3	−0.0154 (3, −12)	1.081 (7, 0.14)	4.613 (7, 0.14)	2.273 (7, 0.28)
	M4	0.0240 (5, 5)	0.822 (2, 0.13)	3.771 (3, 0.16)	1.114 (2, 0.25)
	M5	0.0220 (4, 6)	0.825 (3, 0.13)	3.761 (2, 0.16)	1.122 (3, 0.26)
	M6	0.0265 (6, 6)	0.914 (5, 0.13)	4.080 (4, 0.15)	1.502 (5, 0.29)
	M7	0.0317 (7, 4)	0.810 (1, 0.13)	3.737 (1, 0.16)	1.082 (1, 0.26)

For model fitting and validation data, numbers in brackets are the rankings of the seven models with ranking 1 being the smallest and 7 being the largest in absolute values. Lower ranking numbers indicate better models, with ranking 1 being the best model. For the 5-fold, LOP, 0.632+ and LOT methods, two numbers are presented in each pair of brackets: the first one is the model ranking number as for model fitting and validation data, and the second one is the relative standard error of the corresponding statistics.

methods. All ME values were close to zero for model fitting data and the three CV methods, and the model rankings based on these values was not very meaningful. The absolute values of the relative standard errors for ME were very large, especially for the 5-fold and LOP methods where all absolute values were greater than 100 per cent, and as large as 240 per cent for M2 from the LOP method. These large relative standard errors were caused mainly by the small ME values as the denominators for calculating relative standard errors. Therefore, ME was not used for further evaluation of the CV methods.

On the validation data, MAE and MSE gave identical rankings for all seven models, from the best to the worst: M7–M4–M5–M6–M1–M2–M3. MAPE produced similar rankings, but model M5 was better than M4. Identical model rankings were also observed for MAE, MSE and MAPE on model fitting data as those on model validation data. The main difference was that the magnitudes of these statistics were larger on validation data, which was expected. These statistics are different measures of prediction errors, and prediction errors are in general smaller on model fitting data than those on model validation data.

The 5-fold, LOP and 0.632+ CV methods also produced identical model rankings for MAE, MSE and MAPE as those on model validation data, indicating that the three CV methods were reliable for examining the predictive capabilities of the selected models. The MAE, MSE and MAPE values for all three CV methods were in between those values calculated on model fitting and validation data, but much closer to those calculated on model fitting data. This was also expected since all three CV methods were based on the original model fitting data. Among the three CV methods, 0.632+ bootstrap method produced the closest MAE, MSE and MAPE values to those on model validation data, followed by LOP and 5-fold methods. This was true for all seven models (Table 4). Therefore, the 0.632+ method was the best one for validating the selected nonlinear mixed height–diameter models, either the best models with one, two and three random parameters or the best model of all seven models.

When relative standard errors were examined, it was obvious that the 0.632+ method produced more precise validation statistics than the 5-fold and LOP methods. All relative standard errors associated with MAE, MSE and MAPE were <0.2 per cent from the 0.632+ method, compared with ~1.6 per cent for MAE, ~1.8 per cent for MAPE and ~3.3 per cent for MSE from the 5-fold and LOP methods. It was also clear that the 5-fold and LOP methods produced similar relative standard errors associated with MAE, MSE and MAPE, an indication that a smaller *k* value was good enough for efficient evaluation of the predictive capabilities of the seven nonlinear mixed models, and the LOP method could be replaced by the 5-fold method.

For the LOT method, the mean errors were very small for all seven models with relative large standard errors. But the relative standard errors associated with the other three statistics were small (<0.3 per cent). Unlike the other three CV methods, the model rankings by MAE, MSE and MAPE for the LOT method did not follow the model rankings on model fitting and validation data. For example, the LOT method ranked M1 better than M6 based on MAE and MSE, which was opposite on model validation data. Notice that the MAE, MSE and MAPE values from the LOT method were much smaller than the corresponding values on validation data. They were also smaller than the corresponding values on model fitting data. Therefore, using the LOT method for validating a selected model may give false indication that the selected model provides reliable predictions, which may or may not be true. Therefore, the LOT method was not suitable for evaluating the predictive performance of the candidate models, and an independent validation dataset was preferred.

Discussion

Regression models in forestry are often developed for making predictions on new data not used for model fitting. Typically several candidate models are fitted and a final model is chosen based on some model fitting statistics. The selected model is sometimes further evaluated for its predictive ability on an independent dataset. If we are in a data-rich situation, the best approach is to randomly divide a dataset into three parts: a training set, a validation set and a test set. The first two datasets are used for model selection, and the last one is used to evaluate the predictive ability of the final model (Hastie *et al*. 2009). More often than not, however, modellers have to deal with situations where there is not enough data for such data partition. CV provides an alternative for performance evaluation of a selected model in situations where there is not enough data by efficient sample re-use. It splits a model fitting dataset to model training and model testing datasets numerous times to mimic the data partition as suggested by Hastie *et al*. (2009). It is well known that training an algorithm and evaluating its statistical performance on the same data leads to an overly optimistic result. CV was developed to address this issue based on the concept that testing the output of an algorithm on new data would lead to a better estimate of its performance (Stone 1974). The basic idea is that if a prediction model is valid, it should also predict effectively in a second sample from the population (Wildman 2011).

This study examined five CV methods for their suitability for evaluating the predictive performance of seven candidate nonlinear mixed models based on a height–diameter relationship. The five CV methods were grouped into two categories based on how the two-level data structure was considered. The first category included the *k*-fold, LOP, 0.632 and 0.632+ methods, where plot was used as the basic unit for data resampling. The second category included only the LOT method, where tree was used as the basic unit for data resampling.

It was found that the four CV methods in the first category could be used for reliable evaluation of the predictive performance of the selected models through MAE, MAPE and MSE and their associated relative standard errors. For this evaluation, the rankings of the candidate models on model validation data were considered as proper, and the magnitudes of the validation statistics were important. Of the four methods, the 0.632 and 0.632+ methods were best suited for evaluating the predictive performance of the selected models, followed by LOP and *k*-fold methods.

Although the 0.632 and 0.632+ bootstrap methods produced almost identical results in this study, the 0.632+ method is generally preferred. Instead of using a constant value of 0.632 for bootstrap estimate of prediction errors of the data not included in the bootstrap sample (equation 5), the 0.632+ method takes into account the amount of over-fitting by including the relative over-fitting rate R in the calculation of ω (equation 6), which can be used to check if a model is over fitted. Of the candidate models evaluated, the most complicated model was M7 with three random parameters. The relative over-fitting rates associated with the four validation statistics were all close to zero (<0.004), indicating that M7 was not over-fitted. The other six models had even smaller over-fitting rates.

When evaluated by the LOT method, much smaller mean absolute errors, mean absolute percent errors and mean square errors were observed for all seven candidate models compared with corresponding statistics calculated on model validation data. They were also smaller than those calculated on model fitting data. We suspected that part of the reason for the smaller statistics was that only one tree per plot was reserved as testing data. To test if it was true, several other scenarios were evaluated where two, three and four trees per plot were randomly selected and set aside as testing data. The results indicated that as the number of trees per plot reserved as testing data increased, the MAE, MSE and MAPE values also increased (results not presented), confirming that the smaller statistics were partly due to the fact that only one tree per plot was reserved as testing data. However, the increased validation statistics were still smaller than the corresponding values on model fitting data. We thought it was caused by the data resampling approach adopted by the LOT and associated methods. Trees from the same plot were in both model training and model testing datasets, and as a result, the two datasets were not independent.

Kozak and Kozak (2003) examined the usefulness of several CV procedures on evaluating the predictive capabilities of three forestry models fitted by ordinary least squares. They concluded that prediction errors and other fit statistics based on these procedures provided little, if any, incremental information compared with calculating prediction errors and other fit statistics directly on model fitting data. Robinson and Wykoff (2004), on the other hand, reported that CV led to conclusions quite different from, and arguably more reliable than, the results from the fitted models. As a result, they were reluctant to reject the use of CV as a model comparison tool. This study demonstrated that the *k*-fold, LOP, 0.632 and 0.632+ methods were suitable for evaluating the predictive performance of the selected models, but the LOT method was not. These results, however, were based on a case study and the outcome was local. Further studies are needed to see if the conclusions still hold for other datasets and/or other nonlinear mixed models.

Conclusions

We examined five CV methods for their suitability in performance evaluation of nonlinear mixed-effects height–diameter models, while considering the two-level hierarchical data structure (trees nested in plots). Plot was used as the basic unit for data resampling for the *k*-fold, LOP, 0.632 and 0.632+ methods, while tree was used as the basic unit for data resampling for the LOT method. It was found that the *k*-fold, LOP, 0.632 and 0.632+ methods were suitable for evaluating the predictive performance of the selected model(s), and the 0.632+ method was the best one. All four methods can be applied to situations where predictions are needed for all trees in a plot not used for model fitting.

The LOT method was fundamentally different from the other four methods in that tree was used as the basic unit for data resampling. It can be applied to situations where predictions are needed for a portion of trees in a plot not used for model fitting. It was found that the LOT method was not suitable for evaluating the predictive performance of the selected model(s).

Acknowledgements

This research was supported by the Alberta Government. Mr Darren Aitkin reviewed an early draft and provided many constructive suggestions. We thank the three anonymous reviewers and the Editor for their thoughtful comments and valuable suggestions.

Funding

This work was supported by Alberta Environment and Sustainable Resource Development.

Conflict of interest statement

None declared.

References

Afshartous, D. and de Leeuw, J. 2004 An application of multilevel model prediction to NELS:88. *Behaviormetrika* **31**, 43–66.

Arlot, S. 2010 A survey of cross-validation procedures for model selection. *Stat. Surv.* **4**, 40–79.

ASRD. 2005 *Permanent sample plot (PSP) field procedures manual*. Forest Management Branch, Alberta Sustainable Resource Development, Edmonton, Alberta, 30 pp.

Budhathoki, C.B, Lynch, T.B and Guldin, J.M 2008 A mixed-effects model for the dbh-height relationship of shortleaf pine (*Pinus echinata* Mill.). *South. J. Appl. For.* **32**, 5–11.

Efron, B. 1979 Bootstrap methods: another look at jackknife. *Ann. Stat.* **7**, 1–26.

Efron, B. 1983 Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Stat. Assoc.* **78**, 316–331.

Efron, B. and Tibshirani, R. 1997 Improvements on cross-validation: the 632+ bootstrap method. *J. Am. Stat. Assoc.* **92**, 548–560.

Gelfand, A.E., Dey, D.K and Chang, H. 1992 Model Determination Using Predictive Distributions with Implementation via Sampling Based Methods. Technical Report No. 462, Department of Statistics, Stanford University, Stanford, California, 38 pp.

Greenland, S. 2000 Principles of multilevel modeling. *Int. J. Epidemiol.* **29**, 158–167.

Hastie, T., Tibshirani, R. and Friedman, J. 2009 *The elements of statistical learning: data mining, inference, and prediction*, 2nd ed. Springer Series in Statistics, 745 pp.

Huang, S., Meng, S.X. and Yang, Y. 2009 Prediction implications of nonlinear mixed-effects forest biometric models estimated with a generalized error structure. In *Proceedings of Joint Statistical Meetings, Section on Statistics and the Environment, American Statistical Association*. August 1–6, , D.C pp. 1174–1188.

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.

Lappi, J. 1991 Calibration of height and volume equations with random parameters. *For. Sci.* **37**, 781–801.

Larson, S. 1931 The shrinkage of the coefficient of multiple correlation. *J. Educ. Psychol.* **22**, 45–55.

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D. and Schabenberber, O. 2006 *SAS for Mixed Models*, 2nd ed. SAS Institute Inc. 814 pp.

Mosteller, F. and Turkey, J.W. 1968 Data Analysis, Including Statistics. In *Handbook of Social Psychology*. Addison-Wesley, pp. 601–720.

Ren, S., Lai, H., Tong, W., Aminzadeh, M., Hou, X. and Lai, S. 2010 Nonparametric bootstrapping for hierarchical data. *J. Appl. Stat.* **37**, 1487–1498.

Robinson, A.P. and Wykoff, W.R. 2004 Imputing missing height measures using a mixed-effects modeling strategy. *Can. J. For. Res.* **34**, 2492–2500.

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.

Stone, M. 1974 Cross-validatory choice and the assessment of statistical predictions. J. Roy. Stat. Soc. Ser B. **36**, 111–133.

Temesgen, H., Monleon, V.J. and Hann, D.W. 2008 Analysis and comparison of nonlinear tree height prediction strategies for Douglas-fir forests. *Can. J. For. Res.* **38**, 553–565.

Trincado, G. and Burkhart, H.E. 2006 A generalized approach for modeling and localizing stem profile curves. *For. Sci.* **52**, 670–682.

Vonesh, E.F. and Chinchilli, V.M. 1997 *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. Marcel Dekker, 560 pp.

Wildman, T. 2011 Factors that influence cross-validation of hierarchical linear models. Educational Policy Studies Dissertations, Department of Educational Policy Studies, Georgia State University.

Yang, Y. and Huang, S. 2011 Estimating a multilevel dominant height-age model from nested data with generalized errors. *For. Sci.* **57**, 102–116.

Yang, Y., Huang, S., Meng, S.X., Trincado, G. and VanderSchaaf, C.L. 2009 A multilevel individual tree basal area increment model for aspen in boreal mixedwood stands. *Can. J. For. Res.* **39**, 2203–2214.

Zhang, L. 1997 Cross-validation of non-linear growth functions for modelling tree height–diameter relationships. *Ann. Bot.* **79**, 251–257.