Available online at www.sciencedirect.com
SCIENCE DIRECT®
Forest Ecology and Management 214 (2005) 201–211
Forest Ecology and Management
www.elsevier.com/locate/foreco

Evaluation of spatial predictions of site index obtained by parametric and nonparametric methods—A case study of lodgepole pine productivity

Yonghe Wang [a,*], Frédéric Raulier [b], Chhun-Huor Ung [c]

[a] *Canadian Forest Service, Northern Forestry Centre, 5320-122 Street, Edmonton, Alta., Canada T6H 3S5*
[b] *Faculty of Forestry and Geomatics, Laval University, Sainte-Foy, Que., Canada G1K 7P4*
[c] *Canadian Forest Service, Laurentian Forestry Centre, Sainte-Foy, Que., Canada G1V 4C7*

Received 1 October 2004; received in revised form 12 April 2005; accepted 12 April 2005

Abstract

We demonstrate the potential of using least-squares regression, generalized additive model, tree-based model, and neural network model on layers of environmental data grids for mapping site index in a case study. Grids of numerical environmental variables represented layered data, and a sparse site index plot network was located in the grids. Site index data were based on stem analysis (observed height at the index age of 50 years) of 431 lodgepole pine trees in 88 sample plots. The plots were established in a 17,460 km^2 boreal mixedwood forest of Alberta, Canada dominated by mature and over-mature stands. The generalized additive model presented a better fit and better adaptability to extreme data (i.e., mature stands) than the least squares nonlinear and other nonparametric techniques, such as the tree-based model and neural network model. Among the four models tested, nonlinear regression is of the data modeling culture, which assumes a stochastic data to relate productivity to environmental variables, and such models are optimized for estimation. Other three models belong to the algorithm modeling culture, which treat the relationship between productivity and independent variables as an unknown black box and try to find a function between them; therefore, these models are more suitable for prediction purpose. Implications for biophysical site index modelling with extreme data are discussed.

Keywords: Generalized additive model; Tree-based model; Neural network model; Lodgepole pine; Nonlinear regression; Mature stands; Over-mature stand

1. Introduction

Although there has been a paradigm shift from simple, even-aged silviculture and growth modelling to gap-based silviculture and uneven-aged growth modelling, a quantified surrogate of potential wood production is still required for forest management decision-making. Because of its operational importance, site index (SI), defined as the average height of a certain number of the largest trees per hectare at a particular reference age, is a broadly accepted surrogate of

* Corresponding author. Tel.: +780 435 7237; fax: +780 435 7359.
E-mail address: ywang@nrcan.gc.ca (Y. Wang).

0378-1127/$ – see front matter.
doi:10.1016/j.foreco.2005.04.025

potential productivity. There are many cases where site index cannot be adequately estimated because it is a species-specific and phytometric index of site productivity. For example, it cannot be estimated in treeless areas or in stands where the concerned species is not present. Its estimation in young stands is difficult because a slight error in estimation can result in a large error on predicted yields.

Various studies have thus attempted to relate SI to biophysical factors (Hunter and Gibson, 1984; Kabzems and Klinka, 1987; Ung et al., 2001; McKenney and Pedlar, 2003), but observed correlations are generally low, and the relationships differ among studies because of no definitive connection between SI and the biophysical variables. Almost all of these studies have used a parametric approach that implied solving two problems: the relationship definition between SI and the biophysical variables and collinearity existing between the explanatory variables in the regression model. Given that a relationship needs to be determined before parameter estimation and that no definite relationship exists, the use of the least square regression technique implies an unknown source of error that could be bypassed with a nonparametric approach. Also, collinearity often exists among biophysical variables when some of these can be expressed as linear combinations of other predictor variables. For example, climate moisture index, which is used in this study, is derived directly from temperature and precipitation, which in turn are related to latitude and elevation. Collinearity affects the statistical estimation of the parameters as it inflates the variance of at least one of the estimated regression coefficients, and consequently also inflates the estimation of the confidence interval around the predicted values (Belsley et al., 1980).

To avoid these problems, alternative methods must be identified and assessed. For instance, McKenney and Pedlar (2003) have successfully applied a tree-based regression method (TREE) for relating soil, topographic, and climatic attributes to site productivity. TREE has the potential of producing discontinuous SI values; other nonparametric methods such as the generalized additive model (GAM) and the neural network model (NNT) could generate satisfactory results because their outputs are continuous. TREE, GAM, and NNT are considered nonparametric because no functional structure between predictor and response variables is pre-specified. GAM and NNT have been used as exploratory tools in the analysis of species distribution with respect to climatic factors in a landscape study (Yee and Mitchell, 1991) and as tree growth and mortality models (Guan and Gertner, 1991a,b; Sironen et al., 2003), but, as far as we know, they have not yet been used to predict SI in spatial terms.

The objective of this study was to assess the usefulness of the TREE, GAM, and NNT nonparametric techniques, and to compare them with the least-squares nonlinear regression model (NLIN) in developing a spatial SI model for mature stands of lodgepole pine (*Pinus contorta* var. *latifolia*) in Alberta, Canada.

2. Materials and methods

2.1. Study site and data sources

The Wapiti region of Alberta, Canada, encompasses an area of approximately 17,460 km^2 between 118°W and 120°W and between 54°N and 55°N. It is a mosaic of four natural ecological subregions: Boreal Mixedwood, Lower Foothills, Upper Foothills, and Sub Alpine (Beckingham et al., 1996). Lodgepole pine is the dominant timber species across the region (Corns, 1978). The nine predictor variables used in this study consisted of grid data from various sources. They include three geographical factors: the easting (m) grid (x_1) and the northing (m) (x_2) of the Universal Transverse Mercator (UTM) ordination, along with elevation (m) (x_3). In addition, six biophysical factors are considered to influence variability in site conditions within the study region: climate moisture index (cm) (x_4), growing degree days (°C) (x_5), annual precipitation (mm) (x_6), soil sand fraction (%) (x_7), January monthly mean temperature (°C) (x_8), and July monthly mean temperature (°C) (x_9).

Grid x_3 was created from the National Topographic Database 1:50,000 contour lines projected to UTM Zone 11, NAD 83. Grid x_4 was obtained as the difference between x_6 and the annual sum of monthly potential evapotranspiration (PE), which was computed from relevant data grids in the NatGRID database (Hogg, 1994; McKenney et al., 1996). Monthly PE was determined by means of the

Jensen-Haise method (Bonan, 1989; Jensen et al., 1990). Grids x_5, x_6, x_8, and x_9 came from the NatGRID database, with x_5 calculated according to the method of Moran and Morgan (1997). Grid x_7 was from a dominant soil texture grid created with the Soil Landscapes of Canada, version 2.0 (Shields et al., 1991; SCDWG, 1993; Schut et al., 1994).

The original resolutions were 10 km for x_6 and x_7 and 1 km for x_4, x_5, x_8, and x_9. For spatial prediction purpose, all these grids were converted to a resolution of 100 m for consistency with the resolution of grid x_3. Each of the nine grids consisted of 1149 rows and 1326 columns of cells to cover the Wapiti region.

Stem analysis data for 431 lodgepole pine trees were gathered from a plot network of 88 sample plots distributed across the Wapiti region, along with other stand information such as the mean annual increment (MAI) of volume and the current stand density. The plots were selected to relate vegetation variables with tree growth, soil, and landform types within uniform, even-aged normally stocked stands ranging from 45 to over 200 years old (Corns, 1978; Corns and Pluth, 1984). SI, defined in this study as the average height of all trees on each plot at an index age of 50 years, was adopted as the measure of productivity.

2.2. Model description

Linear and nonlinear regression modelling are parametric methods. However, linear regression was excluded from this analysis, because its goodness-of-fit in trial runs was much worse than that of other models. The parametric NLIN model constructed by Ung et al. (2001) states that SI is the product of the polynomials of the predictor variables:

$$\mathrm{SI} = c_0 \prod_{i=1}^{9} \left[1 + c_{i1} \left(\frac{x_i - \bar{x}_i}{\bar{x}_i} \right) + c_{i2} \left(\frac{x_i - \bar{x}_i}{\bar{x}_i} \right)^2 \right], \quad i = 1, 2, \ldots, 9 \tag{1}$$

where c_0, c_{i1}, and c_{i2} are parameters, and $\bar{x}$ is the mean value of x. In this equation, each variable is expressed as a second-order polynomial with its value centred and reduced relative to its mean.

Detailed discussion on the TREE model can be found in Venables and Ripley (1994); therefore, only a brief overview is provided here. With a binary partitioning algorithm, the TREE technique produces a decision tree by recursively splitting the data until the observations of the response variable in the groups are either homogeneous or the groups contain a user-defined minimum number of observations. Then, the deviance is the sum over leaves, which is the corrected sum of squares for cases within that node, and the value of a split is the reduction in the residual sum of squares. The probability model for SI within each leaf of the tree takes a normal distribution, and the tree-construction process is a hierarchical refinement of probability models, similar to the forward variable-selection method in conventional regression (McKenney and Pedlar, 2003).

The GAM model has the following form:

$$\mathrm{SI} = b_0 + \sum_{i=1}^{9} f_i(x_i) + \varepsilon \tag{2}$$

where $f_i(x_i)$ is a smoothing function for variable i, usually a cubic spline. b_0 and ε are parameter and random variable, respectively. A back-fitting algorithm was used to fit the equation; given Eq. (2), for any predictor variable k, there exists the following relationship:

$$E\left[\mathrm{SI} - b_0 - \sum_{j \neq k} f_j(x_j) | x_k \right] = f_k(x_k) \tag{3}$$

An iterative algorithm was used for computing all functions (Hastie and Tibshirani, 1990).

NNT is an information processing system that consists of many simple processors or neurons and many weighted interconnections among them (Fausett, 1994). Among the NNT methods, the backward propagation method (Russell and Dobbins, 1990) was used in this study. This model has three layers: an input layer with the nine predictor variables as input nodes (or neuron), a "hidden" layer with several hidden nodes and an output layer with only one node, the predicted SI. The node in the output layer is connected with all nodes in the hidden layer, and every node in the hidden layer is linked with all nodes in the input layer; however, nodes in the same layer are not connected. Each connection between a pair of nodes is associated with a weight value ranging from 0 to 1, indicating the magnitude of the signal sent from the lower-layer node and received by the upper-layer node. Since theoretical guidance on how to choose the

number of hidden nodes is lacking, a series of trials from five to nine hidden nodes was conducted. For the final modelling, the number of hidden nodes was five, because increasing the number above five did not significantly improve the fit. The NNT model was repeatedly ''trained'' with the field data until a convergence criterion was met or the limit of iteration was reached (Russell and Dobbins, 1990).

The NNT model has a potential for over-learning. As the number of hidden nodes and/or number of iterations increases, the sum of the squared errors (SSE) decreases. Such an improvement in fit could result in an ''over-learned'' NNT, which would produce larger errors if the model was applied to the population predictions from which the field data were derived (Zhang et al., 2000). The over-learning problem can be solved by splitting the field data into two subsets: one for training and one for validation. Because the field data in this study were relatively few, a simple empirical approach was adopted: select a series of convergence criteria, train the NNT several times with a different number of iterations for each criterion, then analyse the errors to choose a combination of convergence criterion and iteration that gives a reasonable fit.

Eq. (1) was fitted to the field data set by means of SAS software (SAS Institute Inc., 1989). The TREE and GAM models were fitted by means of S-PLUS (Statistical Science, 1993). The NNT model was trained by means of the algorithm and program provided by Russell and Dobbins (1990). The four fitted models were used to predict SI spatially for the Wapiti region with the nine environmental data layers as input.

2.3. Model evaluation

Two approaches were adopted to evaluate four models. To select the ''best'' model, it is necessary to obtain a good approximation of the generalization error of each model, being the average error that the model would make on an unknown test set independent from the fitting data (Vanclay and Skovsgaard, 1997; Simon et al., 2003). The four models could have been compared with such independent data; however, because of the low number of plots used in this study, we first used the bootstrap method for model evaluation. Bootstrapping is advantageous when the number of samples is limited (Simon et al., 2003), which involves repeated sampling from the fitting data, with replacements, to create a series of samples with the same size of the original data. As suggested by Efron and Tibshirani (1993), the procedure was repeated 200 times to generate 200 bootstrap samples. The four models were then fitted to each of the bootstrap samples either to estimate the parameters (for NLIN) or to establish the relationship between SI and the predictors (for TREE, GAM, and NNT). The SSE for the residuals (SSE_r), referred to as ''apparent errors'', was then computed for each bootstrap sample. Alternatively, the fitted models for the 200 samples could be applied to the original data to produce another 200 SSE_rs, or ''original errors''. The difference between the averaged original errors and the averaged apparent errors is ''optimism''. The optimism was added to the SSE_r based on the original data to obtain a better estimate of the generalization errors (Efron and Tibshirani, 1993).

In the second approach, SI was used as input for estimation of volume yield to evaluate the possible impacts of different models on volume estimation. MAI is often used in economic analyses because it is associated with value and rate of return on investment (Fries et al., 1998; Avery and Burkhart, 2002; Woods et al., 2000). Three sources of MAI were used in the evaluation based on the fitting data: the observed MAI from sample plots and two MAIs derived from GYPSY, a growth and yield model used operationally for lodgepole pine in Alberta (Huang et al., 2001).

Firstly, the observed SI and the observed current stem density were used as input to GYPSY to predict MAI. Secondly, four estimated SIs from the NLIN, TREE, GAM, and NNT models were used as input to GYPSY to predict MAI. The MAIs predicted by GYPSY were then evaluated against the observed MAI by calculating bias and residual mean squared error (MSE_r). A linear relationship was found between the predicted and measured MAI values such that an R^2 value could be computed. Thus, the bias, MSE_r, and R^2 could be used to evaluate the models. Fig. 1 is a schematic depiction on the process of model fitting, evaluation, and comparison.

3. Results and discussion

Most lodgepole pine trees were from mature or over-mature stands (Fig. 2). Preliminary statistics for

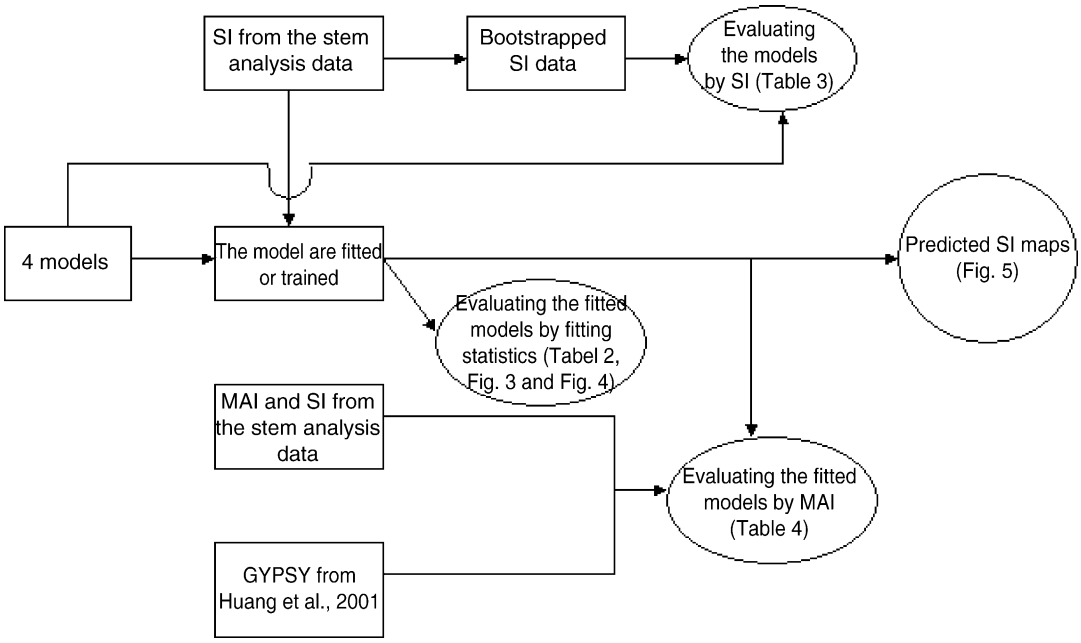

Fig. 1. The flow chart of the process of model fitting, evaluation, and comparison.

variables from both field and grid data were summarized (Table 1). The scatterplots between SI and grids x_1 through x_9 suggested a positive correlation between SI and x_2, x_5, and x_9 and negative correlation between SI and x_3, x_4, x_6, and x_8. No other relationships were apparent. The counterintuitive correlations between SI and x_2 and x_6 are due to the orography of the Wapiti region. Precipitation is positively correlated with elevation (rain shadow),

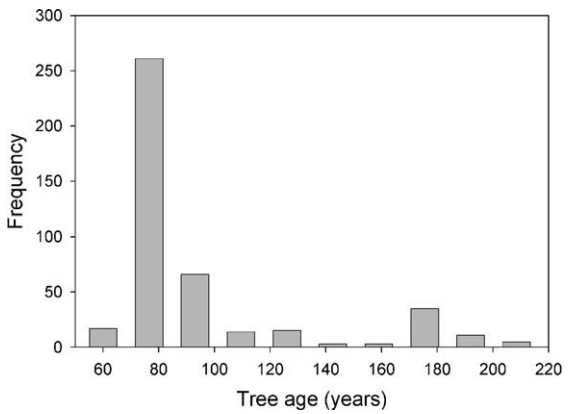

Fig. 2. Age distribution of 431 lodgepole pine trees in the Wapiti region of Alberta, Canada.

and elevation is higher in the southwest of the region, such that lower elevations, even though they are drier, are more productive in terms of SI because of the milder temperatures (Corns, 1978). According to the plot of SI versus x_3, productivity was high at low elevation and decreased with increasing elevation.

According to Corns (1978), the Boreal Mixedwood subregion is primarily in the north of Wapiti at elevations lower than 900 m, the Lower Foothills subregion runs across the centre of the region from west to east, at elevations from 900 to 1200 m, and the Upper Foothills subregion is primarily located in the south and southwest portions of the area at elevations from 1200 to 1400 m. The Sub Alpine subregion, with elevations from 1400 to 2000 m, is in the southwestern corner of Wapiti. Therefore, the spatial distribution of SI should be high in the north, medium in the middle and low in the southwestern part.

Most of the stands in this analysis are mature or over-mature (Smith and Resh, 1999), with the older stands being largely beyond the expected fire-return interval (Wei et al., 2003). Predicting productivity for senescent stands is challenging because the dominant height during that period is also influenced by the mortality of the dominant stems; therefore, it does not

Table 1
Summary statistics of for site index and nine predictor variables of field and grid data[a]

Variable	Mean	S.D.	CV	Minimum	Maximum
Fitting data set[b]					
Site index (m)	12.8	3.2	25.3	6.3	19.8
x Coordinate (m), x_1	358,544.8	34,883.7	9.7	305,206.6	429,982.2
y Coordinate (m), x_2	6,037,628.2	26,839.9	0.4	5,985,620.5	6,097,547.1
Elevation (m), x_3	1,123.7	197.2	17.5	749.0	1,570.0
Climate moisture index (cm), x_4	19.3	8.7	45.3	8.0	50.0
Growing degree days (°C), x_5	861.1	134.8	15.7	480.0	1,069.0
Annual precipitation (mm), x_6	581.3	53.6	9.2	501.0	741.0
Soil sand fraction (%), x_7	32.0	16.9	52.5	10.0	64.0
January mean temperature (°C), x_8	−12.4	1.0	−8.4	−14.7	−10.1
July mean temperature (°C), x_9	13.5	0.9	6.7	10.9	14.9
Grid data set[c]					
x Coordinate (m), x_1	369,665.6	38,278.3	10.4	303,415.6	435,915.6
y Coordinate (m), x_2	6,041,421.5	33,168.8	0.5	5,984,021.5	6,098,821.5
Elevation (m), x_3	1,068.6	306.1	28.6	520.8	2,434.1
Climate moisture index (cm), x_4	19.0	12.1	63.9	3.0	63.0
Growing degree days (°C), x_5	882.7	193.8	22.0	332.0	1,197.0
Annual precipitation (mm), x_6	582.6	83.5	14.3	464.0	933.0
Soil sand fraction (%), x_7	31.9	19.5	61.2	5.0	81.0
January mean temperature (°C), x_8	−12.7	1.5	−11.6	−16.9	−10.0
July mean temperature (°C), x_9	13.6	1.3	9.6	9.7	15.7

[a] Standard deviation (S.D.), coefficient of variation (CV).
[b] $n = 88$ for fitting data.
[c] $n = 1{,}523{,}574$ for grid data.

totally reflect site productivity. For instance, Robichaud and Methven (1993) observed that within break-up black spruce (*Picea mariana* [Mill.] BSP) stands, maximum heights no longer vary with site index. This difficulty is related to the well-known forest decline after stands have reached canopy closure (Kira and Shidei, 1967). Because of the more complex stand dynamics occurring during stand senescence, the stem analysis results may contain extreme data, which make productivity modelling more difficult.

Although all nine variables can be used in fitting, it is possible that fewer variables might yield a similar or even better fit. Therefore, in fitting the NLIN and GAM models, the procedure of all possible regressions (Draper and Smith, 1981) was applied for selecting predictor variables, in which an adjusted R^2 was used to determine the "best" predictors (Fig. 3). Thus, the number of predictor variables in the models would be five, or only x_1, x_2, x_4, x_5, and x_7 would be included in both models. The estimated parameters for NLIN are $b_0 = 10.4792$, $b_{11} = -1.5098$, $b_{12} = 1.5964$, $b_{21} = -16.0546$, $b_{22} = 3652.6$, $b_{41} = 0.3587$, $b_{42} = 0.9538$, $b_{51} = 2.1664$, $b_{52} = 0.7811$, $b_{71} = -0.1870$, and $b_{72} = 0.1650$. The plots in Fig. 3 indicate that GAM always outperforms the NLIN model by achieving smaller residuals when the number of parameters is seven or fewer. However, their performances are similar when eight parameters are

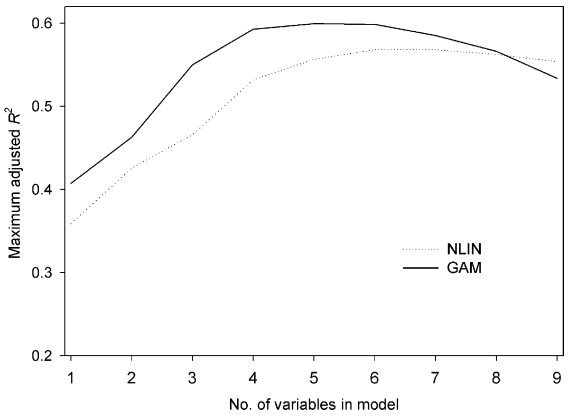

Fig. 3. Maximum adjusted R^2 based on all possible combinations of predictor variables for the nonlinear regression model (NLIN) and the generalized additive model (GAM).

used, and GAM lags behind NLIN when all nine predictor variables are included. This is because each predictor variable in GAM introduces four "parameters" into the model (Statistical Science, 1993), whereas each variable in NLIN uses only two parameters; the adjusted R^2 is calculated from the unadjusted R^2, the number of observations and the number of parameters in the model (Draper and Smith, 1981). TREE has a variable selection mechanism built in, and the selected variables are x_1, x_2, x_3, x_5, x_7, and x_8. Variable was not selected for NNT because the computation required would have been prohibitive; hence, we simply fitted the NNT model on all nine variables.

The four fitted models were first graphically assessed through a comparison of predicted SI versus observed SI and predicted error versus predicted SI (Fig. 4). These plots suggest that NNT and GAM should be better than other models in terms of goodness-of-fit. From good to poor fitting, the four models could be ranked as NNT, GAM, TREE, and NLIN (Table 2). On the basis of the generalization error obtained from bootstrap (Table 3), the models were ranked, from good to poor, as GAM, NLIN (with five variables), NNT, TREE, and NLIN (with nine variables). The change in ranking order is due to a rise of "optimism". Optimism increases roughly linearly with the number of parameters (Simon et al., 2003). Too many parameters lead

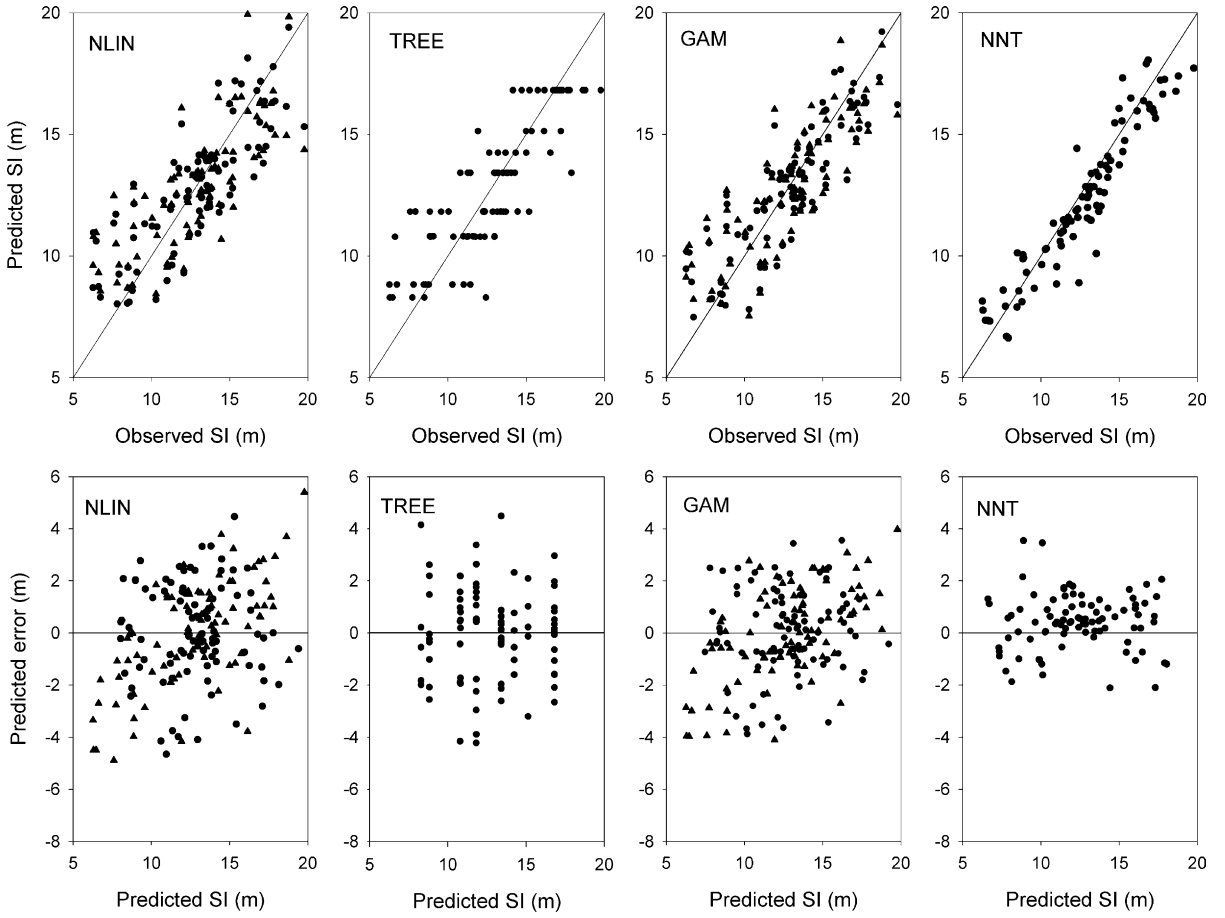

Fig. 4. Scatterplots of predicted site index (SI) values from the four models vs. observed SI (top panels), and of predicted errors vs. predicted SI from the four models (bottom panels). For both NLIN and GAM models, circles and triangles represent values for nine and five predictor variables, respectively. NLIN: nonlinear regression model; TREE: tree-based model; GAM: generalized additive model; NNT: neural network model.

Table 2
SSE_r, R^2, and preliminary statistics of the prediction errors from the four models[a]

Model[b]	SSE_r	R^2	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
NNT	223.53	0.75	−3.91	−0.63	0.66	0.45	1.60	3.52
GAM(9)	249.30	0.73	−3.88	−0.81	0.05	0.00	1.23	3.55
TREE	275.34	0.70	−4.49	−0.99	−0.15	0.00	1.19	4.22
GAM(5)	284.39	0.69	−4.10	−0.93	0.08	0.00	1.37	3.98
NLIN(9)	322.31	0.65	−4.66	−1.19	−0.10	0.01	1.55	4.47
NLIN(5)	357.50	0.61	−4.88	−1.12	0.20	−0.01	1.35	5.40

[a] Minimum value of the errors (Min.); value of the first quartile of the errors (1st qu.); value of the third quartile of the errors (3rd Qu.); maximum value of the errors (Max.).
[b] Values in parentheses indicate number of predictor variables.

Table 3
Results of bootstrap validation of the four models[a]

Model[b]	SSE_1	SSE_2	SSE_3	Optimism	Bootstrap SSE
GAM(5)	284.39	352.50	219.96	132.54	416.93
GAM(9)	249.30	359.35	169.59	189.59	438.89
NLIN(5)	357.50	427.92	307.42	120.50	478.00
NNT	223.53	392.97	136.27	256.70	480.23
TREE	275.34	413.12	151.35	261.77	537.11
NLIN(9)	322.31	957.40	237.16	720.24	1042.55

[a] SSE_r from Table 2 (SSE_1), averaged original error (SSE_2), averaged apparent error (SSE_3), optimism = $SSE_2 - SSE_3$, bootstrap SSE = SSE_1 + optimism.
[b] Nonlinear regression model (NLIN); tree-based model (TREE); generalized additive model (GAM); neural network model (NNT); values in parentheses indicate number of predictor variables.

to overfitting, or in other words, to a low prediction error on the training data set but large prediction errors on independent test sets (excessive optimism). Thus, preference should be given to the GAM and NLIN (with five variables) models. The ranking order is also altered if evaluated by MAI: NNT, TREE, GAM, and NLIN (from good to poor) (Table 4). Disregarding NNT and TREE because of excessive optimism, the preference should be given to GAM when compared with NLIN. A significant bias is observed when MAI is predicted by means of GYPSY with either observed or predicted SI as input. Such a bias is expected, given that the study area represents only a small part of the calibration area of GYPSY (Huang et al., 2001). The bias from NNT is relatively large, and the other three models have similar bias. The performance of NLIN is conditional on the choice of model, but not many such models are available for constructing biophysical SI model. Eq. (1) presents the latest advance in nonlinear regression method applied to productivity-environment situations regardless of the characteristics of individual situation.

Table 4
Results of evaluation based on predicted mean annual increment (MAI) vs. observed MAI

Precision measure	Observed MAI	NNT	TREE	GAM (nine variables)	GAM (five variables)	NLIN (nine variables)	NLIN (five variables)
MAI predicted through GYPSY, based on site index from various models[a]							
Bias ($m^3\ ha^{-1}\ year^{-1}$)	0.32[b]	0.46	0.37	0.36	0.36	0.36	0.37
MSE_r[c]	28.9	35.6	38.5	49.6	48.0	54.0	48.5
R^2	0.56	0.46	0.42	0.25	0.28	0.19	0.27

[a] Neural network model (NNT); tree-based model (TREE); generalized additive model (GAM); nonlinear regression model (NLIN).
[b] $P < 0.001$ for all predicted MAIs.
[c] Residual mean squared error (MSE_r).

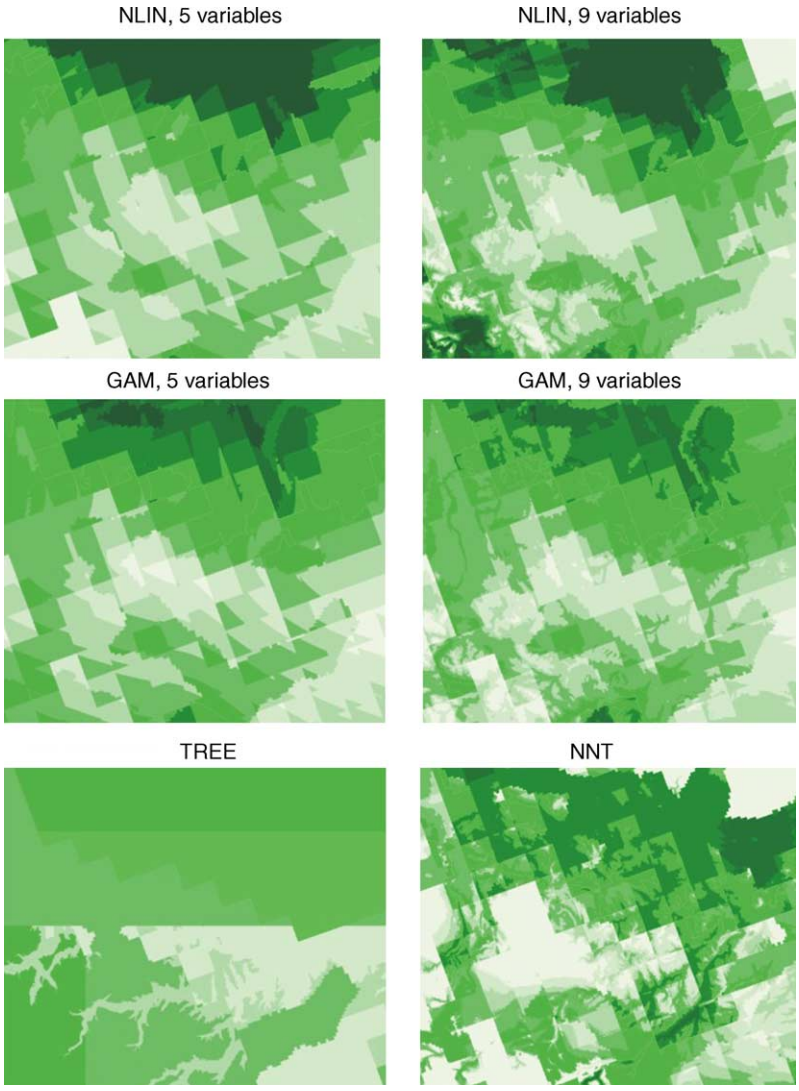

Fig. 5. Predicted site index maps for lodgepole pine in the Wapiti region generated by the four models. From light to dark, the shading represent 0–7 m, 7–9 m, 9–11 m, 11–13 m, 13–15 m, 15–17 m, 17–19 m, 19–21 m, and >21 m. NLIN: nonlinear regression model; TREE: tree-based model; GAM: generalized additive model; NNT: neural network model.

Six predicted SI maps were created, showing high productivity in the north, which reflected reality (Fig. 5). The southwest corner has the highest elevations, and SI should be the lowest there. The GAM with nine variables, the NLIN model with five variables and the NNT model clearly captured this feature spatially, but other models were not very successful in this corner of the region. Other than in this small area, the general productivity patterns depicted by the NLIN, GAM, and NNT models looked similar. Notably, the predicted SIs were continuous for the NLIN, GAM, and NNT models, but the same prediction from the TREE model appeared discrete.

The three nonparametric models improved the fit by about 10% of R^2 compared with the NLIN model (Table 2 and Fig. 3), which indicates that they are more flexible than the parametric models for the data set in our study. However, more vigorous evaluation has to be done on the SI spatial prediction for the whole Wapiti region. Although the predictor variables in this

study were all numerical, the nonparametric models are also capable of using non-numerical variables. This is a huge advantage in dealing with biophysical productivity studies, because the data collected are often a combination of numerical and categorical measurements. However, one should be careful with overfitting when using NNT.

The outcome from TREE consisted of SI classes. Because the number of classes in this study is relatively low (only nine), the spatial prediction by TREE is fragmented. The situation would be improved if more observations were available. Although for both NLIN and GAM five predictor variables would fit the data as well as nine variables, the predicted productivity maps with more variables appear to show more details than the maps based on fewer variables (Fig. 5). General spatial patterns from the two GAM maps looked similar, but the two maps from NLIN showed significant differences in the southwestern corner.

Breiman (2001) asserts that there are two cultures in statistical modeling: the data modeling culture and the algorithm modeling culture. NLIN belongs to the first one, which assumes a stochastic data model to relate SI to environmental variables. Such models are optimized for estimation. GAM, TREE, and NNT are of the second culture, which treats the relation between SI and independent variables as an unknown black box and tries to find a function between the variables. These models are more suitable for prediction purpose. The advantage of algorithm modeling is that it can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data set (Breiman, 2001). Therefore, given that our purpose is to make spatial prediction, GAM, TREE, and NNT are more appropriate than NLIN for our study. Although ours is only a case study for a small region, the results have implications for similar studies in mature or over-mature boreal forest.

4. Conclusions

This study compared four modelling techniques for prediction of biophysical productivity. NLIN can be characterized as a parametric method or of the data modeling culture, since a functional relationship between response and predictor variables must be specified. TREE, GAM, and NNT are regarded as nonparametric methods or of the algorithm modeling culture, since there is no need to specify the mathematical form of the model before estimating the parameters. After fitting and evaluating the models on the basis of a set of field data of mature and over-mature stands, we concluded that GAM produced the best fit to the data. Hence, spatial productivity predicted by this model should accurately summarize the broad spatial distribution of lodgepole pine productivity in Wapiti. The performance of TREE was also impressive in terms of fit; unfortunately, the quality of the data did not allow TREE to produce enough SI classes to make the map smooth. NLIN was the only parametric model tested, and its performance was poorer than that of the other models. However, this does not mean that the NLIN model is inferior, because in developing an appropriate biophysical productivity model, model type, data, and goals must be taken into consideration simultaneously. This study has demonstrated the potential of some nonparametric models in stand productivity prediction, and these models could be considered for use under similar circumstances.

Acknowledgements

The authors thank Eric Arsenault, Harinder Hans and Marty Siltanen of the Northern Forestry Centre (NFC), Canadian Forest Service (CFS), for their help in preparing the grid data. They also thank Xiaojing Guo of the Laurentian Forestry Centre (CFS) for her calculations on model ranking when evaluated by MAI, David Price (NFC-CFS) and two anonymous reviewers for their constructive comments on earlier drafts.

References

Avery, T.E., Burkhart, H.E., 2002. Forest Measurements, fifth ed. McGraw-Hill, New York.

Beckingham, J.D., Corns, I.G.W., Archibald, J.H., 1996. Field Guide to Ecosites of West-Central Alberta. Special Report No. 9. Natural Resources Canada, Canadian Forest Service, Northern Forestry Centre, Edmonton, Alta.

Belsley, D.A., Kuh, E., Welsch, R.E., 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons Inc., New York.

Bonan, G.B., 1989. A computer model of the solar radiation, soil moisture, and soil thermal regimes in boreal forests. Ecol. Model. 45, 275–306.

Breiman, L., 2001. Statistical modeling: the two cultures. Stat. Sci. 16, 199–231.

Corns, I.G.W., 1978. Tree growth prediction and plant community distribution in relation to environmental factors in lodgepole pine, white spruce, black spruce, and aspen forests of western Alberta Foothills. Dissertation, University of Alberta, Edmonton, Alta.

Corns, I.G.W., Pluth, D.J., 1984. Vegetational indicators as independent variables in forest growth prediction in west-central Alberta, Canada. For. Ecol. Manage. 9, 13–25.

Draper, N.R., Smith, H., 1981. Applied Regression Analysis. John Wiley & Sons Inc., New York.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Fausett, L., 1994. Fundamentals of Neural Networks: Architecture, Alogorithms, and Applications. Prentice-Hall, Englewood Cliffs, NJ.

Fries, A., Ruotsalainen, S., Lindgren, D., 1998. Effects of temperature on the site productivity of *Pinus sylvestris* and lodgepole pine in Finland and Sweden. Scand. J. For. Res. 13, 128–140.

Guan, B., Gertner, G., 1991a. Using a parallel distributed processing system to model individual tree mortality. For. Sci. 37, 871–885.

Guan, B., Gertner, G., 1991b. Modeling red pine tree survival with an artificial neural network. For. Sci. 37, 1429–1440.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall, London.

Hogg, E.H., 1994. Climate and the southern limit of the western Canadian boreal forest. Can. J. For. Res. 24, 1835–1845.

Huang, S., Morgan, D., Klappstein, G., Heidt, J., Yang, Y., Greidanus, G., 2001. GYPSY, a growth and yield projection system for natural and regenerated stands within an ecologically based, enhanced forest management framework. Yield tables for seed-origin natural and regenerated lodgepole pine stands. Alberta Sustainable Resource Development, Edmonton, Alta.

Hunter, I.R., Gibson, A.R., 1984. Predicting *Pinus radiata* site index from environmental variables. N. Z. J. For. Sci. 14, 53–64.

Jensen, M.E., Burman, R.D., Allen, R.G. (Eds.), 1990. Evapotranspiration and Irrigation Water Requirements. American Society of Civil Engineers, New York, Man. Rep. Eng. Pract. 70.

Kabzems, R.D., Klinka, K., 1987. Initial quantitative characterization of soil nutrient regimes. II. Relationships among soil, vegetation, and site index. Can. J. For. Res 17, 1565–1571.

Kira, T., Shidei, T., 1967. Primary production and turnover of organic matter in different forest ecosystems of the western Pacific. Jpn. J. Ecol. 17, 70–87.

McKenney, D.W., Pedlar, J.H., 2003. Spatial models of site index based on climate and soil properties for two boreal tree species in Ontario, Canada. For. Ecol. Manage. 175, 497–507.

McKenney, D.W., Mackey, B.G., Sims, R.A., 1996. Primary databases for forest ecosystem management—examples from Ontario and possibilities for Canada: NatGRID. Environ. Monit. Assess. 39, 399–415.

Moran, J.M., Morgan, M.D., 1997. Meteorology: The Atmosphere and the Science of Weather. Prentice-Hall, Simon & Schuster, Aviacom Company, New Jersey.

Robichaud, E., Methven, I.R., 1993. The effect of site quality on the timing of stand breakup, tree longevity, and the maximum attainable height of black spruce. Can. J. For. Res. 23, 1514–1519.

Russell, C.E., Dobbins, R.W., 1990. Neural Network PC Tools, a Practical Guide. Academic Press, New York.

SAS Institute Inc., 1989. SAS/STAT® User's Guide, version 6, vol. 2, fourth ed. SAS Institute Inc., Cary, NC.

Schut, P., Shields, J., Tarnocai, C., Coote, D., Marshall, I., 1994. Soil landscapes of Canada—an environmental reporting tool. In: Proceeding of the Canadian Conference on GIS, Ottawa, Ont., pp. 953–965.

Shields, J.A., Tarnocai, C., Valentine, K.W.G., MacDonald, K.B., 1991. Soil Landscapes of Canada—Procedures Manual and User's Handbook. LRRC Contribution No. 88-29. Land Resource Research Centre, Research Branch, Agriculture Canada, Ottawa, Ont.

Simon, G., Lendasse, A., Wertz, V., Verleysen, M., 2003. Fast approximation of the bootstrap for model selection. In: Proceedings of the European Symposium on Artificial Neural Networks, Bruges, Belgium, 23–25 April 2003, pp. 475–480.

Sironen, S., Kangas, A., Maltamo, M., Kangas, J., 2003. Estimating individual tree growth with nonparametric methods. Can. J. For. Res. 33, 444–449.

Soil Carbon Database Working Group, 1993. Soil Carbon Data for Canadian Soils. CLBRR Contribution No. 92-179, Centre for Land and Biological Resources Research, Research Branch, Agriculture Canada, Ottawa, Ont.

Smith, F.W., Resh, S.C., 1999. Age-related changes in production and below-ground carbon allocation in *Pinus contorta* forests. For. Sci. 45, 333–341.

Statistical Science, 1993. S-PLUS Guide to Statistical and Mathematical Analysis, version 3.2. StatSci, a division of MathSoft Inc., Seattle.

Ung, C.-H., Bernier, P.Y., Raulier, F., Fournier, R.A., Lambert, M.-C., Régnière, J., 2001. Biophysical site indices for shade tolerant and intolerant boreal species. For. Sci. 47, 83–95.

Vanclay, J.K., Skovsgaard, J.P., 1997. Evaluating forest growth models. Ecol. Model. 98, 1–12.

Venables, W.N., Ripley, B.D., 1994. Modern Applied Statistics with S-Plus. Springer, New York.

Wei, X., Kimmins, J.P., Zhou, G., 2003. Disturbances and the sustainability of long-term site productivity in lodgepole pine forests in the central interior of British Columbia—an ecosystem modeling approach. Ecol. Model. 164, 239–256.

Woods, A.J., Nussbaum, A., Golding, B., 2000. Predicted impacts of hard pine stem rusts on lodgepole pine stands in central British Columbia. Can. J. For. Res. 30, 476–481.

Yee, T.W., Mitchell, N.D., 1991. Generalized additive models in plant ecology. J. Veg. Sci. 2, 587–602.

Zhang, Q.-B., Hebda, R., Zhang, Q.-J., Alfaro, R.I., 2000. Modeling tree-ring growth responses to climatic variables using artificial neural networks. For. Sci. 46, 229–239.