Journal of Animal Ecology

British Ecological Society

Journal of Animal Ecology 2008, **77**, 802–813 doi: 10.1111/j.1365-2656.2008.01390.x

A working guide to boosted regression trees

J. Elith[1]*, J. R. Leathwick[2] and T. Hastie[3]

[1]*School of Botany, The University of Melbourne, Parkville, Victoria, Australia 3010;* [2]*National Institute of Water and Atmospheric Research, PO Box 11115, Hamilton, New Zealand; and* [3]*Department of Statistics, Stanford University, CA, USA*

Summary

1. Ecologists use statistical models for both explanation and prediction, and need techniques that are flexible enough to express typical features of their data, such as nonlinearities and interactions.
2. This study provides a working guide to boosted regression trees (BRT), an ensemble method for fitting statistical models that differs fundamentally from conventional techniques that aim to fit a single parsimonious model. Boosted regression trees combine the strengths of two algorithms: regression trees (models that relate a response to their predictors by recursive binary splits) and boosting (an adaptive method for combining many simple models to give improved predictive performance). The final BRT model can be understood as an additive regression model in which individual terms are simple trees, fitted in a forward, stagewise fashion.
3. Boosted regression trees incorporate important advantages of tree-based methods, handling different types of predictor variables and accommodating missing data. They have no need for prior data transformation or elimination of outliers, can fit complex nonlinear relationships, and automatically handle interaction effects between predictors. Fitting multiple trees in BRT overcomes the biggest drawback of single tree models: their relatively poor predictive performance. Although BRT models are complex, they can be summarized in ways that give powerful ecological insight, and their predictive performance is superior to most traditional modelling methods.
4. The unique features of BRT raise a number of practical issues in model fitting. We demonstrate the practicalities and advantages of using BRT through a distributional analysis of the short-finned eel (*Anguilla australis* Richardson), a native freshwater fish of New Zealand. We use a data set of over 13 000 sites to illustrate effects of several settings, and then fit and interpret a model using a subset of the data. We provide code and a tutorial to enable the wider use of BRT by ecologists.

Key-words: data mining, machine learning, model averaging, random forests, species distributions

Introduction

Ecologists frequently use models to detect and describe patterns, or to predict to new situations. In particular, regression models are often used as tools for quantifying the relationship between one variable and others upon which it depends. Whether analysing the body weight of birds in relation to their age, sex and guild; the abundance of squirrels as it varies with temperature, food and shelter; or vegetation type in relation to aspect, rainfall and soil nutrients, models can be used to identify variables with the most explanatory power, indicate optimal conditions and predict to new cases.

The past 20 years have seen a growing sophistication in the types of statistical model applied in ecology, with impetus from substantial advances in both statistics and computing. Early linear regression models were attractively straightforward, but too simplistic for many real-life situations. In the 1980s and 1990s, generalized linear models (GLM; McCullagh & Nelder 1989) and generalized additive models (GAM; Hastie & Tibshirani 1990) increased our capacity to analyse data with non-normally distributed errors (presence–absence and count data), and to model nonlinear relationships. These models are now widely used in ecology, for example for analysis of morphological relationships (Clarke & Johnston 1999) and population trends (Fewster *et al.* 2000), and for predicting the distributions of species (Buckland & Elston 1993).

Over the same period, computer scientists developed a wide variety of algorithms particularly suited to prediction, including neural nets, ensembles of trees and support vector machines. These machine learning (ML) methods are used less frequently than regression methods in ecology, perhaps

*Correspondence author. E-mail: j.elith@unimelb.edu.au

partly because they are considered less interpretable and therefore less open to scrutiny. It may also be that ecologists are less familiar with the modelling paradigm of ML, which differs from that of statistics. Statistical approaches to model fitting start by assuming an appropriate data model, and parameters for this model are then estimated from the data. By contrast, ML avoids starting with a data model and rather uses an algorithm to learn the relationship between the response and its predictors (Breiman 2001). The statistical approach focuses on questions such as what model will be postulated (e.g. are the effects additive, or are there interactions?), how the response is distributed, and whether observations are independent. By contrast, the ML approach assumes that the data-generating process (in the case of ecology, nature) is complex and unknown, and tries to learn the response by observing inputs and responses and finding dominant patterns. This places the emphasis on a model's ability to predict well, and focuses on what is being predicted and how prediction success should be measured.

In this paper we discuss a relatively new technique, boosted regression trees (BRT), which draws on insights and techniques from both statistical and ML traditions. The BRT approach differs fundamentally from traditional regression methods that produce a single 'best' model, instead using the technique of boosting to combine large numbers of relatively simple tree models adaptively, to optimize predictive performance (e.g. Elith *et al*. 2006; Leathwick *et al*. 2006, 2008). The boosting approach used in BRT places its origins within ML (Schapire 2003), but subsequent developments in the statistical community reinterpret it as an advanced form of regression (Friedman, Hastie & Tibshirani 2000).

Despite clear evidence of strong predictive performance and reliable identification of relevant variables and interactions, BRT has been rarely used in ecology (although see Moisen *et al*. 2006; De'ath 2007). In this paper we aim to facilitate the wider use of BRT by ecologists, demonstrating its use in an analysis of relationships between frequency of capture of short-finned eels (*Anguilla australis* Richardson), and a set of predictors describing river environments in New Zealand. We first explain what BRT models are, and then show how to develop, explore and interpret an optimal model. Supporting software and a tutorial are provided as Supplementary material.

EXPLANATION OF BOOSTED REGRESSION TREES

BRT is one of several techniques that aim to improve the performance of a single model by fitting many models and combining them for prediction. BRT uses two algorithms: regression trees are from the classification and regression tree (decision tree) group of models, and boosting builds and combines a collection of models. We deal with each of these components in turn.

DECISION TREES

Modern decision trees are described statistically by Breiman *et al*. (1984) and Hastie, Tibshirani & Friedman (2001), and for ecological applications by De'ath & Fabricius (2000). Tree-based models partition the predictor space into rectangles, using a series of rules to identify regions having the most homogeneous responses to predictors. They then fit a constant to each region (Fig. 1), with classification trees fitting the most probable class as the constant, and regression trees fitting the mean response for observations in that region, assuming normally distributed errors. For example, in Fig. 1 the two predictor variables X_1 and X_2 could be temperature and rainfall, and the response Y, the mean adult weight of a species. Regions Y_1, $Y_{2, \text{etc.}}$ are terminal nodes or leaves, and t_1, $t_{2, \text{etc.}}$ are split points. Predictors and split points are chosen to minimize prediction errors. Growing a tree involves recursive binary splits: a binary split is repeatedly applied to its own output until some stopping criterion is reached. An effective strategy for fitting a single decision tree is to grow a large tree, then prune it by collapsing the weakest links identified through cross-validation (CV) (Hastie *et al*. 2001).

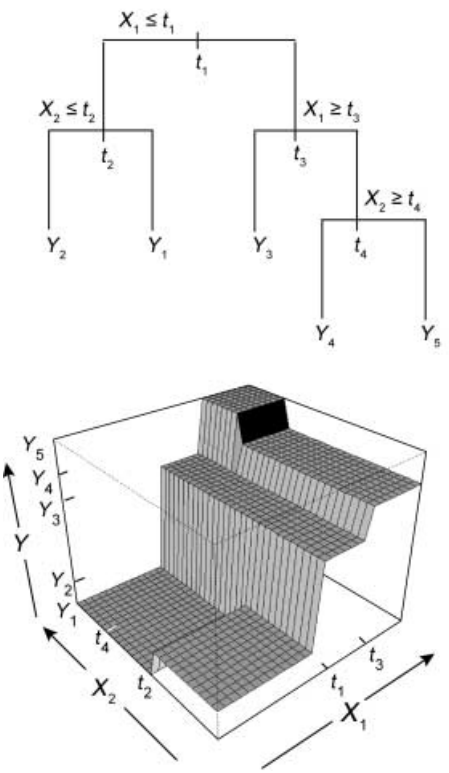

Fig. 1. A single decision tree (upper panel), with a response Y, two predictor variables, X_1 and X_2 and split points t_1, t_2, etc. The bottom panel shows its prediction surface (after Hastie *et al*. 2001)

Decision trees are popular because they represent information in a way that is intuitive and easy to visualize, and have several other advantageous properties. Preparation of candidate predictors is simplified because predictor variables can be of any type (numeric, binary, categorical, etc.), model outcomes are unaffected by monotone transformations and differing scales of measurement among predictors, and irrelevant predictors are seldom selected. Trees are insensitive to outliers, and can accommodate missing data in predictor variables by

using surrogates (Breiman *et al*. 1984). The hierarchical structure of a tree means that the response to one input variable depends on values of inputs higher in the tree, so interactions between predictors are automatically modelled. Despite these benefits, trees are not usually as accurate as other methods, such as GLM and GAM. They have difficulty in modelling smooth functions, even ones as simple as a straight-line response at 45° to two input axes. Also, the tree structure depends on the sample of data, and small changes in training data can result in very different series of splits (Hastie *et al*. 2001). These factors detract from the advantages of trees, introducing uncertainty into their interpretation and limiting their predictive performance.

BOOSTING

Boosting is a method for improving model accuracy, based on the idea that it is easier to find and average many rough rules of thumb, than to find a single, highly accurate prediction rule (Schapire 2003). Related techniques – including bagging, stacking and model averaging – also build, then merge results from multiple models, but boosting is unique because it is sequential: it is a forward, stagewise procedure. In boosting, models (e.g. decision trees) are fitted iteratively to the training data, using appropriate methods gradually to increase emphasis on observations modelled poorly by the existing collection of trees. Boosting algorithms vary in how they quantify lack of fit and select settings for the next iteration. The original boosting algorithms such as AdaBoost (Freund & Schapire 1996) were developed for two-class classification problems. They apply weights to the observations, emphasizing poorly modelled ones, so the ML literature tends to discuss boosting in terms of changing weights.

Here, though, we focus on regression trees (including logistic regression trees), and the intuition is different. For regression problems, boosting is a form of 'functional gradient descent'. Consider a loss function – in this case, a measure (such as deviance) that represents the loss in predictive performance due to a suboptimal model. Boosting is a numerical optimization technique for minimizing the loss function by adding, at each step, a new tree that best reduces (steps down the gradient of) the loss function.For BRT, the first regression tree is the one that, for the selected tree size, maximally reduces the loss function. For each following step, the focus is on the residuals: on variation in the response that is not so far explained by the model. [Technical aside: For ordinary regression and squared-error loss, standard residuals are used. For more general loss, the analogue of the residual vector is the vector of negative gradients. Deviance is used as the loss function in the software we use. The negative gradient of the deviance in a logistic regression BRT model or a Poisson BRT model is the residual $y - p$, where y is the response and p the fitted probability or fitted Poisson mean. These are fitted by a tree, and the fitted values are added to the current logit(p) or log(p).] For example, at the second step, a tree is fitted to the residuals of the first tree, and that second tree could contain quite different variables and split points compared with the first. The model is then updated to contain two trees (two terms), and the residuals from this two-term model are calculated, and so on. The process is stagewise (not stepwise), meaning that existing trees are left unchanged as the model is enlarged. Only the fitted value for each observation is re-estimated at each step to reflect the contribution of the newly added tree. The final BRT model is a linear combination of many trees (usually hundreds to thousands) that can be thought of as a regression model where each term is a tree. We illustrate the way in which the trees combine and contribute to the final fitted model in a later section, 'How multiple trees produce curvilinear functions'. The model-building process performs best if it moves slowly down the gradient, so the contribution of each tree is usually shrunk by a learning rate that is substantially less than one. Fitted values in the final model are computed as the sum of all trees multiplied by the learning rate, and are much more stable and accurate than those from a single decision tree model.

Similarly to GLM, BRT models can be fitted to a variety of response types (Gaussian, Poisson, binomial, etc.) by specifying the error distribution and the link. Ridgeway (2006) provides mathematical details for available distributions in the software we use here, including calculations for deviance (the loss function), initial values, gradients, and the constants predicted in each terminal node. Some loss functions are more robust to noisy data than others (Hastie *et al*. 2001). For example, binomial data can be modelled in BRTs with several loss functions: exponential loss makes them similar to boosted classification trees such as AdaBoost, but binomial deviance is more robust, and likely to perform better in data where classes may be mislabelled (e.g. false negative observations).

From a user's point of view, important features of BRT as applied in this paper are as follows. First, the process is stochastic – it includes a random or probabilistic component. The stochasticity improves predictive performance, reducing the variance of the final model, by using only a random subset of data to fit each new tree (Friedman 2002). This means that, unless a random seed is set initially, final models will be subtly different each time they are run. Second, the sequential model-fitting process builds on trees fitted previously, and increasingly focuses on the hardest observations to predict. This distinguishes the process from one where a single large tree is fitted to the data set. However, if the perfect fit was a single tree, in a boosted model it would probably be fitted by a sum of identical shrunken versions of itself. Third, values must be provided for two important parameters. The learning rate (lr), also known as the shrinkage parameter, determines the contribution of each tree to the growing model, and the tree complexity (tc) controls whether interactions are fitted: a tc of 1 (single decision stump; two terminal nodes) fits an additive model, a tc of two fits a model with up to two-way interactions, and so on. These two parameters then determine the number of trees (nt) required for optimal prediction. Finally, prediction from a BRT model is straightforward, but interpretation requires tools for identifying which variables and interactions are important, and for visualizing fitted functions. In the following sections, we use a case study to show how to manage these features of BRT in a typical ecological setting.

Table 1. Environmental variables used to model fish occurrence

Variable	Description	Mean and range
Reach scale predictor		
LocSed	Weighted average of proportional cover of bed sediment: 1 = mud, 2 = sand, 3 = fine gravel; 4 = coarse gravel; 5 = cobble; 6 = boulder; 7 = bedrock	3·77, 1–7
Segment scale predictors		
SegSumT	Summer air temperature (°C)	16·3, 8·9–19·8
SegTSeas	Winter air temperature (°C), normalized with respect to SegJanT	0·36, –4·2–4·1
SegLowFlow	Segment low flow ($m^3 s^{-1}$), fourth root transformed	1·092, 1·0–4·09
Downstream predictors		
DSDist	Distance to coast (km)	74, 0·03–433·4
DSDam	Presence of known downstream obstructions, mostly dams	0·18, 0 or 1
DSMaxSlope	Maximum downstream slope (°)	3·1, 0–29·7
Upstream/catchment scale predictors		
USAvgT	Average temperature in catchment (°C) compared with segment, normalized with respect to SegJanT	–0·38, –7·7–2·2
USRainDays	Days per month with rain >25 mm	1·22, 0·21–3·30
USSlope	Average slope in the upstream catchment (°)	14·3, 0–41·0
USNative	Area with indigenous forest (proportion)	0·57, 0–1
Fishing method		
Method	Fishing method in five classes: electric, net, spot, trap, mixture	NA

THE CASE STUDY

We demonstrate use of BRT with data describing the distribution of, and environments occupied by, the short-finned eel (*Anguilla australis*) in New Zealand. We aim to produce a model that not only identifies major environmental determinants of *A. australis* distribution, but also can be used to predict and map its occurrence in unsampled rivers. The model will be a form of logistic regression that models the probability that a species occurs, $y = 1$, at a location with covariates $\boldsymbol{X}$, $P(y = 1 \mid \boldsymbol{X})$. This probability will be modelled via a logit : logit $P(y = 1 \mid \boldsymbol{X}) = f(\boldsymbol{X})$.

Anguilla australis is a freshwater eel native to south-eastern Australia, New Zealand and western Pacific islands. Within New Zealand it is a common freshwater species, frequenting lowland lakes, swamps, and sluggish streams and rivers in pastoral areas, and forming a valuable traditional and commercial fishery. Short-finned eels take 10–20 years to mature, then migrate – perhaps in response to rainfall or flow triggers – to the sea to spawn. The eels spawn at considerable depth, then larvae are brought back to the coast on ocean currents and metamorphose into glass eels. After entering freshwater, they become pigmented and migrate upstream. They tend not to penetrate as far upstream as long-finned eels (*Anguilla dieffenbachii*), probably because there is little suitable habitat further inland rather than because they are unable to do so (McDowall 1993).

The data set, developed for research and conservation planning in New Zealand, is described in detail by Leathwick *et al*. 2008). Briefly, species data were records of species caught from 13 369 sites spanning the major environmental gradients in New Zealand's rivers. *Anguilla australis* was caught at 20% of sites. Because this is a much larger data set than is often available in ecology, here we subsample the 13 369 sites, usually partitioning off 1000 records for modelling and keeping the remainder for independent evaluation.

The explanatory variables were a set of 11 functionally relevant environmental predictors (Table 1) that summarize conditions over several spatial scales: local (segment and reach) scale, upstream catchment scale, and downstream to the sea. Most were available as GIS data for the full river system of New Zealand, enabling prediction to all rivers. The exception was one variable describing local substrate conditions (LocSed) that had records at only 82% sites. The 12th variable was categorical, and described fishing method (Table 1). Given these records and covariates, the logistic regression will be modelling the joint probability of occurrence and capture of *A. australis*.

SOFTWARE AND MODELLING

All models were fitted in R (R Development Core Team 2006) version 2.3-1, using gbm package version 1·5–7 (Ridgeway 2006) plus custom code written by J.L. and J.E. Our code is available with a tutorial (Supplementary material). There is also a growing range of alternative implementations for boosted trees, but we do not address those here. The following two sections explain how to fit, evaluate and interpret a BRT model, highlighting features that make BRT particularly useful in ecology. For all settings other than those mentioned, we used the defaults in gbm.

OPTIMIZING THE MODEL WITH ECOLOGICAL DATA

Model development in BRT is best understood in the context of other model-fitting practices. For all prediction problems, overfitting models to training data reduces their generality, so regularization methods are used to constrain the fitting

procedure so that it balances model fit and predictive performance (Hastie *et al.* 2001). Regularization is particularly important for BRT because its sequential model fitting allows trees to be added until the data are completely overfitted. For most modelling methods, model simplification is achieved by controlling the number of terms. The number of terms is defined by the number of predictor variables and the complexity of fitted functions, and is often determined using stepwise procedures (for a critique of these see Whittingham *et al.* 2006) or by building several models and comparing them with information theoretical measures such as Akaike's information criterion (Burnham & Anderson 2002). Controlling the number of terms implies a prior belief that parsimonious models (fewer terms) provide better prediction. Alternatively, more terms can be fitted and their contributions downweighted using shrinkage (Friedman 2001). In conventional regression, this is applied as global shrinkage (direct, proportional shrinkage on the full model) using ridge or lasso methods (Hastie *et al.* 2001; Reineking & Schröder 2006). Shrinkage in BRT is similar, but is incremental, and is applied to each new tree as it is fitted. Analytically, BRT regularization involves jointly optimizing the number of trees (*nt*), learning rate (*lr*), and tree complexity (*tc*). We focus on trade-offs between these elements in the following sections, after explaining the role of stochasticity.

BOOSTING WITH STOCHASTICITY

Introducing some randomness into a boosted model usually improves accuracy and speed and reduces overfitting (Friedman 2002), but it does introduce variance in fitted values and predictions between runs (Appendix S1, see Supplementary material). In gbm, stochasticity is controlled through a 'bag fraction' that specifies the proportion of data to be selected at each step. The default bag fraction is 0·5, meaning that, at each iteration, 50% of the data are drawn at random, without replacement, from the full training set. Optimal bag fractions can be established by comparing predictive performance and model-to-model variability under different bag fractions. In our experience, stochasticity improves model performance, and fractions in the range 0·5–0·75 have given best results for presence–absence responses. From here on we use a bag fraction of 0·5, but with new data it is worth exploration.

NUMBER OF TREES VS. LEARNING RATE

The *lr* is used to shrink the contribution of each tree as it is added to the model. Decreasing (slowing) *lr* increases the number of trees required, and in general a smaller *lr* (and larger *nt*) are preferable, conditional on the number of observations and time available for computation. The usual approach is to estimate optimal *nt* and *lr* with an independent test set or with CV, using deviance reduction as the measure of success. The following analysis demonstrates how performance varies with these parameters using a subsample of the data set for model fitting, and the remaining data for independent evaluation.

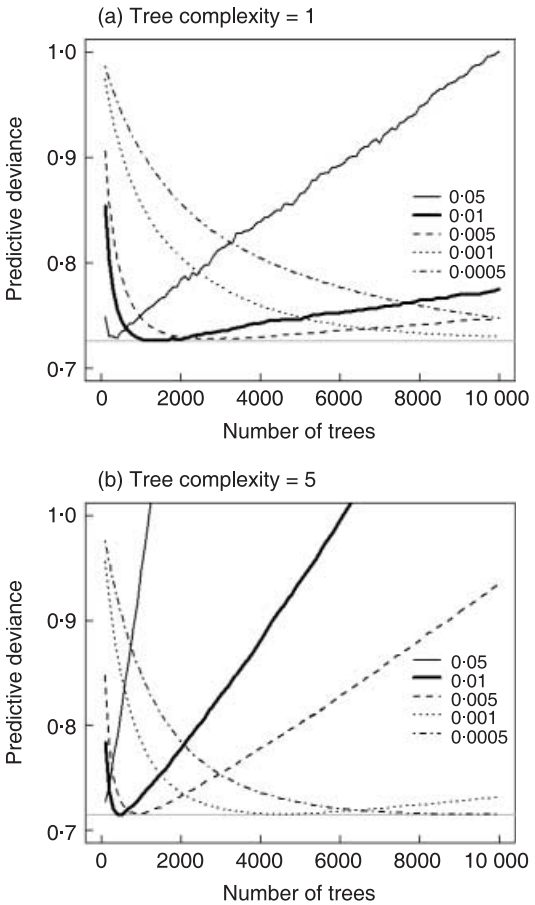

Fig. 2. The relationship between number of trees and predictive deviance for models fitted with five learning rates and two levels of tree complexity. Deviance was calculated from models fitted to a data set of 1000 sites, and predicted to a data set of 12 369 sites. The lowest predictive deviance achieved for each panel is indicated by a dotted horizontal line; the line for learning rate achieving this minimum is shown in bold.

Using a set of 1000 sites and 12 predictor variables, we fitted BRT models with varying values for *nt* (100–20 000) and *lr* (0·1–0·0001), and evaluated them on 12 369 excluded sites. Example code is given in the online tutorial (see Supplementary material). Results for up to 10 000 trees for a *tc* of 1 and 5 are shown in Fig. 2. Our aim here is to find the combination of parameters (*lr*, *tc* and *nt*) that achieves minimum predictive error (minimum error for predictions to independent samples). A value of 0·1 for *lr* (not plotted) was too fast for both *tc* values, and at each addition of trees above the minimum 100 trees, predictive deviance increased, indicating that overfitting occurred almost immediately. The fastest feasible *lr* (0·05) fitted relatively few trees, did not achieve minimum error for *tc* = 1 (see horizontal dashed line) or *tc* = 5, and in both cases predicted poorly as more trees were added (the curves rise steeply after they have reached a minimum, indicating overfitting). In contrast, the smallest values for *lr* approached best predictive performance slowly, and required thousands of trees to reach minimum error. There was little

gain in predictive power once more than 500 or so trees were fitted. However, slower *lr* values are generally preferable to faster ones, because they shrink the contribution of each tree more, and help the final model to reliably estimate the response. We explain this further in Appendix S1, and as a rule of thumb recommend fitting models with at least 1000 trees.

TREE COMPLEXITY

Tree complexity – the number of nodes in a tree – also affects the optimal *nt*. For a given *lr*, fitting more complex trees leads to fewer trees being required for minimum error. So, as *tc* is increased, *lr* must be decreased if sufficient trees are to be fitted (*tc* = 5, Fig. 2b). Theoretically, the *tc* should reflect the true interaction order in the response being modelled (Friedman 2001), but as this is almost always unknown, *tc* is best set with independent data.

Sample size influences optimal settings for *lr* and *tc*, as shown in Fig. 3. For this analysis, the full data set was split into training sets of various sizes (6000, 2000, 1000, 500 and 250 sites), plus an independent test set (7369 sites). BRT models of 30 000 trees were then fitted over a range of values for *tc* (1, 2, 3, 5, 7, 10) and *lr* (0·1, 0·05, 0·01, 0·005, 0·001, 0·0005). We identified, for each parameter combination, the *nt* that achieved minimum prediction error, and summarized results as averages across *tc* (Fig. 3a) and *lr* (Fig. 3b). If the minimum was not reached by 30 000 trees, that parameter combination was excluded.

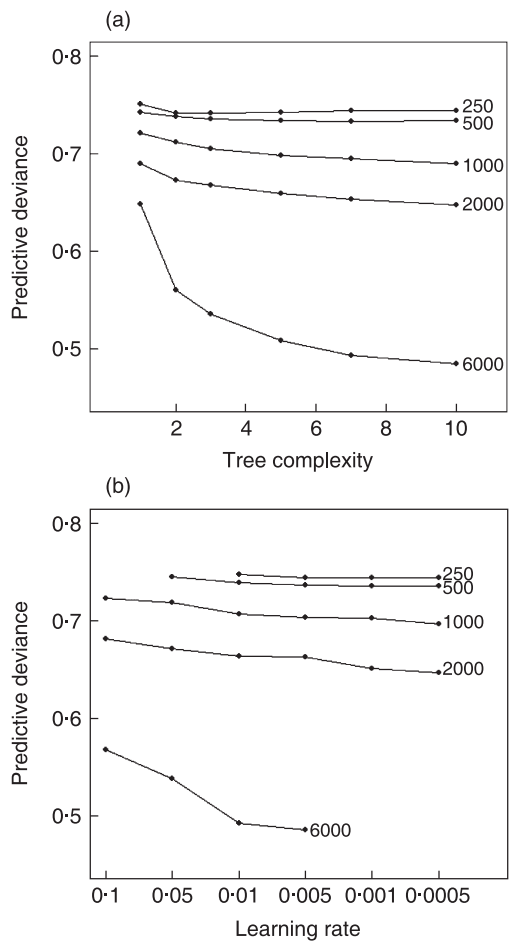

Fig. 3. Predictive deviance as a function of data set size and (a) tree complexity; (b) learning rate. Models were run on data sets of 250–6000 sites, and minimum predictive deviance was estimated on an excluded data set of 7369 sites. No result indicates that no model of those up to 30 000 trees achieved a minimum predictive deviance with that parameter combination.

Predictive performance was influenced most strongly by sample size and, as expected, large samples gave models with lower predictive error. Gains from increased *tc* were greater with larger data sets, presumably because more data provided more detailed information about the full range of sites in which the species occurs, and the complexity in that information could be modelled better using more complex trees. Decision stumps (*tc* 1) were never best (they always had higher predictive deviance), but for small samples there was no advantage – but also little penalty – for using large (higher-*tc*) trees. The reason for not using the highest *tc*, though, is that the model would have to be learnt very slowly to achieve enough trees for reliable estimates. So, small samples here (e.g. 250 sites) would be best modelled with simple trees (*tc* 2 or 3) and a slow enough *lr* to allow at least 1000 trees.

As a general guide, *lr* needs to be decreased as *tc* increases, usually inversely: doubling *tc* should be matched with halving *lr* to give approximately the same *nt*. While the results here suggest that using higher *tc* and very slow *lr* is the best strategy (for samples >500 sites the curves keep descending), the other trade-off is computing time. For example, fitting BRT models on the 1000-site data set on a modern laptop and using our online code took 0·98 min for *tc* 1 and *lr* 0·05 (500 trees), but 3·85 min for *tc* 1 and *lr* 0·01 (2500 trees), 2·36 min for *tc* 5 and *lr* 0·01 (850 trees), and 7·49 min for *tc* 1 and *lr* 0·005 (4600 trees). Where many species are modelled, or many models are required for other reasons (e.g. bootstrapping), using the fastest *lr* that achieves more than, say, 1000 trees is a good strategy. We note, too, that for presence–absence data such as these, optimal settings also vary with prevalence of the species. A very rare or very common species provides less information to model given the same total number of sites, and will generally require slower learning rates.

IDENTIFYING THE OPTIMAL SETTINGS

In many situations, large amounts of data are not available, so techniques such as CV are used for model development and/or evaluation. Cross-validation provides a means for testing the model on withheld portions of data, while still using all data at some stage to fit the model. Use of CV for selecting optimal settings is becoming increasingly common (Hastie *et al*. 2001), led by the ML focus on predictive success. Here we demonstrate a CV implementation that first determines the optimal *nt*, then fits a final model to all the data. The CV process is detailed in Fig. 4, and code is available (function gbm.step) in the Supplementary material.

We use a data set of 1000 sites to develop and test a model via CV, also evaluating it on the withheld 12 369 sites. Our

(1) Randomly divide available data into *n* subsets (from here on for this example, *n* = 10; this is a setting we commonly use);

(2) Make 10 different **training** sets each comprising a unique combination of 9 subsets. Therefore for each training set there is a unique omitted subset that is used for **testing**;

(3) Starting with a selected number of trees (*nt*) , say 50, develop 10 BRT models simultaneously on each training set, and test predictive performance on their respective omitted data. Both mean performance and standard errors are recorded;

(4) Step forward and increase the *nt* in each model by a selected and constant amount, and repeat step 3;

(5) Repeat step 4 and after 10 steps start comparing the predictive performance of the 6th to 10th previous iterations against that of the current to 5th previous ones. Once the average of the more recent set is higher than the average of the previous set, the minimum has been passed;

(6) Stop and record the minimum; this is the optimal *nt*.

Fig. 4. Cross-validation method for identifying the optimal number of trees in a boosted regression tree model.

selected settings are *lr* of 0·005, *tc* of 5 and bag fraction of 0·5; note that all 1000 sites can be used despite missing data for LocSed at 222 sites. As trees are added, there is an initial steep decline in prediction error followed by a more gradual approach to the minimum (Fig. 5, solid circles). With a slow enough *lr*, the CV estimates of *nt* are reliable and close to those from independent data (Fig. 5).

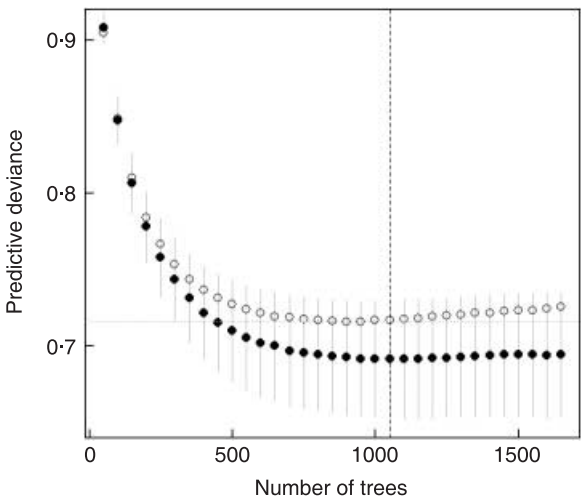

Fig. 5. Cross-validation (CV) model-fitting example. Data set of 1000 observations, with estimates of predictive deviance from the CV (mean as solid circles and SE) and an independent estimate based on 12 369 excluded sites (open circles). Initial number of trees = 50, step size = 50. With a learning rate of 0·005 and a tree complexity of 5, the step procedure identified the optimal number of trees as 1050, whereas with independent data the minimum was 950.

SIMPLIFYING THE PREDICTOR SET

Variable selection in BRT is achieved because the model largely ignores non-informative predictors when fitting trees. This works reasonably well because measures of relative influence quantify the importance of predictors, and irrelevant ones have a minimal effect on prediction. However, unimportant variables can be dropped using methods analogous to backward selection in regression (Miller 1990); these are sometimes referred to as recursive feature elimination. Such simplification is most useful for small data sets where redundant predictors may degrade performance by increasing variance. It is also useful if users are uncomfortable with inclusion of unimportant variables in the model. We detail our methods for simplification in Appendix S2 (see Supplementary material).

UNDERSTANDING AND INTERPRETING THE MODEL

A recognized advantage of individual decision trees is their simplicity, but boosting produces a model with hundreds to thousands of trees, presenting a challenge for understanding the final model. Nevertheless, BRT does not have to be treated like a black box, and we show here how the models can be summarized, evaluated and interpreted similarly to conventional regression models.

RELATIVE IMPORTANCE OF PREDICTOR VARIABLES

Formulae developed by Friedman (2001) and implemented in the gbm library estimate the relative influence of predictor variables. The measures are based on the number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees (Friedman & Meulman 2003). The relative influence (or contribution) of each variable is scaled so that the sum adds to 100, with higher numbers indicating stronger influence on the response.

For the *A. australis* model developed on 1000 sites through CV, the six most important variables described the importance of various reach, segment, upstream and downstream conditions, and fishing method (Table 2).

PARTIAL DEPENDENCE PLOTS

Visualization of fitted functions in a BRT model is easily achieved using partial dependence functions that show the

Table 2. Summary of the relative contributions (%) of predictor variables for a boosted regression tree model developed with cross-validation on data from 1000 sites using tree complexity of 5 and learning rate of 0·005

Predictor	Relative contribution (%)
SegSumT	24·7
USNative	11·3
Method	11·1
DSDist	9·7
LocSed	8·0
DSMaxSlope	7·3
USSlope	6·9
USRainDays	6·5
USAvgT	5·7
SegTSeas	5·7
SegLowFlow	2·9
DSDam	0·1

effect of a variable on the response after accounting for the average effects of all other variables in the model. While these graphs are not a perfect representation of the effects of each variable, particularly if there are strong interactions in the data or predictors are strongly correlated, they provide a useful basis for interpretation (Friedman 2001; Friedman & Meulman 2003). The partial responses for *A. australis* for the six most influential variables (Fig. 6) indicate a species occurring in warm, lowland rivers that have gentle downstream slopes and substantial clearing of upstream native vegetation. They demonstrate that short-finned eels often occur close to the coast, but are able to penetrate some distance inland, and prefer reaches with fine sediments. The species is most commonly caught using electric fishing, with lower success from nets, spotlighting and traps.

IDENTIFYING IMPORTANT INTERACTIONS

Even if a decision tree has several nodes, it may not be modelling interactions between predictors because they will be fitted only if supported by the data. In the absence of interactions, in a multinode tree the same response would be fitted to each side of splits below the first node. In effect, *tc* controls the maximum level of interaction that can be quantified, but no information is provided automatically on the nature and magnitude of fitted interaction effects. To quantify these, we use a function that creates, for each possible pair of predictors, a temporary grid of variables representing combinations of values at fixed intervals along each of their ranges. We then form predictions on the linear predictor scale for this grid, while setting values for all other variables to their respective means. We use a linear model to relate these temporary predictions to the two marginal predictors, fitting the latter as factors. The residual variance in this linear model indicates the relative strength of interaction fitted by BRT, with a residual variance of zero indicating that no interaction effects are fitted. Code and examples are available in the Supplementary material.

For *A. australis*, six of the seven most important pairwise interactions all included the most influential predictor,

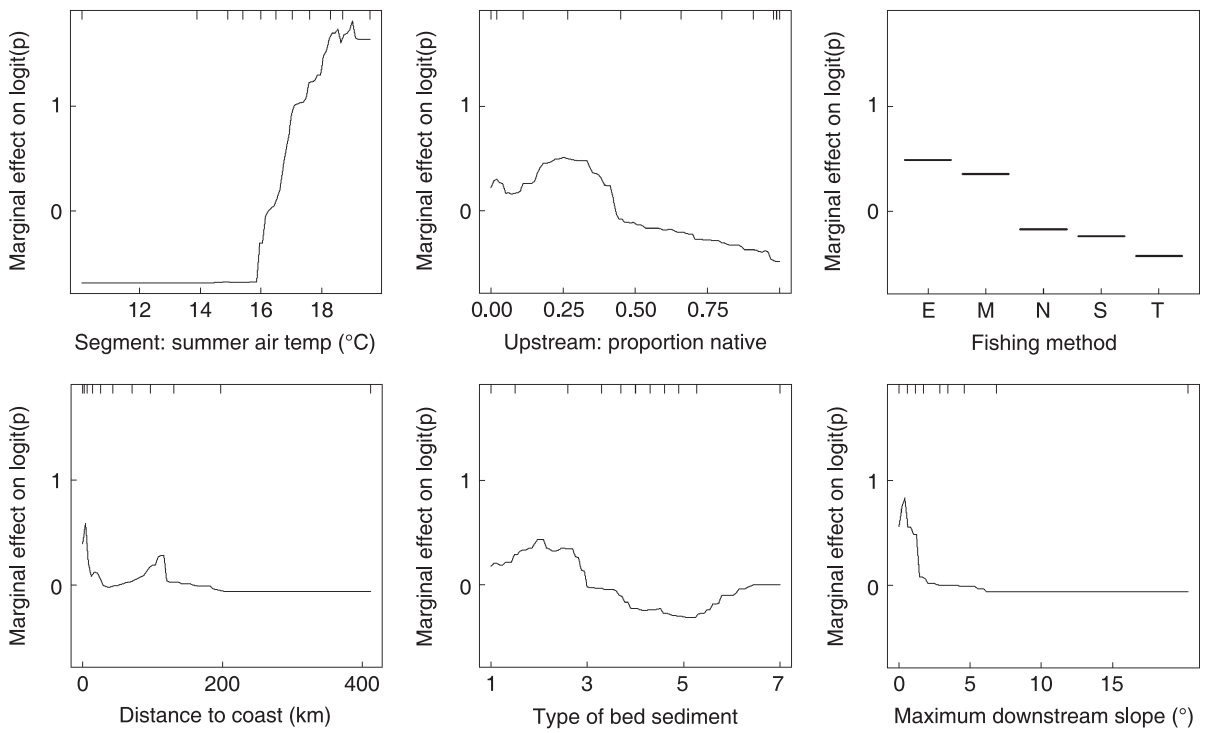

Fig. 6. Partial dependence plots for the six most influential variables in the model for short-finned eel. For explanation of variables and their units see Table 1. *Y* axes are on the logit scale and are centred to have zero mean over the data distribution. Rug plots at inside top of plots show distribution of sites across that variable, in deciles.

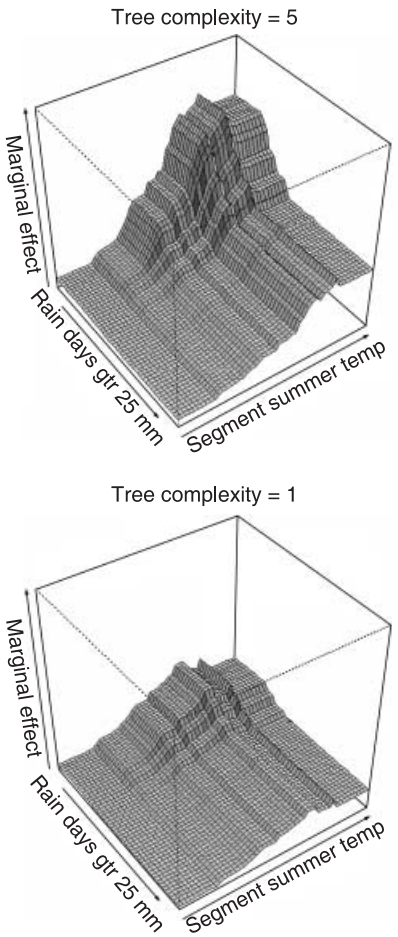

Fig. 7. Three-dimensional partial dependence plots for the strongest interaction in the model for short-finned eel (top), compared with that in a model without interactions (bottom). All variables except those graphed are held at their means. For explanation of variables and their units see Table 1; both plots have the same scaling for each axis.

SegSumT. Once identified, interactions can be visualized with joint partial dependence plots. The most important interaction for *A. australis* is shown in Fig. 7 (top panel), compared to the response predicted if interactions were not allowed ($tc = 1$). In this case, allowing interactions reinforces the suitability of environments that combine warm temperatures with low frequency of floods caused by high-intensity rain events in the upstream catchment. With interactions modelled, fitted values for such environments are more than twice those fitted by a model in which no interaction effects are allowed.

PREDICTIVE PERFORMANCE

BRT models can be used for prediction in the same way as any regression model, but without additional programming their complexity requires predictions to be made within the modelling software (in this paper, R) rather than in a GIS. Where predictions are to be mapped over many (e.g. millions) of points, scripts can be used to manage the process; examples are given in the Supplementary material. Prediction to any given site uses the final model, and consists of the sum of predictions from all trees multiplied by the learning rate. Standard errors can be estimated with bootstrap procedures, as demonstrated by Leathwick *et al*. (2006).

Where BRT models are developed with CV, statistics on predictive performance can be estimated from the subsets of data excluded from model fitting (see Fig. 4 and Supplementary material). For the model presented previously, the CV estimate of prediction error was close to that on independent data, although slightly overoptimistic (Fig. 5, compare solid and open circles; Table 3, see estimates on independent data compared with CV). This is a typical result, although the ability of CV to estimate true performance varies with data set and species prevalence. In small data sets, CV estimates of predictive performance may be erratic, and repeated and/or stratified cross-validation can help stabilize them (Kohavi 1995).

Predictive performance should not be estimated on training data, but results are provided in Table 3 to show that BRT overfits the data, regardless of careful model development (Table 3; see difference between estimates on training and independent data). While overfitting is often seen as a problem in statistical modelling, our experience with BRT is that prediction to independent data is not compromised – indeed, it is generally superior to other methods (see e.g. comparisons with GLM, GAM and multivariate adaptive regression splines, Elith *et al*. 2006; Leathwick *et al*. 2006). The flexibility in the modelling that allows overfitting also enables an accurate description of the relationships in the data, provided that overfitting is appropriately controlled.

HOW MULTIPLE TREES PRODUCE CURVILINEAR FUNCTIONS

Finally, having explored important features of a BRT model, we return to the question of how multiple shrunken trees can, in combination, fit a nonlinear function. In gbm it is possible to view the structure of each tree in a BRT model, and to plot the partial response to any variable over any chosen number

Table 3. Predictive performance of a BRT model, as evaluated on three different data sets (the model, comprising 1050 trees, is the same as reported in Table 2)

	Independent (12 369 sites)	Cross-validation* (1000 sites)	Train (1000 sites)
Percentage deviance explained	28·3	31·3 (0·96)	52·6
Area under the receiver operating characteristic curve	0·858	0·869 (0·015)	0·958

*Mean and SE estimated within model building.

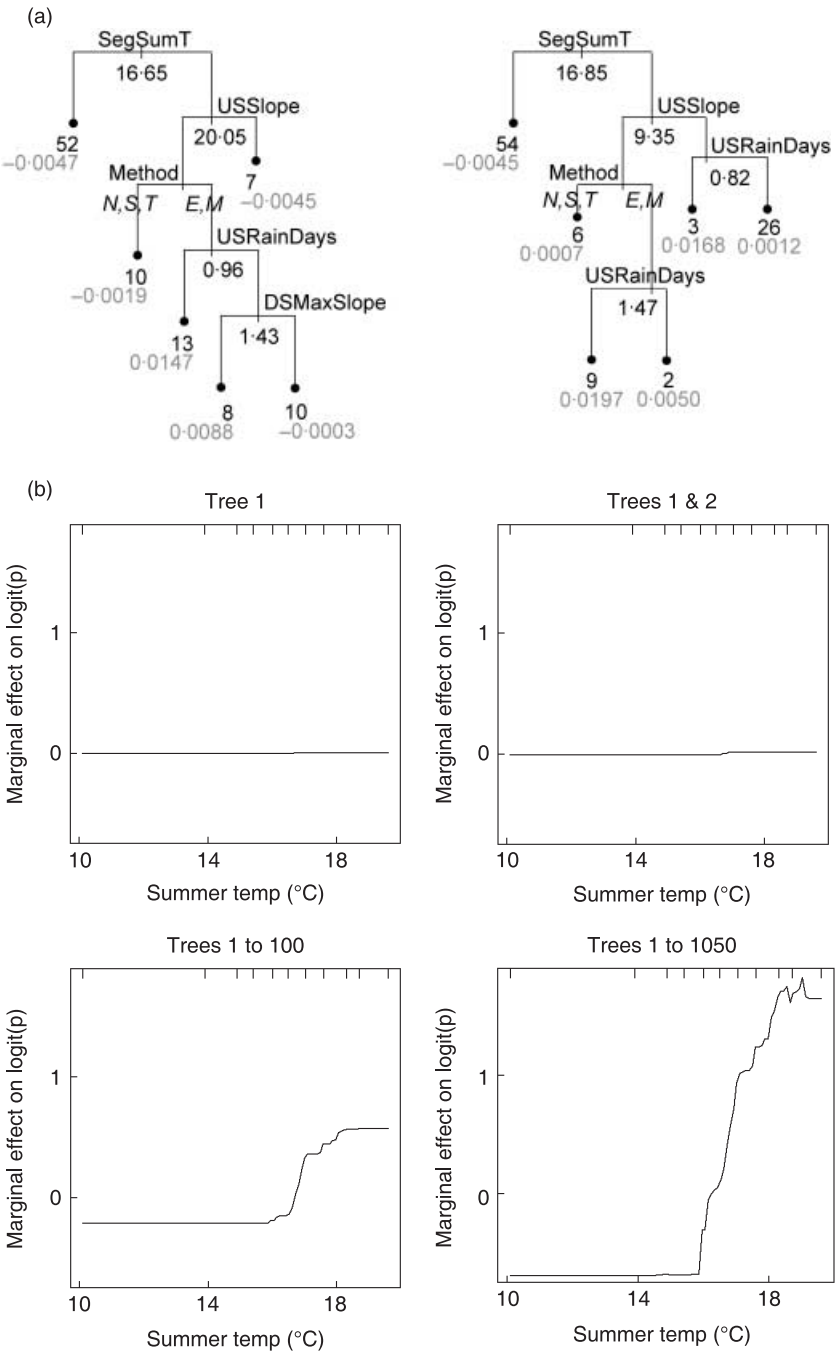

Fig. 8. (a) The first two trees in the boosted regression tree (BRT) model developed on 1000 sites with cross-validation. Variable names and units and codes for 'Method' are in Table 1. Split values are displayed under the split, and terminal nodes show percentage of sites in that node (black) and prediction in logit space (grey). For all splits on continuous variables, the lower values for the variable are to the left; (b) Partial plots from this BRT model for segment summer air temperature (SegSumT), using varying numbers of trees from 1 (top left) to 1050 (bottom right).

of shrunken trees (Fig. 8). Our final CV model contained 1050 trees. The first two trees had four out of five variables in common, with the first split in both trees on the same variable but at slightly different values (Fig. 8a). The first tree split on summer temperature at 16·65 °C, showing as a tiny step in the partial plot constructed only from the first tree (Fig. 8b, top left). The step is small in comparison with the final amplitude of the response because the contribution of each tree in the boosted model is shrunk by the learning rate. Adding information from the second tree (Fig. 8a, right) adds a second step at 16·85 °C (Fig. 8b, top right). Summer temperature was the most influential variable in this model, and occurred in 523 of the trees. Gradually, as more trees are included in the partial plot, the response to summer temperature becomes more complex and curvilinear (Fig. 8b, bottom row).

Discussion

BRT is a flexible regression modelling technique that gives important benefits for modelling ecological data, provided that care is exercised in model fitting. We have focused on both explaining the technique and demonstrating how to fit and evaluate the models, because this information is presented elsewhere in rather technical language, is rarely given an

ecological context, and sometimes portrays boosting as a 'black-box' procedure. The models developed for *A. australis* are consistent with the known ecology of the species, and accurately describe a species occurring in warm, lowland rivers in agricultural landscapes, often close to the coast but also penetrating inland, and preferring reaches with fine sediments. The modelled interactions highlight the suitability of habitats combining low flood frequencies and warm temperatures. Further applications of BRT to these data, analysing the contrasting distributions of 30 fish species, are provided by Leathwick *et al*. 2008).

BRT models are able to select relevant variables, fit accurate functions and automatically identify and model interactions, giving sometimes substantial predictive advantage over methods such as GLM and GAM. A growing body of literature quantifies this difference in performance (Elith *et al*. 2006; Leathwick *et al*. 2006; Moisen *et al*. 2006). Efficient variable selection means that large suites of candidate variables will be handled better than in GLM or GAM developed with stepwise selection. Additionally, in contrast to single decision trees that handle continuous gradients by fitting them in large steps (Figure 1), boosted trees model a much smoother gradient, analogous to the fit from a GAM. Admittedly, the BRT fitted functions can be rather noisy; this is mostly in regions of the data space that are sparsely sampled, but does not seem to adversely affect overall performance. Unlike GAM, BRT can handle sharp discontinuities, an important feature when modelling the distributions of species that occupy only a small proportion of the sampled environmental space. BRT models can be fitted to varying amounts of data, similar to other regression models, but settings need to be carefully controlled in smaller samples, particularly where using stochastic boosting.

Boosting has features that differentiate it from other model aggregation methods, and brief comment may help place BRT into that broader context. One of the most popular ensemble methods, bootstrap aggregation or bagging, underpins methods such as bagged trees and random forests (BT and RF, Prasad *et al.* 2006). These averaging techniques also improve the performance of single tree models by making many trees and, in the case of RF, randomly selecting a subset of variables at each node. While BT and RF reduce variance more than single trees, they cannot achieve any bias reduction, because each tree is based on a bootstrap sample that will be distributed in much the same way as the original training set. As a consequence, the resulting average bias is identical to the bias of any one tree. By contrast, boosting grows the suite of trees by sequentially modelling the residuals throughout all parts of the data space, including those for atypical observations that depart from the dominant patterns explained by the initial trees. In this way, it reduces both bias (through forward stagewise fitting) and variance (through model averaging). Random forest models are starting to be applied in ecology (e.g. Prasad *et al*. 2006), but we know of no comparisons of RF and BRT with ecological data, and comparisons in other disciplines have so far been across a restricted set of data characteristics (Segal 2004). One potential advantage of BRT is that different types of response variable (e.g. binomial, count, normal) are handled explicitly using appropriate and robust loss functions. A presence-only implementation of BRT (one that deals with species records where only presence is recorded, such as museum data) will soon be available (G. Ward, T. Hastie, S.C. Barry, J. Elith & J.R. Leathwick, unpublished data).

We acknowledge that exploiting the advantages of BRT requires some reorientation in thinking. Compared with conventional regression models, there are no *P* values to indicate the relative significance of model coefficients, degrees of freedom in the model are hard to determine, and the paradigm is quite different from one focusing on selecting a single 'best' model containing few parameters. These aspects can be viewed either as problems or as opportunities. There is a vigorous debate in the literature about the use and abuse of *P* values in models (Fidler *et al*. 2004), and a strong case can be made that alternatives provide more insight and are less often misunderstood. Model selection in BRT, consistent with many modern techniques that focus on regularization through shrinkage (see discussion of ridge regression and lasso, L1 and L2 penalties by Hastie *et al*. 2001), provides a coherent and robust alternative to traditional approaches such as stepwise variable selection (Whittingham *et al*. 2006). Although the lack of a single simple model may be regarded as disadvantageous from a traditional perspective, we have demonstrated a range of methods for both interpretation and prediction, and these provide functional equivalence to many of the techniques used with conventional regression.

In conclusion, we have found that moving from a background of using only conventional statistical models to including ML methods as an analytical option has brought with it considerable advantages. In particular, it has provided familiarity with alternative notions of model selection, prompted use of methods for model tuning and evaluation through CV or other resampling methods, and extended our ability to ask questions of both models and predictions in new and informative ways. Motivated by this, and by the analytical gains we have made through use of these models, we provide the online tutorial (Supplementary material) as one starting point for using BRT for typical ecological analyses.

Acknowledgements

J.E. was funded by ARC grants LP0667891 and DP0772671, and the Australian Centre of Excellence for Risk Analysis; J.L.'s contribution was funded by New Zealand's Foundation for Research, Science and Technology under contract C01X0305. We appreciated insightful comments on the manuscript from Mark Burgman, Yung En Chee, and reviewers and editors.

References

Breiman, L. (2001) Statistical modeling: the two cultures. *Statistical Science*, **16**, 199–215.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984) *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, USA.

Buckland, S.T. & Elston, D.A. (1993) Empirical models for the spatial distribution of wildlife. *Journal of Applied Ecology*, **30**, 478–495.

Burnham, K.P. & Anderson, D.R. (2002) *Model Selection and Inference: A Practical Information-Theoretic Approach*, 2nd edn. Springer-Verlag, New York.

Clarke, A. & Johnston, N.M. (1999) Scaling of metabolic rate with body mass and temperature in teleost fish. *Journal of Animal Ecology*, **68**, 893–905.
De'ath, G. (2007) Boosted trees for ecological modeling and prediction. *Ecology*, **88**, 243–251.
De'ath, G. & Fabricius, K.E. (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, **81**, 3178–3192.
R Development Core Team (2006) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
Elith, J., Graham, C.H., Anderson, R.P. *et al*. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.
Fewster, R.M., Buckland, S.T., Siriwardena, G.M., Baillie, S.R. & Wilson, J.D. (2000) Analysis of population trends for farmland birds using generalized additive models. *Ecology*, **81**, 1970–1984.
Fidler, F., Thomason, N., Cumming, G., Finch, S. & Leeman, J. (2004) Editors can lead researchers to confidence intervals but they can't make them think. Statistical reforms from medicine. *Psychological Science*, **15**, 119–126.
Freund, Y. & Schapire, R.E. (1996) Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference*, July 3–6, 1996, Bari Italy. pp. 148–156. Morgan Kaufman, San Francisco, CA, USA.
Friedman, J.H. (2001) Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, **29**, 1189–1232.
Friedman, J.H. (2002) Stochastic gradient boosting. *Computational Statistics and Data Analysis*, **38**, 367–378.
Friedman, J.H. & Meulman, J.J. (2003) Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, **22**, 1365–1381.
Friedman, J.H., Hastie, T. & Tibshirani, R. (2000) Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, **28**, 337–407.
Hastie, T. & Tibshirani, R.J. (1990) *Generalized Additive Models*. Chapman & Hall, London.
Hastie, T., Tibshirani, R. & Friedman, J.H. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York.
Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (ed. C.A. San Mateo), pp. 1137–1143. Morgan Kaufmann.
Leathwick, J.R., Elith, J., Francis, M.P., Hastie, T. & Taylor, P. (2006) Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology Progress Series*, **321**, 267–281.
Leathwick, J.R., Elith, J., Chadderton, W.L., Rowe, D. & Hastie, T. (2008) Dispersal, disturbance, and the contrasting biogeographies of New Zealand's diadromous and non-diadromous fish species. *Journal of Biogeography*, in press.
McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models*, 2nd edn. Chapman & Hall, London.
McDowall, R.M. (1993) Implications of diadromy for the structuring and modelling of riverine fish communities in New Zealand. *New Zealand Journal of Marine and Freshwater Research*, **27**, 453–462.
Miller, A.J. (1990) *Subset Selection in Regression*. Chapman & Hall, London.
Moisen, G.G., Freeman, E.A., Blackard, J.A., Frescino, T.S., Zimmermann, N.E. & Edwards, T.C. (2006) Predicting tree species presence and basal area in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. *Ecological Modelling*, **199**, 176–187.
Prasad, A.M., Iverson, L.R. & Liaw, A. (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, **9**, 181–199.
Reineking, B. & Schröder, B. (2006) Constrain to perform: regularization of habitat models. *Ecological Modelling*, **193**, 675–690.
Ridgeway, G. (2006) Generalized boosted regression models. *Documentation on the R Package 'gbm', version 1·5–7*. http://www.i-pensieri.com/gregr/gbm.shtml, accessed March 2008.
Schapire, R. (2003) The boosting approach to machine learning – an overview. *MSRI Workshop on Nonlinear Estimation and Classification, 2002* (eds D.D. Denison, M. H. Hansen, C. Holmes, B. Mallick & B. Yu). Springer, New York.
Segal, M.R. (2004) Machine learning benchmarks and random forest regression. *eScholarship Repository*. University of California. http://repositories.cdlib.org/cbmb/bench_rf_regn
Whittingham, M.J., Stephens, P.A., Bradbury, R.B. & Freckleton, R.P. (2006) Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, **75**, 1182–1189.

Received 16 October 2007; accepted 25 January 2008
Handling Editor: Bryan Manly

Supplementary material

The following supplementary material is available for this article.

Appendix S1. Explanation of the effect of learning rate on predictive stability in boosted regression trees.

Appendix S2. Simplifying the predictor set.

Appendix S3. Online tutorial with code and data.

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/full/10.1111/j.1365-2656.2008.01390.x
(This link will take you to the article abstract.)

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.