Ecosystems (2006) 9: 181–199
DOI: 10.1007/s10021-005-0054-1

ECOSYSTEMS

Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction

Anantha M. Prasad,[1]* Louis R. Iverson,[1] and Andy Liaw[2]

[1]*Northeastern Research Station, USDA Forest Service, 359 Main Road, Delaware, Ohio 43015, USA;* [2]*Biometrics Research Department, Merck Research Laboratories, Rahway, New Jersey, USA*

Abstract

The task of modeling the distribution of a large number of tree species under future climate scenarios presents unique challenges. First, the model must be robust enough to handle climate data outside the current range without producing unacceptable instability in the output. In addition, the technique should have automatic search mechanisms built in to select the most appropriate values for input model parameters for each species so that minimal effort is required when these parameters are fine-tuned for individual tree species. We evaluated four statistical models—Regression Tree Analysis (RTA), Bagging Trees (BT), Random Forests (RF), and Multivariate Adaptive Regression Splines (MARS)—for predictive vegetation mapping under current and future climate scenarios according to the Canadian Climate Centre global circulation model. To test, we applied these techniques to four tree species common in the eastern United States: loblolly pine (*Pinus taeda*), sugar maple (*Acer saccharum*), American beech (*Fagus grandifolia*), and white oak (*Quercus alba*). When the four techniques were assessed with Kappa and fuzzy Kappa statistics, RF and BT were superior in reproducing current importance value (a measure of basal area in addition to abundance) distributions for the four tree species, as derived from approximately 100,000 USDA Forest Service's Forest Inventory and Analysis plots. Future estimates of suitable habitat after climate change were visually more reasonable with BT and RF, with slightly better performance by RF as assessed by Kappa statistics, correlation estimates, and spatial distribution of importance values. Although RTA did not perform as well as BT and RF, it provided interpretive models for species whose distributions were captured well by our current set of predictors. MARS was adequate for predicting current distributions but unacceptable for future climate. We consider RTA, BT, and RF modeling approaches, especially when used together to take advantage of their individual strengths, to be robust for predictive mapping and recommend their inclusion in the ecological toolbox.

Key words: predictive mapping; data mining; classification and regression trees (CART); Regression Tree Analysis (RTA); decision tree; Multivariate Adaptive Regression Splines (MARS); Bagging Trees; Random Forests; Kappa; fuzzy Kappa; Canadian Climate Centre (CCC); global circulation model (GCM); eastern United States.

Introduction

Ecosystem scientists frequently need tools to extrapolate findings discovered via intensive local research across a larger area. Many management

Received 22 April 2004; accepted 16 December 2004; published online 15 March 2006.
*Corresponding author; e-mail: aprasad@fs.fed.us

decisions, ranging from protecting potentially rare species or communities to estimating potential commodity yields, rely on reliable, landscape-level maps produced from sample locations (Miller and others 2004). Many such decisions also rely on predictions of possible future conditions, such as the impacts of climate change. These maps are generally produced via simulation or statistical modeling from samples and predictor variables, current and potential future, that are already mapped. There are numerous techniques in place for such mapping; in this paper, we evaluate four of them, including two techniques new to ecosystem scientists.

We employ statistical approaches to model species distributions, using relevant predictors to estimate current abundances as well as plausible future abundances resulting from climate change. Typically, regression-like techniques have been used, but traditional parametric methods do not always yield satisfactory results for continental-scale mapping because the same variables may not operate in the same way throughout a species' range (Moore and others 1991; Franklin 1995). Newer computer intensive data-mining techniques based on recursion, resampling, averaging, and randomizations can uncover hidden structures in the data and yield better predictive models.

Estimating species range shifts under changing climate conditions, especially through a fragmented landscape, has become a key area of research for several groups, including ours (for examples, Davis 1989; Pitelka and Plant Migration Group 1997; Clark 1998; Hobbs 1994; Iverson and others 1999b; Schwartz and others 2001; Malcolm and others 2002; Higgins and others 2003). Previously, we were among the first to use Regression Tree Analysis (RTA) to spatially model the distribution of tree importance values for 80 tree species in the eastern United States for future climatic scenarios (Iverson and Prasad 1998; Iverson and others 1999a; Prasad and Iverson 2000a). Our model, DISTRIB, was run at county scale and incorporated 33 climatic, edaphic, and land-use variables in predicting how tree-habitat distributions might change under a doubled carbon dioxide scenario as estimated by five global circulation models (GCMs) (Iverson and Prasad 2002). The results were the basis for a climate-change tree atlas (Iverson and others 1999a; Prasad and Iverson 2000a) for the eastern United States. RTA provided a satisfactory overall prediction model for our data set, although we were aware of its limitations (discussed in section on Regression Trees). In this paper, we evaluate three other techniques, with the purpose of building better predictive models: Bagging Trees (BT), Random Forests (RF), and Multivariate Adaptive Regression Splines (MARS). The aim is to test the four modeling techniques—RTA, BT, RF, and MARS—on four tree species with different distributions and characteristics. Our emphasis throughout will be on the modeling techniques and not on the ecological details of the tree species.

Statistical Models

All of the statistical models tested here are computer-intensive algorithms that have been used in data-mining and large-scale predictions, particularly in the field of machine learning. All but MARS use a classification or regression tree (also known as ''CART'') approach to recursively partition predictor variables. Classification trees are used in applications dealing with categorical or remote sensing data. We use regression trees because our response variable is continuous. The differences between classification and regression trees pertain to the techniques used for data splitting and aggregating. We briefly discuss each technique in the following sections. A comparison summary of the four modeling techniques is presented in Table 1.

Regression Tree Analysis

Unlike classical regression techniques for which the relationship between the response and predictors is pre-specified (for example, straight line, quadratic) and the test is performed to prove or disprove the relationship, RTA assumes no such relationship. It is primarily a method of constructing a set of decision rules on the predictor variables (Breiman and others 1984; Verbyla 1987; Clark and Pregibon 1992). The rules are constructed by recursively partitioning the data into successively smaller groups with binary splits based on a single predictor variable. Splits for all of the predictors are examined by an exhaustive search procedure and the best split is chosen. For regression trees, the selected split is the one that maximizes the homogeneity of the two resulting groups with respect to the response variable (the split that maximizes the between-groups sum of squares, as in analysis of variance [ANOVA], although other options may be available. The output is a tree diagram with the branches determined by the splitting rules and a series of terminal nodes that contain the mean response. The procedure initially grows maximal trees and then uses techniques such as cross-validation to prune the overfitted tree to an optimal size (Therneau and Atkinson 1997).

Table 1. Comparison of the Four Modeling Techniques

Model	Method	Strengths	Limitations
RTA	Recursively partitions data based on a single, best predictor to form a binary tree. Creates a series of decision rules based on the predictor variables.	Better than traditional linear techniques in allowing for interactions and nonlinearities when numerous predictors are present. Easy to interpret and also allows spatial mapping of predictors with the greatest influence on distribution.	Linear functions are highly approximated and the output tree can be highly variant to small perturbations of data.
BT	Creates multiple boot-strapped regression trees without pruning and averages the outputs.	Very effective in reducing variance and error in high dimensional data sets. The data not used in the training set, termed ''out-of-bag'' data, can be used to provide better estimates of errors.	Because large number of trees (30–50, typically) are averaged, interpreting the results is not easy. Also, the bias component of the error is marginally better than single regression trees.
RF	Similar to BT except that each tree is grown with a randomized subset of predictors. Typically, 500 to 2,000 trees are grown and the results aggregated by averaging.	Growing large numbers of trees does not overfit the data, and random predictor selection keeps bias low. Provides better models for prediction.	Even more of a ''black box'' approach than BT. Can be very demanding in terms of time and computer resources.
MARS	Builds localized regression models by fitting separte splines using basis functions to distinct intervals of predictor variables.	Because splitting rules are replaced by continuous smooth functions, MARS is better at detecting global and linear data structure. Also, the output is smoother and less coarse-grained.	Tends to be excessively guided by the local nature of the data, making prediction with new data very unstable. Selecting values for input parameters can be cumbersome.

RTA, Regression Tree Analysis; BT, Bagging Trees; RF, Random Forest; MARS, Multivariate Adaptive Regression Splines.

Regression Tree Analysis has clear advantages over classical statistical methods. It is effective in uncovering structure in data with hierarchical or nonadditive variables. Because no a priori assumptions are made about the nature of the relationships among the response and predictor variables, RTA allows for the possibility of interactions and nonlinearities among variables (Moore and others 1991). RTA's splitting rules enable mapping of predictors with the greatest influence on distributions, providing greater insight into the spatial influence of the predictors (Iverson and Prasad 1998). Yet there are certain disadvantages with RTA compared to conventional regression modeling: (a) simple linear functions are highly approximated; (b) for certain data sets, it is difficult to constrain the model by selecting the optimum pruning parameter through cross-validation; (c) the output can be a discontinuous, coarse-grained response for some species, depending on the threshold established in the regression trees; and (d) the output can be unstable—that is, small changes in data can produce highly divergent trees.

Classification and regression trees have found favor among researchers for several biological applications, including remote sensing (for example, Lees and Ritman 1991; Hansen and others 1996; De'ath and Fabricius 2000; Stoppianea and others 2003), assessing potential for tree mortality (Baker 1993; Dobbertin and Biging 1998), vegetation mapping (for example, Michaelsen and others 1994; Franklin 1998; Iverson and Prasad 1998), and predicting species invasions or pest outbreaks (Hernandez and others 1997; Reichard and Hamilton 1997).

Bagging Trees

The basic idea underlying BT is the recognition that part of the output error in a single regression tree is due to the specific choice of the training data set. Therefore, if several similar data sets are created by resampling with replacement (that is, bootstrapping) and regression trees are grown without pruning and averaged, the variance component of the output error is reduced (Breiman 1996a; Buhlmann and Yu 2002). When a bootstrap resample is drawn, about 37% of the data is excluded from the sample, but other data are replicated to bring the sample to full size. The portion of the data drawn into the sample in a replication is known as the ''in-bag'' data, whereas the portion not drawn is the ''out-of-bag'' data. The latter are not used to build or prune any tree but provide better estimates of node error and other generalization errors for bagged predictors (Breiman 1996b). The result is a slightly perturbed version of the data set with each replication. If the separate analyses differ considerably from each other, trees exhibit instability, so averaging improves results. The primary disadvantage of BT is that it requires averaging 30–80 trees, so interpreting multiple individual trees becomes nearly impossible. For some species for which the multiple trees vary little, the most influential splits are the same and a single RTA tree can be used for interpretation. By contrast, if the multiple trees vary widely, a single RTA tree may be only one of several possible interpretations of the modeled relationship, and the uncertainty of interpretation is higher. Species with this higher uncertainty are typically those whose present distribution is a result of complex natural historical factors that are not adequately captured by the predictor variables.

Bagging has been applied frequently in fields such as biostatistics and remote sensing (for example, Hothorn and others 2004; Chan and others 2001), but its use is uncommon in the field of ecology. A related technique called ''boosting'' (Freund 1995; Schapire and others 1998) has recently gained popularity. In boosting, bias is reduced by repeatedly readjusting the weights of the training samples, by focusing on ''difficult'' examples from previous samples. Boosting is competitive with bagging and is used primarily in classifying data with large training sample sizes (Skurichina and Duin 2002). Because our primary goal was regression and not classification, we did not include boosting in our comparisons.

Random Forests

Random Forests is a new entry to the field of data-mining and is designed to produce accurate predictions that do not overfit the data (Breiman 2001, 2002). RF is similar to BT in that bootstrap samples are drawn to construct multiple trees; the difference is that the each tree is grown with a randomized subset of predictors, hence the name ''random'' forests. A large number of trees (500 to 2,000) are grown, hence a ''forest'' of trees. The number of predictors used to find the best split at each node is a randomly chosen subset of the total number of predictors. As with BT, the trees are grown to maximum size without pruning, and aggregation is by averaging the trees. Out-of-bag samples can be used to calculate an unbiased error rate and variable importance, eliminating the need for a test set or cross-validation. Because a large number of trees are grown, there is limited generalization error (that is, the true error of the popu-

lation as opposed to the training error only), which means that no overfitting is possible, a very useful feature for prediction.

By growing each tree to maximum size without pruning and selecting only the best split among a random subset at each node, RF tries to maintain some prediction strength while inducing diversity among trees (Breiman 2001). Random predictor selection diminishes correlation among unpruned trees and keeps the bias low; by taking an ensemble of unpruned trees, variance is also reduced. Another advantage of RF is that the predicted output depends only on one user-selected parameter, the number of predictors to be chosen randomly at each node.

Random Forests seems more of a ''black box'' approach than BT because we cannot examine the individual trees separately. However, it provides several metrics that aid in interpretation. Variable importance is evaluated based on how much worse the prediction would be if the data for that predictor were permuted randomly. The resulting tables can be used to compare relative importance among predictor variables. As such, the procedure is much more interpretable than methods such as neural networks, and might be better termed a ''gray box'' approach.

We have not seen the use of RF reported in the ecological literature. The only biological application associated with RF was an investigation of risk-mapping of tick-borne encephalitis in landscape epidemiology (Furlanello and others 2003).

Multivariate Adaptive Regression Splines

Multivariate Adaptive Regression Splines (Friedman 1991) is well known in the data-mining field and purportedly addresses some limitations of RTA with respect to continuous variables. The MARS procedure builds flexible regression models by using basis functions to fit separate splines to distinct intervals of the predictor variables. Both the variables to use and the end points of the intervals, or knots, are found by an exhaustive search procedure using a special class of basis functions (Abraham and Steinberg 2001). This approach differs from classical splines where the knots are predetermined and spaced evenly. MARS finds the location and number of required knots in a forward/backward stepwise fashion. First, the model is overfitted by generating more knots than needed, and the resulting knots that contribute least to the overall fit are removed. Basis functions used in MARS are similar to principal components because they re-express the relationship of the predictor variables with the response variable (Steinberg and others 1999).

Multivariate Adaptive Regression Splines has an advantage over RTA in that RTA's discontinuous branching at tree nodes is replaced with continuous smooth functions that are guided by the local nature of the data. Therefore, MARS is better at detecting global and linear data structure so that its output is smoother and not as coarse-grained and discontinuous as RTA. However, MARS also has its limitations: (a) its basis functions are sometimes excessively guided by the local nature of the data, resulting in inappropriate outcomes; and (b) selecting the correct values for the parameters can be cumbersome and may entail multiple trial-and-error steps. MARS also does not lend itself well to interpretation of species–environment relationships and has been used infrequently in ecosystem science except for mapping certain vegetation characteristics (Prasad and Iverson 2000b; Moisen and Frescino 2002; Munoz and Felicisimo 2004).

Methods

We analyzed nearly 3 million tree records generated by the USDA Forest Service's Forest Inventory and Analysis (FIA) program to derive tree importance values (IV) for each species. The species resided in about 100,000 plots across the 37 states within the United States east of the 100th meridian. Four of 135 tree species being modeled were selected as representative examples to report the range of model behavior in this paper: loblolly pine (*Pinus taeda*), sugar maple (*Acer saccharum*), American beech (Fagus grandifolia), and white oak (*Quercus alba*) (Hansen and others 1992).

Importance value was calculated based equally on relative basal area and the number of stems contained within each plot, with a maximum value of 100 in monotypic stands. The plot-level IVs were then averaged over each of 9,782 20×20 km cells for the entire study area. The averaged IVs for each cell were rounded to whole numbers with one exception. If the IV was greater than 0 but less than 1, it was assigned to 1 because rounding would have falsely turned species-present cells to species-absent cells.

Our predictor dataset consisted of 36 variables, including climate, soil, land-use, landscape, and topographic variables from various sources (Table 2). The future climate data set is from Canadian Climate Centre's (CCC) GCM (Boer and others 2000; Kittel and others 2000). CCC is a transient model in which 30-year climatic averages

Table 2. Variables Used to Predict Current and Future Tree Distributions

Abbreviation	Variable
AGRICULT	Cropland (%)
ALFISOL	Alfisol (%)
ARIDISOL	Aridisol (%)
AVGT	Mean Annual Temperature (°C)
BD	Soil Bulk Density (g/cm^3)
CLAY	Percent Clay (<0.002 mm)
ELV_CV	Elevation Coefficient of Variation
ELV_MAX	Maximum Elevation (m)
ELV_MEAN	Average Elevation (m)
ELV_MIN	Minimum Elevation (m)
ELV_RANGE	Range of Elevation (m)
ENTISOL	Entisol (%)
FOREST	Forest land (%)
FRAG	Fragmentation Index (Riitters and others 2002)
HISTOSOL	Histosol (%)
INCEPTSOL	Inceptisol (%)
JANT	Mean January Temperature (°C)
JULT	Mean July Temperature (°C)
KFFACT	Soil Erodibility Factor, Rock Fragments Free (susceptibility of soil erosion to water movement)
MAYSEPT	Mean May–September Temperature (°C)
MOLLISOL	Mollisol (%)
NO10	Percent Passing Sieve No. 10 (coarse)
NO200	Percent Passing Sieve No. 200 (fine)
NONFOREST	Nonforest land (%)
OM	Organic Matter Content (% by weight)
ORD	Potential Soil Productivity (m^3 of timber/ha)
PERM	Soil Permeability Rate (cm/h)
PH	Soil pH
PPT	Annual Precipitation (mm)
ROCKDEP	Depth to Bedrock (cm)
ROCKFRAG	Percent Weight of Rock Fragments (8–25 cm)
SLOPE	Soil Slope (%) of a Soil Component
SPODOSOL	Spodosol (%)
TAWC	Total Available Water Capacity (cm, to 152 cm)
ULTISOL	Ultisol (%)
VERTISOL	Vertisol (%)

were estimated for the period 2071 to 2100, and data were obtained as half-degree cells from the USDA Forest Service Laboratory at Corvallis, Oregon (Neilson and Drapek, personal communication). We first ran the models with current climate and then with CCC.

Map Similarity Measures

Visual evaluation, Kappa statistics, and Pearson's correlation were used to compare actual species abundances with model outputs. Comparing maps visually for similarities and differences is the most comprehensive method of comparison because patterns, coherence, and local and global similarities are grasped intuitively. However, automated methods are more effective if procedures can be clearly defined so that repeatability and objectivity are maintained (Hagen 2002). A software comparison usually captures one of the comparison components, but it lacks the flexibility to look at others when the data demand it. To minimize this problem, we used several map similarity measures, including Kappa, K_{loc}, K_{histo}, and fuzzy Kappa, for a pixel-by-pixel comparison of map classes, based on subdivisions of the continuous variable IV.

Kappa and its Variants

With the Kappa statistic (Monserud and Leemans 1992), the level of agreement between maps is based on the contingency table, which details how the distribution of categories in map A differs from map B. Kappa is the proportion of agreement after chance agreement, or the percentage of agreement expected after randomly relocating all cells in the maps, has been removed. Kappa can range from −1 (no agreement) to 1 (perfect agreement between maps).

Two other variants to Kappa together define the locational and quantitative similarities (Pontius 2000; Hagen 2002). These are calculated by first defining the maximum fraction of agreement that could be attained if the locations of the cells in one of the maps were to be rearranged, $P(\max)$.

K_{loc} describes the spatial allocation of categories and compares the actual to expected success rate relative to the maximum success rate given that the total number of cells of each category does not change. It is calculated as:

$$K_{\text{loc}} = \frac{P(A) - P(E)}{P(\max) - P(E)}$$

Note that while K_{loc} gives an indication of the similarity of the spatial distribution of categories, it makes no distinction between a category that is dislocated by a distance of one cell versus another that is dislocated by a long distance.

K_{loc} values near 1 indicate that further improving the spatial allocation of the categories results in little overall improvement and therefore can be taken to be the upper limit of similarity that can be achieved. K_{loc} values near 0 indicate the lower limit

of similarity. Negative K_{loc} values indicate a major fundamental difference in location pattern of the two maps, larger than can expected from random selection of locations.

Hagen (2002) introduced a measure of quantitative similarity that can be calculated from the histograms of the two maps:

$$K_{histo} = \frac{P(\max) - P(E)}{1 - P(E)}$$

Kappa can now be defined as the product of K_{loc} and K_{histo}. The former is a measure of the spatial allocation of categories of the two maps; the latter is a measure of the quantitative similarity of the two maps (for details, see Hagen 2002; and Pontius 2000).

$$\text{Kappa} = K_{loc} \times K_{histo}$$

Kappa statistics provide insight into the nature of the predictions and facilitate comparison among models. However, Kappa tests usually are more appropriate in remote sensing applications—that is, comparing a field-sample-based land-use map with a satellite-classified map where a pixel-by-pixel comparison is more meaningful. In our case, the relative differences in Kappa values among the four models are more important than the actual values because we are using Kappa primarily as a comparison index.

We reclassified the predicted IV maps into eight categories to obtain the Kappa statistics: 0 or ''not present'' class (which includes model values up to 0.499); 1–3 (0.5 to 3.499), 4–6 (3.5 to 6.499), 7–10 (6.5 to 10.499), 11–20 (10.5 to 20.499), 21–30 (20.5 to 30.499), 31–50 (30.5 to 50.499), and 51–100 (50.5 to 100).

Fuzzy Kappa

Often in the real world, system properties are not crisp, that is, there are grades of similarity. Likewise, in the map world, there are grades of similarity between pairs of cells in two maps—that is, similarity between categories and/or location. The fuzzy set approach to calculating Kappa takes into account the fuzziness of categories as well as location. The fuzzy set approach expresses similarity of each cell in a value between 0 (distinct) and 1 (identical). The degree of uncertainty or ''vagueness'' among categories can be set with the fuzzy-category matrix. The fuzziness of location can be set with a function (usually an exponential, linear, or constant decay) that defines the level to which the neighboring cells influence the target cell (Power and Simms 2001; Hagen 2003). The fuzzy Kappa statistic is calculated similarly to the regular Kappa statistic, with the difference that the expected fraction of agreement $P(E)$ takes into account the fuzziness of location and categories.

In our case, there is some degree of class overlap because we have subjectively categorized a quasi-continuous distribution of IVs into eight logical classes. We created a fuzzy-category matrix to recognize that the closer classes tend to be fuzzy whereas categories far apart (for example, IV = 1–3 versus IV = 31–50) are distinct. We also must take into account the model output fuzziness, which occurs mostly between the 0 and the 1–3 classes. These factors are reflected in the fuzzy category matrix we used: the similarity between the 0 and 1–3 class was set to 0.6, between 1–3 and 4–6 was 0.5, and between 4–6 and 7–10 was 0.4. The remaining class similarity matrix with IV values greater than 10 was set to 0 because those comparisons had classes where misrepresentation by the models is less likely.

When estimating the locational fuzziness of a class, we wanted the level at which the neighboring cells influenced the target cell to be small because we are dealing with large, 20-km cells. After testing various parameters, we chose the exponential decay function with a radius of 1 and halving distance of 1; this was a closer approximation to reality because the IV of the 20-km cells were aggregated from the FIA plots that fell within that individual cell.

Software Used

We used the R statistical software (R Development Core Team 2004), which is based on the S language (Chambers and Hastie 1993; Chambers 1998). R is freeware that was developed by researchers who have contributed novel statistical techniques in the form of packages that can be plugged into R. However, the MARS package was deficient in R, so we used Salford System's (Steinberg and others 1999) MARS software. We used a package in R called ''rpart'' for RTA (Therneau and Atkinson 1997), ''ipred'' for BT (Peters and others 2002), and ''randomForest'' for RF (Liaw and Wiener 2002). GIS analysis was performed using ArcView 3.2a and ArcInfo 8.1.2 (including Grid) [Environmental Systems Research Institute 2001]. Kappa and fuzzy Kappa statistics were obtained via the Map Comparison Kit (Map Comparison Kit 2003) software.

Model Specifics

In RTA, one must decide when to stop pruning because tree size is not limited in the growing

process. Allowing the tree to grow unpruned will result in overfitted models because they fit noise along with data. From several options we chose a complexity parameter (cp) equal to 0.002 for pruning splits. Any split that does not decrease the overall lack of fit by a factor of cp at each step is not attempted. Essentially, any split that does not improve the fit by cp will likely be pruned by cross-validation and is not pursued. The recommended number of cross-validations is between 10 and 20; we chose 15. In BT, we combined 50 trees based on observations that bagging gives satisfactory results after 25 trees. The output of RF depends primarily on one input parameter: the number of predictors to be chosen randomly at each tree node. The default advocated by Breiman (Breiman and Cutler 2003) is one-third the number of predictors, (in our case, 12). But if the predictors contain noise variables (which is true in our case because some variables that are important for one species could be noise for another), a higher number is better. After some testing, we chose 15 and increased the number of trees from a default of 500 to 1,000 to further stabilize the errors. MARS purportedly selects the optimal model by automating several aspects of model development. In reality, several control parameters must be chosen carefully to obtain this optimal model. Selecting the number of basis functions and interactions is of prime importance. Increasing the number of interactions increased computer time needed substantially, with marginal improvement in model fit. In addition, the model's prediction under future climate was highly distorted. We therefore constrained the number of interactions to two to strike a balance between fit and prediction. We chose 100 as the number of basis functions after several trials because this value proved adequate for the number of predictors and interactions.

Correlation

We chose not to use individual model error estimates when comparing models because they are calculated differently for each technique. Instead we use correlation of FIA IVs to the current predicted IVs using Pearson's correlation test. In addition, we decided not to use the out-of-bag predictions even though they tend to be more ''honest'' for depicting the correlation (the non–out-of-bag predictions use the models built from the training data to predict the training data). Our decision was based on the fact that there is no equivalent of out-of-bag in RTA and MARS. Because our primary goal was to compare model predictions, we decided to keep the comparison at the same level to show how BT and RF perform compared to RTA and MARS. As a check, we also compared the out-of-bag correlation of BT and RF with the non–out-of-bag correlation of RTA and MARS; in this case, BT and RF values were at least as high, and mostly higher, than RTA or MARS.

Variable Importance

The variables predicted to be important in the model help us to understand what variables are driving the distribution of species. Some species distributions are strongly driven by climate, whereas others are driven primarily by edaphic or land-use variables. In addition, variable importance enables us to determine what set of variables is deemed important for each of the four models and to compare them to see whether the sets are similar.

For RTA, importance of a variable is simply the total reduction in sum of squares achieved by all splits on that variable. Variable importance in BT is derived similarly because it uses the RTA package ''rpart'' to grow the individual trees. However, in BT, the importance values for all the 50 trees are averaged to obtain an overall measure of the variable importance. RF has two measures of variable importance. The first is based on mean squared error (MSE) and relates to the prediction accuracy of the out-of-bag portion of the data after permuting each predictor variable. The difference between the two MSEs are then averaged over all trees and normalized by the standard error. The second measure is the same as that for BT and is computed on the data used to grow the trees. Thus the conclusion is based on overfitted models. Here we report the second metric to enable a fair comparison among RTA, BT, RF, and MARS. The variable importance list between the two methods differed little, particularly among the first six predictors. We normalized the variable importance measures of RTA, BT, and RF to 0–100 to facilitate comparison among models. MARS calculates variable importance scores by refitting the model after dropping all terms involving the variable in question and calculating the reduction in goodness-of-fit and normalizing the results. The least important variable is the one with the smallest impact on the model quality; similarly, the most important variable is the one that, when omitted, degrades the model fit the most (Steinberg and others 1999).

Data Distribution

A prominent feature of the data is the distribution of IV. It is typically right-skewed with many zeroes.

Nonconstant error variance usually is a ''symptom'' of such data; the usual remedy is transformation;—for example, square root or log. Because of the large number of zeroes in IV, we also built two models as part of a hybrid approach. The first is a classification model that predicts whether or not the data point has IV equal to 0. If not, it is given to the second regression model, which gives a numerical prediction for the IV. We also examined combinations of the hybrid with transformations. Although these models gave superior results for the current distribution compared to our original untransformed, nonhybrid approach, the future distributions tended to be biogeographically unrealistic. We concluded that although the transformation-hybrid approach is highly suitable for predicting the current scenario and for certain other applications, the original approach is more appropriate for predicting future habitat distributions. Also, the errors in the untransformed, nonhybrid approach were mostly model-prediction artifacts where the predictions between $0 < IV < 1$ are in reality $IV = 0$. As a result, we rounded these values to zero to obtain satisfactory current predictions and eliminate the need for more complicated hybrid models.

Model Limitations

An obstacle in our modeling environment is that the response variable distribution may not be comprehensive because we are relying solely on FIA data to capture the current spatial distribution of the species. Although there are more than 100,000 forested plots in our study area, there remain spatial gaps in the FIA data. This sampling intensity is especially problematic in sparsely forested areas with few plots per 20-km cell and in areas where the environmental conditions are particularly heterogeneous within the 20-km cell. Further, of the four species, only the range of loblolly pine does not extend into Canada (mainly the province of Ontario). FIA data stops at the border, so we are not modeling the full range of the species distribution. Range maps (Little 1971, 1977; Prasad and Iverson 2003) superimposed on actual FIA distributions indicate the current range of the species (Figures 1, 2, 3, 4). However, we believe that a preferred modeling technique might predict reasonably well in the gaps based on the existing strength of the relationships between the response and the predictors. Therefore, it should be noted that any agreement between the actual and predicted, although still a strong measure of model predictive ability, might not necessarily be a comprehensive test of the superiority of the prediction.

Results

We compare the four techniques by assessing the outputs of the four species: correlation, Kappa and its variants, variable importance, and the output maps. Each is presented separately, with distinctions among the four species also noted.

Correlation

It is apparent looking at predicted correlations (Table 3) that for all four species, there is a clear distinction between RTA and MARS versus BT and RF, with the latter pair showing much better correlation. Actual versus predicted distributions of loblolly pine have correlations of at least 0.85 for all models, whereas the other species, especially white oak, have lower correlations for RTA and MARS. RF also has a slight edge over BT with all the correlations.

Kappa

The distinction between RTA and MARS versus BT and RF is also apparent with all variants of the Kappa statistics (Table 4), with BT and RF showing much better conformity with the actual values for all four species. Among the four Kappa variants, BT produces the best match of histograms (K_{hist}) for all species except American beech, whereas RF produces the best match of pixel similarity (K_{loc}) for all four species. Thus, RF better preserves locational similarity, whereas BT better preserves categorical similarity. When both are considered in the overall Kappa statistic, RF has slightly higher values than BT, as it does for fuzzy Kappa.

Variable Importance

As expected, the predictors deemed important by the models are different for the four species (Table 5). Each of the species is predominantly climate-driven to some degree. For loblolly pine, all models use potential soil productivity (ORD) as the most important variable in predicting IV. RF and BT agree on the ranking of the first four variables. After agreeing with other models on the importance of ORD and mean January temperature (JANT), MARS departs from the rest.

For sugar maple, all four models agree that mean July temperature (JULT) is the most important variable, with ORD and precipitation (PPT) also important. Four of the first five variables in RF are climate-related, indicating that sugar maple is primarily climate-driven.

American beech has percent of land use in agriculture (AGRICULT) as the most important vari-

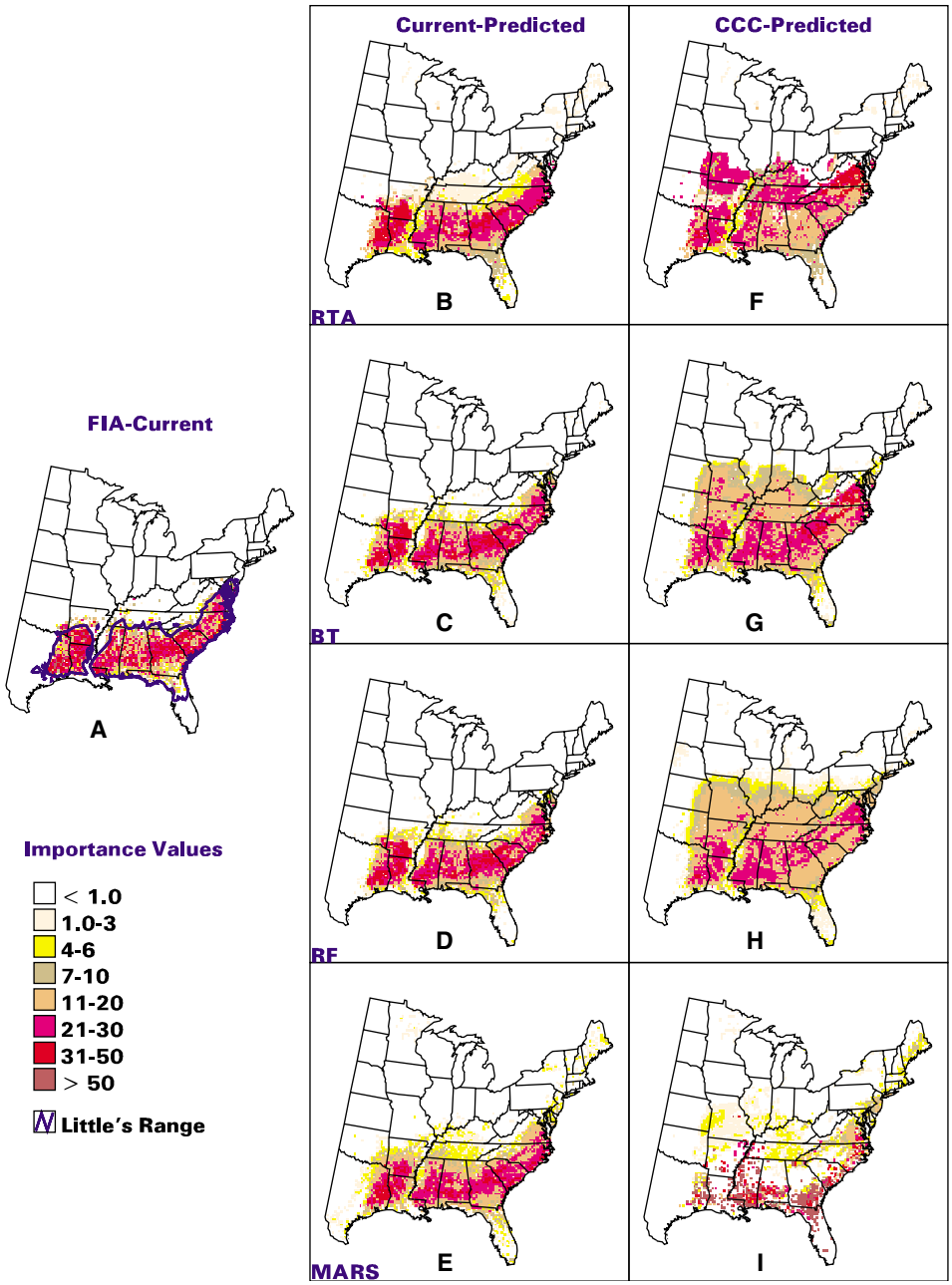

Figure 1. Loblolly pine. **A** Current distribution according to Forest Inventory and Analysis (FIA) and Little (1971) boundaries. **B–E** Predictions of current distribution according to the four models. **F–I** Predictions of potential future suitable habitat according to the four models. *RTA*, Regression Tree Analysis; *BT*, Bagging Trees; *RF*, Random Forest; *MARS*, Multivariate Adaptive Regression Splines.

able, followed closely by JULT in all models except MARS. MARS elevates PPT to the first position followed by minimum elevation (ELV_MIN). RF and BT agree for the first five variables, except that BT places more importance to ELV_MIN versus percent forest (FOREST). Topographic variables that seem to drive American beech in the Appalachians are important in all four models.

The RF and BT models for white oak are similar for four of the top five variables. BT puts heavier weight on FOREST, whereas mean growing season temperature (MAYSEPT) is weighted more heavily in RF. The high-ranking variables for RF are much closer in weight as compared to the other species, indicating that all of these variables could be similarly important in predicting distribution; this characteristic may be an important indicator of a generalist species.

Map of Current Distributions

The maps of the four species (Figures 1, 2, 3, 4) reflect the correlations of Table 3 and the Kappa statistics of Table 4. For loblolly pine (Figure 1), it is clear that compared to the actual FIA distribution, BT-Current and RF-Current match better than RTA

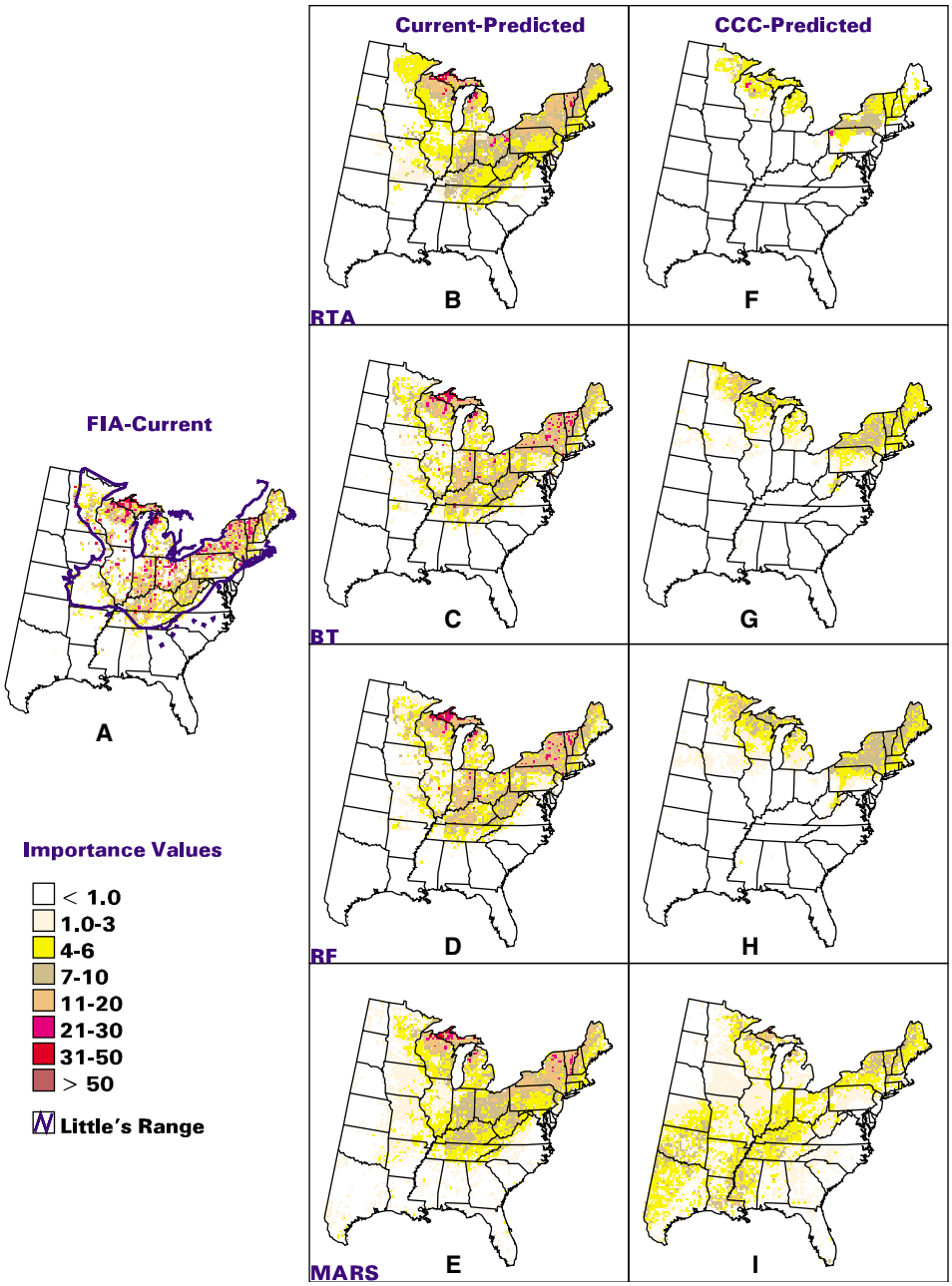

Figure 2. Sugar maple. **A** Current distribution according to FIA and Little (1971) boundaries. **B–E** Predictions of current distribution according to the four models. **F–I** Predictions of potential future suitable habitat according to the four models. *RTA*, Regression Tree Analysis; *BT*, Bagging Trees; *RF*, Random Forest; *MARS*, Multivariate Adaptive Regression Splines.

and MARS, which show false presence (albeit low IVs) in Tennessee and Kentucky and in the northeastern and north-central states. BT-Current shows a better match, although there are some low values in the northeastern states.

The current distribution of sugar maple is captured in all models except MARS, which shows an anomalous southwestward distribution as well as a large drop in abundance in the central states (Figure 2). BT-Current and RF-Current are similar, whereas RTA-Current shows slightly more erroneous distributions in the central portion of the study area.

There is a large difference between RTA and MARS versus BT and RF for American beech. Visually, BT-Current and RF-Current are similar, although RF shows more smoothing between classes (Figure 3). RTA fails to show presence in the south. MARS shows an anomalous westward spread to Missouri.

For white oak, the predictions of current abundance are well matched for BT and RF and are similar, except that RF smoothes the abundances slightly more in some areas (Figure 4). The RTA and MARS models do not do as well because they

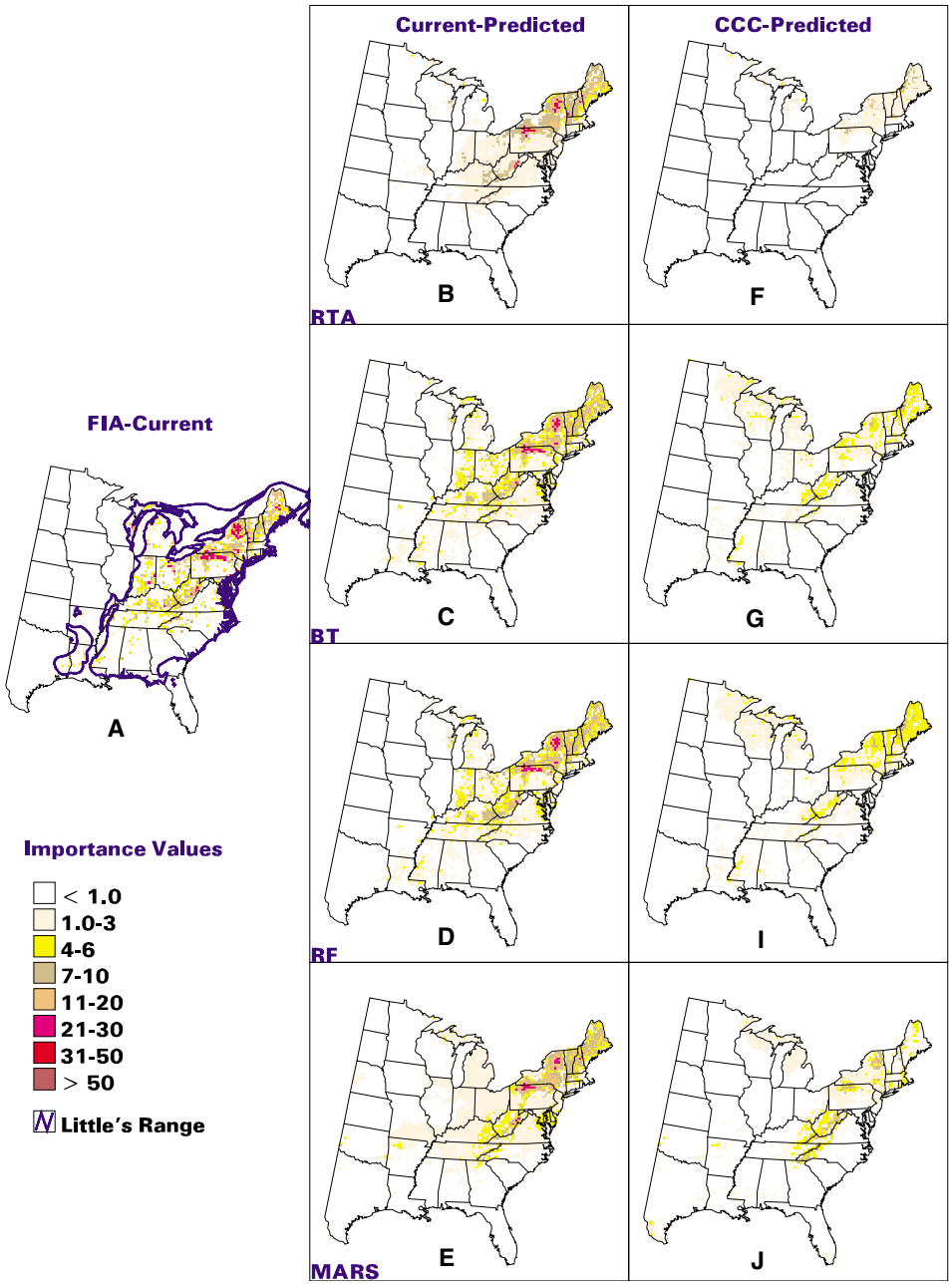

Figure 3. American beech. **A** Current distribution according to FIA and Little (1971) boundaries. **B–E** Predictions of current distribution according to the four models. **F–I** Predictions of potential future suitable habitat according to the four models. *RTA*, Regression Tree Analysis; *BT*, Bagging Trees; *RF*, Random Forest; *MARS*, Multivariate Adaptive Regression Splines.

show a westward presence that is not visible even on Little's range map.

Maps under Future CCC Scenario

It is important to ask how reasonable the maps of future climate scenarios are because one of the main goals in predictive vegetation mapping is to assess the performance of models when predicting under changed climatic conditions. Obviously, it is impossible to reliably validate such models, but in the following discussion, we will attempt to assess the reasonableness of such models based on biological expectations.

One striking feature that emerges when comparing among models for all four species is that, for the MARS model, the future climate predictions are counter to biogeographical expectations (Figures1, 2, 3, 4). The reason for this erratic behavior is addressed in the Discussion section.

Loblolly pine shows an expansion of potential suitable habitat northward in the RTA and especially the BT-CCC and RF-CCC models (Figure1). Compared to BT-CCC, RF-CCC pushes the species habitat farther north (mainly in Ohio) and shows a smoother grading of classes as the suitable habitat spreads northward. The abundance suitability

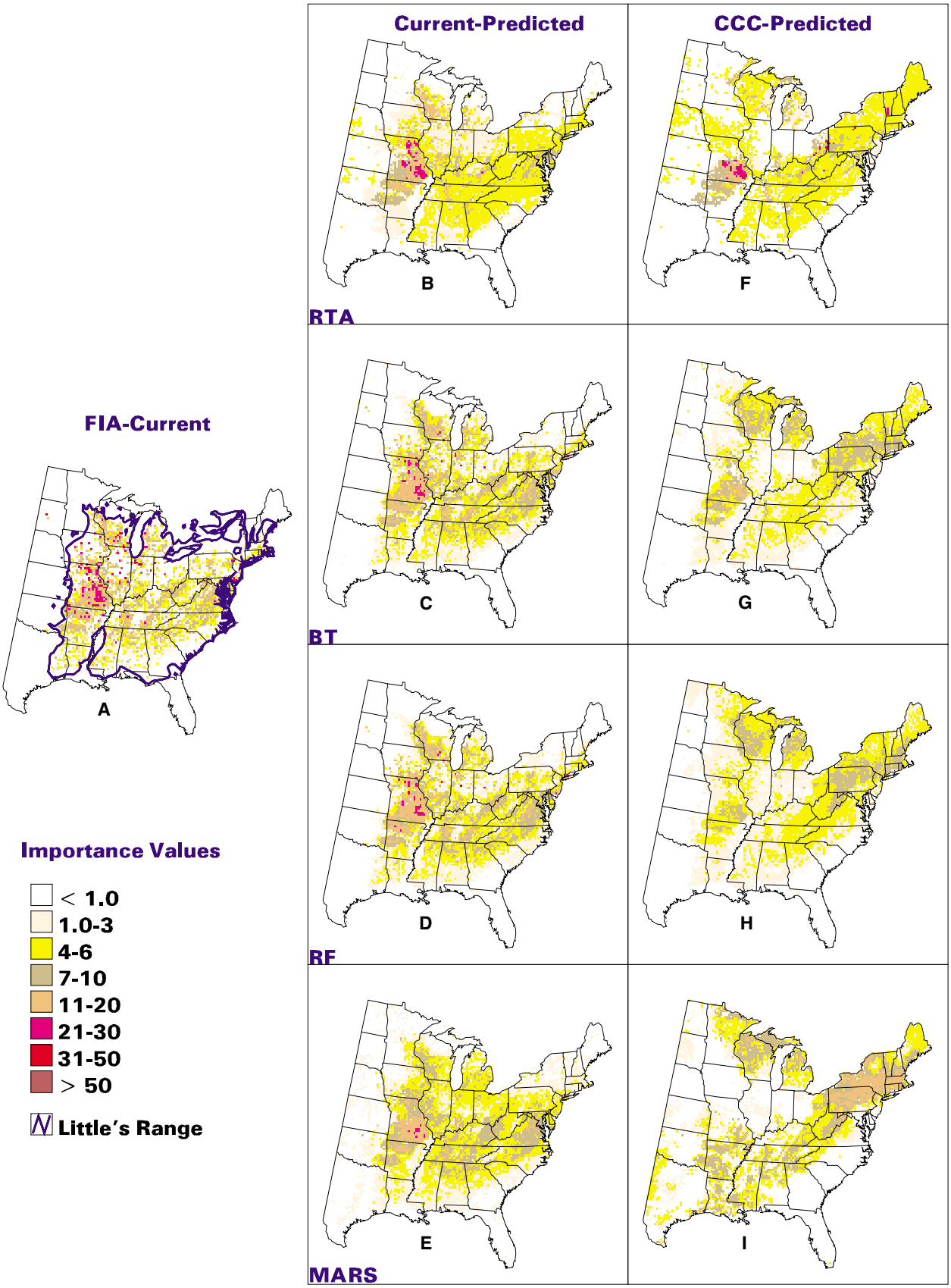

Figure 4. White oak. **A** Current distribution according to FIA and Little (1971) boundaries. **B–E** Predictions of current distribution according to the four models. **F–I** Predictions of potential future suitable habitat according to the four models. *RTA*, Regression Tree Analysis; *BT*, Bagging Trees; *RF*, Random Forest; *MARS*, Multivariate Adaptive Regression Splines.

Table 3. Pearson's Correlation Coefficients between Current FIA and Current Model Predictions

Species	RTA	BT	RF	MARS
Loblolly Pine	0.85	0.96	0.97	0.86
Sugar Maple	0.70	0.92	0.93	0.69
American Beech	0.75	0.92	0.93	0.71
White Oak	0.66	0.90	0.92	0.62

RTA, Regression Tree Analysis; BT, Bagging Trees; RF, Random Forest; MARS, Multivariate Adaptive Regression Splines; FIA, Forest Inventory and Analysis. All correlations are highly significant, with P value < 2.2e−16 for n = 9,782.

Table 4. Kappa Statistics Comparing the Agreement of Current Importance Values Based on FIA and Model Predictions

	RTA	BT	RF	MARS
Loblolly Pine				
Kappa	0.448	0.664	0.676	0.422
K_{loc}	0.648	0.785	0.806	0.607
K_{hist}	0.692	0.845	0.838	0.695
Fuzzy Kappa	0.497	0.717	0.725	0.478
Sugar Maple				
Kappa	0.395	0.593	0.598	0.309
K_{loc}	0.533	0.705	0.718	0.487
K_{hist}	0.741	0.841	0.832	0.635
Fuzzy Kappa	0.393	0.638	0.643	0.349
American Beech				
Kappa	0.378	0.650	0.668	0.365
K_{loc}	0.469	0.686	0.696	0.421
K_{hist}	0.807	0.947	0.960	0.867
Fuzzy Kappa	0.413	0.677	0.692	0.399
White Oak				
Kappa	0.354	0.623	0.630	0.333
K_{loc}	0.438	0.711	0.726	0.462
K_{hist}	0.808	0.877	0.867	0.720
Fuzzy Kappa	0.352	0.652	0.656	0.342

RTA, Regression Tree Analysis; BT, Bagging Trees; RF, Random Forest; MARS, Multivariate Adaptive Regression Splines; FIA, Forest Inventory and Analysis.

gradually decreases northward. This smoothing feature may give RF an advantage over BT for predicting reasonable future climate scenarios. MARS creates an unrealistic picture in that large values appear in the south while low values appear in the north.

The CCC scenario for sugar maple is strikingly similar for BT and RF; RTA shows more thinning (reduction of importance value), whereas MARS is highly distorted with wide expansions to the center-west, southwest, and south (Figure 2). Both BT and RF retain potential future sugar maple habitat more than the RTA model presented here and especially more than the RTA (DISTRIB) outputs from previous work (Iverson and Prasad 1998; Iverson and others 1999a; Prasad and Iverson 2000a). As mentioned, these differences result from updated response and predictor variables, as well as the refined spatial resolution of this work.

As with sugar maple, the RTA-CCC map for American beech shows a vast reduction in potential future habitat within the United States, whereas RF-CCC and BT-CCC retain more habitat and are similar (Figure 3). With RF-CCC, there is a smoother transition of classes (rather than ''jumping,'' for example, from an IV of 1–3 adjacent to an IV of 11–20), giving that prediction an edge over the other models. MARS-CCC does show a bioge-

Table 5. Variable Importance (Top 12 Variables) Predicted by the Four Models for Loblolly Pine, Sugar Maple, American Beech, and White Oak

Predictors	RTA (Rank)	RF (Rank)	BT (Rank)	MARS (Rank)
Loblolly Pine				
ORD	100 (1)	100 (1)	100 (1)	100 (1)
JANT	18 (2)	74 (2)	27 (2)	56 (2)
PPT	8 (5)	37 (3)	14 (3)	41 (5)
ULTISOL	16 (3)	29 (4)	12 (4)	—
AVGT	9 (4)	26 (5)	6 (7)	16 (9)
ELV_MEAN	—	18 (6)	—	54 (3)
FOREST	4 (7)	16 (7)	7 (5)	1 (12)
PH	—	14 (8)	—	—
AGRICULT	6 (6)	13 (9)	6 (6)	43 (4)
ELV_MAX	—	13 (10)	4 (11)	—
NO200	2 (11)	11 (11)	4 (10)	17 (8)
CLAY	2 (10)	10 (12)	5 (9)	—
ELV_MIN	3 (9)	—	5 (8)	—
ELV_RANGE	1 (12)	—	4 (12)	—
OM	3 (8)	—	—	19 (6)
SLOPE	—	—	—	18 (7)
JULT	—	—	—	15 (10)
ELV_CV	—	—	—	14 (11)
Sugar Maple				
JULT	100 (1)	100 (1)	100 (1)	100 (1)
ORD	31 (3)	63 (2)	39 (3)	44 (4)
PPT	32 (2)	58 (3)	47 (2)	62 (3)
MAYSEPT	4 (7)	58 (4)	13 (10)	1 (11)
JANT	2 (11)	38 (5)	—	1 (12)
NO10	7 (5)	34 (6)	14 (6)	—
ELV_MEAN	6 (6)	30 (7)	11 (11)	—
ELV_CV	—	30 (8)	16 (4)	37 (6)
AVGT	—	29 (9)	—	31 (9)
ELV_RANGE	2 (12)	27 (10)	13 (9)	—
ELV_MAX	—	27 (11)	14 (8)	42 (5)
ELV_MIN	—	25 (12)	16 (5)	—
ALFISOL	—	—	14 (7)	—
FOREST	3 (10)	—	8 (12)	—
KFFACT	8 (4)	—	—	71 (2)
NO200	—	—	—	36 (7)
SPODOSOL	—	—	—	35 (8)
SLOPE	—	—	—	27 (10)
CLAY	3 (8)	—	—	—
BD	3 (9)	—	—	—
American Beech				
AGRICULT	100 (1)	100 (1)	100 (1)	55 (6)
JULT	54 (2)	58 (2)	43 (2)	79 (3)
FOREST	12 (5)	50 (3)	16 (7)	42 (10)
PPT	6 (10)	47 (4)	22 (5)	100 (1)
ELV_MAX	9 (8)	41 (5)	23 (4)	1 (11)
MAYSEPT	—	36 (6)	14 (9)	59 (5)
ELV_MIN	24 (3)	35 (7)	27 (3)	99 (2)
ELV_RANGE	—	34 (8)	17 (6)	—
ELV_MEAN	15 (4)	31 (9)	13 (12)	44 (8)
JANT	5 (12)	29 (10)	13 (11)	43 (9)
ELV_CV	—	28 (11)	15 (8)	1 (12)

(continued)

Table 5. Continued

Predictors	RTA (Rank)	RF (Rank)	BT (Rank)	MARS (Rank)
AVGT	12 (6)	27 (12)	13 (10)	62 (4)
ORD	—	—	—	47 (7)
INCEPTIS	9 (7)	—	—	—
CLAY	8 (9)	—	—	—
White Oak				
SLOPE	100 (1)	100 (1)	100 (1)	1 (11)
PPT	39 (4)	98 (2)	62 (2)	55 (6)
ORD	23 (7)	84 (3)	44 (4)	62 (4)
MAYSEPT	18 (10)	73 (4)	34 (11)	79 (3)
AVGT	46 (2)	72 (5)	47 (3)	42 (10)
JULT	41 (3)	66 (6)	40 (7)	100 (1)
PH	—	64 (7)	—	43 (9)
AGRICULT	17 (11)	61 (8)	39 (8)	—
FOREST	20 (9)	59 (9)	41 (5)	99 (2)
NO10	38 (5)	57 (10)	40 (6)	47 (7)
ELV_MEAN	—	56 (11)	—	—
ELV_MIN	—	56 (12)	38 (9)	—
ALFISOL	36 (6)	—	35 (10)	—
ELV_RANGE	—	—	32 (12)	—
JANT	22 (8)	—	—	59 (5)
MOLLISOL	—	—	—	44 (8)
ELV_MAX	—	—	—	1 (12)
KFFACT	15 (12)	—	—	—

RTA, Regression Tree Analysis; BT, Bagging Trees; RF, Random Forest; MARS, Multivariate Adaptive Regression Splines. Variables are listed in Table 2.

ographically reasonable future output, although the presence of American beech in Texas and Florida is unexpected.

The RTA, BT, and RF models for white oak are similar in predicting a northward expansion of suitable habitat (Figure 4). However, RTA retains higher values in the south as well as in the Missouri Ozarks. The BT-CCC and RF-CCC differ for this species in that RF tends to produce reduced levels of IV in the Missouri Ozarks and in other areas scattered throughout the range of white oak. The MARS map is not biogeographically reasonable because it predicts wildly, with a large area of high values in the northeast and a large westward presence.

Discussion

Model Comparisons

It is clear from the maps and statistics presented here that BT and RF have a distinct advantage over MARS and RTA in predictive mapping. BT and RF are similar and both are more effective than single regression tree outputs. The ''multiple-perturbed'' trees in BT and ''multiple-perturbed-randomized'' trees in RF had better predictive capabilities. Although the error rates of BT and RF were similar in our analysis of four tree species, RF proved superior for this type of application because it provides a smoother response surface in that the IVs grade smoothly from lower to higher values and there is no jumping of classes. Because species IVs often are highly variable between nearby FIA plots due to local variations in environment or land-use history, the smoother output that RF generates minimizes this influence and hence is appropriate in regional models.

We realize that evaluating whether a particular species distribution is biogeographically realistic under future climate is primarily a subjective process because numerous factors can influence the final distribution. Here, in addition to our own limited insights into the nature of species distributions, we have to place more confidence in RF than in BT as evaluated by other studies in which the two models are compared (Svetnik and others 2003; Hawkins and Musser 1999; Meyer and others 2003).

Erratic Behavior of MARS

Multivariate Adaptive Regression Splines performed unrealistically under future climate for all four species, with major distortions for loblolly pine, sugar maple, and white oak (Figures 1G, 2G, 4G). In our earlier effort to compare the RTA and MARS models (Prasad and Iverson 2000b), we concluded that MARS was superior for predicting current tree distributions. This conclusion resulted from using a high number of interactions in the MARS software until we achieved a better fit to current distributions. However, we did not evaluate the impact of such a model for predicting the distribution of suitable habitat under future climate. For our current effort, we discovered that increasing the number of interactions highly distorts the model of future distribution. This was true even when we constrained the number of interactions to two. The need to constrain the number of interactions to two is the reason why RTA outperformed MARS in this study.

The reason for these ''wild'' predictions under changed climate is that MARS is highly sensitive to extrapolation caused by the local nature of the basis functions. A change in the predictor value toward the end of its range can cause the prediction to go largely off scale (M. Golovnya personal communication). For certain applications, a strength of MARS is that its basis functions are guided by the local nature of the data; in our case, this proved a major disadvantage for predictions under future climate scenarios. Attempts to control such wild behavior in MARS involves tweaking the basis functions and creating a hybrid RTA–MARS model, which defeats our original purpose of making the model intelligently automatic so that we need not fine-tune the model individually for multiple species. Thus, it appears that tree-based models are better when we want to investigate how species habitat could change in the future, given the level of the current response.

Species Comparisons

It is useful to speculate why the models performed better for some species than for others because we want to know which models to use for the numerous tree species in the eastern United States. Loblolly pine likely had the best-fit models (based on correlation, Kappa, and visual comparison) because (a) it has the highest average IVs of all species (Iverson and others 1999a) and is distributed evenly (b) it has a northern limit well within the United States, so absolute range boundaries can be assessed and (c) its northern range matches temperature isotherms in the eastern United States. American beech and sugar maple, which had similar distributions, both border Canada, so the northern limits of their range are not examined fully by the models. White oak produced the least satisfactory model for RTA and MARS, whereas BT and RF showed surprisingly strong predictions. White oak is a generalist species with a scattered distribution of varying abundance across much of its range, making it difficult to model. It is also common in woodlots across the heavily cultivated Midwest and thus contained in small forest patches that are easily missed in FIA sampling. The strong performance of BT and RF for this species is encouraging. Future work will determine whether these trends are consistent with other generalist species.

Interpretation of Models

Although BT and RF are ensemble methods based on regression trees (RTA), they become more of a black (or gray) box when interpreting the model due to the sheer number of trees generated. Because the main purpose of our effort is to develop a superior prediction model, this does not necessarily pose a problem. However, adding interpretability to the black boxes is worthwhile and gives us additional insight into the nature of species distributions. Our results show that RF can be used for the predictions, with interpretations primarily from RTA and BT. Because RTA is just one model-slice of the data and BT is a 50-tree average, we can examine the variation in the 50 trees by computing a statistical summary of the deviances as well as the variation in the variable importance table among the 50 trees. If there are wide differences in the splitting rules among trees, we know that the model is unstable for that species and bootstrap-resampling and averaging (BT) would predict better than a single tree (RTA). For such species, RF would predict even better by the randomization of predictors and the sheer number of trees grown (1,000). However, if the individual trees are similar, a single RTA tree can be used to map what predictors are driving the distribution of the species spatially. This geographic mapping of predictors is a unique aspect of RTA that offers additional insights into species distribution (Iverson and Prasad 1998; Iverson and others 1999a; Prasad and Iverson 2000a).

Conclusions

In conclusion, the RTA, BT, and RF techniques can be used in combination because they provide both a means to accurately map organism distributions

and a mechanism that provides a better understanding of the drivers of current and potential future distributions. The superior predictive capability of RF can be used to map current distributions and potential future suitable habitat, whereas RTA, and to some extent BT, provide interpretive results. We are using this multiple model procedure to interpret and predict the distributions of 135 tree species in the eastern United States under multiple scenarios of future climate conditions.

The potential applications of these methods are numerous. They can be used to classify landscapes into categories, such as vegetation types or land-use classes. They can also be used to identify target locations or probability surfaces for common, rare, or invasive species, whether plant or animal. They can be extended to map pollution levels or nutrient concentrations across a landscape. They will also become valuable in remote sensing studies using customized software.

Essentially, these techniques can be used to extrapolate any response variable collected at sample locations across the landscape and to understand what predictors are driving the distribution with a higher level of confidence than with other methods. We therefore highly recommend this package of statistical modeling tools for use in predictive biological mapping.

REFERENCES

Abraham A, Steinberg D. 2001. MARS: Still an alien planet in soft computing? In: Alexandrov VN, Dongarra JJ, Juliano BA, Renner RS, Tan CJK, Eds. Lecture notes in computer science 2074. Berlin Heidelberg New York: Springer, p 235–244.

Baker FA. 1993. Classification and regression tree analysis for assessing hazard of pine mortality caused by *Heterobasidion annosum*. Plant Dis 77:136–9.

Boer GJ, Flato GM, Ramsden D. 2000. A transient climate change simulation with historical and projected greenhouse gas and aerosol forcing: projected climate for the 21st century. Clim Dyn 16:427–51.

Breiman L. 1996a. Bagging predictors. Mach Learn 24:123–40.

Breiman L. 1996b. Out-of-bag estimation. Technical report, Department of Statistics: University of California, Berkeley.

Breiman L. 2001. Random Forests. Mach Learn 45:5–32.

Breiman L (2002) Using models to infer mechanisms. IMS Wald Lecture 2. [online] URL: http://oz.berkeley.edu/users/breiman/wald2002-2.pdf.

Breiman L, Cutler A. 2003. Setting up, using, and understanding Random Forests v4.0. [online] URL: http://www.stat.berkeley.edu/users/breiman/rf.html.

Breiman L, Freidman J, Olshen R, Stone C. 1984. Classification and regression trees. Belmont (CA): Wadsworth, 358 p.

Buhlmann P, Yu B. 2002. Analyzing bagging. Ann Stat 30:927–61.

Chambers JM. 1998. Programming with data: a guide to the S language. Berlin Heidelberg New York: Springer, 469 p.

Chambers JM, Hastie TJ (1993) Statistical models in New York; S. Chapman & Hall, 608 p.

Chan JCW, Huang C, DeFries R. 2001. Enhanced algorithm performance for land cover classification using bootstrap aggregating (bagging). IEEE Trans Geosci Remote Sens 39(3):693–5.

Clark JS. 1998. Why trees migrate so fast: confronting theory with dispersal biology and the paleorecord. Am Nat 152:204–24.

Clark LA, Pregibon D. 1992. Tree-based models. In: Chambers JM, Hastie TJ, Eds. Statistical models in S. Pacific Grove (CA): Wadsworth, p 377–419.

Davis MB. 1989. Lags in vegetation response to greenhouse warming. Clim Change 15:75–82.

De'ath G, Fabricius KE. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81:3178–192.

Dobbertin M, Biging GS. 1998. Using the non-parametric classifier CART to model forest tree mortality. For Sci 44(4):507–16.

Environmental Systems Research Institute. 2001. Arc ver. 8.1.2. Environmental Systems Research Institute, Redlands (CA).

Franklin J. 1995. Predictive vegetation mapping: geographic modeling of biospatial patterns in relation to environmental gradients. Prog Phys Geogr 19:494–519.

Franklin J. 1998. Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. J Veg Sci 9:733–48.

Freidman JH. 1991. Multivariate adaptive regression splines. Ann Stat 19:1–141.

Freund Y. 1995. Boosting a weak learning algorithm by majority. Inf Comput 121:256–85.

Furlanello C, Neteler M, Merler S, Menegon S, Fontanari S, Donini A, Rizzoli A, Chemini C. 2003. GIS and the Random Forests predictor: integration in R for tick-borne disease risk assessment. In: Hornik K, Leisch F, Zeileis A, Eds. Proceedings of the 3rd international workshop on distributed statistical computing. Vienna, Austria, p 1–11.

Hagen A. 2002. Technical report: comparison of maps containing nominal data. RIVM project: MAP-SOR S/550002/01/RO, order no. 143699. Maastricht (The Netherlands): Research Institute for Knowledge Systems.

Hagen A. 2003. Fuzzy set approach to assessing similarity of categorical maps. Int J Geog Inf Sci 17(3):235–49.

Hansen M, Dubayah R, Defries R. 1996. Classification trees: an alternative to traditional land cover classifiers. Int J Remote Sens 17(5):1075–81.

Hansen MH, Frieswyk T, Glover JF, Kelly JF (1992) The East-wide forest inventory data base: users manual. General technical report NC-151. St. Paul (MM): US Department of Agriculture, Forest Service, North Central Forest Experiment Station, 48 p.

Hawkins DM, Musser BJ. 1999. One tree or a forest? Alternative dendrographic models. Comput Sci Stat 30:534–42.

Hernandez JE, Epstein LD, Rodriguez MH, Rodriguez AD, Rejmankova E, Roberts DR. 1997. Use of generalized regression tree models to characterize vegetation favoring *Anopheles albimanus* breeding. J Am Mosq Control Assoc 13(1):28–34.

Higgins SI, Lavorel S, Revilla EE. 2003. Estimating plant migration rates under habitat loss and fragmentation. Oikos 101:354–66.

Hobbs RJ. 1994. Dynamics of vegetation mosaics: can we predict responses to global change?. Ecoscience 1(4):346–56.

Hothorn T, Lausen B, Benner A, Radespiel-Troger M. 2004. Bagging survival trees. Stat Med 23:77–91.

Iverson LR, Prasad AM. 1998. Predicting abundance of 80 tree species following climate change in the eastern United States. Ecol Mono 68:465–85.

Iverson LR, Prasad AM. 2002. Potential redistribution of tree species habitat under five climate change scenarios in the eastern US. For Ecol Manage 155(1–3):205–22.

Iverson LR, Prasad AM, Hale BJ, Sutherland EK 1999a. An atlas of current and potential future distributions of common trees of the eastern United States. General technical report NE-265. Northeastern Research Station, USDA Forest Service, 245 p.

Iverson LR, Prasad AM, Schwartz MW. 1999b. Modeling potential future individual tree-species distributions in the Eastern United States under a climate change scenario: a case study with Pinus virginiana. Ecol Mod 115:77–93.

Kittel TGF, Rosenbloom NA, Kaufman C, Royle JA, Daly C, Fisher HH, and others. 2000. VEMAP phase 2 historical and future scenario climate database. Oak Ridge (TN): ORNL Distributed Active Archive Center, Oak Ridge National Laboratory. [online] URL: http://www.daac.ornl.gov/.

Lees BG, Ritman K. 1991. Decision-tree and rule-induction approach to integration of remotely sensed and GIS data in mapping vegetation in disturbed or hilly environments. Environ Manage 15:823–31.

Liaw A, Wiener M. 2002. Classification and regression by Random Forests. R News, 2/3:18–22. [online] URL http://CRAN.R-project.org/doc/Rnews/.

Little EL. 1971. Atlas of United States trees; vol 1. Conifers and important hardwoods. Miscellaneous publication 1146. Washington (DC), US Department of Agriculture, Forest Service, 200 p.

Little EL. 1977. Atlas of United States Trees; vol 4. Minor eastern hardwoods. Miscellaneous publication 1342. Washington (DC): US Department of Agriculture, Forest Service, 230 p.

Malcolm JR, Markham A, Neilson RP, Garaci M. 2002. Estimated migration rates under scenarios of global climate change. J Biogeogr 29:835–49.

Map Comparison Kit. 2003. Research Institute for Knowledge Systems, Netherlands. http://www.riks.nl.

Meyer D, Leisch F, Hornik K. 2003. The support vector machine under test. Neurocomputing 55:59–71.

Michaelsen J, Schimel DS, Friedl MA, Davis FW, Dubayah RC. 1994. Regression tree analysis of satellite and terrain data to guide vegetation sampling and surveys. J Veg Sci 5:673–86.

Miller JR, Turner MG, Smithwick EAH, Dent CL, Stanley EH. 2004. Spatial extrapolation: the science of predicting ecological patterns and processes. BioScience 54(4):310–20.

Moisen GG, Frescino T. 2002. Comparing five modelling techniques for predicting forest characteristics. Ecol Model 157:209–25.

Monserud RA, Leemans R. 1992. Comparing global vegetation maps with the Kappa statistic. Ecol Model 62:275–93.

Moore DE, Lees BG, Davey SM. 1991. A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. J Environ Manage 15:59–71.

Munoz J, Felicisimo AM. 2004. Comparison of statistical methods commonly used in predictive modelling. J Veg Sci 15:285–92.

Peters A, Hothorn T, Lausen B. 2002. ipred: Improved predictors. R News, 2(2):22–6 [online] URL http://CRAN.R-project.org/doc/Rnews/.

Pitelka LF, Plant Migration Workshop Group. 1997. Plant migration and climate change. Am Sci 85:464–73.

Pontius RG Jr. 2000. Quantification error versus location error in comparison of categorical maps. Photogram Eng Remote Sens 66(8):1011–16.

Power C, Simms A. 2001. Hierarchical fuzzy pattern matching for regional comparison of land use maps. Int J Geogr Inf Sci 15(1):77–100.

Prasad AM, Iverson LR. 2000a. A climate change atlas for 80 forest tree species of the eastern United States [database]. [online] URL: http://www.fs.fed.us/ne/delaware/atlas.

Prasad AM, Iverson LR. 2000b. Predictive vegetation mapping using a custom built model-chooser: comparison of regression tree analysis and multivariate adaptive regression splines. In: Proceedings CD-ROM. 4th International Conference on Integrating GIS and Environmental Modeling: Problems, Prospects and Research Needs. Banff, Alberta, Canada. [online] URL: http://www.colorado.edu/research/cires/banff/upload/159/index.html.

Prasad AM, Iverson LR. 2003. Little's range and FIA importance value database for 135 eastern US tree species. Northeastern Research Station, USDA Forest Service, Delaware, Ohio. [online] URL: http://www.fs.fed.us/ne/delaware/4153/global/littlefia/index.html.

R Development Core Team. 2004. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. [online] URL: http://www.R-project.org.

Reichard SH, Hamilton CW. 1997. Predicting invasion of woody plants introduced into North America. Conserv Biol 11:193–203.

Schapire RE, Freund Y, Barlett P, Lee W. 1998. Boosting the margin: a new explanation for the effectiveness of voting methods. Ann Stat 26(5):1651–86.

Schwartz MW, Iverson LR, Prasad AM. 2001. Predicting the potential future distribution of four tree species in Ohio, USA, using current habitat availability and climatic forcing. Ecosystems 4:568–81.

Skurichina M, Duin RPW. 2002. Bagging, boosting and the random subspace method for linear classifiers. Pattern Anal Appl 5:121–35.

Steinberg D, Colla PL, Martin K. 1999. MARS user guide. San Diego (CA): Salford Systems.

Stoppiana D, Gregoire J-M, Pereira JMC. 2003. The use of SPOT VEGETATION data in a classification tree appproach for burnt area mapping in Australian savanna. Int J Remote Sens 24:2131–51.

Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. 2003. Random Forests: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 43(6):1947–58.

Therneau TM, Atkinson EJ. 1997. An introduction to recursive partitioning using the RPART routines. Technical report no. 61. Rochester (MM): Mayo Clinic, 52 p.

Verbyla DL. 1987. Classification trees: a new discrimination tool. Can J For Res 17:1150–2.