Ecology, 88(11), 2007, pp. 2783–2792

RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY

D. Richard Cutler,[1,7] Thomas C. Edwards, Jr.,[2] Karen H. Beard,[3] Adele Cutler,[4] Kyle T. Hess,[4] Jacob Gibson,[5] and Joshua J. Lawler[6]

[1]*Department of Mathematics and Statistics, Utah State University, Logan, Utah 84322-3900 USA*
[2]*U.S. Geological Survey, Utah Cooperative Fish and Wildlife Research Unit, Utah State University, Logan, Utah 84322-5290 USA*
[3]*Department of Wildland Resources and Ecology Center, Utah State University, Logan, Utah 84322-5230 USA*
[4]*Department of Mathematics and Statistics, Utah State University, Logan, Utah 84322-3900 USA*
[5]*Department of Wildland Resources, Utah State University, Logan, Utah 84322-5230 USA*
[6]*College of Forest Resources, University of Washington, Seattle, Washington 98195-2100 USA*

Abstract. Classification procedures are some of the most widely used statistical methods in ecology. Random forests (RF) is a new and powerful statistical classifier that is well established in other disciplines but is relatively unknown in ecology. Advantages of RF compared to other statistical classifiers include (1) very high classification accuracy; (2) a novel method of determining variable importance; (3) ability to model complex interactions among predictor variables; (4) flexibility to perform several types of statistical data analysis, including regression, classification, survival analysis, and unsupervised learning; and (5) an algorithm for imputing missing values. We compared the accuracies of RF and four other commonly used statistical classifiers using data on invasive plant species presence in Lava Beds National Monument, California, USA, rare lichen species presence in the Pacific Northwest, USA, and nest sites for cavity nesting birds in the Uinta Mountains, Utah, USA. We observed high classification accuracy in all applications as measured by cross-validation and, in the case of the lichen data, by independent test data, when comparing RF to other common classification methods. We also observed that the variables that RF identified as most important for classifying invasive plant species coincided with expectations based on the literature.

Key words: additive logistic regression; classification trees; LDA; logistic regression; machine learning; partial dependence plots; random forests; species distribution models.

Introduction

Ecological data are often high dimensional with nonlinear and complex interactions among variables, and with many missing values among measured variables. Traditional statistical methods can be challenged to provide meaningful analyses of such data. In particular, linear statistical methods, such as generalized linear models (GLMs), may be inadequate to uncover patterns and relationships revealed by more sophisticated procedures (De'ath and Fabricius 2000). Classification procedures are among the most widely used statistical methods in ecology, with applications including vegetation mapping by remote sensing (Steele 2000) and species distribution modeling (Guisan and Thuiller 2005). In recent years, classification trees (Breiman et al. 1984) have been widely used by ecologists because of their simple interpretation, high classification accuracy, and ability to characterize complex interactions among variables.

A number of highly computational statistical methods, which have potential for ecological data mining, have recently emerged from the machine-learning literature. Random forests (hereafter RF) is one such method (Breiman 2001). RF is already widely used in bioinformatics (e.g., Cutler and Stevens 2006), but has not yet been utilized extensively by ecologists. In the few ecological applications of RF that we are aware of (see, e.g., Lawler et al. 2006 and Prasad et al. 2006), for both classification and regression RF is competitive with the best available methods and superior to most methods in common use. As the name suggests, RF combines many classification trees to produce more accurate classifications. By-products of the RF calculations include measures of variable importance and measures of similarity of data points that may be used for clustering, multidimensional scaling, graphical representation, and missing value imputation.

Potential applications of RF to ecology include (1) regression (Prasad et al. 2006); (2) survival analysis; (3) missing value imputation; (4) clustering, multidimensional scaling, and detecting general multivariate structure through unsupervised learning; and (5) classification. Descriptions of capabilities 1–4 are given in Appendix A; this article is concerned with RF as a classifier, with particular application to species distribution modeling. We highlight some features and strengths of RF compared to other commonly used classification methods.

Manuscript received 30 March 2007; revised 1 June 2007; accepted 4 June 2007. Corresponding Editor: A. M. Ellison.
[7] E-mail: Richard.Cutler@usu.edu

The Random Forests Algorithm

Classification trees

In the standard classification situation, we have observations in two or more known classes and want to develop rules for assigning current and new observations into the classes using numerical and/or categorical predictor variables. Logistic regression and linear discriminant analysis (LDA) accomplish this by determining linear combinations of the predictor variables to classify the observations. Classification trees build the rule by recursive binary partitioning into regions that are increasingly homogeneous with respect to the class variable. The homogeneous regions are called nodes. At each step in fitting a classification tree, an optimization is carried out to select a node, a predictor variable, and a cut-off or group of codes (for numeric and categorical variables respectively) that result in the most homogeneous subgroups for the data, as measured by the Gini index (Breiman et al. 1984). The splitting process continues until further subdivision no longer reduces the Gini index. Such a classification tree is said to be fully grown, and the final regions are called terminal nodes. The lower branches of a fully grown classification tree model sampling error, so algorithms for pruning the lower branches on the basis of cross-validation error have been developed (Breiman et al. 2004). A typical pruned classification tree has three to 12 terminal nodes. Interpretation of classification trees increases in complexity as the number of terminal nodes increases.

Random forests

RF fits many classification trees to a data set, and then combines the predictions from all the trees. The algorithm begins with the selection of many (e.g., 500) bootstrap samples from the data. In a typical bootstrap sample, approximately 63% of the original observations occur at least once. Observations in the original data set that do not occur in a bootstrap sample are called out-of-bag observations. A classification tree is fit to each bootstrap sample, but at each node, only a small number of randomly selected variables (e.g., the square root of the number of variables) are available for the binary partitioning. The trees are fully grown and each is used to predict the out-of-bag observations. The predicted class of an observation is calculated by majority vote of the out-of-bag predictions for that observation, with ties split randomly.

Accuracies and error rates are computed for each observation using the out-of-bag predictions, and then averaged over all observations. Because the out-of-bag observations were not used in the fitting of the trees, the out-of-bag estimates are essentially cross-validated accuracy estimates. Probabilities of membership in the different classes are estimated by the proportions of out-of-bag predictions in each class.

Most statistical procedures for regression and classification measure variable importance indirectly by selecting variables using criteria such as statistical significance and Akaike's Information Criterion. The approach taken in RF is completely different. For each tree in the forest, there is a misclassification rate for the out-of-bag observations. To assess the importance of a specific predictor variable, the values of the variable are randomly permuted for the out-of-bag observations, and then the modified out-of-bag data are passed down the tree to get new predictions. The difference between the misclassification rate for the modified and original out-of-bag data, divided by the standard error, is a measure of the importance of the variable. Additional technical details concerning the RF algorithm may be found in Appendix A.

Application of Random Forests to Ecological Questions

We provide examples of RF for classification for three groups of organisms commonly modeled in ecological studies: vascular plants (four invasive species), non-vascular plants (four lichen species), and vertebrates (three species of cavity-nesting birds). These examples cover a broad range of data characteristics encountered in ecology, including high to low sample sizes, and underlying probabilistic and non-probabilistic sample designs. The lichen data set also includes independent validation data, thereby providing opportunity to evaluate generalization capabilities of RF.

In all the examples that follow, we compare RF to four other classifiers commonly used in ecological studies: LDA, logistic regression, additive logistic regression (Hastie et al. 2001), and classification trees. The accuracy measures used were the overall percentage correctly classified (PCC), sensitivity (the percentage of presences correctly classified), specificity (the percentage of absences correctly classified), kappa (a measure of agreement between predicted presences and absences with actual presences and absences corrected for agreement that might be due to chance alone), and the area under the receiver operating characteristic curve (AUC). Resubstitution and 10-fold cross-validated estimates of these five accuracy measures were computed for all examples and methods. Except for the analyses pertaining to variable importance, no variable selection or "tuning" of the various classification procedures was carried out. To assess variable importance for LDA, backward elimination was carried out and the variables retained in the model ranked by P value. For logistic regression, backward elimination was carried out using the AIC criterion, and as with LDA, the retained variables ranked by P value. The variables split on in the highest nodes in classification trees were deemed to be most important for that procedure. Lists of the software used in our analyses and of available software sources for RF may be found in Appendix A.

We used predictors typically found in ecological classification applications, such as topographic variables, ancillary data (e.g., roads, trails, and habitat

TABLE 1. Accuracy measures for predictions of presence for four invasive plant species in Lava Beds National Monument, California, USA ($N = 8251$ total observations).

	Classification method							
	Random forests		Classification trees		Logistic regression		LDA	
Accuracy metric	Resub	Xval	Resub	Xval	Resub	Xval	Resub	Xval
Verbascum thapsus (common mullein; $n = 6047$ sites)								
PCC	95.3	**92.6**	84.2	83.2	80.6	80.0	79.4	79.2
Specificity	89.5	**84.5**	53.1	51.4	48.0	46.3	49.7	48.6
Sensitivity	97.4	**95.5**	95.5	94.7	92.5	92.3	90.2	90.3
Kappa	0.878	**0.809**	0.546	0.518	0.449	0.430	0.431	0.422
AUC	0.984	**0.940**	0.789	0.797	0.825	0.825	0.838	0.821
Urtica dioica (nettle; $n = 1081$ sites)								
PCC	93.9	**92.9**	91.3	90.5	88.8	88.6	87.1	87.1
Specificity	96.9	96.2	98.1	97.2	98.1	**97.9**	94.2	94.3
Sensitivity	74.6	**70.4**	45.9	45.6	27.1	26.7	40.0	39.0
Kappa	0.729	**0.680**	0.534	0.506	0.336	0.331	0.378	0.360
AUC	0.972	**0.945**	0.863	0.849	0.872	0.856	0.861	0.847
Cirsium vulgare (bull thistle; $n = 422$ sites)								
PCC	96.8	**96.5**	96.6	96.1	95.1	95.0	94.4	94.4
Specificity	98.8	98.7	99.6	**99.4**	99.4	**99.4**	98.0	98.1
Sensitivity	60.2	**56.4**	41.7	36.5	13.0	13.0	26.5	25.8
Kappa	0.643	**0.607**	0.540	0.474	0.209	0.195	0.297	0.296
AUC	0.938	**0.914**	0.772	0.744	0.810	0.784	0.789	0.762
Marrubium vulgare (white horehound; $n = 137$ sites)								
PCC	99.2	**99.1**	99.2	98.9	99.2	98.9	97.3	97.2
Specificity	99.8	**99.7**	99.9	**99.7**	99.9	**99.7**	97.7	97.7
Sensitivity	67.2	60.6	59.8	52.6	59.9	53.3	72.9	**67.9**
Kappa	0.738	**0.678**	0.716	0.621	0.706	0.627	0.463	0.434
AUC	0.988	**0.949**	0.873	0.867	0.972	0.944	0.918	0.906

Notes: LDA denotes linear discriminant analysis, PCC the percentage correctly classified, and AUC the area under the ROC curve. Resub is resubstitution accuracy estimate and Xval is the 10-fold cross-validated accuracy estimate. Sensitivity is the percentage of presences correctly classified. Specificity is the percentage of absences correctly classified. Kappa is a measure of agreement between predicted presences and absences with actual presences and absences corrected for agreement that might be due to chance alone. The largest cross-validated estimate for each classification metric for each species is in boldface type.

types), measured field variables, and down-scaled bioclimatic variables (e.g., Zimmermann et al. 2007). Tables with detailed information on the predictor variables used in each of our examples, and preliminary analyses and preprocessing of the bioclimatic and topographic predictors, may be found in Appendix B.

Predicting invasive species presences in Lava Beds National Monument, California, USA

Background.—Invasions by nonnative species are an increasing problem, especially in national parks. The U.S. National Park Service (NPS) manages its lands with an aggressive policy to control or remove invasive species and prohibit the establishment of new invaders. We used RF, classification trees, logistic regression, and LDA to predict sites of likely occurrence of four invasive plant species in Lava Beds National Monument (NM).

We obtained detection data from 2000 to 2005 for *Verbascum thapsus* (common mullein), *Urtica dioica* (nettle), *Marrubium vulgare* (white horehound), and *Cirsium vulgare* (bull thistle), and GIS layers for roads and trails within the monument. For our analyses, we imposed a 30-m grid over Lava Beds NM and a 500-m buffer outside the park ($N = 244\,733$ total grid sites). Our data included grid sites with one or more invasive species present ($n = 7361$ grid sites) and sites where all four species were absent ($n = 890$ grid sites). The predictor variables for these analyses were 28 topographic and bioclimatic variables, and three variables of distances to the nearest roads and trails.

Results.—For *V. thapsus*, cross-validated sensitivities for the four methods are all relatively high and similar (Table 1). However, specificities differ substantially, with RF performing substantially better than the other classifiers (Table 1). For *U. dioica* and *C. vulgare* the pattern is reversed: specificities are relatively high and similar, while sensitivities differ, with RF performing substantially better than the other classifiers (Table 1). The estimated sensitivities and specificities for *M. vulgare* are roughly the same for all four classifiers. Overall, RF had the highest PCC, kappa, and AUC values for all four invasive species. Classification trees were consistently second best in terms of PCC, suggesting some nonlinear structure that LDA and logistic regression were unable to adequately model.

Three variables in the Lava Beds NM data set concern distances to roads and trails. Because roads and trails are considered natural vectors for entry and spread of invasive species (Gelbard and Belnap 2003), we expected that distances to roads and trails would be important

predictors of presence for all four invasive species. This expectation was met for RF: for each of the four invasive species, these three variables were identified as most important to the classifications (Fig. 1). The results were similar for the other classifiers for *V. thapsus*: each of the other classifiers selected two of the vector variables among the four most important. However, there was little consistency in the variables identified as most important for the remaining three invasive species. For example, none of the distances to roads or trails variables were identified as being among the four most important for any of the other classifiers for *U. dioica*. Even though we cannot say that variables identified as "important" are right or wrong, the results for RF coincide more closely with expectations based on ecological understanding of invasions by nonnative species.

The preceding results also illustrate how the variable importance in RF differs from traditional variable selection procedures. When several variables are highly collinear but good predictors of the response, as are the distances to roads and trails in the Lava Beds NM data, stepwise and criterion-based variable selection procedures will typically retain one or two of the collinear variables, but discard the rest. In contrast, RF "spreads" the variable importance across all the variables, as we observed with the distances to roads and trails. This approach guards against the elimination of variables which are good predictors of the response, and may be ecologically important, but which are correlated with other predictor variables.

Predicting rare lichen species presences in the Pacific Northwest, USA

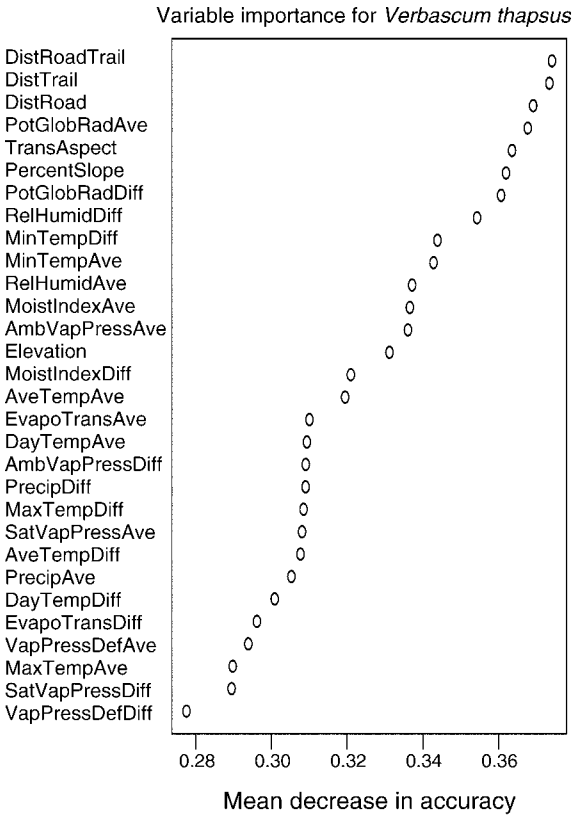
Background.—Our second application of RF involves two data sets on epiphytic macrolichens collected within the boundaries of the U.S. Forest Service's Pacific Northwest Forest Plan, USA. The first data set (hereafter LAQ, $n = 840$ randomly sampled sites) was collected from 1993 to 2000 and the second, independent data set (hereafter EVALUATION, $n = 300$ randomly sampled sites) was collected in 2003 in the same region. We applied RF, classification trees, additive logistic regression, and logistic regression, to the LAQ data and used the EVALUATION data as an independent assessment of the accuracy of the predicted classifications. Design details for the EVALUATION and LAQ surveys and tables of predictor variable descriptions can be found in Appendix B and in Edwards et al. (2005, 2006). Four lichen species in the LAQ and EVALUATION data sets were the subjects of our analyses: *Lobaria oregana*, *L. pulmonaria*, *Pseudocyphellaria anomala*, and *P. anthraspis*. The predictor variables were elevation, aspect, and slope, DAYMET bioclimatic variables, and four vegetation variables: percentage of broadleaf, conifer, and vegetation cover, and live tree biomass.

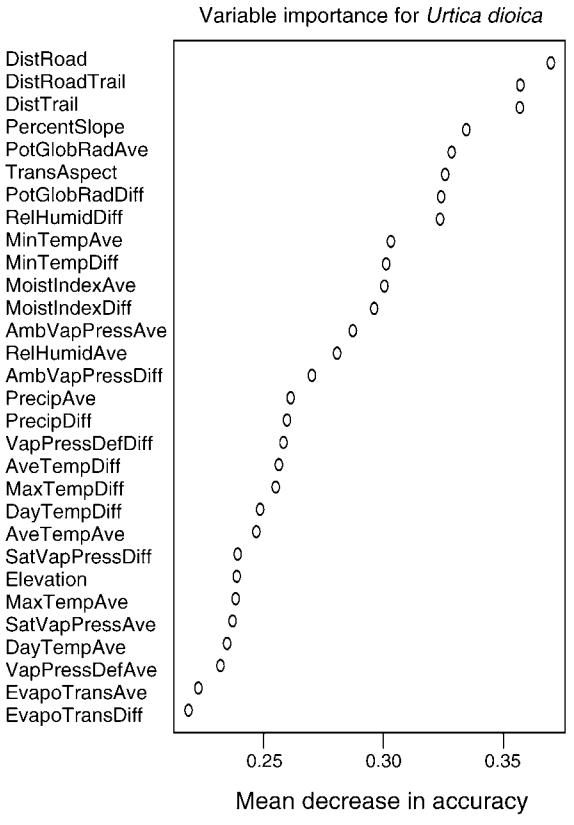
Results.—For all four species, the PCC, kappa, and AUC are highest for RF on the EVALUATION data, and RF generally outperforms the other classification procedures (Table 2). However, differences in accuracies are much smaller than we observed for the Lava Beds NM analyses, and in some cases are negligible. For *L. oregana*, sensitivity and specificity on the EVALUATION surveys for RF is better than the other classifiers, except in the case of sensitivity for additive logistic regression. It is interesting to note that the cross-validated estimates of accuracy for *L. oregana* are essentially the same for all four classifiers, while the EVALUATION estimates differ substantially, suggesting that even when both the training data and test data are collected at randomly selected sites in the same geographical region, cross-validated accuracy estimates may not reflect the true differences among classifiers. For *L. pulmonaria*, the PCC, kappa, and specificity for classification trees and RF are essentially identical, and are somewhat higher to much higher than those for additive logistic regression and logistic regression. For *P. anomala*, RF has somewhat higher EVALUATION accuracy than the other three methods, which are essentially the same. A similar pattern holds for *P. anthraspis*, except that classification trees have the largest sensitivity.

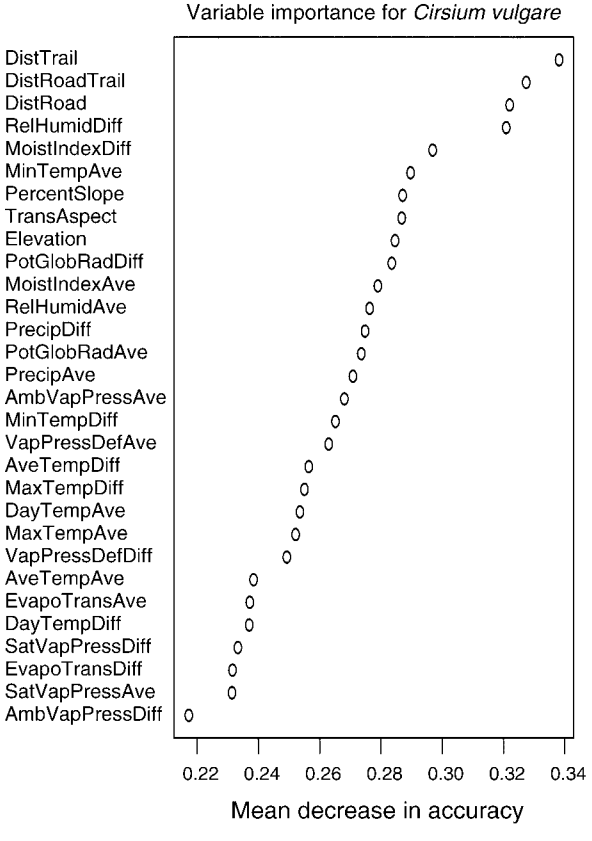
Partial dependence plots (Hastie et al. 2001; see also Appendix C) may be used to graphically characterize relationships between individual predictor variables and predicted probabilities of species presence obtained from RF. For binary classification, the *y*-axis on partial dependence plots is on the logit scale (details in Appendix C). In Fig. 2, there is almost a linear relationship between the logit of the probability of presence for *L. oregana* and the age of the dominant conifer. For *L. oregana*, the logit of predicted probability of presence shows a constant relationship to about 800 m and then decreases sharply. The same plot for *P. anthraspis* shows a more linear decrease between 0 and 1200 m. The logit of estimated probability of presence for *L. pulmonaria* suggests that the presence of this species are associated with sites that have more consistent precipitation over summer and winter.

→

FIG. 1. Variable importance plots for predictor variables from random forests (RF) classifications used for predicting presence of four invasive plant species in the Lava Beds National Monument, California, USA. The mean decrease in accuracy for a variable is the normalized difference of the classification accuracy for the out-of-bag data when the data for that variable is included as observed, and the classification accuracy for the out-of-bag data when the values of the variable in the out-of-bag data have been randomly permuted. Higher values of mean decrease in accuracy indicate variables that are more important to the classification. Variable descriptions are given in Appendix B.


Variable importance for *Verbascum thapsus*
DistRoadTrail
DistTrail
DistRoad
PotGlobRadAve
TransAspect
PercentSlope
PotGlobRadDiff
RelHumidDiff
MinTempDiff
MinTempAve
RelHumidAve
MoistIndexAve
AmbVapPressAve
Elevation
MoistIndexDiff
AveTempAve
EvapoTransAve
DayTempAve
AmbVapPressDiff
PrecipDiff
MaxTempDiff
SatVapPressAve
AveTempDiff
PrecipAve
DayTempDiff
EvapoTransDiff
VapPressDefAve
MaxTempAve
SatVapPressDiff
VapPressDefDiff
0.28
0.30
0.32
0.34
0.36
Mean decrease in accuracy


Variable importance for *Urtica dioica*
DistRoad
DistRoadTrail
DistTrail
PercentSlope
PotGlobRadAve
TransAspect
PotGlobRadDiff
RelHumidDiff
MinTempAve
MinTempDiff
MoistIndexAve
MoistIndexDiff
AmbVapPressAve
RelHumidAve
AmbVapPressDiff
PrecipAve
PrecipDiff
VapPressDefDiff
AveTempDiff
MaxTempDiff
DayTempDiff
AveTempAve
SatVapPressDiff
Elevation
MaxTempAve
SatVapPressAve
DayTempAve
VapPressDefAve
EvapoTransAve
EvapoTransDiff
0.25
0.30
0.35
Mean decrease in accuracy


Variable importance for *Cirsium vulgare*
DistTrail
DistRoadTrail
DistRoad
RelHumidDiff
MoistIndexDiff
MinTempAve
PercentSlope
TransAspect
Elevation
PotGlobRadDiff
MoistIndexAve
RelHumidAve
PrecipDiff
PotGlobRadAve
PrecipAve
AmbVapPressAve
MinTempDiff
VapPressDefAve
AveTempDiff
MaxTempDiff
DayTempAve
MaxTempAve
VapPressDefDiff
AveTempAve
EvapoTransAve
DayTempDiff
SatVapPressDiff
EvapoTransDiff
SatVapPressAve
AmbVapPressDiff
0.22
0.24
0.26
0.28
0.30
0.32
0.34
Mean decrease in accuracy

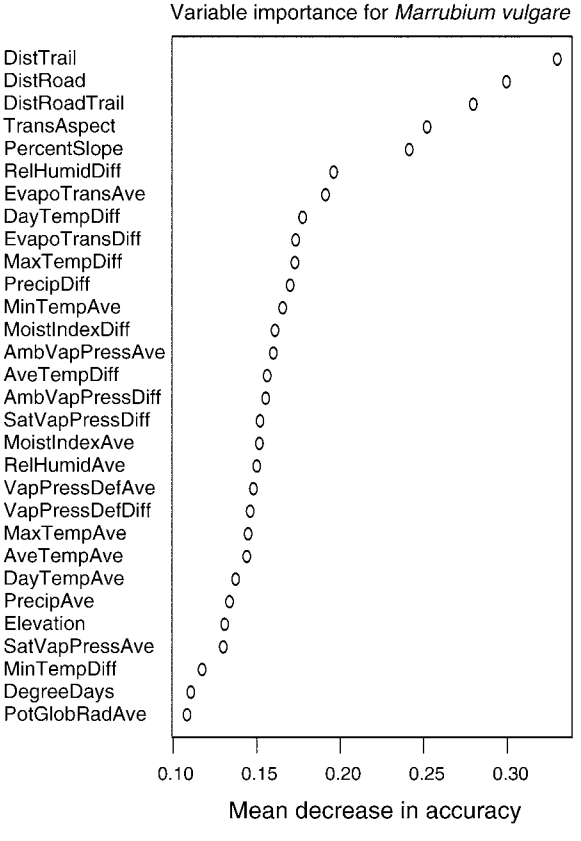

Variable importance for *Marrubium vulgare*
DistTrail
DistRoad
DistRoadTrail
TransAspect
PercentSlope
RelHumidDiff
EvapoTransAve
DayTempDiff
EvapoTransDiff
MaxTempDiff
PrecipDiff
MinTempAve
MoistIndexDiff
AmbVapPressAve
AveTempDiff
AmbVapPressDiff
SatVapPressDiff
MoistIndexAve
RelHumidAve
VapPressDefAve
VapPressDefDiff
MaxTempAve
AveTempAve
DayTempAve
PrecipAve
Elevation
SatVapPressAve
MinTempDiff
DegreeDays
PotGlobRadAve
0.10
0.15
0.20
0.25
0.30
Mean decrease in accuracy

TABLE 2. Accuracy measures for predictions of presence for four lichen species in the Pacific Northwest, USA.

	Classification method											
	Random forests			Classification trees			Additive logistic regression			Logistic regression		
Accuracy metric	Resub	Xval	EVAL	Resub	Xval	EVAL	Resub	Xval	EVAL	Resub	Xval	EVAL
Lobaria oregana ($n = 187$ sites)												
PCC (%)	83.9	85.0	**82.7**	90.8	83.8	71.0	88.3	84.3	77.7	87.0	85.1	74.3
Specificity (%)	93.3	94.0	**90.0**	95.6	90.9	77.3	93.9	90.0	80.9	93.6	91.6	79.1
Sensitivity (%)	51.3	53.5	62.5	74.3	58.8	53.8	68.9	64.2	**68.8**	64.2	62.6	61.3
Kappa	0.489	0.523	**0.542**	0.725	0.516	0.295	0.651	0.544	0.465	0.606	0.557	0.381
AUC	0.889	0.892	**0.867**	0.910	0.817	0.753	0.946	0.897	0.818	0.924	0.904	0.806
Lobaria pulmonaria ($n = 194$ sites)												
PCC (%)	84.7	84.6	**80.3**	88.8	81.3	80.0	88.3	81.2	73.0	85.9	84.6	72.7
Specificity (%)	93.5	93.2	88.5	95.5	91.0	**90.3**	94.4	87.9	75.1	93.2	92.3	76.5
Sensitivity (%)	55.2	56.2	59.0	66.5	48.9	53.0	68.0	58.8	**67.5**	61.8	59.3	62.6
Kappa	0.529	0.533	**0.492**	0.663	0.432	0.464	0.655	0.468	0.387	0.582	0.544	0.364
AUC	0.883	0.885	**0.869**	0.898	0.810	0.818	0.941	0.806	0.776	0.904	0.883	0.759
Pseudocyphellaria anomala ($n = 152$ sites)												
PCC (%)	85.0	85.2	**86.0**	90.0	83.1	83.7	88.9	84.4	83.7	87.0	85.5	83.7
Specificity (%)	95.3	95.0	**95.0**	96.6	91.7	92.5	95.6	91.7	92.5	94.8	93.8	92.9
Sensitivity (%)	38.2	40.8	**49.2**	59.9	44.1	47.4	58.6	51.3	47.4	51.9	48.0	45.8
Kappa	0.398	0.418	**0.499**	0.626	0.386	0.436	0.592	0.449	0.436	0.516	0.460	0.428
AUC	0.865	0.870	**0.861**	0.865	0.794	0.794	0.944	0.865	0.829	0.905	0.878	0.854
Pseudocyphellaria anthraspis ($n = 123$ sites)												
PCC (%)	88.2	87.6	**84.0**	91.7	86.1	80.0	93.2	88.1	78.7	88.1	85.4	81.3
Specificity (%)	97.1	96.6	**93.2**	95.9	93.6	86.0	97.8	95.8	86.0	95.8	94.4	89.4
Sensitivity (%)	36.6	34.9	50.0	66.7	42.3	**57.8**	66.7	43.0	51.6	43.1	32.5	51.6
Kappa	0.416	0.389	**0.476**	0.652	0.392	0.424	0.704	0.449	0.372	0.449	0.315	0.424
AUC	0.875	0.874	**0.816**	0.908	0.822	0.807	0.966	0.682	0.801	0.898	0.862	0.810

Notes: Abbreviations are: Resub, resubstitution accuracy estimates; Xval, 10-fold cross-validated accuracy estimates computed on lichen air quality data ($N = 840$ total observations); EVAL, pilot random grid survey (an evaluation data set with $N = 300$ total observations); PCC, percentage of correctly classified instances; AUC, area under the ROC curve. The largest value for each species and each metric in the EVALUATION data is in boldface type.

Predicting cavity-nesting bird habitat in the Uinta Mountains, Utah, USA

Background.—In this third example, we developed species distribution models relating nest presence to forest stand characteristics in the Uinta Mountains, Utah, USA, for three species of cavity-nesting birds: *Sphyrapicus nuchalis* (Red-naped Sapsucker), *Parus gambeli* (Mountain Chickadee), and *Colaptes auratus* (Northern Flicker). Classifications were developed for each species separately, and for all three species combined. This study is an example of the application of RF to small sample sizes, to a mixture of probabilistic (randomly selected locations) and non-probabilistic (nest cavities) survey data, and shows one simple way in which RF may be used to analyze data on multiple species simultaneously.

The stand characteristics we used consisted of numbers of trees in different size classes, numbers of conifers and snags, and stand type (pure aspen and mixed aspen/conifer), all considered habitat attributes of cavity nesting birds (see Lawler and Edwards 2002 and Appendix B). Within the spatial extent of the birds nest sites for the three species, 106 non-nest sites were randomly selected, and the same information as for the nest sites was collected.

Results.—For *S. nuchalis* and *P. gambeli*, RF has slightly better PCC, kappa and AUC than the other methods, while for *C. auratus* all methods have similar performance (Table 3). According to RF's variable importance measures, two stand characteristics—the numbers of trees of size class 7.5–15 cm (NumTree3-to6in) and 22.5–37.5 cm (NumTree9to15in)—were two of the three most important variables for all three species. Partial plots of these variables (Fig. 3) are interesting for two reasons. First, the plots are nonlinear. For the smaller-sized trees, the probability of a nest cavity drops rapidly with increasing NumTree3to6in, and then levels off. Larger trees (NumTree9to15in) have the opposite effect: the probability of a nest cavity rapidly increases, and then levels off. The second striking feature of the partial dependence plots for cavity nesting birds is that, for these two variables, the plots look very similar for all three species, suggesting that these species may be combined and analyzed as a group. Group results are comparable to the results for the individual species (Table 3). This illustrates how RF is not limited to modeling a single species; it may be used to analyze community data, and to model several species in the same functional group simultaneously. Other approaches to analyzing community data using RF include using models for some

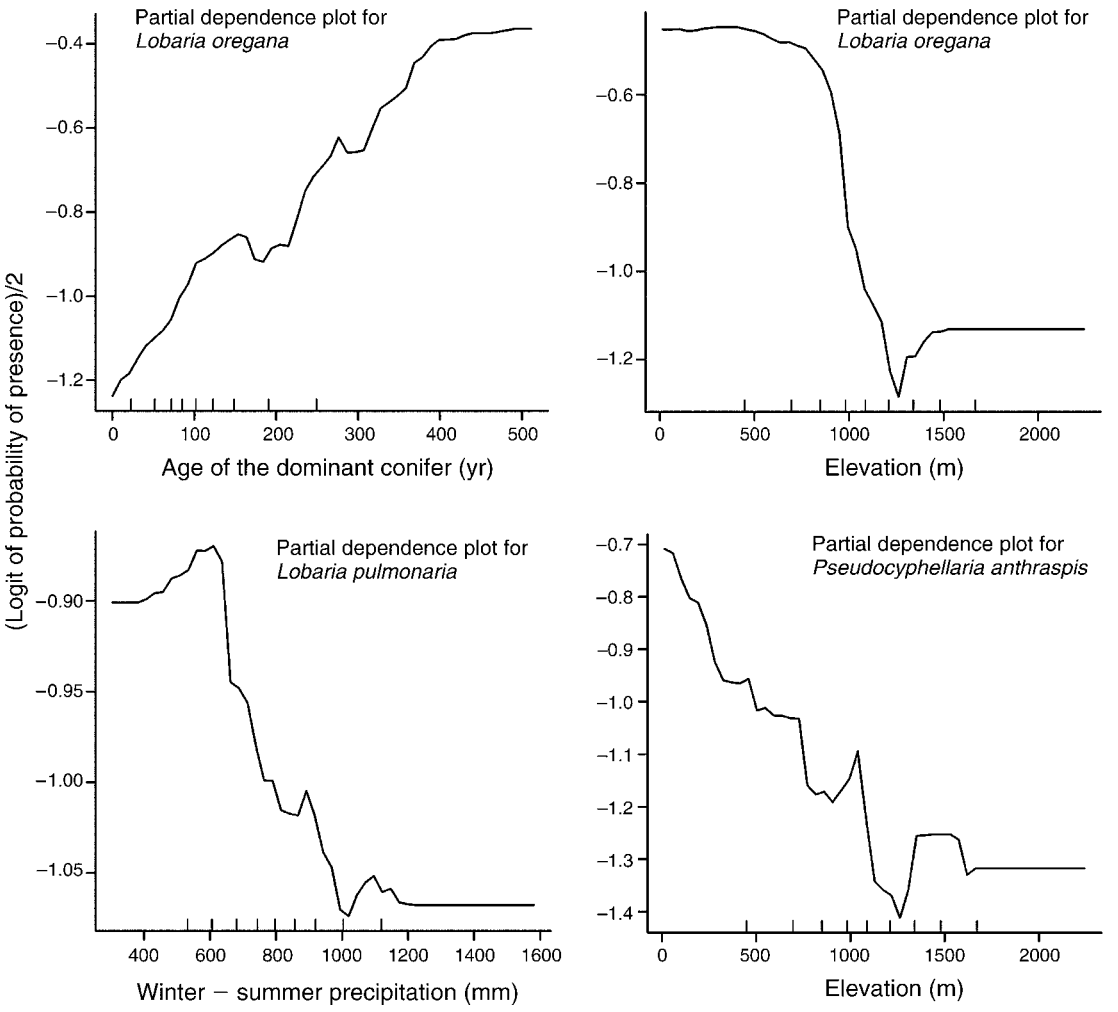


FIG. 2. Partial dependence plots for selected predictor variables for random forest predictions of the presences of three lichen species in the Pacific Northwest, USA. Partial dependence is the dependence of the probability of presence on one predictor variable after averaging out the effects of the other predictor variables in the model. "Winter − summer precipitation" is the total winter precipitation (October–March) minus the total summer precipitation (April–September). An explanation of the *y*-axis metric appears in Appendix C.

species to predict for much rarer, but related, species (Edwards et al. 2005) and modeling combined data with variables that identify different species.

Discussion and Conclusions

In three RF classification applications with presence–absence data, we observed high classification accuracy as measured by cross-validation and, in the case of the lichen data, by using an independent test set. We found a moderate superiority of RF to alternative classifiers in the lichen and bird analyses, and substantially higher accuracy for RF in the invasive species example, which involved complex ecological issues. In general, it is difficult to know in advance for which problems RF will perform substantially better than other methods, but post hoc graphical analyses can provide some insight. In principle, RF should outperform linear methods such as LDA and logistic regression when there are strong interactions among the variables. In Fig. 4, the bivariate partial dependence plot for two variables in the bird analyses shows a nonlinear relationship of the logit of the probability of nest presence, but the effect of each of these variables is approximately the same for each value of the other variable. Thus, the effects of the two variables are approximately additive, and in this case one might expect that RF will only do slightly better than additive methods such as LDA and logistic regression, which is what we observed (Table 3). However, in the partial dependence plot for *U. dioica* in Lava Beds NM, (Fig. 4) there was a complicated interaction in the effects of the distance to the nearest road and the distance to the nearest road or trail. These

TABLE 3. Accuracy measures for nest site classification of three species of cavity nesting bird species in the Uinta Mountains, Utah, USA.

	Classification method							
	Random forests		Classification trees		Logistic regression		LDA	
Accuracy metric	Resub	Xval	Resub	Xval	Resub	Xval	Resub	Xval
Sphyrapicus nuchalis (Red-naped Sapsucker; $n = 42$ nest sites)								
PCC (%)	88.5	**87.8**	87.8	79.7	86.5	83.1	85.1	82.4
Specificity (%)	95.3	**94.3**	98.1	90.6	90.6	86.8	89.6	86.8
Sensitivity (%)	71.4	71.4	61.9	52.4	76.2	**73.8**	73.8	71.4
Kappa	0.702	**0.687**	0.667	0.463	0.668	0.593	0.634	0.574
AUC	0.916	**0.918**	0.848	0.761	0.929	0.879	0.909	0.868
Parus gambeli (Mountain Chickadee; $n = 42$ nest sites)								
PCC (%)	85.8	**85.1**	87.8	78.4	84.5	77.7	86.5	79.1
Specificity (%)	95.3	**93.4**	91.5	85.8	92.5	84.9	91.5	84.9
Sensitivity (%)	61.9	**64.3**	78.6	59.5	64.3	59.5	73.8	**64.3**
Kappa	0.621	**0.612**	0.701	0.460	0.597	0.448	0.663	0.488
AUC	0.872	**0.880**	0.896	0.756	0.890	0.800	0.881	0.803
Colaptes auratus (Northern Flicker; $n = 23$ nest sites)								
PCC (%)	87.6	**86.8**	89.9	82.2	90.7	86.0	89.9	85.3
Specificity (%)	97.2	**96.2**	95.3	92.5	98.1	93.4	98.1	95.3
Sensitivity (%)	43.5	43.5	65.2	34.8	56.5	**52.2**	52.2	39.1
Kappa	0.490	0.469	0.638	0.309	0.632	**0.489**	0.594	0.406
AUC	0.869	**0.885**	0.836	0.731	0.903	0.821	0.882	0.797
All species combined ($n = 107$ nest sites)								
PCC (%)	85.9	**83.1**	85.9	75.1	83.6	77.9	82.6	77.5
Specificity (%)	86.8	**82.1**	86.8	73.6	78.3	72.6	73.6	67.0
Sensitivity (%)	85.0	84.1	85.1	76.6	88.8	83.2	91.6	**87.9**
Kappa	0.718	**0.662**	0.718	0.502	0.671	0.558	0.652	0.549
AUC	0.906	**0.893**	0.883	0.735	0.890	0.816	0.878	0.807

Notes: Abbreviations are: LDA, linear discriminant analysis; PCC, percentage of correctly classified instances; AUC, the area under the ROC curve; Resub, resubstitution accuracy estimate; Xval, 10-fold cross-validated accuracy estimate. There are 106 non-nest sites. The largest cross-validated estimate for each metric and each species is in boldface type.

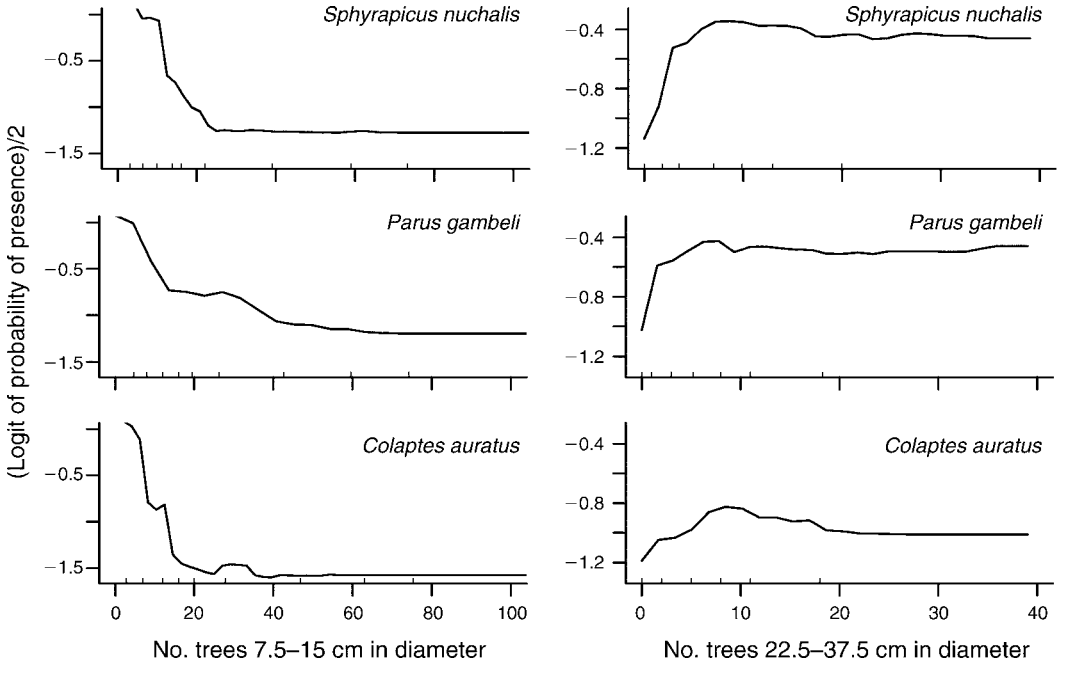


FIG. 3. Partial dependence plots for random forests classifications for three cavity-nesting bird species and two predictor variables. Data were collected in the Uinta Mountains, Utah, USA. Partial dependence is the dependence of the probability of presence on one predictor variable after averaging out the effects of the other predictor variables in the model. An explanation of the *y*-axis metric appears in Appendix C.

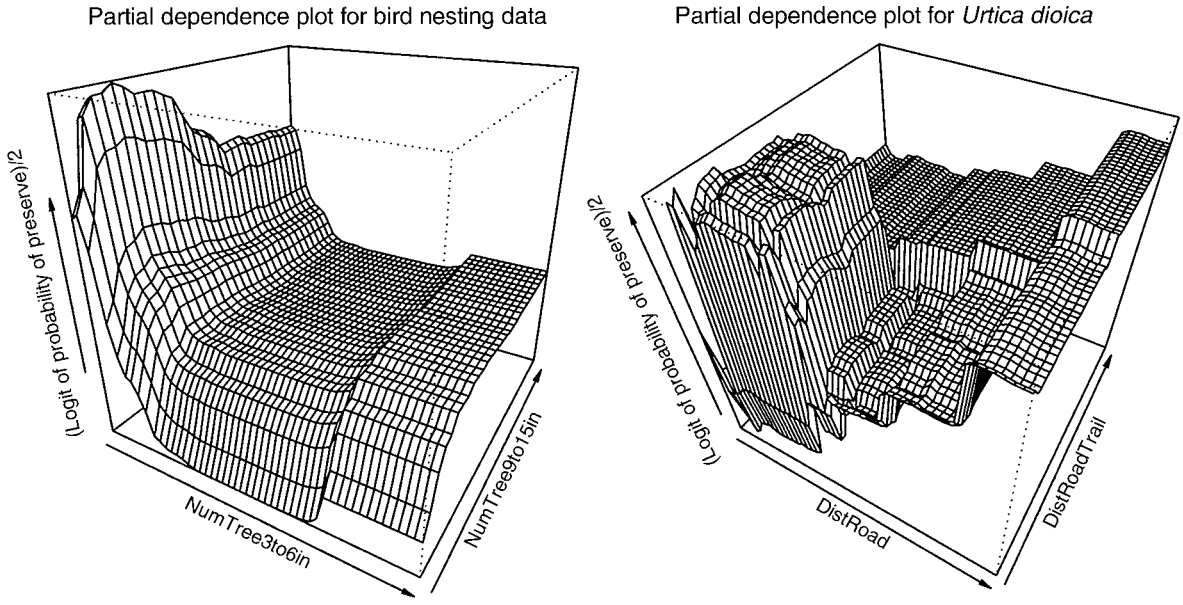


FIG. 4. Bivariate partial dependence plots for bird nesting data (107 nest sites and 106 non-nest sites) in Uinta Mountains, Utah, USA, and for *Urtica dioica* in Lava Beds National Monument, California, USA. Partial dependence is the dependence of the probability of presence on two predictor variables after averaging out the effects of the other predictor variables in the model. Variables are: NumTree3to6in, the number of trees between 7.5 cm and 15 cm dbh; NumTree9to15in, the number of trees between 22.5 cm and 37.5 cm dbh; DistRoad, distance to the nearest road (m); DistRoadTrail, distance to the nearest road or trail (m). An explanation of the *y*-axis metric appears in Appendix C.

kinds of interactions are the likely reason for the clear superiority of the tree-based methods, and RF in particular, in this application.

The original motivation for the development of RF was to increase the classification accuracy and stability of classification trees. In many respects RF supersedes classification trees: it is a more accurate classifier, and it is extremely stable to small perturbations of the data. For the classification situation, Breiman (2001) showed that classification accuracy can be significantly improved by aggregating the results of many classifiers that have little bias by averaging or voting, if the classifiers have low pairwise correlations. RF is an implementation of this idea using classification trees which, when fully grown, have little bias but have high variance. The restriction of the number of predictors available for each node in each tree in a RF ensures that correlations among the resultant classifications trees are small. In practical terms, RF shares the ability of classification trees to model complex interactions among predictor variables, while the averaging or voting of the predictions allows RF to better approximate the boundaries between observations in different classes.

Other classification procedures that have come from the machine learning literature in recent years include boosted trees, support vector machines (SVMs), and artificial neural networks (ANNs). All these methods, like RF, are highly accurate classifiers, and can do regression as well as classification. What sets RF apart from these other methods are two key features. The first of these is the novel variable importance measure used in RF, which does not suffer some of the shortcomings of traditional variable selection methods, such as selecting only one or two variables among a group of equally good but highly correlated predictors. In the invasive species example presented here, we observed that the variables RF identified as most important to the classifications coincided with ecological expectations based on the published literature.

The second feature that distinguishes RF from other competitors is the array of analyses that can be carried out by RF. Most of these involve the proximities—measures of similarity among data points—automatically produced by RF (see Appendix A). Proximities may be used to impute missing data, as inputs to traditional multivariate procedures based on distances and covariance matrices, such as cluster analysis and multidimensional scaling, and to facilitate graphical representations of RF classification results (Appendix C).

As with other highly computational procedures, including boosted trees, ANNs, and SVMs, the relationships between the predictor variables and the predicted values produced by RF do not have simple representations such as a formula (e.g., logistic regression) or pictorial graph (e.g., classification trees) that characterizes the entire classification function, and this lack of simple representation can make ecological interpretation difficult. Partial dependence plots for one or two predictor variables at a time may be constructed for any "blackbox" classifier (Hastie et al. 2001:333). If the

classification function is dominated by individual variable and low order interactions, then these plots can be an effective tool for visualizing the classification results, but they are not helpful for characterizing or interpreting high-order interactions.

RF is not a tool for traditional statistical inference. It is not suitable for ANOVA or hypothesis testing. It does not compute *P* values, or regression coefficients, or confidence intervals. The variable importance measure in RF may be used to subjectively identify ecologically important variables for interpretation, but it does not automatically choose subsets of variables in the way that variable subset selection methods do. Rather, RF characterizes and exploits structure in high dimensional data for the purposes of classification and prediction.

We have focused here on RF as a classification procedure, but RF is a package of fully nonparametric statistical methods for data analysis. Quantities produced by RF may also be used as inputs into traditional multivariate statistical methods, such as cluster analysis and multidimensional scaling. Unlike many traditional statistical analysis methods, RF makes no distributional assumptions about the predictor or response variables, and can handle situations in which the number of predictor variables greatly exceeds the number of observations. With this range of capabilities, RF offers powerful alternatives to traditional parametric and semiparametric statistical methods for the analysis of ecological data.

Acknowledgments

Funding was provided by the U.S. Forest Service Survey and Manage Program, and the USGS National Park Monitoring Program. We thank L. Geiser and her many colleagues for primary data collection of the lichen data, D. Hays, D. Larsen, P. Latham, D. Sarr, for their help and support with the Lava Beds NM analyses, and two anonymous reviewers, for their comments that led to substantial improvements in this manuscript.

Literature Cited

Breiman, L. 2001. Random forests. Machine Learning 45:15–32.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. Classification and regression trees. Wadsworth and Brooks/Cole, Monterey, California, USA.

Cutler, A., and J. R. Stevens. 2006. Random forests for microarrays. Methods in Enzymology 411:422–432.

De'ath, G., and K. E. Fabricius. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81:3178–3192.

Edwards, T. C., Jr., D. R. Cutler, N. E. Zimmermann, L. Geiser, and J. Alegria. 2005. Use of model-assisted designs for sampling rare ecological events. Ecology 86:1081–1090.

Edwards, T. C., Jr., D. R. Cutler, N. E. Zimmermann, L. Geiser, and G. G. Moisen. 2006. Effects of underlying sample survey designs on the utility of classification tree models in ecology. Ecological Modelling 199:132–141.

Gelbard, J. L., and J. Belnap. 2003. Roads as conduits for exotic plant invasions in a semiarid landscape. Conservation Biology 17:420–432.

Guisan, A., and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8:993–1009.

Hastie, T. J., R. J. Tibshirani, and J. H. Friedman. 2001. The elements of statistical learning: data mining, inference, and prediction. Springer Series in Statistics. Springer, New York, New York, USA.

Lawler, J. J., and T. C. Edwards, Jr. 2002. Landscape patterns as predictors of nesting habitat: a test using 4 species of cavity-nesting birds. Landscape Ecology 17:233–245.

Lawler, J. J., D. White, R. P. Neilson, and A. R. Blaustein. 2006. Predicting climate-induced range shifts: model differences and model reliability. Global Change Biology 12: 1568–1584.

Prasad, A. M., L. R. Iverson, and A. Liaw. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9: 181–199.

Steele, B. M. 2000. Combining multiple classifiers: an application using spatial and remotely sensed information for land cover mapping. Remote Sensing of Environment 74:545–556.

Zimmermann, N. E., T. C. Edwards, Jr., G. G. Moisen, T. S. Frescino, and J. A. Blackard. 2007. Remote sensing-based predictors improve habitat distribution models of rare and early successional tree species in Utah. Journal of Applied Ecology, *in press*.

APPENDIX A

Technical details and additional capabilities of random forests (*Ecological Archives* E088-173-A1).

APPENDIX B

Data descriptions and details of data preprocessing (*Ecological Archives* E088-173-A2).

APPENDIX C

Visualization techniques for random forests (*Ecological Archives* E088-173-A3).