RESEARCH ARTICLE

Gradient modeling of conifer species using random forests

Jeffrey S. Evans · Samuel A. Cushman

Abstract Landscape ecology often adopts a patch mosaic model of ecological patterns. However, many ecological attributes are inherently continuous and classification of species composition into vegetation communities and discrete patches provides an overly simplistic view of the landscape. If one adopts a niche-based, individualistic concept of biotic communities then it may often be more appropriate to represent vegetation patterns as continuous measures of site suitability or probability of occupancy, rather than the traditional abstraction into categorical community types represented in a mosaic of discrete patches. The goal of this paper is to demonstrate the high effectiveness of species-level, pixel scale prediction of species occupancy as a continuous landscape variable, as an alternative to traditional classified community type vegetation maps. We use a Random Forests ensemble learning approach to predict site-level probability of occurrence for four conifer species based on climatic, topographic and spectral predictor variables across a 3,883 km^2 landscape in northern Idaho, USA. Our method uses a new permutated sample-downscaling approach to equalize sample sizes in the presence and absence classes, a model selection method to optimize parsimony, and independent validation using prediction to 10% bootstrap data withhold. The models exhibited very high accuracy, with AUC and kappa values over 0.86 and 0.95, respectively, for all four species. The spatial predictions produced by the models will be of great use to managers and scientists, as they provide vastly more accurate spatial depiction of vegetation structure across this landscape than has previously been provided by traditional categorical classified community type maps.

Keywords Predictive modeling ·
Random forests · CART · Gradient

J. S. Evans
The Nature Conservancy – Rocky Mountain Conservation
Region, Fort Collins, CO 80524, USA
e-mail: Jeffrey_evans@tnc.org

S. A. Cushman (✉)
Department of Agriculture, Rocky Mountain Research
Station, US Forest Service, Missoula, MT 59801, USA
e-mail: scushman@fs.fed.us

Introduction

The analysis of landscape pattern to infer process is the underlying tenant in the field of landscape ecology (Forman and Godron 1986; Forman 1995; Turner et al. 2001). One's ability to effectively explain ecological processes therefore depends on correctly representing ecological patterns. Landscape ecology traditionally adopts a patch mosaic model of ecological patterns, implicitly assuming discretely bounded and categorically defined patches are sufficient to explain pattern–process relationships (McGarigal and Cushman 2005;

McGarigal et al. 2009). However, most ecological attributes are inherently continuous and classification of species composition into vegetation communities and discrete patches provides an overly simplistic view of the landscape and limits our ability to explore the continuous nature of plant distributions (McGarigal et al. 2009).

If one adopts a niche-based (Hutchinson 1957), individualistic concept (Gleason 1926; Whittaker 1967) of biotic communities then it would often be more appropriate to represent vegetation patterns as continuous measures of site suitability or probability of occupancy, rather than the traditional abstraction into categorical community types represented in a mosaic of discrete patches (McGarigal and Cushman 2005; Cushman et al. 2007). Although the problem of categorizations of the landscape failing to represent continuous ecological patterns has been identified (McIntyre and Barrett 1992; Manning et al. 2004; McGarigal and Cushman 2005; Cushman et al. 2007), few approaches have been proposed on how to predict gradients in a modeling environment (McGarigal et al. 2009).

Classified, community-level, patch-scale maps of vegetation have long been the foundation of natural resources management and the science of landscape ecology. However, it is unclear the degree to which these maps represent the true spatial structure of underlying biotic processes and patterns. A dominant focus of landscape ecology centers on linking driving processes at appropriate spatial scales to predict ecological patterns. It is essential to utilize methods that are consistent with ecological theory. In complex systems community classifications have been a useful tool for representing high dimensional data in a coherent manner. In the context of vegetation ecology, it is desirable to utilize methods that are consistent with niche-based, individualistic species responses to complex environmental gradients (Gleason 1926; Curtis and McIntosh 1951; Hutchinson 1957; Whittaker 1967). Progress in modeling techniques, computer processing and storage capacities have made individualistic modeling approaches a tractable problem. We can now represent large-numbers of species, individualistically, and readily explore complex relationships and high-dimensionality in a multivariate framework.

Predicting probability of occurrence is one approach for integrating species-level, pixel scale, niche-based theory into landscape mapping and analysis. Instead of representing landscape structure as a mosaic of discrete patches that are implicitly assumed to be categorically discrete and internally homogeneous (McGarigal and Cushman 2005; McGarigal et al. 2009), this approach represents occurrence for each species occurrence as a separate probability surface. This greatly reduces a number of fundamental problems in representing vegetation as classified mosaics, including errors related to the reality and stability of community type definitions, errors in stand delineation and boundary detection, and omission and commission errors.

The niche modeling community has made considerable headway in predicting species probabilities using presence only data with algorithmic and machine learning approaches (Stockwell and Peters 1999; Phillips et al. 2006; Prasad et al. 2006). Machine learning approaches, however, are often considered black-box with little inferential value. Cushman et al. (2007) distinguish between two modeling objectives, described as the "pattern-matching paradigm" and the "driver-response paradigm." In the latter, the goal is to obtain the most parsimonious understanding of the processes driving ecological responses for use in developing ecological theory and making predictions for the future under novel conditions or new locations. In the former case, however, the goal is to obtain the strongest possible prediction for a given data set. Nonparametric procedures like Random Forests (Brieman 2001b) can be used to effectively identify important associations, and graphical tools can be used to characterize relationships between predictor variables and classifications.

In addition, Brieman (2001a) argues that assumptions in parametric models, such as independence and multivariate-normality, are frequently violated whereas algorithmic approaches are not affected by these violations and provide more stable and relevant information. Furthermore, ecological systems often exhibit complex, non-linear relationships, autocorrelation, and variable interaction across temporal and spatial scales. Nonparametric algorithmic classifiers often greatly outperform parametric methods in such cases. The field of ecological informatics is rapidly developing machine learning approaches to explore and quantify complex and nonlinear ecological relationships (Park and Chon 2006). Such ecological informatics methods can then be a starting point for inferential statistics, through which variables are

identified, hypothesis developed, and inferential methods applied post hoc.

Classification and Regression Tree (CART) approaches have gained broad usage in ecological studies (Déath and Fabricius 2000). However, CART suffers from several problems, such as over-fitting and difficulty in parameter selection. Several solutions have been proposed that incorporate iterative approaches (Schapire 1990; Breiman 1996). One approach in particular, Random Forests (Brieman 2001b), has risen to prominence due to its ability to handle large numbers of predictor variables and find signal in noisy data (Cutler et al. 2007). Another advantage of Random Forests is that, by permutation of independent variables, it provides local and global measures of variable importance.

A primary criticism of species distribution models is the lack of incorporation of ecological theory (i.e., expected shape of the species response curve) and influences of model misspecification (Austin 2002; Guisan and Zimmermann 2000). In complex ecological systems, multiple driving factors acting at different scales may have critical effects on processes of interest (Cushman et al. 2007). This hinders model specification and the development of sound hypotheses. With the use of non-parametric, algorithmic modeling approaches these limitations are somewhat mitigated.

The goal of this analysis is to predict the probability of occurrence of common forest trees across a large and complex landscape in northern Idaho, USA. We utilize the nonparametric, algorithmic method Random Forests to predict the occurrence of four tree species based on combinations of multiple topographic, climatic and spectral predictor variables. The purpose is to obtain a highly accurate prediction of current species occurrence for use in management, and related research in wildlife habitat relationships and effects of climate change. Through the prediction of each species' occurrence probability, we demonstrate a method for representing vegetation gradients suitable for an integrated analytical framework for exploring continuous landscape processes.

Methods

Study area

We utilized 411 field plots collected in 2000–2001 as part of a USDA Forest Service pilot project to intensify the Forest Inventory and Analysis grid on the Bonners Ferry Ranger District, Panhandle National Forests. Our study area covers 3,883 km^2 in northern Idaho, USA (Fig. 1), encompassing a wide range of environmental, anthropogenic, and vegetation conditions. Tree species are relatively diverse for temperate conifer forest, with Western redcedar (*Thuja plicata*), Western hemlock (*Tsuga hetrophylla*), and grand fir (*Abies grandis*) at lower elevation sites with high moisture availability;

Fig. 1 Study area orientation map. Cross-hatching is USDA-Forest Service Idaho Panhandle National Forests, Bonners Ferry Ranger District lands
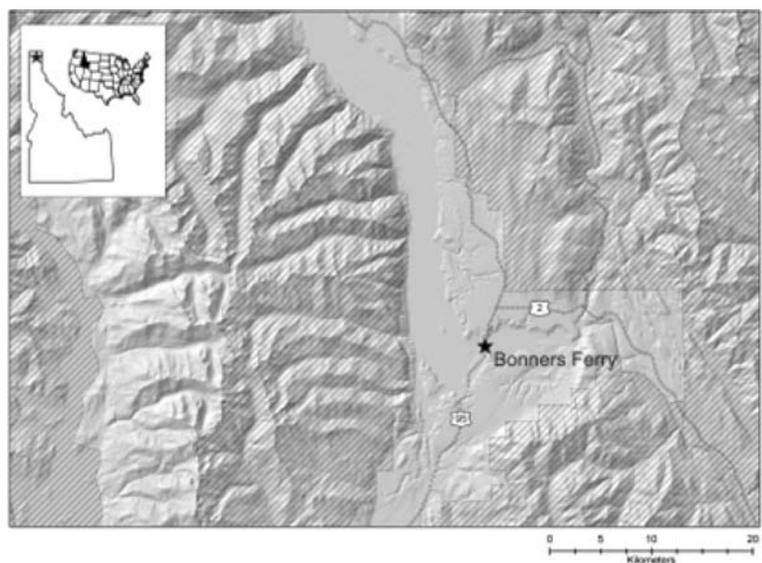
Ponderosa pine (*Pinus ponderosa*), at low elevation dry sites; Douglas fir (*Pseudotsuga menziesii*), lodge-pole pine (*Pinus contorta*), and Western larch (*Larix occidentalis*) at intermediate elevations; and Englemann spruce (*Picea engelmannii*) and subalpine fir (*Abies lasiocarpa*) occupying colder, higher elevation conditions. Management history includes over 100 years of active timber harvest, resulting in a diverse patch mosaic of vegetation across the full range of age and canopy cover (Fig. 2).

Forest inventory and analysis intensification

The FIA program (http://fia.fs.fed.us/) is a national program that gathers annual inventory data on a 4.8 km grid across all forested areas. In our study area, a spatial intensification was appended to the original FIA grid providing a systematic sample of one plot per 1.7 km. The FIA plot design follows a four plot cluster containing one 0.10 ha (17.95 m radius) plot with three 0.01 ha (7.32 m radius) plots

in three directions (360°, 120°, and 240°) and three 120' transects between primary plot and sub-plots. Recorded variables include measured species (spp), crown width (CW), diameter at breast height (DBH), and crown base height (CBH) on all trees >0.9 dbh.

For this analysis we calculated species proportion, utilizing only the primary plot and focusing on four species: *Abies lasiocarpa*, *Pinus ponderosa*, *Thuja plicata*, and *Pseudotsuga menziesii*. These four species were selected to represent the widest range of environmental optima across the temperature/moisture and elevational gradients among species extant in the study area. One generalist (*Pseudotsuga menziesii*) was specifically selected to test the models efficiency in predicting a species that is equally opportunistic across its range of variability.

Independent variables

We selected 40 independent (*x*) variables (Table 1) to represent abiotic (topographic and climate) and
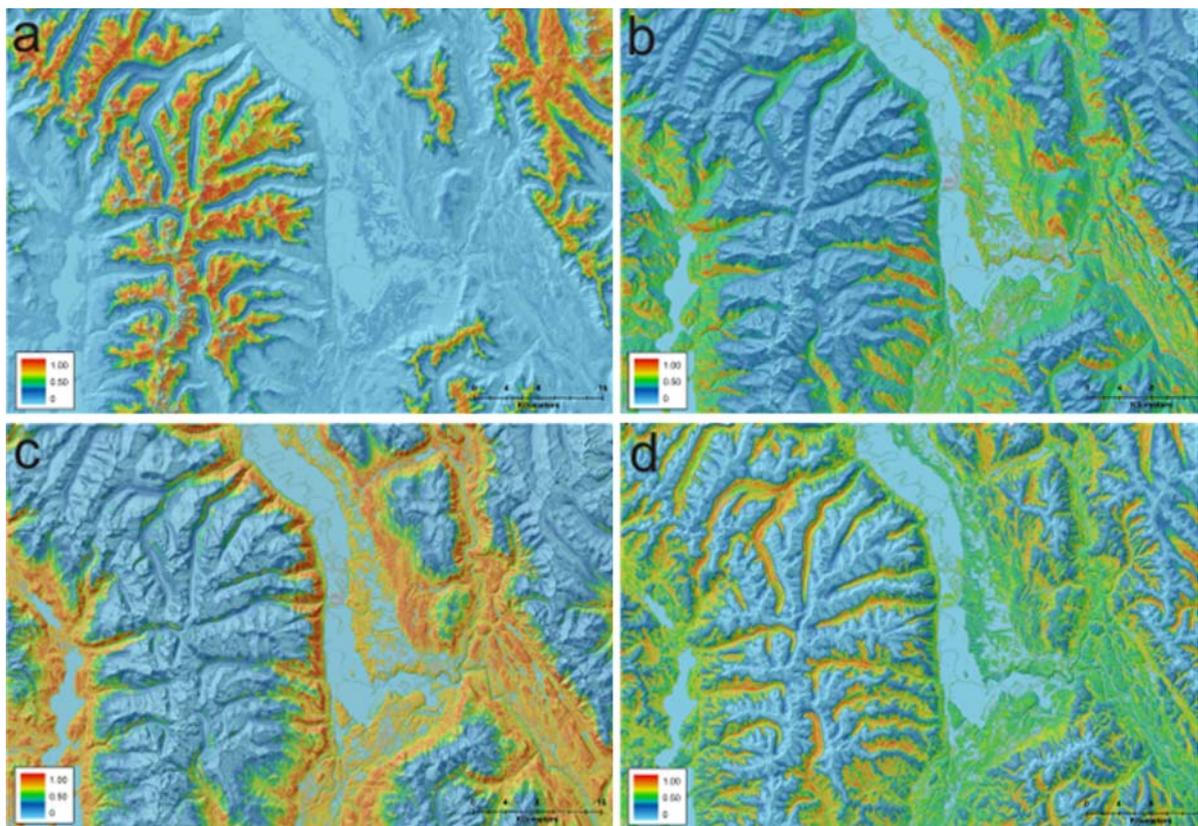
Fig. 2 Maps of predicted probabilities of species occurrences; **a** *A. lasiocarpa*, **b** *P. ponderosa*, **c** *P. menziesii*, **d** *T. plicata*

Table 1 Predictor variable names, descriptions, and references in the ecological literature

Variable	Description	Reference
ELEV	Elevation from shuttle topographic radar mission	Rabus et al. (2003)
SLP	Slope in radians using dinf model	Tarboton (1997)
SPCOSAP	[Slope × COS(Aspect)]	Stage (1976)
SPSINAP	[Slope × SIN(Aspect)]	Stage (1976)
CTI	Compound topographic index (wetness)	Moore et al. (1993)
INSO	Solar insolation (KWH m^2)	Fu and Rich (1999)
HLI	Heat load index	McCune and Keon (2002)
TRASP	Topographic radiation index	Roberts and Cooper (1989)
SP	Relative slope position	Murphy et al. (2009)
HSP	Hierarchical slope position	Murphy et al. (2009)
CRV	Slope curvature	
ROUGH (3, 15, 27)	Variance of elevation (*n* size window)	Murphy et al. (2009)
ERR (3, 15, 27)	Elevation relief ratio (*n* size window)	Evans (1972)
MAT	Mean annual temperature	Rehfeldt et al. (2006)
MTCM	Mean temperature in coldest month	Rehfeldt et al. (2006)
MMIN	Minimum temp in coldest month	Rehfeldt et al. (2006)
MTWM	Mean temperature in warmest month	Rehfeldt et al. (2006)
MMAX	Minimum temp in warmest month	Rehfeldt et al. (2006)
MAP	Mean annual precipitation	Rehfeldt et al. (2006)
GSP	Growing season precipitation	Rehfeldt et al. (2006)
TDIFF	Summer–winter temperature differential	Rehfeldt et al. (2006)
DD5	Number degree-days >5°C	Rehfeldt et al. (2006)
DD0	Number degree-days <0°C	Rehfeldt et al. (2006)
FFP	Length of frost free period	Rehfeldt et al. (2006)
AMI	Annual moisture index [DD5/MAP]	Rehfeldt et al. (2006)
PRATIO	Ratio of summer to total precipitation [GSP/MAP]	Rehfeldt et al. (2006)
B1–B6 and B7	Landsat ETM+7 Bands 1–6 and 7 (reflectance)	Chavez (1988)
NDVIc	MIR corrected normalized difference vegetation index	Nemani et al. (1993)

phenological conditions. The abiotic variables included topographic variables derived from digital elevation models, and climate variables from the spline model presented in Rehfeldt et al. (2006). Phenological variables were derived from atmospherically corrected (Chavez 1988) bands 1–6 and 7 Landsat ETM+7 spectral data (at-sensor reflectance), and mid-infrared corrected normalized vegetation difference index (NDVIc) (Nemani et al. 1993). We calculated topographic-based variables in Workstation ArcInfo using Arc Macro Language (AML) programs developed by the authors. Our elevation source data was from the 30 m^2 Shuttle Topographic Radar Mission (Rabus et al. 2003) downloaded from the USGS national map (http://nationalmap.gov). We

applied atmospheric correction with dark-object subtraction (Chavez 1988) to a 07/28/2000 ETM+7 Landsat image and calculated NDVIc in ERDAS Imagine v9.0 using Spatial Modeler Language (SML). We assigned values from rasters to each plot location in ArcInfo using AML to create a database of y, $x_1 \ldots x_n$ used in the model.

The suite of 40 independent variables (Table 1) together include the major factors likely to influence vegetation response, such as temperature/moisture gradient, direct topographic effects (i.e., slope, slope position, elevation), climate influences (i.e., precipitation, temperature), and vegetation phenology. The Random Forests method utilized in this study is robust in dealing with large

numbers of independent variables (Brieman 2001b; Cutler et al. 2007).

Random forests

We predicted occurrence probabilities for the four selected species using the Random Forests method (Brieman 2001b; Cutler et al. 2007) as implemented in R (R development core 2007; Liaw and Wiener 2002). Random Forests is a classification and regression tree (CART) (Déath and Fabricius 2000) based bootstrap method that corrects many of the known issues in CART, such as over-fitting (Brieman 2001b; Cutler et al. 2007), and provides very well-supported predictions with large numbers of independent variables (Cutler et al. 2007). We ran 5,000 bootstrap replicates (k) with replacement using a 36% data-withhold [out-of-bag (OOB)] sample. The number of bootstrap replicates was initially selected based on the number of replicates where the OOB error ceases to improve. In our analysis, this OOB error stabilization occurs between $k = 1,200$ and $k = 2,500$ replicates. However, variable interaction is thought to stabilize at a slower rate than OOB error (Adele Cutler, personnel communication). A heuristic to account for variable interaction with a large set of independent variables was defined as [$2 \times (k_y$ for OOB stabilization)]. Since the only loss in running more k than necessary is processing time (Brieman 2001b), we selected $k = 5,000$ as an adequate number to account for stabilization in both error and interaction. The m parameter, number of variables permutated at each node, was defined as $m = $ [SQRT(number of x variables)], with a minimum of $m = 2$ (Brieman 2001b).

The response variable (y) was defined as a binary response, presence (1)/absence (0), by applying the rule spp = [IF proportion $>0.10 = 1$ ELSEIF 0]. Previous studies (Chawla et al. 2003; Chen et al. 2004) demonstrated that imbalance between the proportion of presence and absence classes can cause bias in the prediction and model-fit. When an imbalanced sample is present the bootstrap of the data is biased towards the majority class, thus over-predicting the majority-class and under-predicting the minority. The resulting model fit can be deceptive, exhibiting very small overall OOB error due to very small errors in the majority as a result of extremely high cross-classification error from the minority-class. An alternative solution that is often used when

there are many more absences than presences in a classification dataset is to shift the cutoff for the probability of present from 0.5 to something smaller. However, this approach does not work in our case because of our interest in stable and comparable probabilities of presence among species.

To correct for the imbalance between number or presences and absences we developed an approach that iteratively down-sampled the majority class by randomly drawing $2 \times$ [n of minority] and running a new Random Forests model using different random subsets while holding the sample-size of the minority-class constant. To ensure that the sample distributions of the independent variables were captured, we calculated a covariance matrix of the independent variables in the full data. As data was subset for each model, we tested the covariance of the cumulative subset data covariance matrix against the covariance matrix of the full data's covariance until equalevancy (*P*-value of <0.0001) was satisfied, at which point we cease iterating models. The convergence of the covariance matrices was tested using the equivalence statistic introduced in Morrison (2002). Since the underlying theory of Random Forests is ensemble learning, it is possible to combine trees from different models that are based on the same underlying data (Brieman 2001b). We ran a new Random Forests model, with the model parameters defined above, for each random sub-sample of the majority class. The final model was an ensemble derived from combining trees from all the independent models of randomly sub-sampled majority data.

In most cases, it is not necessary to retain all variables in a given model. Often, removing variables can not only result in more parsimonious models that exhibit less noise, but can also improve OOB error (Murphy et al. 2009). To identify the most parsimonious model we applied the Model Improvement Ratio (MIR) (Murphy et al. 2009). The MIR uses the permuted variable importance, represented by the mean decrease in OOB error, standardized from zero to one. The variables are subset using 0.10 threshold increments, with all variables above the threshold retained for each model. This subset is always performed on the original model's variable importance to avoid over-fitting (Svetnik et al. 2004). We compare each subset model and select the model that exhibits the lowest total OOB and lowest maximum within-class error. We nested this procedure in the

down-sampling approach. The selected model is assessed using the median error across the ensemble of all down-sampled models.

We made model predictions in two ways. First, a presence/absence prediction was based on majority votes across all trees. Second, species probabilities were predicted using a ratio of the Random Forests majority votes-matrix to create a probability distribution. Random Forests makes predictions based on the plurality of votes across all bootstrap trees and not on a single rule set. This vote's matrix can be scaled and treated as a probability given the error distribution of the model. A function was added to GridAsciiPredect (Crookston and Finley 2008) that uses the votes-probability function to write the probabilities to ASCII grid(s).

Validation

We approached model validation utilizing model fit, randomization test, sensitivity (proportion of observed positives correctly predicted), specificity (proportion of observed negatives correctly predicted), Kappa, and the area under the ROC curve (AUC) (DeLong et al. 1988). The model fit was assessed using the OOB error estimate. We addressed model significance (P) by running the model 1,000 times with a randomization of y (Murphy et al. 2009).

To achieve a balanced sample, the down-sampling method randomly sub-samples the majority data. A random sub-sample does not ensure a well represented sample of the majority class, thus potentially degrading predictive-power. The OOB error is the median taken from an error distribution across the randomized trees. Multiple trees with extremely high error can change the variance of the error distribution. To more accurately assess the error in the prediction we performed a n independent data withhold of 10% (for each class). Using the final set of subset independent variables identified in the final selected model we ran Random Forests 1,000 times and, at each replicate, made a prediction to the (10%) withheld data. Error was quantified as the cumulative error rate across bootstrap replicates. Using the bootstrapped observed versus predicted we calculate model error (percent incorrectly classified), AUC, sensitivity, specificity, and Kappa values using the PresenceAbsence package in R (Freeman and Moisen 2008).

Results

All four models were very well supported (Table 2) and significant at $P < 0.001$ with very low model-fit error rates. All four models had kappa statistics over 0.86, and very high sensitivity/specificity values (Table 2), indicating excellent predictions with very little cross-classification error. In addition, the area under the ROC curve (AUC) was over 0.98 in all cases (Table 2). As a general rule, AUC values over 0.9 indicate excellent model performance; the values reported here indicate that these models very successfully predict the occurrence patterns of the four focal species.

Models for all four species performed well based on model error (Table 2). The *P. menziesii* and *T. plicata* models were the best of the four models based on model-fit error (with 0.1 and 0.96% respectively) with *A. lasiocarpa* (1%) exhibiting very low error as well. *P. ponderosa* exhibited the highest model-fit error (8.3%).

For all four species, climatic variables are the most important predictors of occurrence. The models are primarily influenced by variables that measure the temperature regime. The variables in the *A. lasiocarpa* model, with the exception of Landsat band-4, are all indicators of cold, moist, high-elevation environments (Table 2). The near-infrared spectral range of band-4 is strongly representative of vegetation condition and is indicative of moisture-stress in colder environments. The model predicts that the probability of *A. lasiocarpa* occurrence is highest at high elevations with high precipitation and very-low average temperatures.

Conversely, the *P. ponderosa* model is strongly related to high temperatures and low water availability, with TRASP being the top variable (Table 2). The TRASP variable represents the effect of aspect on incoming solar radiation, where large values of TRASP reflect steep south-facing slopes which are both hot and dry. The model predicts the highest probability of ponderosa pine occurrence at low elevation hot sites with high incident solar radiation.

Factors driving the occurrence of *T. plicata* are similar to *P. ponderosa*, only with opposite effects. Geomorphology also plays a role in this model. The HSP, CTI, and GSP variables are all indicators of moisture availability. We hypothesize that ROUGH3 reflects topographic influence on cold-air drainage and ROUGH15 represents the general geomorphology

Table 2 Selected variables (in order of importance), model error, sensitivity, specificity, Kappa (k), ROC area under curve (AUC), and significance (P), for each model

Model	Selected variables	Model error (%)	Sensitivity	Specificity	k	AUC	P
A. lasiocarpa	SPSINAP, B4, GSP, MTCM, FFP, MAT, ELEV	1.11	0.98	1	0.97	0.98	0.001
P. ponderosa	TRASP, MTCM, FFP, ELEV, MAT	8.30	1	0.85	0.86	0.98	0.001
P. menziesii	GSP, HSP, ERR15, SCOSA, DISS27, ROUGH27, FFP, MTCM, ELEV, MAT, HLI	0.1	0.99	0.99	0.99	0.99	0.001
T. plicata	ROUGH3, HSP, ROUGH15, SPCOSAP, CTI, B5, NDVIc, GSP, B7, MTCM, MAT, ELEV, FFP	0.96	1	0.97	0.98	0.99	0.001

(alluvial deposits in drainage bottoms) where *T. plicata* is most prevalent. The model predicts that *T. plicata* is most common on mesic sites in broad alluvial valleys at middle elevations.

Distribution of *P. menziesii* is controlled primarily by temperature/moisture gradients in a mid-elevation range. Variables indicating moisture are GSP and FFP. Temperature is indicated by HLI, MAT, MTCM, HSP, and SCOSA. Elevation (ELEV) is most likely redundant with some of the direct measures of the temperature gradient, however, *P. menziesii* in this study area is limited to lower to mid-elevations. ERR15 and ROUGH27 represent general geological characteristics such as subsurface water flow and weathering (soil recruitment through erosion of bedrock) (Evans 1972).

Where the *T. plicata* and *P. ponderosa* overlap is where *P. ponderosa* is mixed with *P. menziesii*, in the cooler portion of its distribution. All four species are strongly influenced by the elevational gradient and had elevation in the final selected model(s). The variables in all the final selected models are consistent with the current ecological knowledge of each species.

Discussion

Random Forests was exceptionally effective in predicting the probability of presence in response to complex gradients of topography and inferred microclimate. These results indicate that species-level, gradient prediction of vegetation across complex mountain landscapes can be highly effective. The variables of most importance in all four species' models reflect the primary climatologically limiting factors that one would expect to have dominant

influence on the topographical distribution of tree species. The combination of very high prediction success with inclusion of variables known to be important drivers suggests that these models tightly reflect the realized niche space of each of these four species. This, in turn, indicates that the Random Forests method can be highly effective at describing realized niches and mapping them across complex landscapes, accounting for both local and global effects occurring across scales.

Class imbalance can have a profound effect on model performance (Chawla et al. 2003; Chen et al. 2004). With the development of the down-sampling approach we have found an effective means of addressing the problem and significantly improving our models. Previous approaches have applied oversampling methods (Chawla et al. 2003; Chen et al. 2004), where the minority class is duplicated or synthesized to increase the number of observations. The presence of duplicate observations of the minority data present a potentially serious problem in that the bootstrap no longer represents an independent random draw of data. This leads to both a model bias and a large inflation of accuracy in the minority class. In applying a down sampling approach we avoid this potentially serious issue while still addressing the issues of imbalance in classes.

The model improvement ratio (Murphy et al. 2009) has demonstrated an effective means of identifying a parsimonious set of variables and selecting a model that minimizes noise and improves model performance. The OOB error statistic provided in Random Forests is indicative of model fit, but not necessarily predictive performance or power. We chose not to sacrifice data for an independent validation; instead we conducted a bootstrap of the bootstrap, providing a qusai-independent measure of

model performance. We believe that, even though all the data was used in the final model, the validation statistics provided are a true representation of model performance.

For all the merits of Random Forests in prediction, its interpretability is somewhat limited. In a classification, for instance, one is not provided with an equation that provides the interpretation of slope and intercept coefficients. Unlike CART, Random Forests does not provide a rule-set but rather predicts based on a vote majority. Even though Random Forests solves many of the problems in CART, the interpretability of a set of rules is desirable. Random Forests is somewhat of a black-box, as are many other effective machine learning classifiers, such as support vector machines, artificial neural networks, adaboost and gradient modeling machines. However, Random Forests excels at identifying important independent variables and 2- and 3-dimensional partial dependence plots may be used to visually characterize relationships between predictor variables and predicted classes (Hastie et al. 2001).

The very high model prediction success suggests training very tightly to the nuances of this data set. This makes generalization to these species in other landscapes less likely to succeed. However, our goal in this analysis is to obtain the highest possible predictive accuracy for discriminating locations where each species is present from those where each species is absent, and to produce species-level, pixel scale maps of occurrence probability. Given that goal, we feel this method is highly effective and limitations on the generalization to other landscapes is not a problem given the scope of our objectives.

In our study area, natural resources agencies typically base management decisions on estimates of current vegetation conditions taken from classified maps. These maps represent vegetation as categorical "types", or "communities" in a discrete mosaic of patches. Such maps differ from those presented here in three critical respects. First, they are based on classifications into species community assemblages rather than prediction of individual species. Second, they are based on assigning locations into discrete patches in which a given patch is believed to share the same "community type", rather than a continuous gradient of proportion or suitability. Third, they predict to a discrete group membership rather than a continuous probability of occurrence.

Such classified, community-level, patch-scale maps of vegetation have long been the foundation of natural resources management and the science of landscape ecology. However, the question remains to what degree they reflect the underlying patterns and processes that drive ecological systems. A major rallying cry of contemporary landscape ecology is the central importance of linking key processes at appropriate spatial scales to predict ecological patterns. Thus, in vegetation modeling and mapping it is essential to utilize methods that are most in accord with ecological theory; specifically, it is important to utilize methods that are consistent with niche-based, individualistic species responses to complex environmental gradients, which has been the core of community ecology for decades (Gleason 1926; Curtis and McIntosh 1951; Hutchinson 1957; Whittaker 1967).

The approach adopted here is explicitly niche-based and individualistic. It predicts the occurrence probability of individual species continuously across the landscape based on combinations of limiting environmental gradients. An obvious question is how well this individual species, continuous mapping approach performs in comparison to the classified vegetation maps so commonly used. The most direct comparison is of the stated accuracy of the classified maps to the accuracy of these models. When this comparison is made it is clear that the individual species models produced in this analysis appear to be substantially more effective (Cushman SA, Evans JS, McGarigal K, Do classified maps predict the composition of plant communities? The need for Gleasonian landscape ecology, unpublished data).

Comparison of prediction accuracy between individual-species based maps and classified community type maps is informative, but does not provide a full evaluation. If one's goal is to understand the factors governing the distribution of tree species in a landscape and to produce the most accurate map of their occurrence, then a better comparison would be to evaluate how well the classified maps can predict the occurrence patterns of individual tree species among a large sample of vegetation plots, and compare it to how well species distribution is predicted based on environmental gradients. Comparison of variance explained and model accuracy would provide a means to evaluate how well the classified maps represent the major patterns of tree species distribution and how

well they reflect the action of dominant limiting processes such as climate, disturbance and succession. Such an analysis is beyond the scope of this paper, but would provide an objective measure of the relative success of species-level, pixel scale species models, such as produced here, and patch-scale, community-level maps, such as typically used by managers and landscape scientists, in their ability to describe the patterns of tree species in complex landscapes (see Cushman SA, Evans JS, McGarigal K, Do classified maps predict the composition of plant communities? The need for Gleasonian landscape ecology, unpublished data).

Conclusion

When one adopts a niche-based, individualistic concept of biotic communities it is more appropriate to represent vegetation patterns as continuous measures of site suitability or probability of occupancy, rather than the traditional classification of community types represented in a mosaic of discrete patches Although the problem of categorizations of the landscape failing to represent continuous ecological patterns has been identified, few approaches have been proposed on how to predict gradients in a modeling environment. This analysis shows that a Random Forests ensemble learning approach has very high power to predict site-level probability of occurrence for four tree species based on climatic, topographic and spectral predictor variables. We believe the predictions of these models will be of great use to managers and scientists, as they provide vastly more accurate spatial depiction of vegetation structure across this landscape than has previously been provided by traditional categorical classified community type maps.

References

Austin MP (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol Model 157:101–118

Breiman L (1996) Bagging predictors. Mach Learn 24(2): 123–140

Brieman L (2001a) Statistical modeling: the two cultures. Stat Sci 16(3):199–231. doi:10.1214/ss/1009213726

Breiman L (2001b) Random forests. Mach Learn 45:5–32. doi:10.1023/A:1010933404324

Chavez PS (1988) An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. Remote Sens Environ 24:459–479. doi:10.1016/0034-4257(88)90019-3

Chawla NV, Lazarevic A, Hall LO, Bowyer KW (2003) SMOTEboost: improving prediction of the minority class in boosting. In: 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp 107–119

Chen C, Liaw A, Breiman L (2004) Using random forest to learn imbalanced data. http://oz.berkeley.edu/users/chenchao/666.pdf

Crookston NL, Finley AO (2008) yaImpute: an R package for kNN imputation. J Stat Softw 23:1–16

Curtis JT, McIntosh RP (1951) An upland forest continuum in the prairie-forest border region of Wisconsin. Ecol Monogr 32:476–496

Cushman SA, McKenzie D, Peterson DL, Littell J, McKelvey KS (2007) Research agenda for integrated landscape modelling. USDA Forest Service General Technical Report RMRS-GTR-194

Cutler DR, Edwards TC Jr, Beard KH, Cutler A, Hess KT, Gibson J, Lawler J (2007) Random forests for classification in ecology. Ecology 88:2783–2792. doi:10.1890/07-0539.1

Déath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81:3178–3192

DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the area under tow or more correlated receiver operating characteristics curves: a nonparametric approach. Biometrics 59:837–845. doi:10.2307/2531595

Evans IS (1972) General geomorphometry, derivatives of altitude, and descriptive statistics. In: Chorley RJ (ed) Spatial analysis in geomorphology. Harper & Row, New York, pp 17–90

Forman RTT (1995) Land mosaics: the ecology of landscapes and regions. Cambridge University Press, Cambridge

Forman RTT, Godron M (1986) Landscape ecology. John Wiley & Sons, New York

Freeman EA, Moisen G (2008) Presence absence: an R package for presence absence analysis. J Stat Softw 23(11):31

Fu P, Rich PM (1999) Design and implementation of the solar analyst: an ArcView extension for modeling solar radiation at landscape scales. Proceedings of the 19th Annual ESRI User Conference, San Diego, USA, http://www.esri.com/library/userconf/proc99/proceed/papers/pap867/p867.htm

Gleason HA (1926) The individualistic concept of the plant association. Bull Torrey Bot Club 53:7–26. doi:10.2307/2479933

Guisan A, Zimmermann NE (2000) Predictive habitat distribution model in ecology. Ecol Model 135:147–186. doi:10.1016/S0304-3800(00)00354-9

Hastie T, Tibshirani R, Friedman JH (2001) The elements of statistical learning. Springer, New York

Hutchinson GE (1957) Concluding remarks. Cold Spring Harb Symp Quant Biol 22:415–427

Liaw A, Wiener M (2002) Classification and regression by random forest. R news: the newsletter of the R project (http://cran.r-project.org/doc/Rnews/) 2(3):18–22

Manning AD, Lindenmayer DB, Nix HA (2004) Continua and umwelt: novel perspectives on viewing landscapes. Oikos 104:621–628. doi:10.1111/j.0030-1299.2004.12813.x

McCune B, Keon D (2002) Equations for potential annual direct incident radiation and heat load index. J Veg Sci 13:603–606. doi:10.1658/1100-9233(2002)013[0603:EFPADI]2.0.CO;2

McGarigal K, Cushman SA (2005) The gradient concept of landscape structure. In: Wiens J, Moss M (eds) Issues and perspectives in landscape ecology. Cambridge University Press, Cambridge, pp 112–119

McGarigal K, Tagil S, Cushman SA (2009) Surface metrics: an alternative to patch metrics for the quantification of landscape structure. Landscape Ecol 24:433–450

McIntyre S, Barrett GW (1992) Habitat variegation, an alternative to fragmentation. Conserv Biol 6:146–147. doi:10.1046/j.1523-1739.1992.610146.x

Moore ID, Gessler PE, Nielsen GA, Petersen GA (1993) Terrain attributes: estimation methods and scale effects. In: Jakeman AJ, Beck MB, McAleer M (eds) Modeling change in environmental systems. Wiley, London, pp 189–214

Morrison D (2002) Multivariate statistical methods. McGraw-Hill series in probability and statistics, 4th edn. McGraw-Hill, New York

Murphy MA, Evans JS, Storfer AS (2009) Quantifying Bufo boreas connectivity in Yellowstone National Park with landscape genetics. Ecology (in press)

Nemani R, Pierce L, Running S, Brand L (1993) Forest ecosystem processes at the watershed scale; sensitivity to remotely-sensed leaf-area index estimates. Int J Remote Sens 14:2519–2534. doi:10.1080/01431169308904290

Park Y-S, Chon T-S (2006) Biologically inspired machine learning implemented to ecological informatics. Ecol Inform 203:1–7

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. Ecol Model 190:231–259. doi:10.1016/j.ecolmodel.2005.03.026

Prasad AM, Iverson LR, Liaw A (2006) Random forests for modeling the distribution of tree abundances. Ecosystems (N Y, Print) 9:181–199. doi:10.1007/s10021-005-0054-1

R Development Core Team (2007) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Rabus B, Eineder M, Roth A, Bamler R (2003) The shuttle radar topography mission—a new class of digital elevation models acquired by spaceborne radar. Photogramm Eng Remote Sens 57:241–262. doi:10.1016/S0924-2716(02)00124-7

Rehfeldt GE, Crookston NL, Warwell MV, Evans JS (2006) Empirical analyses of plant–climate relationships for the western United States. Int J Plant Sci 167(6):1123–1150. doi:10.1086/507711

Roberts DW, Cooper SV (1989) Concepts and techniques of vegetation mapping, land classifications based on vegetation: applications for resource management. GTR-INT-257, USDA Forest Service Intermountain Research Station, Ogden, UT, pp 90–96

Schapire R (1990) Strength of weak learnability. J Mach Learn 5:197–227

Stage AR (1976) An expression of the effects of aspect, slope, and habitat type on tree growth. For Sci 22(3):457–460

Stockwell DRB, Peters DP (1999) The GARP modeling system: problems and solutions to automated spatial prediction. Int J Geogr Inf Syst 13:143–158. doi:10.1080/136588199241391

Svetnik V, Liaw A, Tong C, Wang T (2004) Application of Breiman's random forest to modeling structure–activity relationships of pharmaceutical molecules. In: Roli F, Kittler J, Windeatt T (eds) Lecture notes in computer science, vol 3077. Springer, Berlin, pp 334–343

Tarboton DG (1997) A new method for the determination of flow directions and contributing areas in grid digital elevation models. Water Resour Res 33(2):309–319. doi:10.1029/96WR03137

Turner MG, Gardner RH, O'Neill RV (2001) Landscape ecology in theory and practice. Springer-Verlag, New York

Whittaker RH (1967) Gradient analysis of vegetation. Biol Rev Camb Philos Soc 42:207–264. doi:10.1111/j.1469-185X.1967.tb01419.x