

Chapter 8

Modeling Species Distribution and Change Using Random Forest

Jeffrey S. Evans, Melanie A. Murphy, Zachary A. Holden,
and Samuel A. Cushman

8.1 Introduction

Although inference is a critical component in ecological modeling, the balance between accurate predictions and inference is the ultimate goal in ecological studies (Peters 1991; De'ath 2007). Practical applications of ecology in conservation planning, ecosystem assessment, and bio-diversity are highly dependent on very accurate spatial predictions of ecological process and spatial patterns (Millar et al. 2007). However, the complex nature of ecological systems hinders our ability to generate accurate models using the traditional frequentist data model (Breiman 2001a; Austin 2007). Well-defined issues in ecological modeling, such as complex non-linear interactions, spatial autocorrelation, high-dimensionality, non-stationary, historic signal, anisotropy, and scale contribute to problems that the frequentist data model has difficulty addressing (Olden et al. 2008). When one critically evaluates data used in ecological models, rarely do the data meet assumptions of independence, homoscedasticity, and multivariate normality (Breiman 2001a). This has caused constant reevaluation of modeling approaches and the effects of reoccurring issues such as spatial autocorrelation. Model misspecification problems such as the modifiable aerial unit (MAUP) (Cressie 1996; Dungan et al. 2002) and ecological fallacy (Robinson 1950) have also arisen as clearly defined challenges to ecological modeling and inference.

Expert knowledge and well-formulated hypotheses can lend considerable insight into ecological relationships. However, given the complexities in ecological systems, it may be difficult to develop hypotheses or to select variables without first knowing if there is a correlative relationship. These correlative relationships may be highly non-linear, exhibit autocorrelation, be scale dependent, or function as an interaction with another variable. To uncover these relationships non-parametric data mining approaches provide obvious advantages (Olden et al. 2008). Scale can

J.S. Evans (✉)

The Nature Conservancy, North America Science, Fort Collins, CO 80524, USA
e-mail: jeffrey_evans@tnc.org

generate complex interactions in space and time that are inherently unobservable, given standard sample designs and modeling approaches (Wiens 1989; Dungan et al. 2002). Machine learning provides a framework for identifying these variables, building accurate predictions, and exploring mechanistic relationships identified in the model. We advocate performing a critical evaluation of variables used in a model, with careful *a priori* selection of those variables believed to directly relate a proposed explanatory hypothesis linking mechanisms to responses. Developing theories on mechanistic relationships that can account for non-linear variable interaction and process across scale is a particular challenge. Machine learning can provide a starting point for these investigations. The unique advantage of machine learning is that complex relationships and spatial patterns can be discovered more readily than in the traditional probability data model that assumes normality. A scientist should not stop at the identification of key predictor variables using a machine learning approach, but treat this as a starting point to develop and test new theory in an experimental framework.

The issue of machine learning in ecology is as contentious as the frequentists/Bayesian debate (Cressie et al. 2009; Lele and Dennis 2009). We do not wish to inflame this argument, but unfortunately due to the nature of proposing a fundamental paradigm shift in ecological modeling, it is difficult not to offend certain sensibilities. We are not recommending the abandonment of well established methods; quite the contrary. If the goal of an analysis is prediction rather than formal explanation of hypotheses, machine learning provides a set of tools that can dramatically improve results. We hope to provide insight into ways that machine learning can augment our current toolbox and hope that broader application of algorithmic modeling will lead to increased understanding of ecological system and the development of new ecological theory. The goal of this chapter is to illustrate the emerging and potential role that machine learning approaches can assume in ecological studies and to introduce a powerful new model, Random Forest (Breiman 2001b; Cutler et al. 2007; Rogan et al. 2008), that is becoming an important addition to ecological studies. We provide a case study of species distribution modeling using the Random Forest model. As well, we illustrate the utility of Random Forest for exploring the impact of climate change by projecting the model into new climate space.

8.2 Ecological Theory and Statistical Framework

An emerging consensus in quantitative ecology is that spatial complexity across scale fundamentally alters pattern–process relationships (Wiens 1989; Dungan et al. 2002). This has major implications for both research and management. Effective and informed management decisions depend on accurate and precise estimates of current ecological conditions and reliable predictions of future changes (Austin 2007). Given the complexity across space and time inherent in high-dimensional ecological data, there is a critical need for an expanded statistical framework for ecological analysis based on ecological informatics (Park and Chon 2007).

The fusion of individualistic community ecology (Gleason 1926; Curtis and McIntosh 1951; Whittaker 1967) with the Hutchinsonian niche concept (Hutchinson 1957) and its extension to spatially complex and temporally dynamic systems provides an excellent example of the application of powerful non-parametric analytical methods to this new paradigm. Each species responds to local environmental and biotic conditions and the biotic community is an emergent collective of species that are occurring together at a particular place and a particular time due to overlapping tolerances of environmental conditions and vagaries of history, rather than an integrated and deterministic mixture (Whittaker 1967; McGarigal and Cushman 2005). The natural level of focus of such analyses is the species, not community type, assemblage, or patch type; the natural focal scale for such analyses is the location, rather than the stand or patch (McGarigal and Cushman 2005; Cushman et al. 2008).

Adopting a species-level, gradient paradigm based on application of powerful and flexible algorithmic models greatly improves predictions for management, thereby allowing for improved decision making (Evans and Cushman 2009). For example, accurate prediction of current species distributions is a foundation for many management decisions (see Chap. 14). Decisions based on these highly accurate predictions are facilitated by flexible, algorithmic modeling approaches based on combinations of topographic and climatic limiting factors (Evans and Cushman 2009). In contrast with traditional approaches that classify data into categorical community assemblages (i.e., remote sensing) or presence/absence (niche models), predictions that are continuous in nature (i.e., proportion, probability) provide considerable more sensitivity to inference. The very nature of many classification schemas ignores scale, convolves the results, and potentially introduces aggregation errors, reducing the flexibility of data use for meaningful ecological inference (Chesson 1981; Cushman et al. 2008). Given the rapid changes in climate, disturbance regimes, and human impact on ecosystems, reliable prediction of future ecological change is equally important as an understanding of current conditions. The prediction of future changes faces the additional complexity of predicting the decoupling of species responses as responding on an individual level to alterations in limiting factors and disturbance regimes. Species-level predictions are necessary to address this non-equilibrium (Cushman et al. 2008; Evans and Cushman 2009). Whereas reliably predicting complex interactions between species and environmental change over space and time is extremely difficult (Dungan et al. 2002; Austin 2007), algorithmic modeling approaches applied to multi-scale gradient databases provide an effective tool, in that they can address complex interactions, non-linearity, and are ideally suited to modeling individualistic species responses in dynamic, high-dimensional systems (Olden et al. 2008).

A great strength of algorithmic methods, such as Random Forest, is their ability to identify and explore non-intuitive relationships. In traditional inferential approaches the scientist is trained to start by proposing an *a priori* hypothesis relating mechanisms to responses and then using inferential statistics to determine if the hypothesis can be rejected given the observed patterns in data. Whereas this has certain advantages, it does assume that the scientist is able to formulate the correct hypothesis prior to exploring the data (Cook and Campbell 1979). However, nature is full of surprises

and complex ecological systems often behave in non-intuitive ways that defy our *a priori* expectations. Algorithmic methods provide a tool that facilitates development of ecological theory through an iterative process of exploration and discovery, followed by hypothesis generation and testing. Specifically, the flexibility of algorithmic approaches – such as Random Forest – to handle complex, high-dimensional interactions allows them to discover relationships that are hidden in traditional parametric analysis and are unlikely to be proposed *a priori* by a non-omniscient observer. The patterns and relationships identified then provide fertile material that the scientist may use to fashion new explanations and develop new theories.

8.3 Random Forest

8.3.1 *Classification and Regression Trees*

Classification and Regression Trees (CARTs) (Breiman et al. 1983) have gained prominence in both ecology and remote sensing due to their easy of interpretation and ability to address data that interacts in a non-linear or hierarchical manner (Breiman et al. 1983; De'ath and Fabricius 2000; Rogan et al. 2008). CARTs are a binary recursive partitioning approach where the response is iteratively partitioned into nodes based on a measure of impurity (e.g., Gini index, sum of squares, entropy information). Each partition represents exclusive portions (groups) of the variance that are as homogeneous as possible in relation to the response variable. The algorithm identifies the best candidate split (parent node) that minimizes the mean impurity of the two derived child nodes. When no improvement can be made in a given partition a terminal node is defined, thus ending that branch. Splitting is terminated when all observations within each node have the identical distribution of independent variables, making splitting impossible. Cumulatively, nodes represent a set of rules that can be run down the tree to make a prediction.

The hierarchical nature of CARTs is an attractive quality of derived predictions. Local variation is represented within each branch, whereas the global trend is accounted for when the prediction is voted down the tree (Breiman et al. 1983). Although not empirically demonstrated in the literature, we postulate that this balance between global and local trend can account for non-stationarity and anisotropy. There are many additional advantages of CARTs, including the fact that CARTs are non-parametric and not subject to distributional assumptions; do not require transformations; can use categorical, ordinal, and continuous data simultaneously; are invariant to outliers; are capable of identifying and incorporating complex variable interaction and; are capable of handling high-dimensional data. With these advantages come three major drawbacks: (1) CARTs are subject to severe over-fit, (2) due to the sensitivity of CARTs to the complexity (α) parameter, the final tree may not be the optimal solution (Sutton 2005), and (3) CARTs can exhibit high variance and small changes in the data can result in different splits making interpretation somewhat unstable (Hastie et al. 2009). In addition to these

two issues, colinearity between independent variables can cause difficulties in interpretation and bias in the Gini index (Sutton 2005). To address these limitations, ensemble learning approaches including, Bagging (Breiman 1996), Boosting (Freund and Schapire 1996; Friedman 2001; De'ath 2007), and Random Forest (Breiman 2001b) were developed.

8.3.2 Random Forest Algorithm

Random Forest (Breiman 2001b) is an algorithm that developed out of CART and bagging approaches. This algorithm is gaining prominence in remote sensing (Lawrence et al. 2006), forestry (Falkowski et al. 2009), ecology (Cutler et al. 2007; Evans and Cushman 2009; Murphy et al. 2010), and climate change (Prasad et al. 2006; Rehfeldt et al. 2006). By generating a set of weak-learners based on a bootstrap of the data, the algorithm converges on an optimal solution while avoiding issues related to CARTs and parametric statistics. Breiman (2001b) defines Random Forest as a collection of tree-structured weak learners comprised of identically distributed random vectors where each tree contributes to a prediction for x . Ensemble-based weak learning hinges on diversity and minimal correlation between learners. Diversity in Random Forest is obtained through a Bootstrap of training, randomly drawing selection of M (independent variables) at each node (defined as m), and retaining the variable that provides the most information content. To calculate variable importance, improvement in the error is calculated at each node for each randomly selected variable and a ratio is calculated across all nodes in the forest (Fig. 8.1).

The algorithm can be explained by:

1. Iteratively construct N Bootstraps (with replacement) of size n (36%) sampled from Z , where N is number of Bootstrap replicates (trees to grow) and Z is the population to draw a Bootstrap sample from.
2. Grow a random-forest tree T_b at each node randomly select m variables from M to permute through each node to find best split by using the Gini entropy index to assess information content and purity. Grow each tree to full extent with no pruning (e.g., no complexity parameter).
3. Using withheld data (OOB, out-of-bag) to validate each random tree T_b (for classification OOB Error; for regression pseudo R^2 and mean squared error).
4. Output ensemble of random-forest trees

$$\{T_b\}_1^B$$

5. To make a prediction for a new observation x_i :

Regression:

$$\hat{f}_{\text{rf}}^B(x) = \frac{1}{B} \sum_{b=1}^B T_b(x)$$

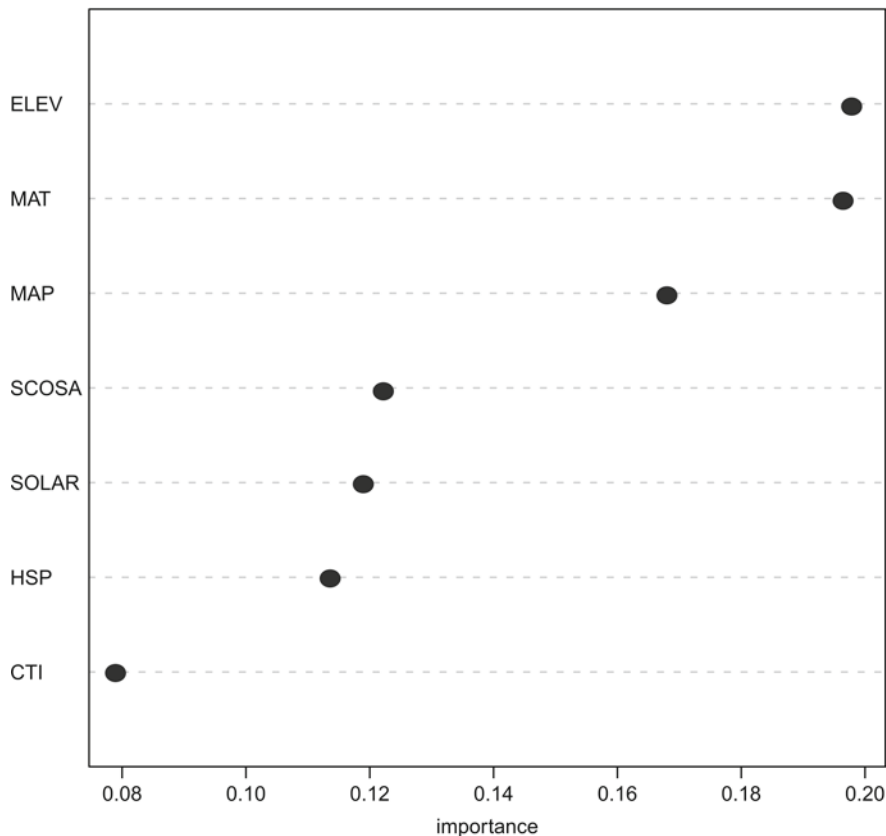


Fig. 8.1 Example of Random Forest scaled variable importance plot using permuted variable mean increase in error for all x variables in *A. lasiocarpa* model

Classification: Let $\hat{C}_b(x)$ be the class prediction of the B th random-forests tree then

$$\hat{C}_{rf}^B(x) = \text{majorityvote} \left\{ \hat{C}_b(x) \right\} \frac{1}{B}$$

Even though the performance of an individual tree will improve given an increase of variables randomly permuted through a node (m), the correlation of trees is increased, reducing the overall performance of the model. Commonly, the optimal m is defined for classification problems as \sqrt{M} ; and for regression $M/3$, where M is a pool of independent variables. It has been demonstrated that Random Forest is robust to noise even given a very large number of independent variables (Breiman 2001a; Hastie et al. 2009). Hastie et al. (2009) showed that with six relevant and 100 noise variables the probability of selecting a relevant variable at any given split is $p=0.46$. Because the algorithm's power is not affected by degrees of freedom it is possible to specify considerable more independent variables (x) than

observations of the dependent variable (y). Whereas Random Forest is not subject to over-fit (Breiman 2001a), caution should be made to not over correlate or inflate the variance of a Random Forest ensemble.

8.3.3 Model Selection

Parsimony is an underlying requirement in the frequentist modeling effort; however, it is rarely addressed in machine learning. The primary motivations for parsimony in frequentist approaches are maintaining well defined hypothesis and reducing the risk of over-fit; the impetus for parsimony in machine learning is motivated by model performance and interpretability (Murphy et al. 2010). Models with fewer variables are much easier to interpret and easier to apply *post-hoc* exploratory techniques to. We have observed another reason for seeking parsimony in a Random Forest models – model performance. In applying model selection we have seen a marked improvement in model fit and predictive performance. There are two explanations for this. First, when Random Forest is run with a large M , but the number of variables that actually provide signal to the data is relatively small, Random Forest is likely to perform poorly with a small m (Hastie et al. 2009). However, if you arbitrarily increase m you risk correlating the ensemble learners. By reducing M to a subset of variables with a signal you improve the overall performance of the model. Second, as spurious variables are removed, trees become much shallower (simpler). This in turn reduces the size of the plurality vote matrix by reducing votes that account for noise, resulting in a higher signal to noise ratio and overall reduction in error (Evans and Cushman 2009; Falkowski et al. 2009; Murphy et al. 2010). Colinearity and multi-colinearity problems can also influence model performance and interpretability (Murphy et al. 2010).

We draw a distinction between variable and model selection. For example, in studies aimed at gene expression, the final goal of utilizing Random Forest is to identify a subset of genes that best describe a particular trait (Díaz-Uriarte and Alvarez de Andrés 2006). In short, the results of the analysis are the final selected variables and not inference or prediction. In ecological models we are not only interested in describing a process but also in inference and prediction. The distinction is drawn around the fact that variable selection approaches are not seeking parsimony but rather reductionism. These approaches are overly aggressive and often result in too few variables to explain a process comprehensively. If a variable has very high explanatory power, Random Forest can exhibit a good fit given a single variable. It is important to seek a parsimonious set of variables, that when predicted to a landscape not only will provide a good fit but also adequately represent the complexities of spatial pattern.

Murphy et al. (2010) developed a model selection approach that uses the permuted variable importance measures and model optimization to select a parsimonious model. The procedure standardizes the importance values to a ratio and iteratively subsets variables within a given ratio, running a new model for each subset of variables. Each resulting model is compared with the original model,

which is held fixed. Model selection is achieved by optimizing model performance based on a minimization of both “out-of-bag” error and largest “within-class” error for classification or maximizing variance explained and minimizing mean squared error in regression. There is also an optional penalty for the number of parameters that will select the model with the fewest number of parameters from closely competing models. There are other simple approaches, including a “leave one variable out” and test performance of sub-models, or simply sub-setting variables based on the importance values and running a single new model with the strongest variables. This may not be as empirically driven as a formalized model selection procedure, but will often result in a more interpretable model.

8.3.4 *Imbalanced Data*

One issue with Random Forest arises within classification problems when classes in categorical response variables are imbalanced (Chen et al. 2004). Imbalances in the response variable result in biased classification accuracy. This is due to the bootstrap over-representing the majority class, leading to under-prediction of the minority class. The resulting model fit is deceptive – exhibiting very small overall OOB error due to very small errors in the majority class as a result of extremely high cross-classification error from the minority-class. With highly skewed data there is a possibility that this same problem could arise in regression problems, but to date there is no published work that has tested this. However, Jiménez-Valverde and Lobo (2006) imply that unbalanced samples are not as serious a problem as historically thought. Despite this, the Bootstrap approach to generate weak-learners in Random Forest causes additional issues not seen in other modeling approaches. Due to minority samples not being drawn with the same frequency as the majority class, a prediction bias is given to the majority class, thus an adequate picture of model fit is not provided. Historically, there are three common ways to address imbalanced data: (1) assign a high cost to misclassification of the minority class (2) down-sample the majority class (Kubat et al. 1998), and (3) over-sample the minority class (Chawla et al. 2002). Chen et al. (2004) proposed the addition of class weights, making Random Forest cost sensitive.

To address this problem, Evans and Cushman (2009) developed a novel approach to balance the response variable that iteratively down-samples the majority class by randomly drawing $2*[n \text{ of minority}]$ from the majority class and running a new Random Forest model iteratively using different random subsets while holding the sample-size of the minority-class constant. To ensure that the distributions of the independent variables in each sub-sample matched distribution in the full data, the covariance of each sub-sampled model is tested against the covariance in full data using a matrix equality test (Morrison 2002). The final ensemble model is built by combining trees from all the resulting down-sampled Random Forest models. Because the underlying theory of Random Forest is ensemble learning, it is possible to combine trees from different models to make a prediction using the combined plurality votes-matrix of the full ensemble (Breiman 2001a).

8.3.5 Model Validation

Two common methods for evaluation models in machine learning classification problems are the Kappa statistic and the Area Under the Curve (AUC) of a Receiver Operator Characteristic (ROC) (Fawcett 2006). ROC is defined as the sensitivity plotted against $[1 - \text{specificity}]$. Sensitivity indicates the proportion of true positives $[a/(a+c)]$ and specificity the proportion of false negatives (commission error) $[b/(b+d)]$. The balance between sensitivity and specificity is a indication of model performance at a class level. The AUC indicates the area under a ROC, ranging from 0 to 1 (so that 0.5 indicates no discrimination and 1.0 perfect classification). Caution should be used when interpreting the ROC/AUC because; (1) error components are equally weighted (Peterson et al. 2008), (2) models can over-value models of rare species (Manel et al. 2001), and (3) certain models do not predict across the spectrum of probabilities violating the assumption that the specificity spans the entire range of probabilities (0–1). Peterson et al. (2008) proposed modifications to ROC by calculating a partial ROC that limits the x -axis to the domain to each specific model. Manel et al. (2001) recommends the Kappa statistic as an alternative to ROC/AUC. The Kappa (Cohen 1960; Monserud and Leemans 1992) evaluates the agreement between classes (binary or multiple) adjusting random chance agreement. Because the Kappa does not account for the expected frequency of a class and does not make distinctions among various types and sources of disagreement the weighted Kappa was developed (Cohen 1968) The incorporation of a weight allows for near agreement and adjusts for expectation in the frequency of observations.

It has continually been stated that one compelling component of Random Forest is that there is no need for independent validation. The model error is assessed against the OOB data in each Bootstrap replicate, providing an error distribution. Although we agree with the robustness of this approach to test model sensitivity to sample distribution, we also advocate the addition of simple data-withhold cross validation techniques. An additional validation procedure we commonly apply is a randomization procedure where the independent data are randomized, and the model is run and the error tabulated. This is performed a large number of times (i.e., 1,000), providing an error distribution that the model under scrutiny can be compared to, thus providing a significance value (Evans and Cushman 2009; Murphy et al. 2010).

8.3.6 Visualization

There are many visualization techniques that can be employed to explore mechanistic relationships, variable interaction, and model performance. Conditional density estimates (Hall et al. 1999; Falkowski et al. 2009) (Fig. 8.2a), partial dependence plots (Friedman 2001) (Fig. 8.2b), bivariate kernel density estimates (Simonoff 1998) (Fig. 8.3), and multi-dimensional scaling (Cox and Cox 1994) plots to explore mechanistic relationships, both before and after analysis. With discrete data, conditional density plots can be used to explore the influence of a given independent variable on a set of responses. In our example we use elevation

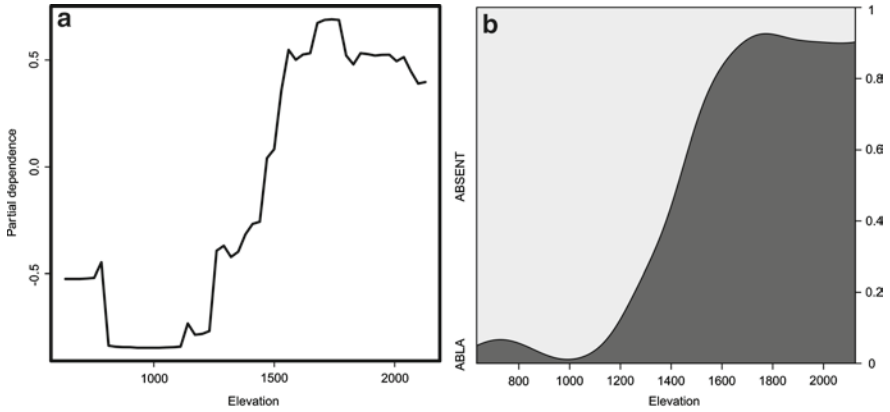


Fig. 8.2 Methods for exploring mechanistic relationships and variable interaction in categorical data. (a) Partial plot of elevation and presence of *A. lasiocarpa* and (b) conditional density plot of elevation of presence/absence for *A. lasiocarpa*

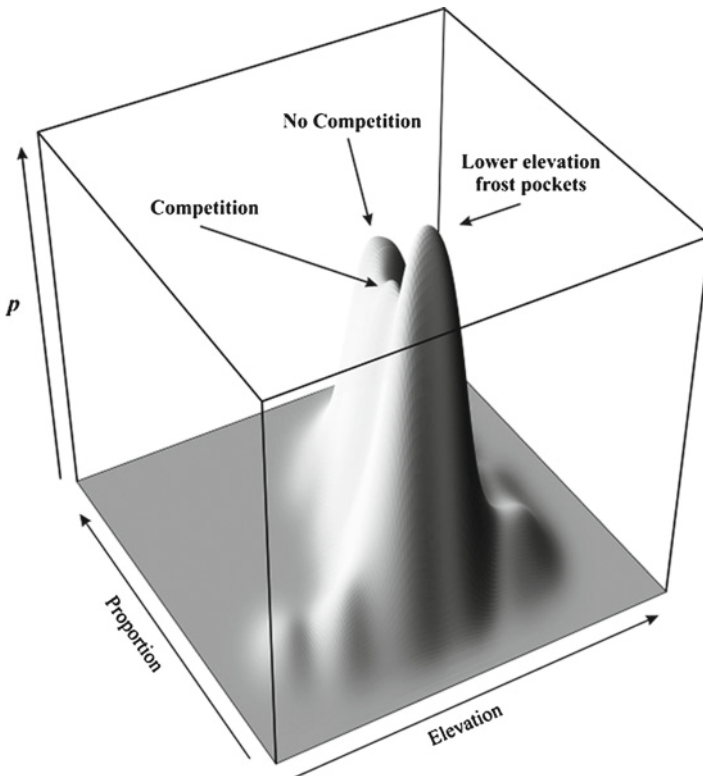


Fig. 8.3 Method for exploring the mechanistic relationships in continuous data. Non-parametric bivariate Kernel Density Estimate (KDE) plot of proportion of *A. lasiocarpa* and elevation, illustrating competition with *Picea engelmannii* (Engelmann spruce) across an elevational gradient

and examine its influence on presence/absence of *Abies lasiocarpa* (subalpine fir) (Fig. 8.2b). With continuous independent and response variables, nonparametric bivariate and multivariate kernel density estimates can be used (Fig. 8.3). This is an effective exploratory method for visualizing gradient relationships.

By partialing out the average effect of all other variables we can explore the influence of a given variable on the probability of occurrence of this species (Fig. 8.2a). This method can also be applied to multiple independent variables to explore variable interaction (Cutler et al. 2007). A proximity matrix is created within Random Forest by running in-bag and out-of-bag down each tree, if cases 1 and 2 both result in the same terminal node, then the proximity is increased by one.

The normalized proximity is calculated by dividing by the number of trees in the ensemble. The pairwise values in the proximity matrix can be treated as a dissimilarity or distance measure. This measure of dissimilarity can be used to visualize the separability of classes using multi-dimensional scaling plots. Crookston and Finley (2008) derived a method for using the scaled proximities to perform nearest neighbor multiple-imputation.

8.3.7 *Spatial Structure*

It should be noted that although non-parametric models do not assume independence and thus, are not affected by spatial-autocorrelation, they also do not explicitly incorporate spatial structure. There are a few proposed methods to account for spatial structure that range from naïve trend to direct incorporation of spatial structure. The most simple approach is to incorporate geographic coordinates or a variable that indicate the trend of geographic space (Legendre and Legendre 1998; Chefaoui and Lobo 2007). More complex approaches include incorporating a function into a model that allows for a spatial lag, effectively acting as a die-off function (Mouer and Riemann 1999), or adding a distance matrix of observed samples to act as a spatial weight (Allouche et al. 2008).

8.4 Data Suitability

8.4.1 *Dependent Variable*

Statistical issues relating to spatial aggregated data, such as photo-interpreted stands or landcover classifications based on remote sensing efforts, have been emerging (Chesson 1981; Cressie 1996; Dungan et al. 2002; Cushman et al. 2008), justifying a trend toward individual species gradient models. Attempting to use polygon data to build niche models overly smoothes or misrepresents the underlying variation of variables used to construct the niche hypervolume. To support a gradient modeling approach, spatially-referenced plot level data are necessary

(Evans and Cushman 2009). A sample design that captures spatial and statistical variability in both dependent and independent variables is critical to ensure that the model provides an adequate representation of the ecological niche and is capable of landscape level prediction. Unfortunately, few data-collection efforts are comprehensive or designed to capture fine-scale spatial variability over entire landscapes, making it necessary to plan extensive data collection efforts into a study.

8.4.2 Independent (Predictor) Variables

In mountainous terrains patterns of precipitation and topography are the primary drivers of species occurrence and of the formation of plant communities (Whittaker and Niering 1975; Costa et al. 2008). Precipitation and timing interact with a variety of processes (i.e., temperature, solar radiation) to determine temperature and moisture regimes. These, in combination with biotic factors, competition, and disturbance, largely determine vegetation composition and structure.

Independent variables that influence species occurrence can be grouped into direct and indirect predictors (Guisan and Zimmermann 2000). Direct predictors include soil characteristics, temperature, precipitation and solar radiation – which directly influence characteristics of the physical environment and its suitability for individual species. Indirect predictors include geomorphometric surrogates such as elevation, aspect, slope, and slope position and are effective surrogates for some of the driving physical variables that influence vegetation distribution and abundance. Direct predictors have several critical advantages over indirect measures. First, using indirect geomorphometric surrogates rather than limiting variables such as temperature add an extra inferential step in interpreting species–environment relations. Second, and most importantly; the use of direct measures allow for projection into future climate space.

8.5 Prediction of Current and Future Species Distributions

In this case study, we demonstrate the application of the Random Forest algorithm for predicting current and potential future distribution of plant species (Prasad et al. 2006; Rehfeldt et al. 2006). We focus on two plant species, *A. lasiocarpa* (subalpine fir) and *Pseudotsuga menziesii* (Douglas fir). We used 30-year normalized mean annual temperature (MAT) and precipitation (MAP) predictions from the spline climate model presented in Rehfeldt et al. (2006). We combine the two climate variables with variables describing slope position (HSP, Murphy et al. 2010), long-wave solar radiation (INSO, Fu and Rich 1999), a slope/aspect transformation (SCOSA, Stage 1976), and a wetness index (CTI, Moore et al. 1993) to develop a limiting-factor niche model for these two species. Finally, we apply these niche models across a topographically complex landscape to predict the occurrence of

these species. We characterize the structure of the realized niche of these species and develop fine-scale species distribution maps based on predicted climate–species relationships. We focus on species-level predictions because multi-species, community level analyses fail to optimally predict any given species and may not be useful for extrapolation into novel climates where communities plausibly disassemble due to the trajectories of individual species responses.

8.5.1 Study Area

We use species occurrence and abundance data from 411 vegetation plots systematically spaced at 1.7-km intervals across the Bonners Ferry Ranger District, on the Panhandle National Forest in Northern Idaho, USA. These plots were established by the US Forest Service as part of a pilot project aimed to intensify the USFS Forest Inventory and Analysis grid. Our 3,883 km² study area encompasses portions of the Selkirk and Purcell mountain ranges, with elevations ranging from 630 to 2,600 m. *Pinus ponderosa* and *P. menziesii* occur at lower elevations and south-facing slopes and *Picea engelmannii* and *A. lasiocarpa* occupy higher elevation sites. The study area and vegetation sampling methods are described in detail by Evans and Cushman (2009).

8.6 Methods

We built our Random Forest model using the Random Forest (Liaw and Wiener 2002) package available in R (R Development Core Team 2009). This allowed us the flexibility to program customized model selection (Murphy et al. 2010), validation (Evans and Cushman 2009), and visualization routines. We focused directly on the key limiting environmental gradients of temperature, solar energy, and water availability, rather than variables that would provide indirect measures of process. We predicted the distribution of the two tree species at fine spatial scales (30 m²) corresponding to the dominant scale at which species interact with limiting environmental resources. By constructing species-level, fine-grain, limiting-factor predictive models, we hope to contribute to improved accuracy of predicting the distributional shifts in species occurrence with the advance of changing climate.

To predict the distribution of species under potential future climates, we fixed the contemporary (2000) climate-niche model and predicted the model into new climate space (2040 and 2080). We perturbed our two climate variables using a weighted mean from 20 GCMs (General Circulation Models) (McGuffie and Henderson-Sellers 1997). For the 2040 climate we increased the temperature by 2°C and precipitation by 2%, for the 2080 climate we increased temperature by 3.39°C and precipitation by 4.3%. We are, in effect, shifting the climate niche space and projecting the resulting changes in distribution.

8.6.1 Uncertainty

There is very large variation in GCMs ranging from positive to negative change. This provides considerably uncertainty in future climate spaces. A common approach to address this variation is to derive a weighted mean across all 20 GCMs. Although justifiable, the weighted mean collapses variability among model scenarios, thus affecting model sensitivity. To evaluate model sensitivity in future climate spaces, we implemented a Monte Carlo simulation, generating a simulation envelope through randomization of climate variables. We used the range of 20 GCMs to parameterize the simulation for each year; 2040 min-temp=+1.26, 2040 max-temp=+2.9, 2040 min-precipitation=-8.9%, 2040 max-precipitation=+9%, 2080 min-temp=+1.91, 2080 max-temp=+5.93, 2080 min-precipitation=-9.9%, 2080 max-precipitation=+19.6%. The climate data were randomized ($n=9,999$) with the GCM ensemble range for each year and the Random Forest model representing the contemporary climate re-predicted into the randomized climate space. The simulation envelope was created using the predicted probability from each randomization.

8.7 Results

The results for the contemporary distribution model (Fig. 8.4a, b) provided very good model fits for both species. The OOB error for the *A. lasiocarpa* model was 10.11% with classification error equally balanced between presence and absence classes. Back-prediction to the data demonstrated an almost perfect fit with AUC of 0.99 and a 1% error rate. The *P. menziesii* model exhibited a 14.83% OOB error with a slightly higher (4%) error rate for the presence class. Back-prediction provided an AUC of 0.98 and a 3% error rate. One-thousand cross-validations with a 10% data-withhold provided <2% error rates for both models.

The two future-climate projections provided intuitive results very consistent with the ecological gradient of each species (Fig. 8.4). Both species demonstrate contraction of the predicted distribution (Fig. 8.4). The 2040 *A. lasiocarpa* projection (Fig. 8.4c) exhibited a 27.45% contraction and a 36.40% contraction for 2080 (Fig. 8.4e). Both time-steps show *A. lasiocarpa* receding into higher elevations. *A. lasiocarpa* is highly influenced by both temperature and moisture, as it prefers cold temperatures with high levels of precipitation. With increased temperature, *A. lasiocarpa* is constrained to higher elevations than its current range. The *P. menziesii* results are considerably more dramatic, showing a 66% decrease in 2040 (Fig. 8.4d) and an 80.92% decrease in 2080 (Fig. 8.4f). Across its range *P. menziesii* is fairly opportunistic, capable of occupying a wide range of environment conditions. In our study area *P. menziesii* occurs in a very hot portion of its gradient. With increased temperatures, *P. menziesii* exceeds its temperature tolerance in short order, showing considerable loss by 2040 (Fig. 8.4d).

The Monte Carlo simulation indicated that there is considerable uncertainty using the weighted-mean of the GCMs in the 2040 time period (Fig. 8.5a, c).



Fig. 8.4 (a) *A. lasiocarpa* contemporary (2000) climate presence, (b) *P. menziesii* contemporary (2000) climate presence, (c) Projected *A. lasiocarpa* 2040 climate presence, (d) Projected *P. menziesii* 2040 climate presence, (e) Projected *A. lasiocarpa* 2080 climate presence, (f) Projected *P. menziesii* 2080 climate presence

The 2040 predicted probability distribution is very poorly matched for the *P. menziesii* model (Fig. 8.5c) exhibiting an extremely mismatched distribution well outside the simulation envelope. The 2040 *A. lasiocarpa* model exhibited the correct distributional shape but falls outside the simulation envelope (Fig. 8.5a). In the 2080 *P. menziesii* projection, the simulation envelope shows considerable stochasticity with the projected probability distribution matching the shape but falling outside the envelope (Fig. 8.5d). In contrast, the *A. lasiocarpa* model for 2080 shows a projection well matched with the simulation envelope (Fig. 8.5b). Overall results demonstrate that there is a certain amount of uncertainty in using the weighted-mean GCM values in all but the *A. lasiocarpa* 2080 model (Fig. 8.5b).

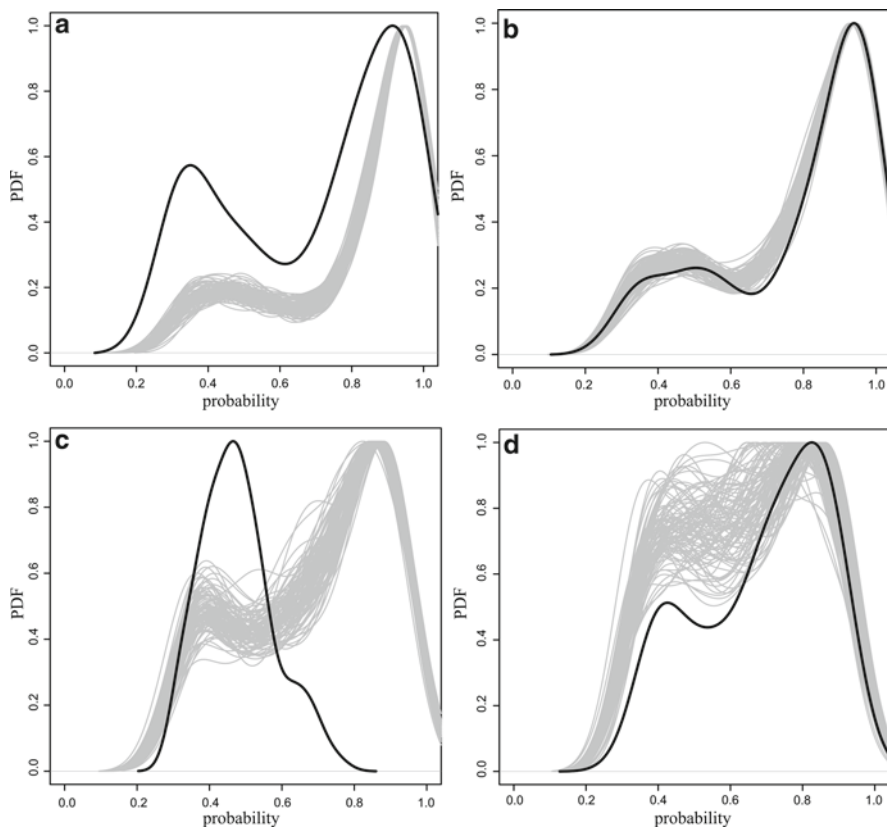


Fig. 8.5 Monte Carlo simulation of GCM's to quantify uncertainty. (a) *A. lasiocarpa* 2040 simulated probabilities, (b) *A. lasiocarpa* 2080 simulated probabilities, (c) *P. menziesii* 2040 simulated probabilities, (d) *P. menziesii* 2080 simulated probabilities. The black line represents the probability distribution using the weighted-mean GCM and the grey lines represent the simulation envelope probability distributions from the Monte Carlo

This uncertainty is acceptable in all but the *P. menziesii* 2040 model (Fig. 8.5c) that exhibits a very poorly matched distribution. This simple implementation of Monte Carlo simulation allowed us to examine the stability of our projections into future climate space. These results illustrate the necessity of examining the effect of selected values used to perturb climate.

To illustrate the utility of nonparametric bivariate kernel density estimates we examined an environmental gradient of *A. lasiocarpa* proportion and elevation (Fig. 8.3). It is clear that elevation highly influences the proportion of *A. lasiocarpa*; however, there is an obvious discontinuity in the mid-range of the elevational gradient. Exploration of this discontinuity uncovered a competition with *Picea engelmannii* (Engelmann spruce). Occurrence of mid-elevation frost pockets exclude *P. engelmannii* in the lower elevation limits of *A. lasiocarpa*. The conditional density and partial plots (Fig. 8.2) both support this inference.

8.8 Discussion

The details of landscape structure influence relationships between forest ecosystems, climate, and disturbance regimes in complex and interacting ways. Fine-scale environmental structure has a strong influence on species distribution, dominance, and succession (Whitaker 1967; Tilman 1982; ter Braak and Prentice 2004). The biophysical context of a location within a landscape also strongly influences growth rates and regeneration (Bunn et al. 2005). Furthermore, the probability of disturbances (Runkle 1985; Risser 1987), and the patterns of recovery (Finegan 1984; Glenn and Collins 1992) are strongly dependent on the pattern of environmental variation across the landscape. In addition, each species has an ecological response to variations in these environmental conditions, characterized by its ecological niche.

In this case study, we adopt a gradient perspective in lieu of hierarchical models of system organization (McGarigal et al. 2009). With a gradient perspective, a system of hierarchically organized aggregate subsystems is not assumed. Rather, emphasis is on directly measuring the response variables and the factors that drive their behavior, and modeling the relationships between them across space at the dominant scale of their operational environment.

By focusing directly on fine-scale, species-level responses to limiting ecological factors, we can describe interesting details of the niche structure of these species. For example, our analysis shows that, contrary to much of classic ecological theory, many species exhibit multimodal niche structure along limiting environmental gradients. In addition, our analysis enables quantification of the degree of niche overlap and environmental partitioning among species. By focusing on the species-level to identify limiting ecological variables, we feel these results allow for more reliable projection into future environments with altered climate regimes. Had we only focused on measures of the *abiotic* substratum rather than *biotic* limiting factors such as temperature, it would have been virtually impossible to perturb these variables to represent future climate space. It should be noted that our analysis was designed to illustrate the utility of nonparametric modeling methods. There are severe limitations to envelope approaches that are related to scale mismatches between fine-scale species responses and coarse-scale GCM climate models (Randin et al. 2009; Willis and Bhagwat 2009). These limitations are somewhat mitigated by the inclusion of geomorphometric variables that represent micro-topographic influences that interact with climate so as to better represent fine-scale spatial variation.

Resource management decisions should be made in the context of potential future climate change impacts. Dynamic, climate-adapted, spatial predictions of vegetation, based on the physical variables that ultimately drive species occurrence, are needed to understand and predict changes in future vegetation. The analysis presented here represents an important step toward developing tools that will allow us to accurately predict changes in species distribution with continuing climate change. Vegetation distribution is already changing (Rehfeldt et al. 2006; McKenney et al. 2007; Iverson et al. 2008) in response to changes in climate. However, precisely how an individual species will move across the landscape at local scales while interacting with the

environment and disturbance is unknown. A key step in this direction is developing high-resolution models that can integrate highly accurate, fine-scale, species-level vegetation predictions with associated future climate projections. Whereas the spatial distributions of vegetation will certainly shift, many variables may interact in non-linear ways to influence the potential suitability of the biophysical environment for an individual plant species. In the two species presented in this chapter, we observed that responses to climate change can be highly variable and species specific (Willis and Bhagwat 2009), contradicting the notion that vegetation communities respond to climate change in unison. Machine learning approaches that can account for the many complexities in these systems will allow us to explore the potential impacts of changing climates on future tree species distributions at scales that match the dominant biological governing processes, such as dispersal, initiation, and growth.

Acknowledgments Funding for this research was provided by the USDA Forest Service, Rocky Mountain Research Station and The Nature Conservancy. The authors would like to thank G. Rehfeldt, A. Hudak, N. Crookston, L. Iverson, and A. Cutler for valuable discussion on Random Forest and species distribution modeling and A. Prasad, J. Kiesecker and two anonymous reviewers for comments that strengthened this chapter. Additionally we would like to thank the editors for their patience and perseverance in seeing this book published.

References

- Allouche O, Steinitz O, Rotem D, Rosenfeld A, Kadmon R (2008) Incorporating distance constraints into species distribution models. *J Appl Ecol* 45:599–609.
- Austin M (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecol Modell* 200:1–19.
- Breiman L, Friedman JH, Olshen RA, Stone CJ (1983) *Classification and regression trees*. Wadsworth, London.
- Breiman L (1996) Bagging predictors. *Mach Learn* 24:123–140.
- Breiman L (2001a) Statistical modeling: the two cultures. *Stat Sci* 16:199–231.
- Breiman L (2001b) Random forests. *Mach Learn* 45:5–32.
- Bunn AG, Graumlich LJ, Urban DL (2005) Trends in twentieth-century tree growth at high elevations in the Sierra Nevada and White Mountains, USA. *Holocene* 15:481–488.
- Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority oversampling technique. *J Artif Intell Res* 16:321–357.
- Chefaoui RM, Lobo JM (2007) Assessing the conservation status of an Iberian moth using pseudo-absences. *J Wildl Manage* 71:2507–2516.
- Chen C, Liaw A, Breiman L (2004) Using random forest to learn imbalanced data. Technical Report 666. Statistics Department, University of California, Berkeley.
- Chesson PL (1981) Models for spatially distributed populations: the effect of within-patch variability. *Theor Popul Biol* 19:288–325.
- Cohen J (1960) A coefficient of agreement for nominal scales. *Educ Psychol Meas* 20:37–46.
- Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 70:213–20.
- Cook TD, Campbell DT (1979) *Quasi-experimentation: design and analysis issues for field settings*. Houghton Mifflin, Boston.
- Costa GC, Wolfe C, Shepard DB, Caldwell JP, Vitt LJ (2008) Detecting the influence of climate variables on species distribution: a test using GIS niche-based models along a steep longitudinal environmental gradient. *J Biogeogr* 35:637–646.

- Cox TF, Cox MAA (1994) Multidimensional scaling. Chapman and Hall, Boca Raton.
- Cressie N (1996) Change of support and the modifiable areal unit problem. *Geogr Syst* 3:159–180.
- Cressie N, Calder CA, Clarke JS, Ver Hoef JM, Wikle CK (2009) Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecol Appl* 19:553–570.
- 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. *Ecology* 32:476–496.
- Cushman SA, McKelvey K, Flather C, McGarigal K (2008) Do forest community types provide a sufficient basis to evaluate biological diversity? *Front Ecol Environ* 6:13–17.
- 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.
- De'ath G (2007) Boosted trees for ecological modeling and prediction. *Ecology* 88:243–251.
- De'ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* 81:3178–3192.
- Díaz-Uriarte R, Alvarez de Andrés SA (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7:3.
- Dungan JL, Pery JN, Dale MRT, Legendre P, Citron-Pousty S, Fortin MJ, Jakomulska A, Miriti M, Rosenberg MS (2002) A balanced view of scale in spatial statistical analysis. *Ecography* 25:626–240.
- Evans JS, Cushman SA (2009) Gradient modeling of conifer species using random forests. *Landsc Ecol* 24:673–683.
- Falkowski MJ, Evans JS, Martinuzzi S, Gessler PE, Hudak AT (2009) Characterizing forest succession with lidar data: an evaluation for the inland Northwest, USA. *Remote Sens Environ* 113:946–956.
- Fawcett T (2006). An introduction to ROC analysis. *Pattern Recognit Lett* 27:861–874.
- Finegan B (1984) Forest succession. *Nature* 312:109–114.
- Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Saitta L (ed) *Machine learning: proceedings of the thirteenth international conference*. Morgan Kaufmann, San Francisco.
- Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Ann Stat* 29:1189–1232.
- Fu P, Rich PM (1999) Design and implementation of the Solar Analyst: an ArcView extension for modeling solar radiation at landscape scales. In: *Proceedings of the 19th annual ESRI User Conference*, San Diego.
- Gleason HA (1926) The individualistic concept of the plant association. *Bull Torrey Bot Club* 53:7–26.
- Glenn RH, Collins SL (1992) Effects of scale and disturbance on rates of immigration and extinction of species in prairies. *Oikos* 63:273–280.
- Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. *Ecol Modell* 135:147–186.
- Hall P, Wolff RCL, Yao Q (1999) Methods for estimating a conditional distribution function. *J Am Stat Assoc* 94:154–163.
- Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical learning: data mining, inference, and prediction*. 2nd edition. Springer, New York.
- Hutchinson GE (1957) Concluding remarks. *Cold Spring Harb Symp Quant Biol* 22:415–427.
- Iverson LR, Prasad AM, Matthews SN, Peters M (2008) Estimating potential habitat for 134 eastern US tree species under six climate scenarios. *For Ecol Manage* 254:390–406.
- Jiménez-Valverde A, Lobo JM (2006) The ghost of unbalanced species distribution data in geographic model predictions. *Divers Distrib* 12:521–524.
- Kubat M, Holte RC, Matwin S (1998) Machine learning for the detection of oil spills in satellite radar images. *Mach Learn* 30:195–215.
- Lawrence RL, Wood SD, Sheley RL (2006) Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomForest). *Remote Sens Environ* 100:356–362.

- Legendre P, Legendre L (1998) Numerical ecology. Elsevier, Amsterdam.
- Lele SR, Dennis B (2009) Bayesian methods for hierarchical models: are ecologists making a Faustian bargain? *Ecol Appl* 19:581–584.
- Liaw A, Wiener M (2002) Classification and regression by Random Forest. *R News* 2:18–22.
- Manel S, Williams HC, Ormerod SJ (2001) Evaluating presence-absence models in ecology: the need to account for prevalence. *J Appl Ecol* 38:921–931.
- 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.
- McGarigal K, Tagil S, Cushman SA (2009) Surface metrics: an alternative to patch metrics for the quantification of landscape structure. *Landsc Ecol* 24:433–450.
- McGuffie K, Henderson-Sellers A (1997) A climate modelling primer. John Wiley & Sons, Chichester.
- McKenney DW, Pedlar JH, Lawrence K, Campbell K, Hutchinson MF (2007) Potential impacts of climate change on the distribution of North American trees. *BioScience* 57:939–948.
- Millar CI, Stephenson NL, Stephens SL (2007) Climate change and forests of the future: managing in the face of uncertainty. *Ecol Appl* 17:2145–2151.
- Monserud RA, Leemans R (1992) Comparing global vegetation maps with the Kappa statistic. *Ecol Modell* 62:275–293.
- Moore ID, Gessler P, Nielsen GA, Peterson GA (1993) Terrain attributes: estimation and scale effects. In Jakeman AJ, Beck MB, McAleer M (eds) Modelling change in environmental systems. John Wiley & Sons, Chichester.
- Morrison D (2002). Multivariate statistical methods. 4th edition. McGraw-Hill series in probability & statistics. McGraw-Hill, New York.
- Mouer MH, Riemann R (1999) Preserving spatial and attribute correlation in the interpolation of forest inventory data. In: Lowell K, Jaton A (eds) Spatial accuracy assessment: land information uncertainty in natural resources. Ann Arbor Press, Chelsea.
- Murphy MA, Evans JS, Storfer AS (2010) Quantifying *Bufo boreas* connectivity in Yellowstone National Park with landscape genetics. *Ecology* 91:252–261.
- Olden JD, Lawler JJ, Poff NL (2008) Machine learning methods without tears: a primer for ecologists. *Q Rev Biol* 83:171–193.
- Park YS, Chon TS (2007) Biologically inspired machine learning implemented to ecological informatics. *Ecol Modell* 203:1–7.
- Peters RH (1991) A critique for ecology. Cambridge University Press, Cambridge.
- Peterson AT, Papes M, Soberón J (2008) Rethinking receiver operating characteristic analysis applications in ecological modelling. *Ecol Modell* 213:63–72.
- Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* 9:181–199.
- R Development Core Team (2009). 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>.
- Randin CF, Engler R, Normand S, Zappa M, Zimmermann N, Pearman PB, Vittoz P, Thullier W, Guisan A (2009) Climate change and plant distribution: local models predict high-elevation persistence. *Glob Chang Biol* 15:1557–1569.
- Rehfeldt GE, Crookston NL, Warwell MV, Evans JS (2006) Empirical analyses of plant-climate relationships for the western United States. *Int J Plant Sci* 167:1123–1150.
- Risser PG (1987) Landscape ecology: state of the art. In: Turner MG (ed) Landscape heterogeneity and disturbance. Springer-Verlag, New York.
- Robinson WS (1950) Ecological correlations and the behavior of individuals. *Am Sociol Rev* 15:351–357.
- Rogan J, Franklin J, Stow D, Miller J, Woodcock C, Roberts D (2008) Mapping land-cover modification over large areas: a comparison of machine learning algorithms. *Remote Sens Environ* 112:2272–2283.
- Runkle JR (1985) Disturbance regimes in temperate forests. In: Pickett STA, White PS (eds) The ecology of natural disturbance and patch dynamics. Academic Press, New York.

- Simonoff JS (1998) Smoothing methods in statistics. Springer-Verlag, New York.
- Stage A (1976) An expression for the effect of aspect, slope and habitat type on tree growth. *For Sci* 22:457–460.
- Sutton CD (2005) Classification and regression trees, bagging, and boosting. In: Rao CR, Wegman EJ, Solka JL (eds) *Handbook of statistics: data mining and data visualization*, Volume 24. Elsevier, Amsterdam.
- ter Braak CJF, Prentice IC (2004) A theory of gradient analysis. *Adv Ecol Res* 34:235–282.
- Tilman D (1982) *Resource competition and community structure*. Princeton University Press, Princeton.
- Whittaker RH (1967) Gradient analysis of vegetation. *Biol Rev* 42:207–264.
- Whittaker RH, Niering WA (1975) Vegetation of the Santa Catalina mountains, Arizona. V. biomass, production and diversity along the elevation gradient. *Ecology* 56:771–790.
- Wiens JA (1989) Spatial scaling in ecology. *Funct Ecol* 3:385–397.
- Willis KJ, Bhagwat SA (2009) Biodiversity and climate change. *Science* 326:806–807.