Chapter 6
Machine Learning for Macroscale Ecological Niche Modeling - a Multi-Model, Multi-Response Ensemble Technique for Tree Species Management Under Climate Change

Anantha M. Prasad

6.1 Introduction

Machine learning has come a long way in recent decades due to huge increases in computing power and the availability of robust public platforms for statistical analysis (e.g., R Core Team 2016). Machine learning techniques have benefited from advances in statistical learning and vice versa (Hastie et al. 2009; Slavakis et al. 2014), resulting in impressive applications of big data in imaging, astronomy, medicine, finance and to a lesser extent in ecology (Van Horn and Toga 2014; Zhang and Zhao 2015; Belle et al. 2015; Hussain and Prieto 2016; Hampton et al. 2013). A healthy relationship with computer science and engineering has invigorated the field even more, resulting in a variety of techniques suitable for diverse applications. One successful and frequently used method is ensemble learning, where learning algorithms independently construct a set of classifiers or regression-estimates and classify or regress newer data points by either taking a weighted vote (classifiers) or an average (regression) of their predictions (Zhou 2012).

A majority of the ensemble learning problems deal with classification due to the binary, or in some cases multinomial, response that is of interest. However, in the field of ecology, and especially in tree species abundance modeling, we have access to continuous data thanks to the Forest Inventory Analysis (FIA) in the United States (Woudenberg et al. 2010) that lends itself to a regression approach. Valuable information can be lost if the continuous data are classified *a priori* into classes. Therefore, it is best to solve the problem in a regression context, and classify the results later to retain most of the information in the response. I will choose the regression approach for this reason and also to highlight this less used aspect of statistical learning.

A. M. Prasad (✉)
Research Ecologist, USDA Forest Service, Northern Research Station, Delaware, OH, USA
e-mail: aprasad@fs.fed.us

© Springer Nature Switzerland AG 2018

G. R. W. Humphries et al. (eds.), *Machine Learning for Ecology and Sustainable Natural Resource Management*, https://doi.org/10.1007/978-3-319-96978-7_6

Modeling the abundance response of trees under current and future climates is an exercise fraught with assumptions and uncertainties due to the dynamic nature of the species' range boundaries. We are essentially capturing a slice in the eco-evolutionary history of the species and trying to project it into future climatic space as forecast by the general circulation models (GCMs; McGuffie and Henderson-Sellers 2014). Of the many uncertainties, the non-equilibrium nature of the tree species (they could still be expanding their ranges and not yet have achieved climatic equilibria) (Garcia-Valdes et al. 2013), and inability to capture biotic interactions (Belmaker et al. 2015) are cited most often. These limitations, however, are often due to the scale of analysis; a macroscale analysis will typically include biotic interactions as an emergent phenomenon. Only finer scale analysis can deal with biotic interactions in a more fundamental way. However, the question of species non-equilibrium also affects macroscale studies because of the historical nature of eco-evolutionary processes and can be addressed to some extent by comparing various studies as slices in time (Prasad 2015).

Of the many techniques that have emerged in recent years (Iverson et al. 2016), ensemble techniques based on decision trees have become the most popular among ecologists modeling niche related phenomena (Galelli and Castelletti 2013; Hill et al. 2017; Vincenzia et al. 2011). The transition from more parametric analysis like generalized linear and additive models (glm, gam and shrinkage based regression) to decision tree based techniques has to do mainly with the nature of ecological systems. They tend to be high dimensional and nonlinear with many embedded interactions; all of which are handled well by decision tree based techniques (Guisan et al. 2002; Guisan and Thuiller 2005). Hence a multitude of techniques have evolved, each appropriate for a subset of problems and dealing mostly with various shortcomings arising from more conventional decision-tree based techniques like bagging, randomized trees and boosting (Elith et al. 2010).

As datasets have become larger and easier to acquire (large scale inventories, digital elevation models, satellite imagery, demographic financial data, to name a few) with a corresponding increase in computing power, there has been a movement away from more parametric forms of analysis towards computationally intensive machine learning, such as non-parametric methods that are flexible and data-driven. While older constraints based on limited data and computing power have relaxed, newer ones have emerged because the analysis has moved more into the "prediction" space (e.g., models that overfit because of non-optimal variance-bias ratio). These newer challenges are being addressed via increasingly sophisticated algorithms that combine flexible models with resampling, permuting, shrinkage and regularization techniques (Tibshirani 1996; Zou and Hastie 2005; Hastie et al. 2009).

The focus of this chapter is to show how to tackle these issues when modeling the abundance of tree species at a macroscale (20 km resolution) in the eastern United States (where we have sufficiently large predictor and response data), and also, how to address the problems of model reliability and prediction confidence while interpreting the results. Towards this goal, I develop a multi-response, multi-model ensemble technique that addresses problems of bias, variance and output noise – resulting in more reliable prediction.

6.2 Controlling Bias and Variance

Some ecological projects are fortunate to have large amounts of data at their disposal while other studies fall into the category of designed experiments where data collection can be cumbersome and costly. Large Data projects are typically those that use datasets that are large and complex, of fairly coarse resolution, and are already available (e.g., remotely sensed topography and land-use, climate, soils, national forest inventory plots and bird surveys). Niche based analyses of these data lend themselves well to statistical machine learning techniques, unlike studies that require formal experimental design, which may be more appropriate for parametric statistical analyses. The existence of Large Data begs for a data-driven approach with complex and flexible models that capture nonlinearities and interactions well and can screen out less important predictors. However, this flexibility can result in overfitting and attendant variance; the models may fit the training data well, but not generalize well to newer prediction space (Domingos 2012; Merow et al. 2014). In statistical terms, these models have low bias (good) but high variance (not good). If bias is too high, the models are less likely to fit the underlying data (think straight line fitting curvilinear data), but if we lower bias too much, we risk overfitting and increased variance, making the models poor predictors of newer data (Dietterich and Kong 1995). To understand this a little better, imagine that we are training a flexible model with a data set that yields low training mean square error (MSE). If we use this same model with data set aside for testing, the test MSE will be much higher because it is picking up too many patterns associated with random noise (Hastie et al. 2009). A less flexible model (say a linear model) would have showed lower MSE with the test data even though the training MSE would be higher than the flexible model because it approximates nonlinearity with a linear fit. The quest in statistical learning is to optimize models to achieve a favorable bias-variance ratio, i.e., to simultaneously achieve low bias and low variance (Hastie et al. 2009).

6.3 Ensemble Learning Via Decision Trees

The basic idea of ensemble learning is to construct a mapping function y = F(x), based on the training data $\{(x_1,y_1), \ldots\ldots, (x_n,y_n)\}$, where

$$F(x) = a_o + \sum_{m=1}^{M} a_m f_m(x)$$

Where M is the size of the ensemble and {fm(x)} is an ensemble of functions called base learners (Friedman and Popescu 2008). The base learners are chosen from a function class of predictor variables and can vary with the ensemble methods used. An algorithmic procedure is specified to pick functions and also to obtain linear combination of the parameters {am}0 M based on the minimization of some cost function. This procedure generalizes the framework of ensemble learning to include algorithms like bagging, Random Forests, boosting, RuleFit etc.

The fundamental component of all ensemble learning algorithms that use "ensemble of decision trees" algorithms is the individual decision tree (Breiman et al. 1984). Decision tree is a recursive partitioning algorithm that partitions the response into subsets (left and right child nodes) based on splitting rules of the form $x_j < k$, where x_j is the splitting variable (predictor) and k is the splitting value. The left node gets all the observations (response) that satisfy the splitting rule and the right node gets the rest. The algorithm evaluates all possible splitting rules (for all the predictors) based on the response and selects the one that minimizes a statistical criterion (usually lowest MSE for regression). The observations in the resulting left and right nodes are again subject to the same partitioning scheme, and this goes on recursively until a stopping rule is satisfied (usually, minimum number of observations in the node, or the maximum depth of the tree or some other cost parameter). The end result of the recursive partitioning procedure is a decision tree with splitting rules and fitted values for terminal nodes (for regression, the average of the observations that fall into the terminal node).

Decision trees are intuitive, easy to interpret, capture nonlinearities and interactions very well and are very useful for high dimensional data. These properties make them very attractive for many ecological problems that exhibit these behaviors (Loh 2011; Rokach and Maimon 2015; Iverson and Prasad 1998). However, individual decision trees exhibit high variance and have poor prediction ability. Yet, they are very good building blocks in an ensemble setting where they can be used to build more complex models to achieve good variance bias tradeoffs (Dietterich 2000).

6.4 Ensemble Models

6.4.1 Bagging, Random Forest and Extreme Random Forests

Bagging is a way of reducing variance of decision trees via bootstrapping and aggregation of an ensemble of trees (Breiman 1996). In bagging, a number of decision trees are grown without pruning with a bootstrapped sample (sampling with replacement) and the resulting prediction rules averaged. It is based on the principle that if a single regressor has high variance, an aggregated regressor has smaller variance than the original one (Breiman 1996).

Random forests (RF) is a modification of bagging by taking a step further and randomizing even the predictor space. If along with the bootstrap sample, the predictors are also sampled randomly at each node and the results averaged, it results in further reducing variance (Prasad et al. 2006). This is the technique used in RF (randomForest package in R), where both datasets and predictors are perturbed to slightly increase the independence of each tree and then averaged to reduce variance (Breiman 2001). In RF, because a random subset of predictors are chosen at each split, many dominant predictors may not be present to define a split. This results in more local features defining the split instead of the dominant ones. When a large number of such trees are averaged, this can result in good balance between bias and variance and result in extremely reliable predictions. Another innovation in RF is that instead of computationally costly cross-validation or a separate test set to get unbiased error estimates, the observations not used in the training sample (usually one-third of the observations in the bootstrap sample), called "out-of-bag" (OOB), are used to obtain forecasts from the tree fitted to the remaining two-thirds (Liaw and Wiener 2002).

Extremely randomized trees (ERF) takes RF one step further in randomization (extraTrees package in R). While RF chooses the 'best' split at each node, ERF creates p splits randomly (i.e., independently of the response variable, p being the subset of predictors randomly chosen in each node) and then the split with the best gain (MSE for regression) is chosen. The rationale for ERF is that by randomizing the selection of split, the variance is reduced even further compared to the RF. However, ERF typically uses the entire learning sample instead of the bootstrapped sample to grow the trees in order to reduce bias (Geurts et al. 2006). Bias reduction becomes more important with this form of extreme randomization, because randomization increases bias when the splits are chosen independent of the response (Galelli and Castelletti 2013). ERF can be useful as a robust predictor after initially screening for irrelevant predictors. For example we can use RF to select a parsimonious, but ecologically meaningful set of predictors, and then use this set to predict with ERF.

6.4.2 Boosting Decision Trees

Boosting is a method of iteratively converting weak learners to stronger ones (in our case, using decision trees). Boosting initially builds a base learner after examining the data and then reweights observations that have higher errors. Stochastic gradient boosting (gbm package in R) is a form of optimization algorithm of a loss function with added tools to reduce variance by shrinkage and stochasticity (Ridgeway 1999; Friedman 2002). It optimizes a loss function over function space (as opposed to parameter space in ordinary regression problems) by estimating gradient directions of steepest descent (negative partial derivatives of the loss function called the pseudo-residuals) such that each iteration learns from previous errors (pseudo-residuals) and improves on them

$$F_m\left(x\right)=F_{m-1}\left(x\right)+\nu\cdot\gamma_m h_m\left(x\right)$$

At every stage of gradient boosting $1 < m \leq M$, the weak model Fm is slowly converted to a stronger one by improving on the previous iteration Fm-1 by adding an estimator. The value hm(x) is the decision tree (at the m-th step) with J terminal nodes (the tree partitions the predictor space into J disjoint regions). The goal is to minimize γm as a loss function (typically mean square error for regression), which has its own separate value for each of the J terminal nodes. The depth of the trees (i.e., the number of terminal nodes) J, defines the level of interaction and usually works best between 4 and 8. The shrinkage parameter ν ($0 < \nu \leq 1$) controls the learning rate of the boosting algorithm. If the number of boosting iterations (number of trees grown) is too large, it can lead to overfitting - ν therefore is usually chosen via cross-validation after finding the shrinkage parameter (values between 0.01 to 0.001 works best). In addition, the base learner, instead of using the entire training set, randomly subsamples without replacement (usually set to 50% of the training set), which adds stochasticity and leads to increased accuracy (Friedman 2002).

There is another slightly different approach to boosting that differs in the way the objective function is optimized with separate terms for training loss and regularization (Friedman 2001) called xgboost (Chen and Guestrin 2016). This method (xgboost package in R) differs from gbm in the way regularization is implemented when boosting, improving on its ability to control overfitting. It also handles tree pruning differently; gbm would stop splitting a node if it encounters a negative loss while xgboost splits to the maximum depth specified and then prunes the tree backwards to remove splits with no positive gain. Although boosting with carefully selected parameters can outperform RF, it can overfit noisy datasets due to the iterative learning process and has to be used with caution, or by using algorithms that automatically control overfitting with internal mechanisms (Opitz and Maclin 1999; Hastie et al. 2009).

6.4.3 RuleFit

RuleFit also uses decision tree ensembles to derive rules - however, these rules are used to fit regularized linear models in a flexible way that captures interactions (Friedman and Popescu 2008). It is similar to stochastic gradient boosting in that it combines base learners (decision tree rules) via a memory function with shrinkage to form a strong predictor. A large number of trees are generated from random subsets of the data and numerous rules assembled from a specified subset of terminal nodes. The predictor variables from these nodes allow for the estimation of linear functions where in addition to the rule-based base learner, linear basis functions are included in the predictive model. This is a useful feature because linearity from decision trees are hard to approximate. The large number of rules formed in the rule-generation phase, along with the linear basis functions are then minimized using regularized regression using lasso penalty (Tibshirani 1996; Zou and Hastie 2005). In regularized regression (ridge, lasso or elastic net) an additional penalty is imposed on the coefficients while minimizing the loss function. The final ensemble formed by regularized regression, results in rules, variables and linear coefficients sorted by importance. In contrast with other ensemble methods, RuleFit outputs coefficients in addition to prediction rules, which can be interpreted as regular linear coefficients.

6.5 Multiple Abundances – Habitat Suitability

The response, which in our case is an assessment of the habitat quality of white oak, is typically a measure of species abundance as reflected by its dominance and density (McNaughton and Wolf 1970). Dominance and density together capture many aspects of habitat quality. The measure that we used traditionally (Iverson et al. 2008; Prasad et al. 2016) is the importance value (IV) which captures the relative abundance weighted by other species present in the FIA plot (Woudenberg et al. 2010) as follows for each species X in a FIA plot:

$$IV(x) = \frac{50 * BA(x)}{\sum_{i=1}^{N} BA(i)} + \frac{50 * NS(x)}{\sum_{i=1}^{N} NS(i)}$$

BA is basal area, NS is number of stems (summed for overstory and understory trees) and N is the total number of species in the plot. This measure, which is a blend of dominance and density, reflects the biotic pressure that accounts for the interaction with other species and hence can reflect the realized niche better.

Another measure of species abundance that is proposed here is called mature average diameter (MAD). This dominance measure is derived by averaging the mean diameter of all trees of the target species in the plot after discounting the contribution of juveniles; juveniles are considered ephemeral because their contribution is negligible for this application. Juveniles were defined as: (min (avgdia) + q1(avgdia))/2; where avgdia is the average diameter, min is the minimum and q1 is the first quartile average diameter of all the FIA plots with white oak. This measure of dominance captures the absolute abundance of the species in contrast to the relative importance value (IV).

To capture the density of the species better, I propose another measure of abundance, mature species density (MNT), as the total number of trees of the species in the plot after discounting the juveniles. This measure of abundance denotes how well the species has colonized a site.

All three forms of abundance measures (IV, MAD, and MNT) in FIA plots were aggregated to 20 km cells and scaled from 0–100 (Fig. 6.1). They reflect different

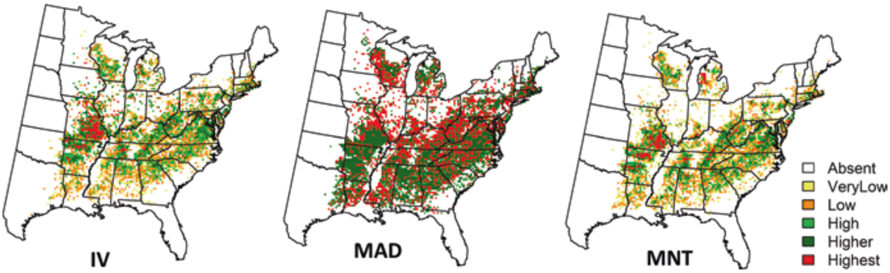

Fig. 6.1 The current maps of abundance for white oak - the importance value (IV), mature average diameter (MAD) and mature number of trees (MNT) per FIA plot aggregated to 20 km cells. The abundance values have been reclassified in the legend for illustrative purposes

aspects of habitat quality and should be modelled separately, with the overall effect spatially summarized similar to the multi-stage ensemble models (Anderson et al. 2012). I expect this approach to provide a better estimate of how the species would respond to climate change at a macro-scale compared to a single measure of abundance. Plurality of outputs and methods are important in gauging the overall response of the species, which has a complex nonlinear relationship with the environment under changing climates (Bowman et al. 2015).

6.6 Explanatory Variables (Predictors)

The explanatory variables represented a blend of climate, soil and topographic variables that were deemed most ecologically relevant after repeated tests (Table 6.1). For sources and other details, refer to Prasad et al. (2016). The current climate data are for the period 1981–2010 (Daly et al. 2008), and the future climate is Hadley Global Environment Model [HAD, Jones et al. 2011] for the greenhouse concentration pathway of RCP 8.5 (Representative Concentration Pathways; Moss et al. 2008) which represents the high emission future scenario (Meinshausen et al. 2011). The future RCP 8.5 climate scenario represents equilibrium conditions of the general circulation model (GCM; McGuffie and Henderson-Sellers 2014) for approximately 2100.

Table 6.1 The explanatory variables (predictors) used in the five models for white oak. These are a parsimonious set of ecologically relevant variables screen selected after repeated modeling

Climate	
tjan	Mean January temperature (°C)
tmaysep	Mean May–September temperature (°C)
pmaysep	May–September precipitation (mm)
gsai	Growing season aridity index (ratio of May–September precipitation by May–September evapotranspiration index)
Elevation	
elvmax	Maximum elevation (m)
elvsd	Elevation standard deviation
Soil	
clay	Percent clay (< 0.002 mm)
om	Organic matter content (% by weight)
ph	Soil pH
sieve10	Percent passing sieve no. 10 (coarse)
sieve200	Percent passing sieve no. 200 (fine)

Climate: Data for the period 1981–2010 from (PRISM Climate Group), GCM data from NEX-DCP30 (Thrasher et al. 2013).
Elevation: From the NASA's Shuttle Radar Topography Mission provided at a resolution of 3" (Guth 2006). We calculated the maximum value and standard deviation at 10 and 20 km^2 grids.
Soil: From Natural Resource Conservation Service's County Soil Survey Geographic (SSURGO) database (NRCS 2009). Data was processed by (Peters et al. 2013) and aggregated to 10 and 20 km^2 grids

6.7 Multi-Model Ensemble Approach

To achieve good bias-variance tradeoff, I used an ‘ensemble-of-trees’ via aggregation, randomization, boosting (randomForest, extraTrees, gbm, xgboost packages in R) and the ruleFit module (http://statweb.stanford.edu/~jhf/R_RuleFit.html). All these five approaches have their strengths and weaknesses depending on the training set. RandomForest and extraTrees have the least number of parameters to manipulate but cannot outperform the carefully tuned gbm and xgboost models. The gbm and xgboost algorithms, however, have more parameters to manipulate although the default settings often perform well. RuleFit in addition to robust prediction, gives linear coefficients and rule-sets. Multi-model ensemble approaches have been used where prediction uncertainty needs to be stabilized to yield more robust predictions (Jones and Cheung 2015; Martre et al. 2015). For the multi-model approach to work well, the models should be based on a similar framework (in this case decision trees) but should adopt structurally different approaches so that the final ensemble averages these heterogeneous approaches (Tebaldi and Knutti 2007). My approach consists of combining the five models (ensemble of models) to obtain two types of predictions: a) where output of all models are averaged (AVGMOD), and b) where they are averaged but only those cells common to these five models (an AND operation) make it to the final model (CAVGMOD). This procedure treats these models as a committee of experts and uses their average and common averaged prediction, improving prediction of single models by averaging out the errors. The overall thrust of the predictions are better captured by this approach for future climates. For this to work most effectively, the parameters for each of these five models need to be optimized via a repeated cross-validation approach in order to obtain a model with the most favorable bias-variance ratio. To do this, I used the caret package in R and repeated the ten-fold cross-validation, five times and chose the parameters with the lowest error (Kuhn 2008).

The multi-model, multi-response ensemble approach for the high emission future climate is illustrated for white oak using the three measures of abundance (IV, MAD and MNT) for the average model (AVGMOD), and the common average model (CAVGMOD) (Fig. 6.2). The CAVGMOD retains all the important habitats, while smoothing out the lower abundance values compared to AVGMOD and is therefore preferred in situations where reducing noise is desirable.

6.8 Results and Interpretation

One of the main goals while modeling future climate habitats of tree species is the need to gauge both model reliability and prediction confidence.

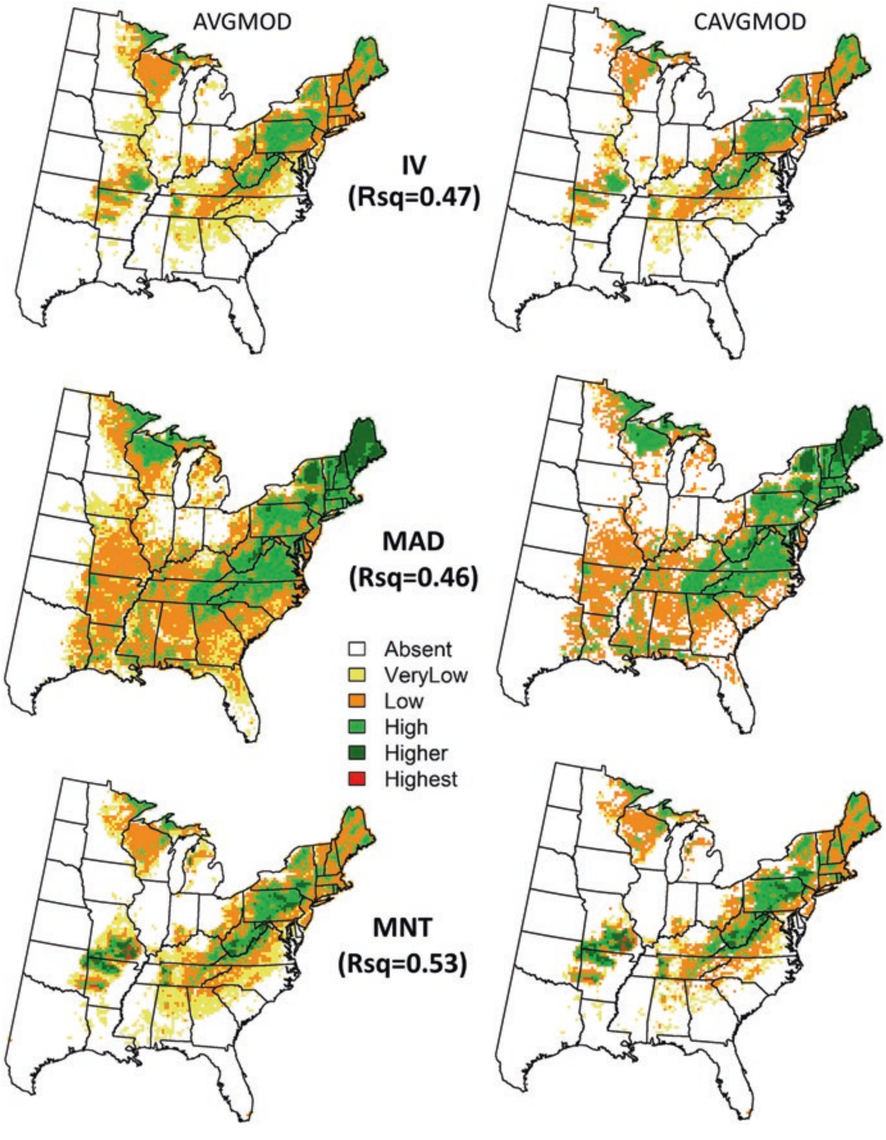

Fig. 6.2 The multi-model predictions for the three responses (importance value (IV), mature average diameter (MAD) and mature number of trees (MNT)) for the future harsh (Hadley, RCP 8.5) climate scenario for white oak. The AVGMOD is the average response across the five models, the CAVGMOD is the average response across the five models restricted to values common to all models. The abundance values have been reclassified in the legend for illustrative purposes

6.8.1 Model Reliability

Model reliability, which measures how well the models fit the data, reflects the vagaries of the training data, depending on whether the tree species is habitat specific, sparse, or a generalist. The sparser species have poor fit due to lack of training data and generally have poor model reliability. The habitat specific trees have the best model fit due to a better correlation with the environmental variables, with higher confidence in future predicted habitats. The model fit of generalists can vary depending on how widely and sparsely the species are distributed spatially. These species-specific vagaries affecting model reliability can be roughly measured via R-square-like measures via OOB, cross-validation or through a separate training and test dataset. For example, the R-square for the IV response of the RF model for the habitat-specific loblolly pine (*Pinus taeda*) was 0.79. In comparison, the R-square measure for our generalist species example of white oak (for the five models and three responses) averaged ~ 0.47.

6.8.2 Prediction Confidence

Even for species with good model reliability, the spatial configuration of the habitat quality in the predicted output (as measured via abundance values) can vary. For example, in Fig. 6.2, the classes 1–3 and 4–7 figure prominently even in CAVGMOD, and are of lower habitat quality than the higher classes. Because we can take advantage of the continuous distribution via regression models (after rescaling the abundances to values between 0 and 100), we have the ability to interpret the predicted habitats in terms of "prediction confidence" by reclassifying the results. The multi-model ensemble method helps mitigate the effects of spurious model artifacts (what can be termed "fuzzy values") at the low end of the abundance spectrum. The CAVGMOD approach further helps us identify only those prediction signals that have been strong in all five of the model predictions. Further, continuous predictions do not lend themselves to easy interpretation. Therefore, reclassifying them with the purpose of identifying the core regions where we have the highest confidence (based on abundance values) becomes useful for interpretation.

6.8.3 Combined Habitat Quality and Prediction Confidence

Using the CAVGMOD approach, we can average the predicted abundances of IV, MAD and MNT to capture the important future habitats as reflected by these three aspects of abundance and then reclassify the output to highlight the prediction confidence of the averaged response (Fig. 6.3). I have classified the future habitats to

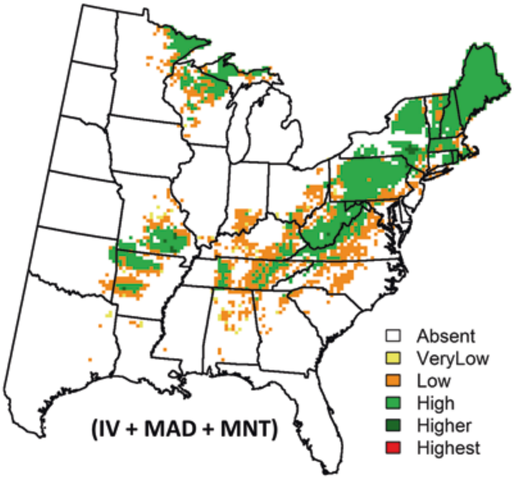

Fig. 6.3 The average of the three predictions (importance value (IV), mature average diameter (MAD) and mature number of trees (MNT)) for CVAGMOD (Fig. 6.2), with values common to the three predictions for white oak

five confidence zones based on the predicted abundance: (1) Very low (1–3); (2) Low (4–7); (3) High (8–15); (4) Higher (16–25); (5) Highest (26–100). Class 1 (Very low) would include many model artifacts (for example values close to zero that were regressed as 1–3) that are of dubious habitats that can be discarded as unreliable. Class 2 (Low) may also contain some regions with dubious habitats and some with low habitat suitability and should be treated with caution. Confidence in the habitat suitability classes increase steadily from Class 3 onwards (High, Higher and Highest).

Compared to the three CAVGMOD responses (Fig. 6.2), the single combined response (Fig. 6.3) highlights those areas (High and Higher classes) where we have the most confidence in the habitat quality of future habitats based on all three aspects of the abundances. For white oak, these areas (green and dark green) are predominantly in the north-east, north-central and south-central regions.

6.8.4 *Predictor Importance*

The importance of the predictors for each of the responses (IV, MAD and MNT) varied among the five models for white oak, although the first three were similar for all five models. These were recorded and averaged across the five models for the three responses (Table 6.2). For IV and MAD, the three most important variables are ph, tmaysep and tjan (Table 6.1), which explain 47.5% (IV) and 48.2% (MAD) of the variation for white oak. For MNT, the order varies with sieve10 and clay becoming important, but the same three variables (ph, tmaysep and tjan) still explain 40.1% of the variation. The predictor importance of the final combined response of the multi-model ensemble is the average for the three individual responses (IV, MAD and MNT) (Table 6.3). Again, the three most important variables (ph, tmaysep and tjan) explain 46.5% of the total variation. Because white oak is a generalist species occupying a vast swath of the eastern US, ph captures variation from east to west, while tjan and tmaysep are more important in capturing the north-south variation, and hence figures prominently in the final response.

6.9 Discussion

The main goal of the multi-model, multi-response approach developed here is to produce more reliable and ecologically interpretable models that can be used to help decision makers in managing tree species (Bell and Schlaepfer 2016). Tree species ranges are dynamic by nature and the additional impact of anthropogenic climate change makes it harder to predict distribution for future climates irrespective of the

Table 6.2 The predictor importance of white oak averaged across the five models for importance value (IV), mature average diameter (MAD) and mature number of trees (MNT). The Percent Gain reflects the proportion of variance explained by the variable

IV		MAD		MNT	
Variables	Percent gain	Variables	Percent gain	Variables	Percent gain
ph	16.7	ph	22.1	ph	17.2
tmaysep	16.6	tmaysep	16.5	sieve10	12.9
tjan	14.2	tjan	10.6	tmaysep	12.3
sieve10	10.2	elvmax	7.3	clay	10.8
clay	7.9	pmaysep	7.0	tjan	10.6
gsai	6.5	om	6.9	sieve200	8.1
elvmax	6.1	sieve10	6.7	elvsd	7.7
pmaysep	6.1	elvsd	6.4	pmaysep	5.6
om	5.5	gsai	6.3	om	5.5
elvsd	5.3	sieve200	6.0	gsai	5.3
sieve200	5.0	clay	4.2	elvmax	4.0

Table 6.3 The average predictor importance of the five models for white oak averaged across the three responses (IV, MAD and MNT in Table 6.2) and sorted by the Percent Gain

AVG	
Variables	Percent gain
ph	18.7
tmaysep	15.1
tjan	11.8
sieve10	9.9
clay	7.6
elvsd	6.5
sieve200	6.4
pmaysep	6.2
gsai	6.0
om	6.0
elvmax	5.8

modeling approaches used (Zurell et al. 2016). However, managers need to be able to target specific areas for facilitating species conservation and other multiple-use management objectives. The first step in accomplishing these goals is to explore where the most probable future suitable habitats will occur. The multi-model, multi-response approach addresses the inherent complexity in tree species response in a systematic and statistically defensible manner. It also provides maps of regions where we have high confidence in the future suitable habitats for tree species that exhibit good model reliability (Hannemann et al. 2015). The tree species that exhibit high model reliability are typically species that are habitat specific, although generalists like white oak can also be adequately modelled. The tree species that typically have poor model reliability are those that are sparse (both closely and widely distributed), which for eco-evolutionary and biogeographic reasons have not extended their range. Models for these species should be treated with caution because their habitats are difficult to predict with environmental variables; biogeographic and eco-evolutionary variables are not easy to incorporate without extensive Gene X Environment studies.

The multi-model, multi-response model I present as an example, demonstrates that suitable future habitats for white oak are most likely to be in the north-east, north-central and south-central regions of the eastern United States (Fig. 6.3). This type of information is important for resource managers dealing with uncertainty and mandates to incorporate climate change in their management portfolios. While suitable habitats lack information on the likelihood of colonization, these can be assessed at a later stage via dispersal models (Prasad et al. 2016). However, to assess the probability of establishment of colonized sites involves finer scale process-based models that account for biotic interactions.

Another challenge when modeling tree species habitats under current and future climates lies in the transfer of ecological space (the niche of the species) to eco-geographic space (the mapped niche), which results in spatial autocorrelation effects. The problem of spatial autocorrelation can become acute with conventional parametric techniques and, while less problematic with non-parametric statistical learning methods, can still manifest in residual errors (Hawkins 2012; Kühn and Dormann 2012). In this study, there was negligible global residual spatial autocorrelation, although local ones were present. However in niche-based spatial modeling, some residual spatially auto-correlated errors have to be tolerated, and interpreted with caution. The alternative is extremely complex, autoregressive, parametric models that in many cases defeat the purpose of a more flexible modeling approach (Merow et al. 2014).

6.10 Conclusion

Predicting habitat quality is the first stage in the analysis of future distribution of tree species because dispersal and site-specific constraints will prevent colonization and establishment in all available suitable habitats (Prasad et al. 2016). Predicting these suitable habitats using robust modeling techniques is the essential first step and I present the multi-model and multi-response ensemble technique as a method for modeling tree species dynamics for better management under changing climates.

Acknowledgements The author would like to thank Louis Iverson for his comprehensive review and also two anonymous reviewers for their valuable suggestions. Thanks to the Northern Research Station, USDA Forest Service, for funding.

References

Anderson BJ, Chiarucci A, Williamson M (2012) How differences in plant abundance measures produce different species-abundance distributions. Methods Ecol Evol 3:783–786

Bell DM, Schlaepfer DR (2016) On the dangers of model complexity without ecological justification in species distribution modelling. Ecol Model 330:50–59

Belle A, Thiagarajan R, Soroushmehr SMR, Navidi F, Beard DA, Najarian K (2015) Big data analytics in healthcare. BioMed Res Int 370194, 16. doi:https://doi.org/10.1155/2015/370194

Belmaker J, Zarnetske P, Tuanmu M-N, Zonneveld S, Record S, Strecker A, Beaudrot L (2015) Empirical evidence for the scale dependence of biotic interactions. Glob Ecol Biogeogr 24:750–761

Bowman DM, Perry GLW, Marston JB (2015) Feedbacks and landscape-level vegetation dynamics. Trends Ecol Evol 30:255–260

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

Breiman L (2001) Random forests. Mach Learn 45:5–32

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. CRC press, Boca Raton

Chen T, Guestrin C (2016) XGBoost: reliable large-scale tree boosting system. ar Xiv: 1603.02754 [cs. LG]. http://arxiv.org/pdf/1603.02754v1

Daly C, Halbleib M, Smith JI, Gibson WP, Doggett MK, Taylor GH, Curtis J, Pasteris PP (2008) Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. Int J Climatol 28:2031–2064

Dietterich TG (2000) An experimental comparison of three methods for constructing ensembles of decision trees. Mach Learn 40:139–157

Dietterich TG, Kong EB (1995) Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Mach Learn 255:0–13

Domingos P (2012) A few useful things to know about machine learning. Commun ACM 55(10):78–87

Elith J, Kearney M, Phillips S (2010) The art of modelling range-shifting species. Methods Ecol Evol 1:330–342

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232

Friedman JH (2002) Stochastic gradient boosting. Comput Stat Data Anal 38:367–378

Friedman JH, Popescu BE (2008) Predictive learning via rule ensembles. Ann Appl Stat 2:916–954

Galelli S, Castelletti A (2013) Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling. Hydrol Earth Syst Sci 17:2669–2684

Garcia-Valdes R, Zavala MA, Araujo MB, Purves DW (2013) Chasing a moving target: projecting climate change-induced shifts in non-equilibrial tree species distributions. J Ecol 101:441–453

Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63:3–42

Guisan A, Edwards TC Jr, Hastie T (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol Model 157:89–100

Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. Ecol Lett 8:993–1009

Guth PL (2006) Geomorphometry from SRTM: Comparison to NED. Photogramm Eng Remote Sens 72:269–277

Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, Batcheller AL, Duke CS, Porter JH (2013) Big data and future of ecology. Front Ecol Environ 11:156–162

Hannemann H, Willis KJ, Macias-Fauria M (2015) The devil is in the detail: unstable response functions in species distribution models challenge bulk ensemble modelling. Glob Ecol Biogeogr 25:26–35

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning, 2nd edn. Springer Science, New York

Hawkins BA (2012) Eight (and a half) deadly sins of spatial analysis. J Biogeogr 39:1–9

Hill L, Hector A, Hemery G, Smart S, Tanadini M, Brown N (2017) Abundance distributions for tree species in Great Britain: a two-stage approach to modeling abundance using species distribution modeling and random forest. Ecol Evol 7:1043–1056

Hussain K, Prieto E (2016) Big data in the finance and insurance sectors. In: Cavanillas JM et al (eds) New horizons for a data-driven economy. Springer Open. https://doi.org/10.1007/978-3-319-21569-3

Iverson LR, Prasad AM (1998) Predicting abundance of 80 tree species following climate change in the eastern United States. Ecol Monogr 68:465–485

Iverson LR, Prasad AM, Matthews SN, Peters M (2008) Estimating potential habitat for 134 eastern US tree species under six climate scenarios. For Ecol Manag 254:390–406

Iverson LR, Thompson FR, Matthews S, Peters M, Prasad AM, Dijak WD, Fraser J, Wang WJ, Hanberry B, He H, Janowiak M, Butler P, Brandt L, Swanston C (2016) Multi-model comparison on the effects of climate change on tree species in the eastern U.S.: results from an enhanced niche model and process-based ecosystem and landscape models. Landsc Ecol. https://doi.org/10.1007/s10980-016-0404-8

Jones MC, Cheung WWL (2015) Multi-model ensemble projections of climate change effects on global marine biodiversity. ICES J Mar Sci 72:741–752

Jones CD, Hughes JK, Bellouin N, Hardiman SC, Jones GS, Knight J, Liddicoat S, O'Connor FM, Andres RJ, Bell C, Boo K-O, Bozzo A, Butchart N, Cadule P, Corbin KD, Doutriaux-Boucher M, Friedlingstein P, Gornall J, Gray L, Halloran PR, Hurtt G, Ingram WJ, Lamarque J-F, Law RM, Meinshausen M, Osprey S, Palin EJ, Parsons Chini L, Raddatz T, Sanderson MG, Sellar AA, Schurer A, Valdes P, Wood N, Woodward S, Yoshioka M, Zerroukat M (2011) The HadGEM2-ES implementation of CMIP5 centennial simulations. Geosci Model Dev 4:543–570

Kühn I, Dormann CF (2012) Less than eight (and a half) mis- conceptions of spatial analysis. J Biogeogr 39:995–998

Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28:1–26

Liaw A, Wiener M (2002) Classification and regression by random forest. R News 2:18–22

Loh W-Y (2011) Classification and regression trees. WIREs Data Min Knowl Discovery 1:14–23. https://doi.org/10.1002/widm.8

Martre P, Wallach D, Asseng S, Ewert F, Boote KJ, Ruane AC, Peter J, Cammarano D, Hatfield JL, Rosenzweig C, Aggarwal PK, Angulo C, Basso B, Bertuzzi P (2015) Multimodel ensembles of wheat growth: many models are better than one. Glob Chang Biol 21:911–925

McGuffie K, Henderson-Sellers A (2014) A climate modelling primer, 4th edn. Wiley, p 456. isbn:978-1-119-94336-5

McNaughton SJ, Wolf LL (1970) Dominance and the niche in ecological systems. Science 167:131–139

Meinshausen M, Smith SJ, Calvin K, Daniel JS, Kainuma MLT, Lamarque JF, Matsumoto K, Montzka SA, Raper SCB, Riahi K, Thomson A, Velders GJM, van Vuuren DPP (2011) The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. Clim Chang 109:213–241

Merow C, Smith MJ, Edwards TC Jr, Guisan A, McMahon SM, Normand S, Thuiller W, Wuest RO, Zimmermann NE, Elith J (2014) What do we gain from simplicity versus complexity in species distribution models? Ecography 37:1267–1281

Moss R, Babiker M, Brinkman S, Calvo E, Carter T et al (2008) Towards new scenarios for analysis of emissions, climate change, impacts, and response strategies. Intergovernmental Panel on Climate Change, Geneva, p 132 http://www.aimes.ucar.edu/docs/IPCC.meetingreport.final.pdf

NRCS (Natural Resources Conservation Service) (2009) Soil Survey Geographic (SSURGO). Available at https://datagateway.nrcs.usda.gov/. Accessed between August 2009 and November 2010

Opitz D, Maclin R (1999) Popular ensemble methods: an empirical study. J Artif Intell Res 11:169–198

Peters MP, Iverson LR, Prasad AM, Matthews SN (2013) Integrating fine-scale soil data into species distribution models: preparing soil survey geographic (SSURGO) data from multiple counties. US Department of Agriculture, Forest Service, Northern Research Station, Newtown Square, p 70

Prasad AM (2015) Macroscale intraspecific variation and environmental heterogeneity: analysis of cold and warm zone abundance, mortality, and regeneration distributions of four eastern US tree species. Ecol Evol 5:5033–5048

Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9:181–199

Prasad AM, Iverson LR, Matthews SN, Peters MP (2016) A multistage decision support framework to guide tree species management under climate change via habitat suitability and colonization models, and a knowledge-based scoring system. Landsc Ecol. https://doi.org/10.1007/s10980-016-0369-7

PRISM Climate Group. Oregon State University, http://prism.oregonstate.edu

Ridgeway G (1999) The state of boosting. Comput Sci Stat 31:172–181

Rokach L, Maimon O (2015) Data mining with decision trees - theory and applications, 2nd edn. World Scientific

R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna URL https://www.R-project.org/

Slavakis K, Giannakis GB, Mateos M (2014) Modeling and optimization for big data analytics. IEEE Signal Process Mag 5:18–31

Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. Phil Trans R Soc A 365:2053–2075

Thrasher B, Xiong J, Wang W, Melton F, Michaelis A, Nemani R (2013) Downscaled climate projections suitable for resource management. Trans Am Geophys Union 94:321–323

Tibshirani R (1996) Regression shrinkage and selection via the Lasso Robert Tibshirani. J R Stat Soc Ser B Stat Methodol 58:267–288

Van Horn JD, Toga AW (2014) Human neuroimaging as a “big data” science. Brain Imaging Behav 8:323–331. https://doi.org/10.1007/s11682-013-9255-y

Vincenzia S, Zucchettab M, Franzoib P, Pellizzato M, Pranovib F, De Leo GA, Torricelli P (2011) Application of a random Forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy. Ecol Model 222:1471–1478

Woudenberg SW, Conkling BL, O’Connell BM, LaPoint EB, Turner JA, Waddell KL (2010) The forest inventory and analysis database: database description and User’s manual version 4.0 for phase 2. General Technical Report RMRS-GTR-245, USDA Forest Service, Rocky Mountain Research Station, Fort Collins, Colorado, 336 p

Zhang Y, Zhao Y (2015) Astronomy in the big data era. Data Sci J 14:11. https://doi.org/10.5334/dsj-2015-011

Zhou ZH (2012) Ensemble methods: foundations and algorithms. CRC press, Boca Raton

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. R Stat Soc Ser B Stat Methodol 67:301–320

Zurell D, Thuiller W, Pagel J, Cabral JS, Münkemüller T, Gravel D, Dullinger S, Normand S, Schiffers KH, Moore KA, Zimmermann NE (2016) Benchmarking novel approaches for modelling species range dynamics. Glob Chang Biol 22:2651–2664