Part III

An overview of the modeling methods

"Any mechanistic process model of ecosystem dynamics should be consistent with a static, quantitative and rigorous description of the same ecosystem"

(Austin 2002, p. 112)

This section addresses the third part of the framework presented by Austin (2002), outlined in Chapter 1 and used as an organizing principle for this book – the statistical part. Austin's statistical modeling framework includes the choice of modeling methods and decision regarding implementation (calibration and validation) of a model. Some appropriate and widely used methods in SDM are not statistical in the strict sense, and so we can more broadly refer to quantitative and rule-based empirical models. In any case the methods included are explicit and the modeling is repeatable.

Guisan and Zimmermann (2000) divided the statistical modeling portion of Austin's framework into four steps: (a) conceptual model formulation, (b) statistical model formulation, (c) calibration (fitting or estimation), and (d) evaluation. Those steps provide a useful outline for this section. Conceptual model formulation in species distribution modeling generally relies on a number of key ecological concepts and was described in Chapter 3. Guisan and Zimmermann emphasized that species distribution models are usually empirical or phenomenological models, designed to condense empirical facts, and are judged on their ability to predict, that is, judged on their precision and reality. This distinguishes them from distinct but complementary mechanistic (process) models, that aim to be general and realistic, and from analytical (theoretical) models, built for generality and precision. The complementarity of these three modeling approaches is the subject of Austin's comment, quoted above.

Guisan and Zimmermann define statistical model formulation as choosing a type of model that is suitable for the data – often, statistical

models are specific to a type of response variable (to use traditional statistical terminology) and its probability distribution (Franklin, 1995). What is the measurement scale of the response variable: interval, ratio, ordinal, counts, or categories? While many of the methods considered in this book are remarkably flexible with regard to the measurement scale and distribution of the predictor and response variables, some have advantages over others when applied to certain kinds of data, and certain forms of the relationship between response and predictors, as summarized in Table III.1. For example, categorical variables can be used as predictors in regression modeling – each category is simply treated as a "dummy" variable, that is, with a value of zero or one for each observation. But, when there are many categories, for example, if soil series is being used to predict the occurrence of a plant species, it can be difficult to estimate and interpret meaningful parameters for all the categories that occur in the data. Decision trees (regression or classification trees) handle this problem very easily. In addition, decision trees automatically identify interactions between variables, while interactions must be specified in advance in regression modeling. Hopefully, these kinds of differences will become clear with more examples in the chapters of this section. Elith and Leathwick (2009; their Table 6.2) also provide a summary of the key features of modeling methods and associated software with respect to types of species data, complexity of fitted functions, and ability to estimate model uncertainty.

As was discussed in Part II (Chapters 4 and 5), the species occurrence and environmental data typically used in SDM have some particular characteristics. The observations are geographically referenced. Species occurrence, abundance, or some other measure, is recorded from surveys, sometimes as an ordinal (none, few, many) or binary (presence or absence) variable. Alternatively, records of species occurrences are available from natural history collections, atlases, or other kinds of observations that are not from probability-based samples, and that only provide information on species occurrence (and not species absence). If there are no geographically located observations, other kinds of information on habitat preference might be used to develop a model (Chapter 8). With multiple environmental predictors, collinearity is an issue to be concerned about, and environmental predictor variables often represent surrogates or indirect gradients, rather than proximal drivers of species abundance or fitness. Those mapped predictors used in SDM often include a mixture of categorical and continuous variables, the relationship between predictors and response is not expected to be linear, and interactions

Table III.1. *Modeling methods that can be applied to quantitative, categorical or binary (presence–absence) response variables*

Modeling method	Best for type of response	Predictors (covariates)	Response function
Statistical			
Discriminant function analysis	Categorical	Quantitative	Linear
Generalized linear model (GLM)	Quantitative, categorical binomial	Quantitative, categorical	Parametric – linear, polynomial, piecewise, interaction terms
Spatial autoregressive models	Quantitative, categorical binomial	Quantitative, categorical	Same as GLM. Includes autocovariate
Generalized additive model (GAM)	Quantitative, categorical binomial	Quantitative, categorical	Smoothing function, estimated using local regression, splines or other method
Multivariate adaptive regression splines (MARS)	Quantitative, categorical	Quantitative, categorical	Adaptive piecewise linear regression (combines splines and binary recursive partitioning)
Machine learning			
Decision tree (DT) – classification and regression trees	Quantitative, categorical multinomial	Quantitative, categorical	Divisive, monothetic decision rules (thresholds) from binary recursive partitioning
Ensemble trees (bagging, boosting, random forests)	Quantitative, categorical multinomial	Quantitative, categorical	Weighted and unweighted model averaging applied to decision trees
Artificial neural network (ANN)	Categorical multinomial	Quantitative, categorical	Non-linear decision boundaries in covariate space

(*cont.*)

Table III.1. (*cont.*)

Modeling method	Best for type of response	Predictors (covariates)	Response function
Hybrid methods (GARP)	Categorical binomial	Quantitative, categorical	Combines decision rules using a genetic algorithm (response functions not visualized)
Maximum entropy	Categorical binomial	Quantitative, categorical	Non-linear response functions can be described
Other			
Distance methods	Categorical "event only"	Quantitative	Similarity to conditions where event occurs: does not estimate response function or importance of predictors
Expert methods	Any	Quantitative, categorical	Response functions and weights defined by expert knowledge

The statistical methods GLM, GAM and MARS can handle dependent variables with Gaussian, binomial, Poisson, ordinal, or other distributions via a link function. Categorical variables can be binomial or multinomial. "Event-only" refers to data documenting the occurrence of a species only, and not its absence, that is, not resulting from a probability based sample design aimed at documenting occurrence and absence (Chapter 3). Multivariate "distance methods" are described in Chapter 8.

between predictors are expected. These are all factors to keep in mind when formulating a statistical model for SDM.

Model estimation or fitting (called calibration by some) is usually taken to mean estimation of model parameters, for example, the coefficients in a regression model, the splitting rules in a decision tree, or the weights in an artificial neural network. Model fit is evaluated, based on some measure of the agreement between the model and the data (Rykiel, 1996), such as variance reduction in traditional regression, deviance reduction in modern regression, or information criteria such as the Akaike Information Criterion (AIC, discussed in the following chapter). Model estimation is also frequently investigated by some kind of analysis of model residuals (the pattern of differences between the modeled cases and the data), including spatial analysis. By its very nature, SDM almost always examines multiple environmental variables (Chapters 3, 5) as potential predictors of species' distributions, and calibration usually involves selecting a subset of candidate explanatory variables. Model fitting also includes the transformation of variables, including polynomials of, and interactions between, candidate explanatory variables, with attention to the expected shape of the species' response curves (Chapter 3).

In the model evaluation step, many criteria could be used for validating the output of a model of species-habitat relations (Chapter 10 in Morrison *et al.*, 1998). Evaluation is distinct from model fitting when the model is used to make predictions based on new or different data. If a strictly independent dataset with suitable attributes is not available, it is common to divide the dataset into "training" and "testing" data prior to modeling, or to use some kind of resampling method (such as bootstrapping) to estimate, from the training data, what the prediction accuracy of the model would be if it were applied to new data. This is the subject of Chapter 9.

My understanding of the quantitative empirical models used in SDM has been greatly influenced by Hastie *et al.* (2001), and some fundamental concepts from that excellent book are a good place to begin this section. (Note that, while I relied on the 2001 edition, a second edition was published in 2009). The organizing principle for their book is the idea of learning from data. In classical hypothesis testing in statistics, relatively small amounts of data result from controlled experiments that enforce certain restrictions in order to meet a set of criteria and assumptions. Although exploratory data analysis has long been part of scientific problem solving (Tukey, 1977), today more than ever many problems in science, engineering, business, marketing and medicine start with very

large datasets – many observations and many variables. The objective is to learn from the data and build a model that can make accurate predictions, for example of the price of a stock, the chances a patient will have a heart attack, or the likelihood that a pixel in a satellite image belongs to a certain land cover class. Hastie *et al.* refer to this as the supervised learning problem (what some of us from remote sensing background call supervised classification). Observations of the outcome (response variable) are used to guide the learning – in other words, to estimate the rules or parameters –in order to calibrate the model. The observations of the predictor and response variables are called the training data, and as noted above, a separate set of test data (or a cleverly fabricated "independent" dataset) is used to validate the ability of the model to predict.

Because there have been such great changes in statistical and computational methods in recent decades, it is somewhat difficult to classify them or group them into chapters, as the distinction between classical statistical approaches, like regression, and various other methods of supervised "learning from data," has become blurred. In this section I have grouped the models into: statistical (emphasizing regression; Chapter 6), machine learning (Chapter 7), and classification and distance methods (Chapter 8). Chapter 8, in particular, addresses methods that are widely used for "presence-only" data on species occurrence. This grouping and progression makes sense to me – however, it is certainly not the only way that these methods can be related one to another. In each chapter an overview of model formulation and estimation is given. An alternative approach to mapping species distributions, direct interpolation of species data, is really a form of modeling done in geographical space, using only location, or proximity to species records, as predictors. This was discussed in Section 2.3, but is also relevant to issues considered in Chapter 8. In Chapter 9 I discuss model validation or verification, because most of the validation concepts and approaches used in SDM can be applied to all types of models.

There are now a number of both general purpose and dedicated software packages and systems available for SDM, as shown in Table III.2 (see also Table 3 in Guisan & Thuiller, 2005 and Table 4 in Elith *et al.*, 2006). Some are proprietary, some are open source, and some are free. Some include environmental data (primarily low-resolution global climate maps). A list is provided here with the caveat that it is undoubtedly incomplete and will be out of date by the time you read it. However, it will help in identifying software resources for SDM.

Table III.2. *Examples of stand-alone or dedicated software for species' distribution modeling*

Software	Algorithm family	URL, reference	Additional software
ANUCLIM	EE	http://cres.anu.edu.au/outputs/anuclim.php	–
BIOCLIM	EE	http://cres.anu.edu.au/outputs/anuclim/doc/bioclim.html; (Busby, 1991)	BIOCLIM ArcView ® extension; DIVA-GIS
BIOMAP	EE	http://cres.anu.edu.au/outputs/anuclim.php	–
BIOMAPPER	Ecological niche factor analysis (ENFA)	www.unil.ch/biomapper (Hirzel *et al.*, 2002)	–
BIOMOD	GLM, GAM, CART, ANN	At the discretion of the author (Thuiller *et al.*, 2003)	–
DIVA	EE	http://www.diva-gis.org (Hijmans *et al.*, 2001)	–
DOMAIN	Gower-similarity; multivariate distance	www.cifor.cgiar.org/scripts/default.asp?ref=research_tools/domain/index.htm (Carpenter *et al.*, 1993) http://uwadmnweb.uwyo.edu/wyndd/	–
ECOSPAT	Logistic regression (GLM, GAM)	http://www.ecospat.unil.ch; with permission of the developer	–
GARP	Genetic algorithms	http://lifemapper.org/desktopgarp (Stockwell & Peters, 1999)	–
GARP/WhyWhere?	WhyWhere	http://landshape.org/enm/whywhere-22-download/ (Stockwell, 2006);	–
GRASP	Logistic regression (GLM, GAM)	https://www.unine.ch/cscf/grasp (Lehmann *et al.*, 2002)	S-Plus ® or R
HyperNiche	Non-parametric multiplicative regression	http://home.centurytel.net/~mjm/hyperniche.htm (McCune, 2006)	–

(*cont.*)

Table III.2. (*cont.*)

Software	Algorithm family	URL, reference	Additional software
MARS	Multivariate adaptive regression splines	www.salford-systems.com/mars.php;	MDA package ®
MaxEnt	Maximum Entropy	http://www.cs.princeton.edu/~schapire/maxent/ (Phillips *et al.*, 2006)	–
NNETW	Artificial neural networks		S-plus ® w/libraries (nnet, NNETW)
OpenModeller	GARP, bioclimate envelopes	http://openmodeller.sourceforge.net/	–
PRESENCE	Logistic regression	http://www.mbr-pwrc.usgs.gov/software/presence.html (MacKenzie *et al.*, 2002)	–
Presence Absence	SDM performance evaluation (Kappa, AUC, etc.)	http://cran.r-project.org/web/packages/PresenceAbsence/index.html	R
SAM	OLS, Autoregression	http://www.ecoevol.ufg.br/sam/ (Rangel *et al.*, 2006)	–
STATMOD ZONE	CART	http://www.gis.usu.edu/~chrisg/avext/	ArcView and S-plus

Table modified from Guisan and Thuiller (2005), courtesy of J. Miller. Note that many methods, including GLM, GAM, CART, ANN, MARS, and Mahalanobis distance, are also available in proprietary or open source general purpose statistical and/or GIS software, either integrated into the software or available as additional modules or packages CART = classification and regression trees; EE = Environmental (climatic) envelope; GAM = generalized additive models; GLM = generalized linear models; OLS = ordinary least squares regression.

7 · *Machine learning methods*

7.1 Introduction

As discussed in the overview of Part III, species distribution modeling can be treated as a supervised learning problem – observations of a response, such as species presence or absence, and associated environmental predictors, are used to develop rules that can be used to classify new observations where the values of the predictors, but not the response, are known. Statistical or machine learning approaches can be used to solve a supervised learning problem. In Chapter 6 it was noted that the linear (regression) model can be thought of as a model-driven or parametric approach to statistical learning, in which certain assumptions are made about the form of the model, and also a "global" method, meaning that all of the data (observations) are used to estimate the parameters. In other words, the problem in supervised learning is to construct a function that "maps" inputs X to outcome Y. In statistical inference the distributional form is chosen by the analyst and its parameters are estimated from the data. Machine learning methods, in contrast, are various kinds of algorithms that are used to learn the mapping function or classification rules inductively, directly from the training data (Breiman, 2001a; Gahegan, 2003).

As we also saw in Chapter 6, GAMs and MARs have been described as "non-parametric" extensions of GLMs because they are not global, but use a local subset of the data (in predictor measurement space) to estimate the response by assuming a particular structured form to that response (e.g., using piecewise linear functions, smoothing splines or polynomials). So, as mentioned in the overview of Part III, it would have been just as valid to put GAMs and MARSs in this chapter as in Chapter 6, and some would disagree with the way I have organized these topics (but I trust the readers to use the table of contents to find what they are looking for). Hastie *et al.* (2001), for example, grouped decision tree methods with GAMs and MARSs because each assumes a different but nonetheless structured form for the unknown regression

relationship. Decision trees, for example, partition feature space into a set of rectangles, and fit a constant to each one. In this chapter decision trees are discussed along with other machine learning methods.

The methods that will be described in this chapter have all been applied in SDM, have been called inductive (supervised) machine learning and have been developed in the fields of artificial intelligence and statistics. They include decision tree-based methods, artificial neural networks, genetic algorithms, maximum entropy and support vector machines. I will begin by focusing on decision trees because they have been widely used in SDM, and then will give an overview of the others. Ensemble forecasting methods will also be discussed in this chapter. I recommend, in addition to Hastie *et al.* (2001), the well-written discussion of statistical learning, machine learning, data mining and inference applied to geographical data by Gahegan (2003), as background reading. Also, an edited book and a recent review paper provide an introduction to and overview of machine learning methods for ecological applications (Fielding, 1999; Olden *et al.*, 2008).

7.2 Decision tree-based methods

7.2.1 How decision trees work

Tree-based methods, referred to as classification and regression trees, or, collectively, decision trees (DT), are the first methods I ever used for SDM and I have learned a lot about seeking patterns in data from using them. I found several book chapters to be useful general introductions to decision tree modeling, and although they are associated with particular software implementations, I still consult them (Clark & Pregibon, 1992; Venables & Ripley, 1994 or the newer edition). A nice description of decision trees is also given in Elith *et al.* (2008). In this book, I have focused on the special case of classification trees with binary outcomes (presence versus absence). However, in the field in which I was first trained, remote sensing, the goal is often to map many categories of land cover, for example, using remotely sensed imagery and other variables. Classification trees are a particularly useful method of supervised classification when the response is a categorical variable with many (more than two) categories and when the predictors include both categorical and continuous variables. Regression examples are relatively uncommon in machine learning, but we expect them to be particularly useful in ecology.

Decision trees are divisive, monothetic, supervised classifiers. What does that mean? The goal in decision tree modeling is to partition (divide) the data into subgroups that are homogeneous, that is, where the response variables have similar values or are members of the same class, based on ranges of values of predictor variables. This takes place in three stages, tree building or growing, tree stopping, and tree pruning or optimal tree selection (Olden *et al.*, 2008). In tree building, first the multivariate data are sorted according to the values of each predictor variable, one at a time, and then every possible threshold value of each predictor is examined. For unranked categorical predictors (nominal variables), every possible grouping of classes is examined. Each potential threshold or grouping is referred to as a "candidate split" that can possibly be used to divide the data.

The two subsets of the data that result from any one of these candidate splits are described in terms of their homogeneity in the response variable. This is easier to show in an example than to describe in words (Table 7.1). The subsets may be of any size. For a continuous response variable, reduction in variance or deviance (sum of squares) is a measure of homogeneity, and the decision tree is called a regression tree. For categorical responses, usually some measure of the homogeneity or "purity" of class membership in the resulting subsets is used – an information or entropy statistic – and the resulting tree is referred to as a classification tree.

The single candidate split – the threshold value (or categorical grouping) of the single predictor – that gives the greatest increase in homogeneity or purity (reduction in deviance or entropy) of the subsets is used to divide the data into those subsets. The single split, resulting in a tree with nested binary decision thresholds (like a dichotomous key), is why the method is called monothetic. Then, in an iterative and nested fashion, the same procedure is applied to the two subsets of the data, again using all predictor variables to search for candidate splits. The subsets of the data in the tree are often referred to as nodes, or sometimes as leaves.

A number of purity measures have been used to select among candidate splits for a classification tree. These include the misclassification error (the proportion of observations misclassified) of the individual split, cross-entropy, the Gini index and the deviance. The deviance or likelihood ratio was described in Chapter 6. If there are 1 through K classes, and p_{mk} is the proportion of class k observations at node m, then the so-called information statistic, entropy statistic, or cross-entropy is:

$$-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

Table 7.1. *A tiny example of recursive partitioning to develop a classification tree*

Y (P/A)	*X*1 (soil)	*X*2 (elevation)
0	A	930
0	A	1100
1	A	760
0	A	880
1	B	545
1	B	650
0	B	750
1	B	700
1	B	590

The dependent variable *Y* is the presence or absence of a species coded 1 or 0. There are two explanatory variables, soil type (*X*1) which is categorical (A or B), and elevation (*X*2), which is continuous. First, nine observations are sorted according to values of the first predictor, *X*1. In this simple example, there is only one possible threshold value or grouping for this categorical variable, shown by the dashed line. The purity of class membership (the two possible classes of *Y*) of the resulting groups is measured by calculating the reduction in entropy or deviance (see text) – the change in value from the unsplit dataset to the two groups. Then the process is repeated by sorting the observations by *X*2, and examining all nine possible splits in that case. Whichever candidate split of whichever variable (*X*1 or *X*2) results in the most pure grouping in the response is used to split the data. The process then is repeated for each subgroup.

The Gini index is:

$$1 - \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

Deviance has been preferred in some software implementations (Venables & Ripley, 1994) and the Gini index in others (Breiman *et al.*, 1984), although the Gini index and entropy have similar properties (Hastie *et al.*, 2001, p. 271).

To summarize thus far, in tree building the data are divided using recursive binary partitions (Hastie *et al.*, 2001). The result is a "decision tree" – a set of nested, binary decision rules that can be used to classify observations into subgroups (nodes) based on threshold values of the predictors. The decision tree can be presented as text (Table 7.2) or

Table 7.2. *Example of a classification tree. Decision rules are organized like a dichotomous key*

```
(1) root 1470 360.100 0 ( 0.973469 0.026531 )
    (2) jan < 6.565 1383 137.800 0 ( 0.991323 0.008677 )
      (4) jan < 4.525 889    0.000 0 ( 1.000000 0.000000 ) *
      (5) jan > 4.525 494 112.900 0 ( 0.975709 0.024291 ) *
    (3) jan > 6.565 87 107.800 0 ( 0.689655 0.310345 )
      (6) precip < 265.5 22    8.136 0 ( 0.954545 0.045455 ) *
      (7) precip > 265.5 65   87.490 0 ( 0.600000 0.400000 )
        (14) precip < 289 49 67.910 1 ( 0.489796 0.510204 ) *
        (15) precip > 289 16  7.481 0 ( 0.937500 0.062500 ) *
```

The dependent variable is the presence/absence (1/0) of *Ceanothus verrucosus*, a rare shrub in coastal southern California, USA. In this dataset, there are 1470 observations, and *C. verrucosus* is present in 39 of them. There are two predictors in this model, average minimum January temperature (jan), and average annual precipitation (precip). The elements in each line are the node number, the splitting rule (variable and threshold), the number of observations at that node, the deviance of the node, the predicted value (0, 1) and the proportions of classes 0 and 1 in that subgroup of observations. For example, *C. verrucosus* is classified as present only at terminal node 14 (* denotes a terminal node; see Fig. 7.1) based on a probability threshold of 0.5. At that node there are 49 observations, the deviance is 67.91, the predicted *Y* value is "1" (*C. verrucosus* present) and the proportions of 0s and 1s are roughly 0.49 and 0.51. A quick calculation shows that 51% of 49 observations at node #14 equals 25 (64% of the 39 total) occurrences of this species correctly classified by this model, or about a 36% omission error rate. This does not account for varying the classification thresholds or for errors of commission (see Chapter 9).

graphically (Figs. 7.1, 7.2). The splitting rules define the branching at the "internal nodes" of the tree, and the composition of the "terminal nodes" or leaves of the tree defines the predicted value. The predicted value is the average value of the training data in that node in the case of regression trees, or the majority class in the case of classification trees.

When do you stop partitioning the data or growing the tree? Taken to its logical extreme, rules could be derived to classify every observation in the training data into a one-member terminal node. These rules would classify the data perfectly. However, the resulting decision rules would probably not be very good at correctly predicting new observations. This is a classic example of "over fitting" the training data. Using the terminology introduced in Chapter 6 we would say that a tree model carried to this extreme has low bias but high variance (it makes very few assumptions and can make accurate predictions for the training

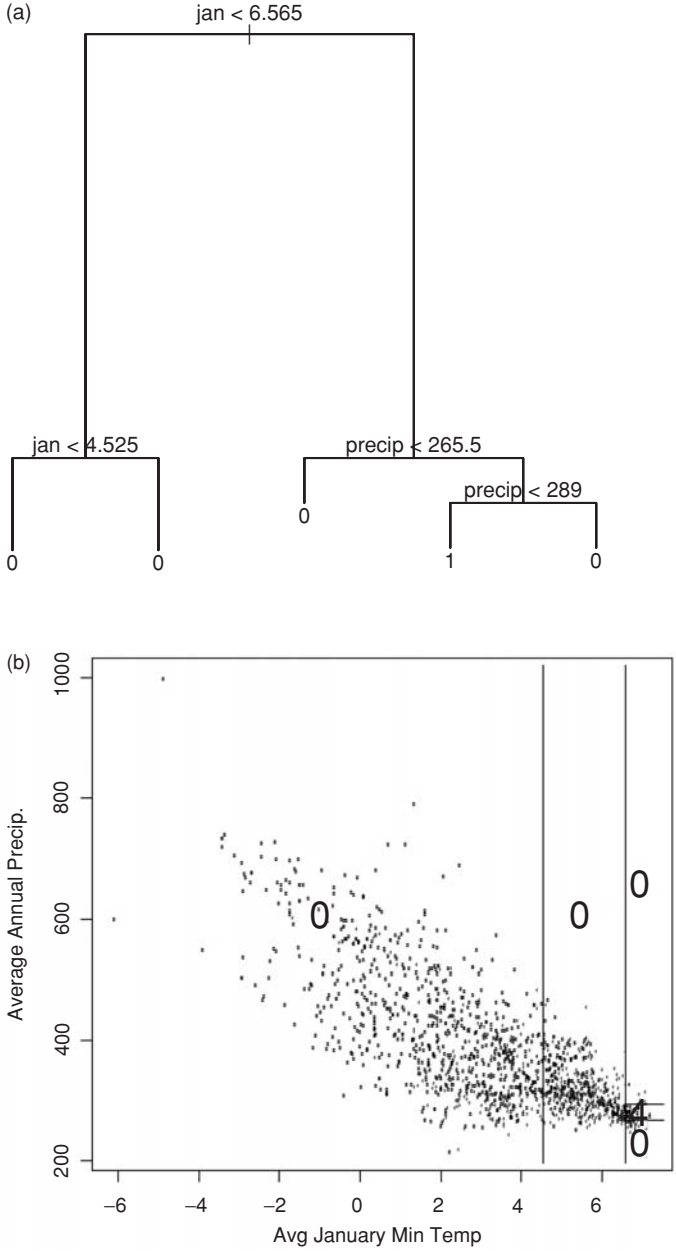

Fig. 7.1. Graphical representations of the decision tree presented as text in Table 7.1: (a) the tree diagram is useful because the length of the branch is proportional to the decrease in impurity or the deviance explained by the split.

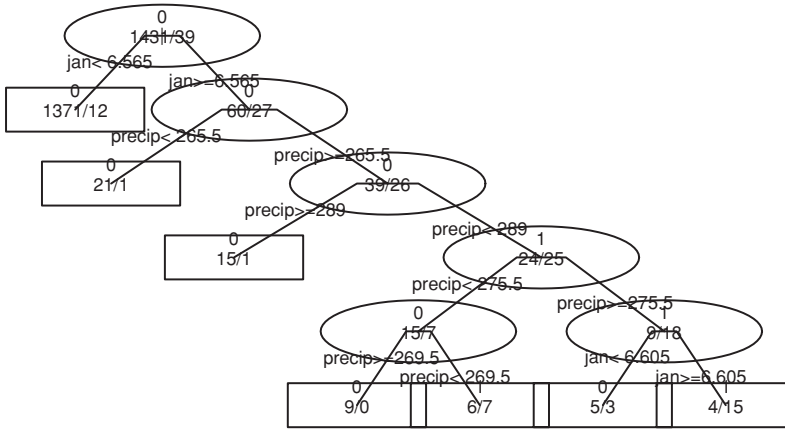

Fig. 7.2. An alternative decision tree model and graphical representation based on the same data used in Table 7.2 and Fig. 7.1. Internal nodes (groupings other than the final groups) are ovals and terminal nodes (the final groups) are rectangles. The predicted class is given (0 or 1), and below it the number of each class (0/1) at each node. In this case *C. verrucosus* presence (1) is the predicted class at two terminal nodes, correctly classifying 7 + 15 = 22 of the 39 cases of this species occurrence. This model was derived using a different software implementation, the rpart package in R.

data but is not very transferable to new data). Usually, partitioning is stopped when the resulting split does not achieve some defined level of increased homogeneity (or explained deviance), or when the resulting subsets would have less than some minimum number of members (say, five). However, because of the nested nature of the process, using a rigorous threshold level of explained deviance is not a good strategy – a split of low value may lead to one of higher value below it (Hastie *et al.*, 2001).

←

Caption Fig. 7.1 continued.
Comparing Table 7.2 with this figure, the convention used is to plot the tree with the decision rule ("January temperature less than 6.565") indicating go left in the decision tree. So, observations would be classified as *C. verrucosus* is present (predicted value = 1) if average January minimum temperature is above ~6.57 °C and average annual precipitation is between about 266 and 289 mm. Because the model only has two predictors, it can also be depicted as shown in the bottom graph, (b) showing the decision thresholds for the two environmental predictors shown on a scatterplot of the data. This tree model was developed using the tree package (Ripley, 1996) in the R software (R Development Core Team, 2004).

Tree-based methods have been developed over the past 25 years (Breiman *et al.*, 1984), and the approach that has proven most effective is to "grow" a large decision tree using fairly liberal stopping criteria, and then to "prune" the tree, which means to remove the splits (that is, collapse the internal nodes) that add the least to overall subgroup homogeneity (according to those same criteria of explained deviance or purity used to define the splits). Often, the change in the misclassification error rate is used as the criterion for pruning nodes from classification trees. The goal is to prune the decision tree to a size (number of final groups or terminal nodes) that is likely to provide robust predictions for new data. That size is usually determined through some form of *k*-fold cross-validation procedure – dividing the training data into *k* subsets, developing a tree model with $(k-1)/k$ of the data, testing it with $1/k$ of the data by calculating error on that $1/k$, and repeating *k* times (Fig. 7.3). Then the error rates for the *k* trees are calculated for all possible sizes as the trees are successively pruned from their terminal to root nodes by successively collapsing nodes that produce the smallest per-node decrease in a purity measure (or increase in error rate). The best tree size is the smallest one that produces an estimated error rate within one standard error of the minimum (Breiman *et al.*, 1984; De'ath & Fabricius, 2000), reflecting a balance between size and error. This procedure is also called cost-complexity pruning (Hastie *et al.*, 2009).

It should be noted that, when a classification tree is used for prediction, the majority (plurality) category can be predicted at a node, or alternatively the proportion of training observations in the majority category at that node (Table 7.2, Fig. 7.2) can be predicted. This is particularly useful for spatial prediction in SDM because that proportion has been interpreted as the probability or "suitability" (Pontius & Schneider, 2001) of the category or event, e.g., occurrence of a species. This makes the output of prediction from a CT analogous to the probability of an event predicted by logistic regression (GLM), or its GAM analog. The other machine learning methods described below also yield probabilistic predictions, e.g., on some kind of continuous scale.

7.2.2 When are decision trees useful?

Why are classification and regression trees useful as an alternative modeling method in SDM? What are their advantages and disadvantages? Decision trees have been used in many fields, including medical diagnosis, and have been developed for "mining" large, messy datasets for patterns and information. SDM is just one of these applications. They

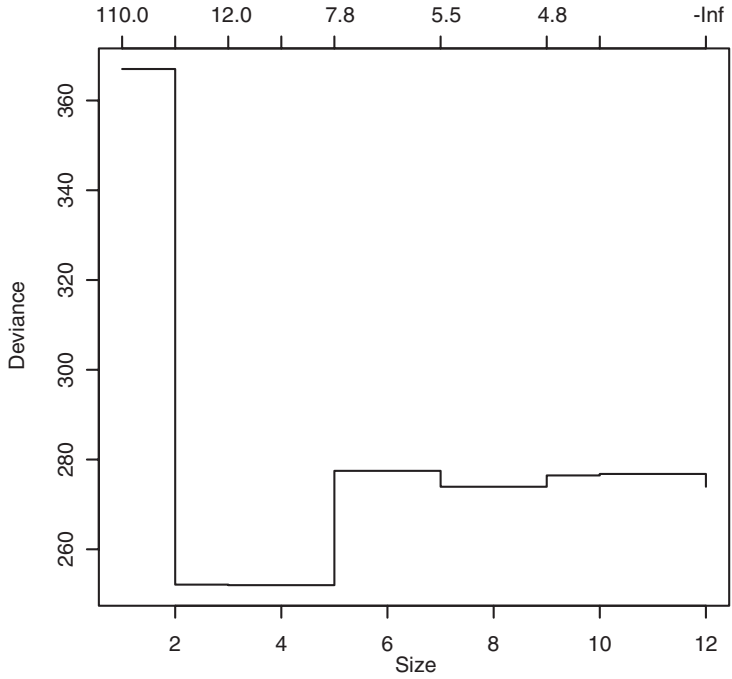

Fig. 7.3. Cross validation to determine the optimal tree size – that is, a tree model that makes robust predictions for unseen data (not used to train), and is therefore not overfitted to the training data. This shows the average deviance explained (*y*-axis), based on ten-fold cross validation, for nested trees of various sizes (number of terminal nodes, *x*-axis), calculated by pruning each tree from its maximum size (12 nodes) to the root (1 node). The minimum unexplained deviance at 2–5 nodes indicated that, for this example, a tree pruned to 2–5 terminal nodes will make robust predictions for unseen data and not be overfitted to the training data. The scale on the upper axis shows the cost-complexity parameter (a measure of the change in fit per number of terminal nodes) associated with the trees of each size (Breiman *et al.*, 1984; Venables & Ripley, 1994). This plot is based on the data for *Ceanothus verrucosus*, described in Table 7.2, and the resulting tree model shown in Table 7.2 and Fig. 7.1 is the best five-node (pruned) tree, not the full tree. Although a two-node tree would be the most parsimonious based on this corss-validation, a five-node tree is shown for illustrative purposes.

are noted to be particularly good at handling some kinds of data and problems:

(1) *Categorical predictors*: Decision trees handle both ordered and categorical predictors "in a simple and natural way" (Breiman *et al.*, 1984, p. 56). In SDM, categorical predictors such as vegetation type, soil

type or land cover can be difficult to parameterize and interpret in a linear model if they have many categories. While it is possible to treat each category as a dummy variable (0/1) in linear models, the results are often difficult to interpret, and it uses up a lot of degrees of freedom. It is best to aggregate categorical variables to as few categories as possible, based on ecological criteria, when using them in GLMs and GAMs (Chapter 6). In contrast, in classification trees, categorical predictors are easily associated with categorical responses in an approach that is analogous to contingency table analysis. This is especially useful with a categorical response such as species occurrence.

(2) *Hierarchical interactions*: Tree-based methods characterize interactions between variables, or " . . . [make] powerful use of conditional information" (Breiman *et al.*, 1984, p. 56). They identify and model non-linear and non-additive relationships between predictors and a response in relatively simple ways. In addition, hierarchical responses (when the response to a predictor is conditional on the values of another predictor) are very naturally described by decision trees. For example, a certain level of soil moisture may be adequate to support a species only if the maximum temperatures in the warmest season are below some threshold level. The species may be able to survive at higher temperatures given more moisture. These kinds of interactions would have to be specified at the outset in a linear model (Michaelsen *et al.*, 1994).

(3) *Threshold responses*: Decision trees characterize threshold effects of predictor variables on response. Multivariate species response functions are likely to be complex, or appear to be because of data limitations (Barry & Elith, 2006), and may be characterized by threshold effects, rather than linear or smooth responses (Chapter 3). The method of recursive partitioning is very effective at characterizing a threshold effect of an environmental variable on species response given that the splitting rules are based on threshold values of ordered response variables.

(4) *Informative output*: Trees are effective at exploring, graphically and quantitatively, complex relationships in multivariate data (De'ath & Fabricius, 2000). For many people, the structure of a decision tree, as it is displayed graphically, is an intuitive and informative way to describe patterns in data ("if elevation is above X, and slope aspect is north-facing, and soil type is A, then the species is present"). On the other hand, very large trees can be difficult to interpret.

(5) *Missing data*: Decision trees handle missing values and outliers in a very robust way. These observations get isolated in the tree without influencing the estimation of other model "parameters," that is, decision rules (Clark & Pregibon, 1992; De'ath & Fabricius, 2000).
(6) *Classifying new data*: Trees were developed to be used for prediction (Breiman *et al.*, 1984; De'ath & Fabricius, 2000). Although applying the classification rules in a GIS for spatial prediction can be cumbersome for large decision trees, the process is conceptually straightforward. Some software applications generate a simplified set of classification rules (C5 or "See5," Quinlan, 1993) and spatial prediction from decision trees has been automated in some software systems (see Table III.1).

There are some drawbacks to classification and regression trees. When a species response is linear or smooth, it is not well characterized by trees because they use rules based on thresholds, and would have to approximate a linear response with a step function (De'ath & Fabricius, 2000). Further, because later splits are based on fewer and fewer observations, trees may require large samples to detect patterns (Vayssiéres *et al.*, 2000), especially when there are many predictors. With species presence/absence data, the limiting factor may not be the overall sample size but the number of observations characterizing a rare class (such as presence of a rare species).

Finally, trees can be very unstable (Benito Garzón *et al.*, 2006; Prasad *et al.*, 2006; Hastie *et al.*, 2009). This means that varying the inputs, either by sampling from a set of observations, or varying the set of explanatory variables used, can result in very different models (predictors, decision rules), although the error rate or predictive ability may be very similar (for example see Scull *et al.*, 2005). This illustrates several points. One is that this method is very data-driven or non-parametric, and a slightly different set of training data can yield a different selection of threshold values or variables. Another is that, given a large set of potential predictors, applying the criterion of maximizing the increase in node purity does not necessarily distinguish well among them. In other words, two variables may give very similar results and the tree model arbitrarily selects the one that gives an ever-so-slightly greater increase in purity. So, paradoxically, although trees are touted as effective and efficient at variable selection, it can be difficult to interpret variable importance from these models. The instability of decision trees is being addressed recently through the improved tree-based methods described in Section 7.3.

7.2.3 A note about multivariate decision trees

Multivariate decision trees (Brodley & Utgoff, 1995), that is, trees with multiple response variables, have also been proposed as a method of predicting species composition based on environmental data (De'ath, 2002). I have not yet seen this approach applied in SDM, that is, for mapping multiple species, although decision trees have been used to map multi-species assemblages or communities (e.g., Lees & Ritman, 1991; Franklin *et al.*, 2000; Ferrier *et al.*, 2002a). Frequently ecological surveys collect information about multiple species (Chapter 4), and given the demonstrated strength of multivariate response modeling of multiple species using MARS (Leathwick *et al.*, 2006b), measures of compositional dissimilarity (Ferrier *et al.*, 2002a) and other approaches (see Section 6.5), I anticipate that these multivariate approaches will become more widely used.

7.2.4 Application of decision trees in species distribution modeling

An excellent "introduction" to classification and regression trees for ecologists was published in 2000 (De'ath & Fabricius, 2000). However, there were applications of decision tree (DT) models in ecological analysis, biospatial prediction and landscape stratification published prior to that (for review, see Franklin, 1995; Guisan & Zimmermann, 2000). Table 7.3 summarizes some of the recent and not so recent studies specifically related to SDM. In these and other studies, DTs have been applied to the SDM problem and used for spatial prediction because of the advantages listed above, or have been compared to other modeling methods. As was discussed, an emerging generality about the efficacy of decision trees for SDM is that they are somewhat unstable and have lower classification (prediction) accuracy than other methods discussed in Chapters 6 and 7. This has led to the development of new decision tree methods described in the next section.

7.3 Ensemble methods applied to decision trees – bagging, boosting, and random forests

New, computationally intensive methods have been developed that address some of the shortcomings of classification and regression trees. Because these all involve estimating a large number of tree models based on subsets of the data and then averaging the results, they are considered

Table 7.3. *Examples of studies using decision trees (DT) and random forests (RF) for species distribution modeling and related applications*

Citation	Response variable	Application	Comments
(Michaelsen *et al.*, 1987)	Oak seedling survival	Ecological data analysis of hierarchical responses	Early application of DTs to ecological data analysis.
(Walker, 1990)	Kangaroo species occurrence	Potential habitat in relation to climate change	DTs can identify local interactions.
(Moore *et al.*, 1991a)	Vegetation class	Vegetation mapping	Early application of DTs for predictive mapping of vegetation classes.
(Lees & Ritman, 1991)	Vegetation class	Vegetation mapping from remotely sensed and GIS data	Combined gradient modeling of species composition with imagery to distinguish land cover using DT.
(Michaelsen *et al.*, 1994)	Spectral vegetation index	Ecological land classification	Noted usefulness of DTs with hierarchical interactions among predictors.
(Lynn *et al.*, 1995)	Vegetation types	Forest mapping	Focus on accuracy assessment.
(Bell, 1996)	Bird species	Potential habitat	Demonstration of DT method.
(O'Connor *et al.*, 1996)	Bird species richness	Predict large-scale biodiversity patterns	DTs can identify hierarchical relationships among predictors.
(Franklin, 1998)	Plant species occurrence	Compare modeling methods	DTs outperformed GAMs and GLMs but on training data.
(Vayssiéres *et al.*, 2000)	Plant species occurrence	Demonstrate method and compare to GLM	DTs outperformed GLMs in 2/3 of cases, selected similar predictors, were more interpretable and better at detecting interactions.
(Meentemeyer *et al.*, 2001)	Plant species abundance	Ecological inference	Used method because of ability to describe hierarchical interactions among predictors.

(Franklin, 2002)	Plant species occurrence	Potential species distribution maps required for landscape simulation model	Refining a regional vegetation map with distribution of dominant species predicted using DTs.
(Miller & Franklin, 2002)	Vegetation class	Compare modeling methods with and without spatial dependence	DTs lower accuracy than logistic regression (illustrating the overfitting problem).
(Flesch & Hahn, 2005)	Invasive bird species (brood parasite)	Compare native and colonizing range of invasive species	Used DT because predictors not expected to have a common effect across the entire sample, and hierarchical interactions expected.
(Fertig & Reiners, 2002)	Plant species	Species range maps for conservation planning and management	DT and logistic regression selected different predictors, and while they had similar commission error rates, DTs had lower omission error.
(Thuiller *et al.*, 2003)	Plant species	Compare modeling methods at multiple scales	DT performance slightly lower than GLMs, GAMs, especially at finer scale.
(Accad & Neil, 2006)	Vegetation class	Map potential distribution of vegetation classes from remnant patches	DTs modeled presence of 28 vegetation types and combine results *a posteriori* for a more realistic representation of vegetation patterns.
(Rehfeldt *et al.*, 2006)	Plant communities and plant species	Evaluate potential climate change impacts	RF used because of high accuracy in preliminary analysis.
(Prasad *et al.*, 2006)	Tree species	Evaluate potential climate change impacts	RF and bagging trees yielded better interpolations and projections than DT and MARS.
(Poulos *et al.*, 2007)	Vegetation types and fuel classes	Predictive mapping for forest and fire management	DTs used because they were able to deal with outliers and categorical predictors, and yield interpretable classification rules.
(Benito Garzón *et al.*, 2008)	Tree species	Predict impacts of climate change	RF used because of better performance than DT and ANN.

(By no means a comprehensive list – see also De'ath & Fabricius, 2000; Olden *et al.*, 2008). The selected studies emphasized spatial prediction.

forms of model averaging or "ensemble modeling" (see Section 7.8) and are beginning to be used in SDM and related ecological applications with great success.

In order to avoid developing a tree model that is over-fit to the training data, bootstrap aggregation or "bagging" (Breiman, 1996) works by repeatedly (say, 30–80 times) sampling the data with replacement (bootstrapping) and developing trees for each dataset using some stopping rule but without pruning. Typically about 1/3 of the data are held out of each sample ("out-of-bag") and used to evaluate the model, while other data are replicated to bring the "in-bag" sample to full size (Prasad *et al.*, 2006). Then the predictions based on all of the trees are averaged (using a plurality voting rule in the case of a categorical response such as species presence/absence). In other words, each of the many models is used to make a prediction for each observation in a new dataset, and these predictions are averaged.

Another variation called "boosting" (Freund & Schapire, 1996; Ridgeway, 1999) is somewhat similar to bagging except that each observation, instead of having an equal probability of being selected in subsequent samples, is weighted to have a higher probability of selection if it is a "problem" observation (tended to be misclassified by previous models). Therefore, among the methods available for developing and then combining the results from many tree models, boosting is unique because it is sequential.

A form of boosting called stochastic gradient boosting (SGB) (Friedman, 2002) builds many small tree models sequentially from the residuals (referred to as stepping down the gradient of the loss function) of the previous tree. The loss function is defined as some measure such as deviance (Chapter 6) that quantifies the lack of fit of a sub-optimal model. Again, at each step, a model is developed with a random subsample of the data. Hundreds to thousands of tree models are developed in this way (Elith *et al.*, 2008).

Boosting, particularly SGB, has been used for SDM and spatial prediction of other ecological variables and has been shown to perform better that ordinary decision trees (Elith *et al.*, 2006; Leathwick *et al.*, 2006a; Moisen *et al.*, 2006; De'ath, 2007; Elith *et al.*, 2008). These recent papers give a useful introduction to boosted tree methods and practical guidance for their application to SDM and similar ecological questions. For example, measures of the relative importance of predictor variables are available for Boosted Regression Trees (based on how many times the variable was used, in how many trees, weighted by the deviance it

explained), and partial dependence plots show the effect of the value of the predictor on the response, yielding a response curve (Elith *et al.*, 2008). With these diagnostic tools, this ensemble modeling method does not have to be a "black box" with regard to ecological interpretation of the model.

Random forests (Breiman, 2001b) is a form of bagging that builds a large number of de-correlated trees (Hastie *et al.*, 2009) and averages them. As in bagging, many trees are developed with subsets of the data, but in addition, each split in each tree model is also developed with a random subset of candidate predictor variables. A large number (500–2000) of trees are "grown" to a maximum size (without pruning) and then the resulting predictions are averaged. The "out-of-bag" (test) sample, the set of observations held back, is used to estimate model error and variable selection or importance.

Variable importance is estimated in two ways for random forests (RF). For each decision tree there is a misclassification error rate (in the case of classification of a categorical outcome) or mean squared error (regression) calculated from the out-of-bag sample. The difference between this error rate and the error rate calculated by randomly assigning the values of a predictor variable, and then passing the test data down the tree to get new predictions, is a measure of the importance of that predictor. This decrease in accuracy if the variable were randomly permuted, the difference between the error rates or mean squares errors, divided by the standard error, is calculated as one measure of variable importance (Cutler *et al.*, 2007). Another measure of variable importance, based on the training (in-the-bag) data, is the reduction in sum of squares (deviance) achieved by all splits in the tree that use that variable, averaged across all the trees (Prasad *et al.*, 2006).

With all of these ensemble tree methods, the tendency to over-fit the data is overcome by averaging the predictions from a large number of models based on subsets of the data. Random Forest modeling is beginning to be used in SDM (e.g., Benito Garzón *et al.*, 2008; Lawler *et al.*, 2009), and, like SGB, has been shown to have higher prediction accuracy than ordinary decision trees in SDM and other applications (Prasad *et al.*, 2006; Cutler *et al.*, 2007). While the methods for determining variable importance are a great improvement over interpreting variable importance from single trees (because of the instability problem), the ease of interpretation of the rule sets for individual decision trees is lost in the ensemble methods (bagging and boosting). That is, these methods do not allow one to determine if "the likelihood of Species A occurring

declines drastically if annual precipitation is less than 200 mm," as from a single classification tree. However, fitted functions can be visualized for all ensemble tree methods using partial dependence plots, as noted above (Elith *et al.*, 2008). Although these graphs do not perfectly represent the effect of each predictor, especially if there is strong multicollinearity (Friedman, 2001), they are extremely useful for interpreting species response functions, and analogous to graphical tools used for GLMs or GAMs.

7.6 Maximum entropy

Maximum entropy or "Maxent" is a general-purpose machine learning method that has been developed in statistical mechanics and is being applied in fields as disparate as finance and astronomy. Maximum entropy is a principle from statistical mechanics and information theory that states that a probability distribution with maximum entropy (the most spread out, closest to uniform), subject to known constraints, is the best approximation of an unknown distribution because it agrees with everything that is known but avoids assuming anything that is not known (Phillips *et al.*, 2006). When applied to the species distribution modeling problem, for example, the distribution being estimated is the multivariate distribution of suitable habitat conditions (associated with species occurrences) in environmental feature-space. The unconstrained distribution is that of all factors in the study area, and the constraint is that the expected value is approximated by an empirical set of observations of species presence. A software application designed specifically for SDM has recently been developed (Phillips *et al.*, 2006; Phillips & Dudík, 2008). This method has generated great interest because in comparisons it has shown higher predictive accuracy than many other methods when applied to "presence-only" species occurrence data (Elith *et al.*, 2006). Because it was specifically developed for use with presence-only data, and to overcome problems of small undesigned samples, applications of Maxent in SDM will be discussed further in Chapter 8.

instead of estimating the probability density function, the range of conditions (set of points) where the probability is greater than zero is modeled by estimating a kernel. In other words, hyperplanes are fit around the data points. The one-class SVM, like Maxent and GARP, has been touted as being particularly effective for presence-only species observations (Drake *et al.*, 2006). It was suggested by Drake *et al.* that, because one-class SVM estimates the support of a distribution based on presence-only observations, its interpretation is consistent with the concept of the species niche as a portion of environmental hyperspace. In contrast, other "presence-only" methods that discriminate occupied versus available habitat using background data (ENFA, Maxent) would be more consistent with the resource selection function concept (discussed in Chapter 3).

However, one-class and two-class SVMs were compared in an SDM application where only presence observations were available to predict the distribution of an invasive forest pathogen (Guo *et al.*, 2005). In the two-class case, pseudo-absence (background) data were generated (Chapter 4). This study found that one-class SVMs tended to predict a much larger area of suitable habitat or potential distribution when compared to other kinds of models. In applications where minimizing omission error is important, e.g., estimating risk of forest disease, this could be an advantage. But the tendency to overestimate species distributions is a characteristic that one-class SVM shares with some other presence-only methods (Chapter 8).

SVM has been applied to the presence-only specie distribution modeling problem only to a limited extent, but the one-class model has been shown to work well with as few as 40 observations (Drake *et al.*, 2006). In another study focused on predicting forest pathogen invasion, two-class SVM was compared with four other methods: expert rule-based (Chapter 8), logistic regression, GARP, and decision trees. SVM yielded similar spatial predictions to the other models and had the lowest omission error, although validation data were very limited (Guo *et al.*, 2007). This machine learning method seems to hold some promise for SDM and similar applications, although it is difficult to tell if it is possible to interpret its classification rules from these models in order to verify its ecological meaning or validity (in terms of the importance of predictors and the form of their relationship with the response). Tools to interpret and visualize the classification rules would be important to develop if SVM is to be more widely and effectively used in SDM.

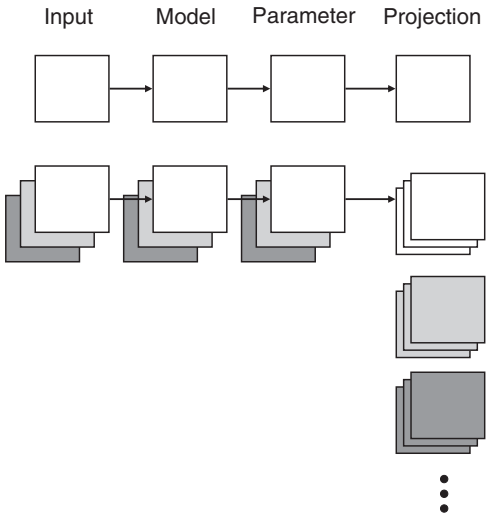

Fig. 7.5. Schematic representation of ensemble model forecasting in SDM (modified from Araújo & New, 2007; Beaumont, 2008). Ensembles of forecasts (projections) can result from varying the input (initial conditions), for example, the species occurrence data and predictors; the model type used in SDM; and the parameters (may vary with model selection and input predictors). New input predictors may result from environmental change (e.g., climate change) scenarios that are derived from multiple realizations of models and inputs.

7.8 Ensemble forecasting and consensus methods

Ensemble forecasting (Fig. 7.5) can be defined as making multiple simulations (projections) across some range of initial conditions (data inputs), model types, parameters (coefficients), and/or boundary conditions (new predictors) (described for SDM by Araújo & New, 2007). An ensemble of predictions can result from different realizations of the same class of model, either because the model has a stochastic element, or because the inputs or other modeling decisions are varied, or an ensemble of predictions can result from different classes of models. In the second case, more than one model is developed and they are used together in some way either to explore the range of predictions, better understand model uncertainty, or get a better projection by using the predictions together. Averaging predictions from different classes of models has also been called model fusion (Elder, 2003).

Several machine learning methods described in this chapter are referred to ensemble modeling methods because many models are estimated, often using subsets of the data, and then their predictions are averaged or composited in some way. For example, bagging, boosting and Random Forests use this approach where the predictions from many decision tree models are averaged. Operationally, GARP and ANN models are often run multiple times, and then a subset of the best models is averaged.

When thinking about how to combine predictions or forecasts from different models we might automatically think of taking their average. "Consensus method" refers to a way of finding the majority view or agreement among different model outputs, and this can involve more than just simple averaging. For example, five consensus methods were compared, in the context of SDM, and these included taking the average, median, weighted average, and median PCA (principal components) of the predictions, as well as using the single best model (Marmion *et al.*, 2009). That study found that average and weighted average predictions were more robust than other consensus methods or single model predictions. Other studies have found that other consensus methods worked equally well or better, and that consensus predictions are not always an improvement over single models. So, while a variety of consensus methods exist, the best approach will vary with the input data and according to whether extrapolations are being made over space or time.

BIOMOD is software dedicated to SDM that typically uses an ensemble modeling approach based on multiple classes of models (e.g., model fusion). Different kinds of models are used to estimate SDM (specifically GLM, GAM, DT and ANN), and the resulting projections are combined using principal components analysis (Thuiller *et al.*, 2003; 2006). Similarly, BIOMOD was used in one case to develop four models per species in a study projecting climate change impacts on species distributions. However, instead of combining the four model outputs using PCA, the entire ensemble of forecasts, using four models, two post-processing rules, and five climate change scenarios, was combined using clustering to detect groups of forecasts making similar projections (Araújo *et al.*, 2006).

Recent SDM studies using an ensemble of forecasts to improve reliability and characterize uncertainty have emphasized temporal extrapolation of species distributions under climate change scenarios (for example,

Thuiller, 2004; Araújo *et al.*, 2005a). These studies are particularly concerned with characterizing SDM variability because, in addition to uncertainty caused by variations in the SDM modeling methods, inputs, and parameters, coupling SDMs with climate change scenarios involves using multiple realizations of different climate models (Araújo & New, 2007; Beaumont, 2008) to establish new inputs (boundary conditions, Fig. 7.5).

However, whether it was explicitly called "ensemble forecasting" or not, a number of recent SDM studies have also fused or averaged predictions across different types of models, for example, to examine the risk of invasive species establishment (Kelly *et al.*, 2007; Crossman & Bass, 2008). Consensus approaches are particularly useful in cases where there are not adequate data to evaluate which model makes the most accurate predictions, as is the case with species invading new areas, projecting future distributions under environmental change scenarios, and in very data-poor regions (e.g., Hernandez *et al.*, 2008).

7.9 Summary

Supervised machine learning methods have been used in SDM and related ecological modeling applications for several decades (Skidmore, 1989; Lees & Ritman, 1991; Stockwell & Noble, 1992, and see Table 7.3). However, the application of newer methods to the SDM problem has greatly increased recently. A pattern is emerging whereby, when they are compared, "classic" single decision trees tend to perform somewhat worse than both the statistical methods discussed in Chapter 6, and the newer machine learning methods discussed in this chapter. Single decision trees can be very useful for exploratory analysis, especially with categorical predictors, because of the interpretable decision rules presented in graphical form. Boosted regression trees and random forests are ensemble decision tree methods that show a great deal of promise for SDM and other complex classification problems. Other machine learning approaches that can be implemented as iterative or ensemble methods, including artificial neural networks and genetic algorithms, tend to perform well given complex classification problems, but can be difficult to use – there can be a steep learning curve for their effective application. Further, interpreting the modeled relationships between predictors and response is not always straightforward in some machine

learning implementations. GARP and Maxent are machine learning-based approaches with dedicated software developed specifically for SDM using presence-only data. They are discussed in detail in Chapter 8. Maxent, in particular, is emerging as a tool for SDM that is able to make robust predictions with very few presence-only species observations, e.g., in remote and poorly studied regions.

8 · *Classification, similarity, and other methods for presence-only data*

8.1 Introduction

Many of modeling methods described in Chapters 6 and 7 require observations of species presence and absence, preferably a lot of them (in order to characterize complex response functions), well distributed in space and along environmental gradients. These data are required to estimate model parameters or to derive decision rules for supervised classification. If presence and absence data are available, the modeling approaches that are designed for binary response variables, discussed in the previous chapters, generally give more accurate predictions than models based on presence-only data (e.g., Brotons *et al.*, 2004), but not always (Hirzel *et al.*, 2001). But what if only observations of species presence (but not absence) are available, or what if there are no georeferenced observations at all, but simply some expert knowledge on species habitat requirements? The methods described in this chapter can be, and have been, applied to SDM in these situations. It is actually a very common predicament, to have species presence data, or no species location data at all, and the reasons for that are reviewed here.

If a species has been recorded as being present in a location, we can be fairly certain that it occurs there (except for taxonomic misidentifications). We then make the assumption the occurrence of an organism indicating habitat *use*, *occupancy* or *suitability* for the purpose of modeling. This assumption may perhaps be more easily made with some kinds of data (abundance, observations of foraging, nesting, maps of home ranges) than others (point counts, trapping, species lists). In some types of surveys, if observations are made at a location that do not include a species, e.g., if it is recorded as "absent" (implying non-use or non-suitability), we are less certain about this absence because of the species detectability issues discussed in Chapter 4. Wide-ranging or highly mobile organisms may potentially use or occupy the location, but not at the time of the survey, cryptic organisms may require a more extensive search, and so forth.

Other types of data provide no information about absence but only include locations where a species was observed. As Boyce *et al.* (2002, p. 282) stated, "in some sampling situations we cannot estimate a sample of unused sites." This includes georeferenced natural history collection records (Graham *et al.*, 2004), opportunistic "sightings" that are recorded, and radiotelemetry data. Even these three examples include widely varying types of data with regard to their ability to identify suitable habitat. In contrast with specimen records and informal sightings, radiotelemetry data are generally collected densely on a small portion of the landscape with the expressed purpose of determining activity, movement or habitat use by an individual.

This chapter describes three general approaches to modeling habitat suitability when only data on species presence are available:

(a) For unvisited (new) sites, calculate some measure of similarity to suitable habitat as described by the values of environmental variables for observed species locations (Section 8.2).
(b) Model presence (use) versus availability of habitat, characterizing the available habitat by a sample of other locations in the landscape, or by complete census of all locations, made possible through GIS (Section 8.3).
(c) Develop a habitat suitability model based on expert opinion to assign weights to, and defined transfer (mapping) functions, for environmental predictors (Section 8.4).

These approaches will be discussed and compared in this chapter. An important concept to keep in mind is that models fit to presence-only data predict the relative likelihood of species presence at a site, or the relative habitat suitability (as articulated by Elith & Leathwick, 2009). This is because presence-only data lack information about species prevalence (Chapter 4). Presence/absence data, on the other hand, can be used to predict the probability of species presence (or of encountering the species if sampling with the same method used to derive the training data) using the methods described in Chapters 6 and 7.

8.2.2 Environmental distance methods

Ecologists may already be familiar with multivariate coefficients of similarity that can be applied to community data (multiple species and sites; see also Section 6.5). For example, Legendre and Legendre (1998, p. 299) list more than 20 metrics. Similarly, there are a number of measures of association among objects (observations) based on multivariate

descriptors of those objects (Legendre & Legendre, 1998, p. 300). For example, the Euclidean multivariate distance (dissimilarity) between two objects can be calculated based on several descriptor variables.

DOMAIN was one of the first implementations of distance-based method for SDM (Carpenter *et al.*, 1993). DOMAIN used "a point-to-point similarity metric to assign a classification value to a candidate site based on the proximity in environmental space of the most similar record" (p. 670–671). A Gower metric is a multivariate measure of similarity (Legendre & Legendre, 1998), and is used in DOMAIN to map predictions of this continuously-varying similarity measure.

The Gower coefficient of similarity (Legendre & Legendre, 1998, p. 259) between two observations, x_1 and x_2, is:

$$G(x_1, x_2) = \frac{1}{p}\sum_{j=1}^{p} s_{12j}$$

or the average, over the p descriptors, of the similarities, s, calculated for each descriptor. For quantitative descriptors, similarity is calculated as the complement of a normalized distance, where R_j is largest distance found across all sites or in a reference population:

$$s_{12j} = 1 - [y_{1j} - y_{2j}]/R_j$$

Another measure of multivariate association is Mahalanobis generalized distance. This measure takes into account the correlations among descriptors by scaling the difference in the mean vectors describing two groups by their covariance. Thus, this distance measure is independent of the scales of the various predictors (say, elevation and mean annual temperature). This is a very useful quality in a measure of multivariate environmental similarity between unsurveyed locations and locations that a species occupies when the descriptors typically show multicollinearity.

Although designed to compare groups of observations or sites (Legendre & Legendre, 1998) (p. 280), when applied to habitat suitability modeling, Mahalanobis distance is usually used to compare a single observation to a group of sites (Farber & Kadmon, 2003). The Mahalanobis distance statistic is a measure of dissimilarity and can be written, in matrix notation:

$$D^2 = (x - \hat{\mu})^T V^{-1}(x - \hat{\mu}),$$

where, as it is applied in SDM, $\hat{\mu}$ is the mean vector of habitat characteristics based on some number of observations of species occurrence, x is the vector of habitat characteristics associated with a new observation or

point, such as a grid cell in a map, T indicates the transpose of the vector, and V^{-1} is the covariance matrix associated with $\hat{\mu}$. While Mahalanobis distance has long been used as a measure of multivariate similarity in ecology (see Greig-Smith, 1983), I believe that its first application to habitat modeling was by Clark *et al.* (1993). From the very first, this dissimilarity measure was applied directly to environmental maps in a GIS – the distance statistic was calculated from each grid cell in the map (the stack of mapped environmental variables) to the mean vector calculated from a sampling of locations where the species was recorded, scaled by the covariance matrix for species presences. The raw Mahalanobis distance values can be depicted in a map, or can be scaled in some way (Fig. 8.2). For example, assuming multivariate normality, the statistic has a χ^2 distribution, and the P-values can be mapped (Clark *et al.*, 1993).

Since it was first presented in the SDM context and used for spatial prediction, Mahalanobis distance has been used as a habitat similarity index (Knick & Rotenberry, 1998) to predict the distribution of suitable habitat in a number of GIS-based studies. Interestingly, this approach seems to be most frequently applied in landscape-scale wildlife management applications of SDM (Knick & Dyer, 1997; Watrous *et al.*, 2006; Hellgren *et al.*, 2007), including species restoration, recovery and reintroductions (Corsi *et al.*, 1999; Thatcher *et al.*, 2006; Thompson *et al.*, 2006; Telesco *et al.*, 2007).

A variation improving on Mahalanobis distance as a measure of habitat similarity, based on partitioning, has been described (Rotenberry *et al.*, 2002, 2006). This approach, instead of using dissimilarity to optimal habitat, as D^2 conceptually measures, identifies the minimum set of basic habitat requirements by partitioning Mahalanobis distance using principal-components analysis (PCA) applied to the multivariate environmental data associated with species occurrence. Although it may seem counterintuitive to those who have used PCA for other kinds of data-reduction applications, it is actually the components (or partitions) with the smallest eigenvalues (smallest variation) that are of interest in this case – these represent the variables that maintain a constant value where the species occurs (Rotenberry *et al.*, 2006). The rationale is that environmental variables whose values vary widely among species locations are not very informative, while those that vary very little represent factors that limit a species distribution. This approach is beginning to be applied in wildlife management (Browning *et al.*, 2005) and global change studies (Preston *et al.*, 2008).

Distance-based approaches have limitations – they work best when organisms are using optimal habitat, are well-sampled in environmental

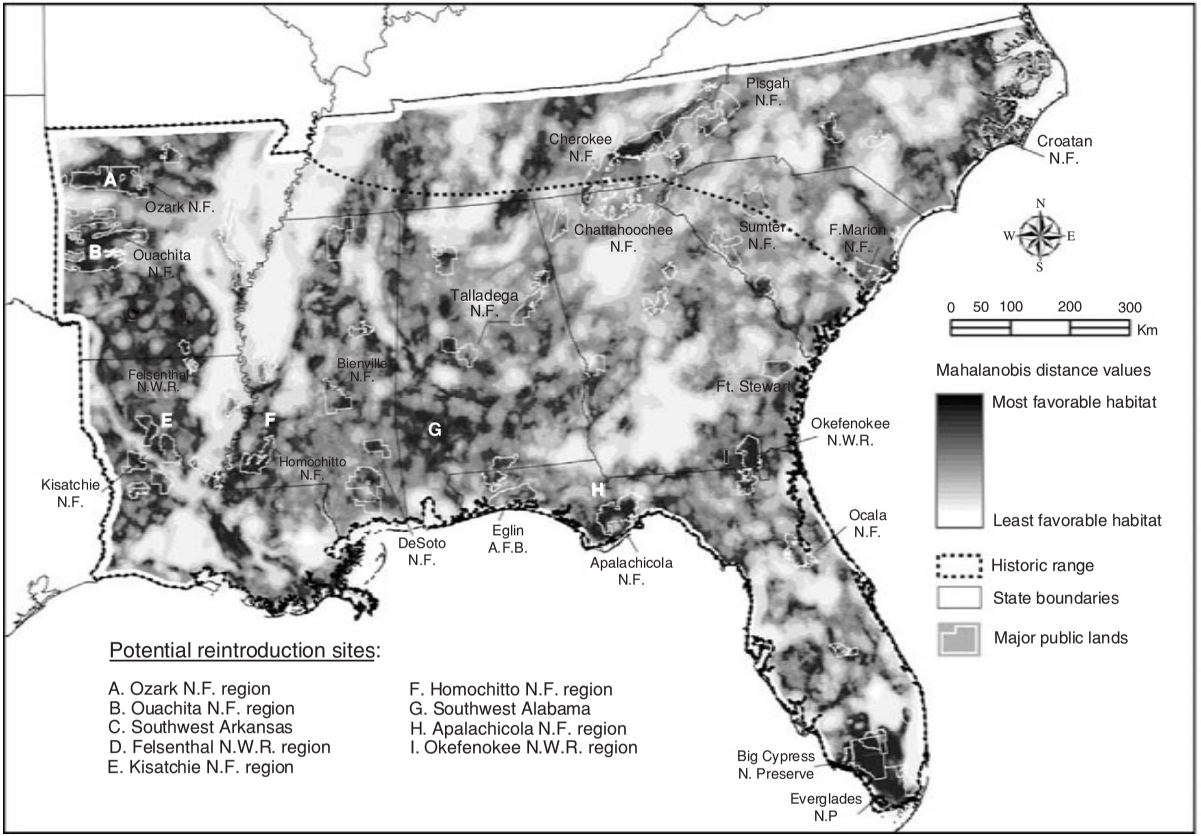

Fig. 8.2. An example of Mahalanobis distance used as a measure of habitat suitability, estimated for an endangered species, the Florida panther (*Puma concolor coryi*), and mapped for the southeastern USA (from Thatcher *et al.*, 2006). Mahalanobis distance is shown on a relative scale (darker tones indicate greater suitability). This analysis was used to identify potential reintroduction sites (indicated by letters A–I). Copyright © The Wildlife Society; Used with permission.

space, and when habitat variables are not dynamic (Knick & Rotenberry, 1998). Mahalanobis distance is more restrictive than other SDM methods in that it assumes that predictors are equally weighted, have normal error distributions and only considers linear relationships of species to environmental predictors. The metric is based only on continuous, quantitative, predictor variables. To overcome this last limitation, categorical predictors that are often associated with wildlife habitat suitability, such as land cover or vegetation type, are typically expressed numerically on a continuous measurement scale as an area or proportion of the area within some radius of a species location (e.g., Knick & Rotenberry, 1998; Johnson & Gillingham, 2005). However, I did find one application that appeared to use categorical predictors including vegetation type (Thompson *et al.*, 2006), perhaps by expressing each category as a 0/1 "dummy variable."

How does Mahalanobis distance compare to other SDM methods applied to presence-only data? Farber and Kadmon (2003) showed that D^2 described the bioclimatic species envelope more accurately than the BIOCLIM hyper-box classifier, while Tsoar *et al.* (2007) found that D^2 was among the most accurate of six presence-only methods compared. In another comparison, D^2 was evaluated using independent observations of caribou occurrence (Johnson & Gillingham, 2005) and performed comparably to a variety of methods. The other methods tested included logistic regression, used to estimate a resource selection function (habitat use versus availability), an HSI expert model, and GARP – these methods are discussed below. Notably, in that study, although prediction accuracies were comparable, the spatial distribution of predicted suitable habitat varied widely among methods. Disadvantages of Mahalanobis distance noted in that study were that it does not provide a measure of the influence of individual environmental variables on the distribution of the species (statistical inference, parameter estimation), nor is it amenable to statistical tests or information theoretic approaches to variable selection (as logistic regression is). Another disadvantage is that predictors are weighted equally. The distance based methods described here do not appear to offer any particular advantages over currently available methods described in Section 8.3. More comprehensive comparisons of presence-only SDM methods are discussed in Section 8.5.

8.3.4 Maximum entropy

Maximum entropy has been called both a general machine learning method and a statistical method, so I think of it as a statistical learning method. As noted in Chapter 7, it has been developed and used in other fields (for example, physics). However, because a software application, Maxent, was specifically developed to develop SDMs with "presence-only" species occurrence data (Dudik *et al.*, 2007), it is discussed in more detail here. Recall that, in maximum entropy, the multivariate distribution of suitable habitat conditions in environmental feature-space is estimated according to the principle of maximum entropy that states that the best approximation of an unknown distribution is the one with maximum entropy (the most spread out) subject to known constraints. The constraints are defined by the expected value of the distribution, which is estimated from a set of species presence observations.

Maxent has been described as a generative modeling approach (as opposed to discriminative; Phillips & Dudík, 2008) that models the species distribution directly by estimating the density of environmental covariates conditional on species presence. The raw output of Maxent is an exponential function that assigns a probability to each site. Raw values are dependent on the number of background and occurrence sites used because they must sum to one, and so they are converted to a cumulative probability (Phillips *et al.*, 2006). In the cumulative format values from 0–100 represent the range of probabilities predicted by the model. The cumulative format, while scale independent, is not necessarily proportional to the probability of presence, for example, if probability values are similar across the entire region, e.g., for a generalist species. Therefore,

newer versions of Maxent have introduced a logistic output format that estimates the probability of presence (Phillips & Dudík, 2008).

An important distinction between Maxent and logistic regression-type models is that in Maxent locations without species occurrence records are not interpreted as absences, but rather as representing the background environment. When the regression models are applied to presence/background data, interpretation of the resulting prediction is not clear cut because, instead of modeling probability of occurrence, the prediction is one of relative suitability (Phillips *et al.*, 2006), or the relative likelihood of habitat use (Boyce *et al.*, 2002). However, Ward (2007) have recently described an expectation-maximization (EM) algorithm to estimate the underlying presence-absence logistic model using presence/background data.

Phillips *et al.* (2006) outlined some advantages and disadvantages of Maxent for SDM compared to other methods. Maxent only requires presence data plus environmental information for the whole study area. The results are amenable to interpretation of the form of the environmental response functions (Fig. 8.5). Maxent has properties that make it very robust to limited amounts of training data (small samples), namely, that it is a density estimation method, not a regression method, and it is well-regularized (Phillips & Dudík, 2008). Because it uses an exponential model for probabilities, it can give very large predicted values for conditions that are outside the range of those found in the data used to develop the model. However, the new logistic output format addresses this problem (Phillips & Dudík, 2008). Extrapolation outside of the range of values used to develop an SDM should be done very cautiously no matter what modeling method is used (Elith & Graham, 2009). Its developers provide freely available Maxent software customized for SDM (http://www.cs.princeton.edu/~schapire/maxent/).

Maxent has been used in studies of species richness (Graham & Hijmans, 2006; Pineda & Lobo, 2009), invasive species (Ficetola *et al.*, 2007; Ward, 2007), climate change effects on species distributions (Hijmans & Graham, 2006; Fitzpatrick *et al.*, 2008), endemism hotspots (Murray-Smith *et al.*, 2009) and to estimate the extent of occurrence (Pearson *et al.*, 2007; Sergio *et al.*, 2007) and quality of protection (DeMatteo & Loiselle, 2008; Thorn *et al.*, 2009) of rare species. Maxent has also been used to investigate the degree to which climate constrains distributions of species (Wollan *et al.*, 2008; Echarri *et al.*, 2009; Cordellier & Pfenninger, 2009), including pathogens and their hosts. Maxent has also been employed as a tool to address ecological questions

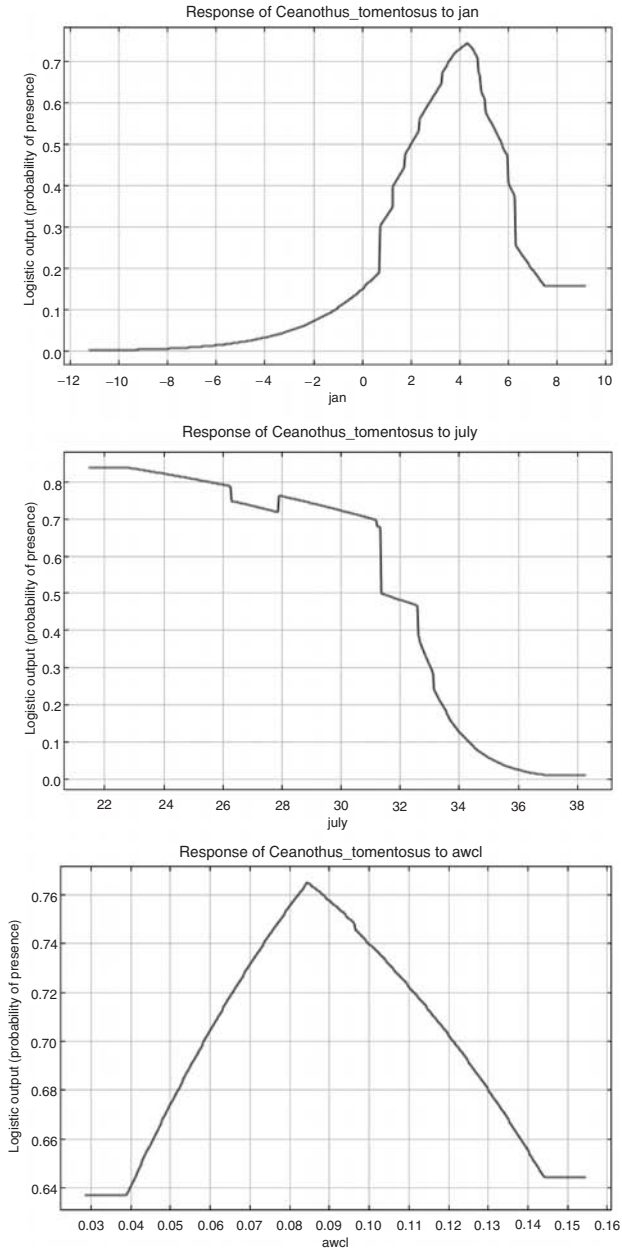

Fig. 8.5. Example of response curves as estimated from Maxent showing the log response of the focal species (*Ceanothus tomentosus*), a woody shrub (data described in Fig. 4.1), to three environmental predictors in a southern California, USA, study area (Fig. 5.1): average January minimum temperature (jan), average July maximum temperature (july) and soil water-holding capacity (awcl, derived from STATSGO data; see Section 5.2.3).

concerning seasonal changes in habitat use (Suárez-Seoane *et al.*, 2008) and the relative importance of abiotic conditions and biotic interactions in determining species distributions (Cunningham *et al.*, 2009).

In several of these examples, Maxent was used in conjunction with, or was compared to, other presence-only modeling methods, primarily GARP, ENFA, BIOCLIM and DOMAIN. In these comparisons, Maxent outperformed GARP in terms of several measures of prediction accuracy (Phillips *et al.*, 2006; Elith & Graham, 2009), especially with small sample sizes (Hernandez *et al.*, 2006; Pearson *et al.*, 2007). Further, in a comprehensive study of a large number of SDM methods applied to presence-only data, Maxent was usually among the top-performing methods in terms of prediction accuracy (Elith *et al.*, 2006). Other novel methods such as MARS, boosted decision trees, and multivariate models (generalized dissimilarity modeling) also fell into the top performing group of models in that study. Maxent performed somewhat better than GLMs and GAMs (applied to presence/background data), and substantially better than envelope methods (BIOCLIM) and GARP. In another comparison, Maxent again performed marginally better than a GLM, and both models suggested similar environmental response curves (Gibson *et al.*, 2007). Elith and Graham (2009) compared the ability of several models to capture known response curves using simulated species data, and found that boosted regression trees (Chapter 7), Maxent and random forests (Chapter 7) performed best in this regard, slightly better than GLMs, and much better than GARP. In particular, GARP was not able to model responses to categorical predictors. GARP models tend to overpredict the extent of species distributions, that is, to have higher commission errors (Hernandez *et al.*, 2006; Elith & Graham, 2009; Phillips, 2008).

A study that addressed the ability of Maxent versus GARP to predict habitat suitability under novel conditions (geographical transferability or extrapolation) claimed that GARP models were better able to estimate potential distributions under novel conditions, that is, in unsampled regions (Peterson *et al.*, 2007). However, that study did not actually adequately address extrapolation because background data from the entire study area were used, and only three species models were examined qualitatively (Phillips, 2008). Rather, it addressed sample bias, also an important issue when using presence-only data.

In one comparison with BIOCLIM and DOMAIN, as well as mechanistic models, both Maxent and GAMs were better able to model species range shifts under future climate change (Hijmans & Graham,

2006). Using SDMs of any kind to extrapolate to novel environmental conditions and locations is a very important and challenging application (Chapter 1) and a topic that needs further investigation (Chapter 10).

8.5 Summary

Modeling species distributions with presence-only data is such an important practical problem that a number of comparisons among the various approaches have been made, and some of those have already been mentioned in the chapter. It seems that different methods developed over time

because information about species distributions was required for different kinds of applications, at different scales, and using various kinds of "presence-only" data. In general, "profile" methods, ENFA, and later the more complex machine learning approaches, GARP and Maxent, were developed to infer species distributions from georeferenced museum records (natural history collections). They are tailored to small numbers of species occurrence records and, when predictions encompass entire species ranges they are frequently used with coarse-scale environmental predictors (Chapter 5) for spatial prediction.

Mahalanobis distance and related distance methods, as well as HSI methods, have a strong tradition in wildlife ecology and management, have frequently been developed for vertebrates, and tend to be applied at ecological or landscape scales. In these applications, environmental predictors are proximally or distally related to the availability of food and shelter and include factors such as vegetation characteristics (type, cover), soil, topographic and hydrological features (landform, proximity to streams). While distance-based methods are derived quantitatively using occurrence data, HSI methods are useful when georeferenced observations are lacking but expert knowledge of species habitat requirements exists. The straightforward HSI approach is now complemented by powerful new methods for eliciting and formalizing expert opinion as decision rules.

There are some clear distinctions among the approaches presented in this chapter. HSI, or more generally, models based on decision rules derived from expert opinion, do not require any georeferenced species occurrence data, but it is also possible for decision rules to be estimated from data in this framework. However, HSI models tend to perform more poorly than quantitatively estimated models when training data are available (but see Pereira & Duckstein, 1993).

Distance-based methods are not well-suited for categorical predictors, they equally weight predictors, and assume linear relationships between environmental variables and habitat suitability. An even more significant drawback is that distance-based methods do not provide information about, or estimates of, the influence of environmental predictors on species occurrences (the ecological response functions). ENFA presumably is best able to discriminate species niche from background when both are normally distributed so that means are the best measures of central tendency (Figs. 8.3 and 8.4), a condition that is rarely met in nature.

In spite of the potential theoretical advantages of the expert, envelope and distance-based methods that only use presence data to describe the species niche, in model comparisons, methods that use presence and background data, including those discussed in Chapters 6 and 7, tend to show greater predictive performance. Specifically, they have less of a tendency to overpredict, and are able to make more discriminating predictions. Comparisons suggest that the statistical learning methods such as Maxent and ensemble tree methods are particularly robust when only a small and biased sample of observations is available.