Integrative Zoology 2013; 8: 124–135 doi: 10.1111/1749-4877.12000

REVIEW

Applying various algorithms for species distribution modelling

Xinhai LI and Yuan WANG

Key Laboratory of the Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

Abstract

Species distribution models have been used extensively in many fields, including climate change biology, landscape ecology and conservation biology. In the past 3 decades, a number of new models have been proposed, yet researchers still find it difficult to select appropriate models for data and objectives. In this review, we aim to provide insight into the prevailing species distribution models for newcomers in the field of modelling. We compared 11 popular models, including regression models (the generalized linear model, the generalized additive model, the multivariate adaptive regression splines model and hierarchical modelling), classification models (mixture discriminant analysis, the generalized boosting model, and classification and regression tree analysis) and complex models (artificial neural network, random forest, genetic algorithm for rule set production and maximum entropy approaches). Our objectives are: (i) to compare the strengths and weaknesses of the models, their characteristics and identify suitable situations for their use (in terms of data type and species–environment relationships) and (ii) to provide guidelines for model application, including 3 steps: model selection, model formulation and parameter estimation.

Key words: algorithms, machine learning, model formulation, model selection, species distribution models

INTRODUCTION

Species distribution models (SDMs) are also known as habitat models, ecological niche models, bioclimatic envelopes and resource selection functions (Elith & Graham 2009). These models use computer algorithms to predict the distribution of species in geographic space on the basis of data (mathematical representations of their known distribution in environmental spaces) (Austin 2007). They rely on statistical correlations between existing species distributions and environmental variables. Levins (1966) points out 3 goals of ecological models: reality, generality and precision. Typically, only 2 out of the 3 desirable model goals can be attained simultaneously, while the third goal has to be sacrificed. This trade-off leads to a distinction of 3 different groups of models (Korzukhin *et al.* 1996). SDMs are static empirical models (i.e. phenomenological and statistical) rather than mechanistic models (i.e. physiological, fundamental and process-based) or general analytical models (i.e. theoretical and mathematical). They relate observed presence or abundance of a species to values of environmental variables at particular sites. In contrast,

Correspondence: Xinhai Li, Key Laboratory of the Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, 1 Beichen West Road, Beijing 100101, China.
Email: lixh@ioz.ac.cn

mechanistic (or process-based) models assess the biophysiological aspects of a species to determine the conditions in which the species can ideally persist, usually based on observations made in controlled field or laboratory studies (Guisan & Zimmermann 2000).

The relationship between species and environment is complex (Pearson *et al.* 2006a; Morin & Thuiller 2009; Wiens *et al.* 2009) and overcoming the uncertainty of model applications is a great challenge (Patt *et al.* 2005; Pearson *et al.* 2006a; Dormann *et al.* 2008; Fuller *et al.* 2008; Cressie *et al.* 2009; Morin & Thuiller 2009; Conroy *et al.* 2011). In the past 3 decades, scientists have developed numerous models to estimate the relationship between species and associated environmental variables. However, sometimes, different models provide diverse predictions (Pearson *et al.* 2006a; Randin *et al.* 2006). Consequently, hybrid or ensemble model frameworks were suggested to make reliable and robust predictions of the potential distribution of species (e.g. del Barrio *et al*. 2006; Araujo & New 2007; McRae *et al*. 2008; Coetzee *et al*. 2009; Morin & Thuiller 2009; Thuiller *et al*. 2009b; Conroy *et al*. 2011). For example, BIOMOD (Thuiller *et al.* 2009b), a package of software R (R Development Core Team 2011), can automatically compare 9 SDMs and is able to suggest the best model (from the 9 models) for any specific species–environment situation. Several studies compare different models to explore SDM characteristics (type of algorithms) and their suitability (regarding interpretation ability and prediction power) for the available data and the intended application (e.g. Moisen & Frescino 2002; Brotons *et al.* 2004; Segurado & Araujo 2004; Thuiller *et al.* 2004; Araujo & Guisan 2006; Elith *et al.* 2006; Pearson *et al.* 2006b; Tsoar *et al.* 2007; Elith & Graham 2009; Aertsen *et al.* 2010, 2011; Kampichler *et al.* 2010).

Species distribution models are powerful tools and are applied in many aspects of ecology, yet their outputs are based on a number of assumptions. For example, in studies on biological consequences of climate change, SDMs have been used extensively to estimate the current distributions and future range shifts of species (e.g. Araujo & New 2007; Thuiller *et al*. 2008; Li *et al.* 2010) on the basis of the following assumptions: (i) the current occurrences of the species are in their favorite habitats; (ii) the species distributions are determined by the explanatory variables (e.g. temperature and precipitation); and (iii) the association between the species distribution and the explanatory variables does not change in the future (no adaptation). SDMs can provide valuable information about how much species might move in future climate conditions, and which climate variables might impact the future ranges the most.

Many SDMs are easy to use, but researchers often find it difficult to select the most appropriate model (or models) for their cases, and there has been significant variability in model performance (Araujo *et al.* 2005). Some review papers provide insightful discussions covering broad aspects of model development and applications (e.g. Guisan & Zimmermann 2000; Guisan & Thuiller 2005; Meynard & Quinn 2007). However, knowledge is still required regarding which methods are best suited to the available data and the intended applications, because the advice enabling informed choice of methods is currently scattered throughout the published literature and is incomplete (Elith & Graham 2009).

In this paper, we review up-to-date studies of SDM applications, and aim to provide intuitive insight into the prevailing SDMs for newcomers in the field of ecological modelling. We compare 11 popular models, including classic regression models, advanced classification tree methods, complex machine learning techniques (e.g. neural network, random forest, maximum entropy [Maxent] and genetic algorithm for rule set production [GARP] approaches) and hierarchical models. Our objectives are: (i) to compare the strengths and weaknesses of the models, their characteristics and identify suitable situations for their use (in regards to data type and species–environment relationships) and (ii) to provide guidelines for model application, including 3 steps: model selection, model formulation and parameter estimation.

MODELS

There are a variety of SDMs available to predict species distributions based on the association of species occurrences and environment variables. In this paper, we list 11 popular models (Table 1), describe their characteristics and highlight their strengths and shortcomings.

Generalized linear models

Generalized linear models (GLMs) are a generalization of general linear models (McCullagh & Nelder 1989). GLMs were introduced in the 1970s (Nelder & Wedderburn 1972) and formulated in the late 1980s (McCullagh & Nelder 1989). The common types of GLMs are linear regression, logistic regression and Poisson regression. Logistic regression is suitable for present and absent species occurrence data and Poisson

Table 1 Popular species distribution models, history, complexity levels, popularity, types of species data and reference papers

Models	History	Complexity	Popularity[†]	Species data[‡]	Reference
Generalized linear model	1972	Low	35158	p/a or abundance	Nelder & Wedderburn 1972
Generalized additive model	1986	Medium	6848	p/a or abundance	Hastie & Tibshirani 1986
Multivariate adaptive regression splines	1991	Medium	278	p/a	Friedman 1991
Mixture discriminant analysis	1996	Medium	569	p/a	Hastie & Tibshirani 1996
Classification and regression tree	1984	Medium	6820	p/a	Breiman *et al.* 1984
Generalized boosting models	1999	Medium	226	p/a	Friedman *et al.* 2000
Random forest	1995	High	12203	p/a	Breiman 2001a
Artificial neural networks	1943	High	66155	p/a	Hopfield 1982
Genetic algorithm for rule set production	1999	High	143	p	Stockwell & Peters 1999
Maximum entropy method	2006	High	646	p	Phillips *et al.* 2006
Hierarchical modeling	1996	Low	1421	p/a or abundance	Wikle 2003

[†]Popularity is the number of references obtained by querying the database of the Web of Knowledge (Thomson Reuters) using the model names as key words for database subject, while publication year is restrained within 2000–2011. [‡]The p/a indicates presence and absence data; p indicates presence only data.

regression is suitable for species count data. The logit link function for logistic regression and the logarithm link function for Poisson regression enable the dependant variable (indicating species presence or abundance) to be linearly related with a number of explanatory variables. The explanatory variables of GLMs can contain interaction terms and polynomial terms, so they are preferable for nonlinear yet simple relationship between species and environment variables. All model parameters of GLMs can be clearly interpreted with ecological meanings. Applying GLMs requires careful calibration: users must check the significance of each explanatory variable and remove the non-significant variables (model selection).

Generalized additive model

Generalized additive models (GAMs) are non-parametric extensions of GLMs, thus providing the potential for better fits to data than GLMs (Hastie & Tibshirani 1986). GAMs use data-defined smoothing functions to fit nonlinear species–environment relationships. The smooth functions are computed independently for each explanatory variable and added to build the final model. The number of smoothing parameters can be specified by the users or default settings of statistical software can be used; this number should be reasonably small to avoid overfitting, and certainly well under the degrees of freedom offered by the data. The interaction of explanatory variables can be fitted by GAMs. The polynomial terms do not need to be considered because the smoothing functions already take into account the multimodal species–environment relationship. GAMs are useful when the relationship between species and environmental variables has a more complex form not easily fitted by GLMs (Yee & Mitchell 1991).

Multivariate adaptive regression splines

Multivariate adaptive regression splines (MARS) are an extension of linear regression models that automatically model nonlinearities and interactions (Friedman 1991).

A major assumption of linear models is that the coefficients are stable across all levels of the explanatory variables. In contrast, MARS allow changes in coefficients, and are suitable when it is suspected that the model's coefficients have different optimal values across different levels of the explanatory variables. The breakpoints or thresholds of coefficients are termed spline knots. The regression with a number of spline knots can be thought of a piecewise regression. In MARS, the spline knots are determined automatically. In addition, first-order interactions between variables can also be specified. MARS are particularly powerful when there are large numbers of explanatory variables and low-order interaction effects (Thuiller *et al*. 2009a).

Mixture discriminant analysis

Mixture discriminant analysis (MDA) is an extension of linear discriminant analysis (Venables & Ripley 2002). Discriminant analysis is used to predict the categories of the observations (e.g. presence or absence of a species), based on the combination of the explanatory variables. Linear discriminant analysis is closely related to general linear model, which also attempts to express 1 dependent variable as a linear combination of explanatory variables. Linear discriminant analysis was developed in the 1930s (Fisher 1936). Advanced discriminant analysis has been developed for nonlinear, nonparametric situations (Hastie *et al.* 1994; Hastie & Tibshirani 1996). MDA is a method for classification based on mixture models. It assumes that the distribution of the class of each environmental variable follows a normal distribution. The mixture of normals is used to obtain a density estimation for each class (e.g. species presence or absence). MDA can control the within-class spread of the subclass centers relative to the between-class spread, by forming multiple normal distributions within 1 class.

Generalized boosting models

Boosting methods are designed to fit many simple models whose predictions are then combined to give more robust estimates of the relationship between species distribution and a set of environmental variables, whereas GLMs seek to fit the single model that best explains the relationship between species and environment (Friedman *et al.* 2000; Friedman 2001). Generalized boosting models (GBM) have 2 algorithms: the boosting algorithm iteratively uses the regression-tree algorithm to construct a combination of trees. GLMs and GAMs can be used in tree building. Boosting is used to overcome the inaccuracies inherent in a single tree model. It can compute a sequence of simple classification trees, where each successive tree is built for the prediction residuals of the preceding tree. GBMs can eventually produce a good fit of the predicted values to the observed values, even if the specific nature of the relationships between the predictor variables and the dependent variable is complex (e.g. nonlinear, interacted or noisy with outliers). Boosting can be understood as a method for developing a model in a forward stage-wise fashion, at each step adding small modifications in parts of the model space to fit the data better (Friedman *et al.* 2000). GBMs can be used for regression as well as classification problems, with continuous and/or categorical predictors.

Artificial neural networks

The concept of artificial neurons was first proposed in 1943 (McCulloch & Pitts 1943). An artificial neural network (ANN) is a complex model system, involving a network of simple processing elements (artificial neurons) that can exhibit complex global behavior (e.g. select a site as habitat based on numerous environmental variables), determined by the links between the neurons and associated functions (Hopfield 1982). The key feature of an ANN is that it contains a hidden layer. Each neuron in the hidden layer receives information from each input, sums the inputs, adds a constant (the bias), then transforms the result using a fixed function. ANNs can operate like multiple regressions when the outputs are continuous variables, or like classifications when the outputs are categorical. The accuracy of ANNs is mainly controlled by the amount of weight decay of the links and the number of hidden neurons. All model parameters can be optimized by new observations. As a whole, ANNs are nonlinear models but with so many parameters they are extremely flexible; flexible enough to approximate any smooth function (Thuiller *et al.* 2009b). ANNs usually require longer time for computers to process than other parallel models.

Classification and regression tree

Classification and regression tree (CART) analysis was designed in the 1980s (Breiman *et al.* 1984). It uses recursive partitioning to split the data into increasingly smaller, homogenous subsets until a termination is reached (Venables & Ripley 2002). In a CART, each group of records (species presence or absence data), represented by a ‘node’ in a decision tree, can only be split into 2 groups. The heterogeneity of a node can be interpreted as a deviance of a Gaussian model (regression tree) or of a multinomial model (classification tree). The best tree is a trade-off between a high decrease of deviance and the smallest number of nodes. CART is able to uncover complex interactions between predictors that may be difficult or impossible to uncover using traditional multivariate techniques (Breiman *et al.* 1984; De’ath & Fabricius 2000).

Random forest

Random forest is an ensemble classifier that consists of many decision trees, implementing Breiman's random forest algorithm for classification and regression (Breiman 2001a). To classify a new object from an input vector, random forest puts the input vector down each of the trees in the forest. Each tree gives a classification (which is usually called the tree 'votes' for that class). The forest chooses the classification with the most votes (over all the trees in the forest).

Random forest is not sensitive to the problem of multicollinearity; it handles correlated variables well (Breiman 2001b). It runs efficiently with large databases, and it can process thousands of input variables without variable deletion; it provides estimates of what variables are important in the classification; it is robust with missing data and unbalanced datasets; it offers an experimental method for detecting variable interactions (Breiman 2001b). Random forest is one of the most accurate learning algorithms with high performance in predicting species distributions (e.g. Iverson *et al.* 2008).

Genetic algorithm for rule set production

The genetic algorithm for rule set production (GARP) is an SDM based on a genetic algorithm for developing rule sets constraining species distribution (Stockwell & Peters 1999). A GARP model is a random set of mathematical rules that can be interpreted as limiting environmental conditions (e.g. range specifications) and certain species–environment relationships (e.g. formed regression patterns). Each rule is considered as a gene, and the set of genes is combined randomly to further generate many possible models describing the potential of species occurrences. GARP is claimed to be robust when dealing with smaller samples of presence-only data (Costa *et al*. 2002; Stockwell & Peterson 2002).

Maximum entropy method

The maximum entropy method (Maxent) is a general-purpose machine learning method for modelling species geographic distributions (Phillips *et al.* 2006). Maxent can make robust predictions with its 'default settings' without much effort in parameter tuning (Phillips & Dudik 2008). The rationale of Maxent is to estimate a target probability distribution by finding the probability distribution of maximum entropy (i.e. most spread out or closest to uniform), subject to a set of constraints that represent the species distribution (Phillips *et al*. 2006). The constraints are that the expected value of each environmental variable should match its empirical average (average value for a set of sample points taken from the species distribution) (Phillips *et al*. 2006). Like GARP, Maxent only uses presence-only datasets.

Hierarchical modelling

Hierarchical modelling combines different processes (e.g. species habitat selection process and human detection process) using their joint distribution model (Wikle 2003).

For example, there are 10 individuals at a site and we observe 5 individuals among them. Such a result can be explained by 2 models: an ecological process model that associates 10 individuals with the environmental variables of the site; and a detection model that describes the detection rate. The observed value (5 individuals) is the product of the ecological process model and the detection model. Thus, the hierarchical modelling framework makes a clear distinction between an observation component of the model (which typically describes nuisance variation in the data) and the process component of the model (which is usually fundamental to the object of inference) (Royle & Dorazio 2008).

Historically, the observation component and the process component have not been distinguished in model formulation (Royle & Dorazio 2008). Hierarchical models have been developed in environmental sciences studies (e.g. Berliner 1996; Wikle *et al.* 1998; Wikle 2003) to combine the observation and the process components, and are especially suitable for estimating the uncertainty of both data recording (in the context of detection probability and recording probability) and species occurrences (in the context of species–environment relationship).

MODEL COMPARISON

The relationship between species and environmental variables is complex and diverse. No modelling technique is best for all situations (Marmion *et al.* 2009). Researchers need to know which modelling technique is suitable for the particular species they study. The SDMs can be categorized as regression models, classification models and complex models.

Regression models

Generalized linear modelling is the classic method to quantify the association between species information (presence/absence or count data) and environmental variables. GAM can further fit data when multimodal relationships exist between species occurrencs and environmental variables, so it is more powerful for complex species–environmental relationships than GLM (Yee & Mitchell 1991). Model selection (selecting the significant environmental variables and their polynomial and interaction terms) is usually the key step for GLM and GAM because insignificant or correlated variables, including high order and interaction terms, influence the effects (the values of variable coefficient and their significance) of other variables.

Multivariate adaptive regression splines, similar to GAM, are more powerful than GLM. MARS provide an alternative regression-based method for fitting nonlinear responses, using piecewise linear fits rather than smooth functions in GAM. MARS runs much faster than GAM (Elith *et al.* 2006). MARS can automatically quantify the interaction effects, whereas GAM and GLM need to define the interaction terms manually in the model equations. In most GLM applications, only first-order interaction terms are considered, which is simple. In contrast, the interaction term in GAM is very complicated when the 2 environmental variables have a multimodal relationship with the dependent variable. The interaction term in MARS is the combination of many simple first-order interactions; the combination is the composition of different interactions at different levels of the 2 environmental variables, not the sum of additive effects of different terms in GLM and GAM at the same levels of environmental variables.

Hierarchical modelling usually combines 2 or 3 regression processes. Essentially, it is a series of GLMs. Hierarchical modelling incorporates the detection rate and estimates its variance, while other models assume that the detection rate is 1 (i.e. the surveyed abundance is the actual abundance). Consequently, the model performance of hierarchical modelling is substantially improved. Since 2000, the term 'hierarchical modelling' has been widely adopted and has become synonymous with Bayesian analysis (Royle & Dorazio 2008). In fact, hierarchical modelling and Bayesian analysis are conceptually different. Hierarchical modelling emphasizes model construction, whereas Bayesian analysis is concerned with technical aspects of inference (Royle & Dorazio 2008). They are just frequently used together because Bayesian analysis is convenient for parameter estimation for hierarchical models. The Bayesian method is used to estimate the model coefficients of hierarchical models because solving a few regressions simultaneously is sometimes difficult.

Classification models

Mixture discriminant analysis (MDA), CART and GBM are classification methods, yet regression algorithms are embedded in these methods. Compared with regression methods, classification models are much more robust with the presence of outliers in the datasets. MDA is advanced discriminant analysis, allowing local fit for every class by a mixture of a few normal distributions.

The CART is one type of classification tree model. Classification tree analysis is similar to traditional discriminant analysis and cluster analysis. Discriminant analysis predictions are reproduced by simultaneous multiple regression of predictor variables. The classification tree predictions are reproduced by separated simple regressions, which have the recursive, hierarchical nature that builds up leaves and stems of the tree. Like GAM, classification trees do not rely on a priori hypotheses about the relationship between species occurrences and environmental variables. The tree is built by repeatedly splitting the data, defined by a simple rule based on a single explanatory variable. At each split, the data are partitioned into 2 exclusive groups, each of which is as homogeneous as possible. CART combines both regression trees and classification trees. Regression trees are like classification trees except that the end point will be a predicted function value rather than a predicted classification. Breiman *et al.* (1984) invented CART and he developed decision trees as computationally efficient alternatives to ANN.

Generalized boosting models combine many simple models, whose predictions are then used to give more robust estimates. MDA, CART and GBM are non-parametric, and, hence, suitable for complex species–environmental relationships. CART and GBM use recursive partitioning for final model prediction; they can handle numerical data that are highly skewed or multi-modal, as well as categorical predictors with either ordinal or nonordinal structure. Over the past 20 years, CART has become one of the most popular techniques for species distribution modelling. Over the past 10 years, the GBM has become one of the most powerful methods for predictive data mining.

Complex models

ANN, random forest, Maxent and GARP are complex models. They formed frames that contain local models (e.g. regression and classification) within their structures. CART, GBM, ANN, random forest, Maxent and GARP have recursive parameter optimization features and, thus, are referred to as machine learning techniques.

The complex models can extract hidden features from the input data. They are suitable when datasets with very complicated correlation structures are to be analyzed or when datasets contain many highly intercorrelated variables. These complex models can reproduce minor details of the training data, which can result in an overfitting problem, causing biased model prediction in other regions or other time periods (Reibnegger *et al*. 1991).

To detect overfitting problems, approaches such as resubstituition (the data used to calibrate models are also used to validate them), bootstrap, data-splitting and independent validation can be used (Araujo *et al*. 2005). To reduce overfitting problems, approaches such as controlling the model complexity by using only the most predictive top ranked variables (e.g. Liu *et al.* 2011), developing new algorithms to improve the generalization capability (e.g. Bramer 2002) or controlling the data distribution skewness (Sun *et al*. 2006) are applicable.

Artificial neural networks are widely used in statistical estimations, classification optimization and control theory (Olden *et al.* 2008). CART is a much more efficient machine learning tool than ANN. Random forest is a more advanced model and consists of many decision trees; CART has only 1 tree. Breiman (2001b) compares the misclassification error of random forest and CART using 10 different datasets by leaving out a random 10% of the data for checking. Random forest was found to be better for all datasets. Breiman indicated that the model is not overfitting using a complex model (a forest) instead of a rather simple model (a tree) (Breiman 2001b).

Maxent and GARP are presence-only approaches that are suitable for most cases in which the true absence data is lacking. Theoretically, they fit models from presence and background data, compared with other presence-only models that either generate pseudo-absence data by sampling points outside the boundary of presence data, or strictly apply presence-only modelling (e.g. BIOCLIM) (Phillips *et al.* 2006; Li *et al.* 2011). Maxent and GARP project the realized niche into a geographical space without giving weight to observed absence information, which could result in a poorer fit to the current observed distribution than the presence–absence approaches, such as random forest and GBM (Pearson *et al.* 2006b). However, based on an extensive comparison of 16 SDMs over 226 species from 6 regions of the world, Elith *et al.* (2006) point out that the presence-only approaches, such as Maxent and GARP, are effective for modelling distributions for many species and regions. Maxent has good penalty functions (i.e. regularization) in parameter estimation to prevent overfitting. Regularization has most impact when sample sizes are small (Phillips *et al.* 2004), enabling reliable predictions in such cases. In most situations, Maxent outperforms GARP in terms of prediction accuracy (Phillips *et al.* 2006).

Cross comparisons

The predictive accuracy of different SDMs, including regression, classification and complex models, have been compared in some studies (e.g. Breiman 2001b; Elith *et al.* 2006; Meynard & Quinn 2007; Olden *et al.* 2008; Phillips & Dudik 2008; Coetzee *et al.* 2009; Elith & Graham 2009; Marmion *et al.* 2009; Morin & Thuiller 2009; Thuiller *et al.* 2009b; Aertsen *et al.* 2010; Kampichler *et al.* 2010). For different cases (different species with different geographical and environmental distributions), model performance usually varies (Segurado & Araujo 2004), and no single model is best for all cases. In some situations, simpler models work best, especially when the study area in small (e.g. at a mountain slope with monotonous elevation gradient). Aertsen *et al* (2010) modeled the distribution of 3 tree species using 5 SDMs, and found that GAM was the best and ANN was the worst, compared with GLM, CART and GBM. However, in general, the complex models have overall better performance than simpler models (e.g. Prasad *et al.* 2006; Meynard & Quinn 2007; Olden *et al.* 2008; Phillips & Dudik 2008), indicating that model complexity contributes to predictive accuracy (Tsoar *et al.* 2007).

All SDMs we discuss can use both numeric and categorical environmental variables. GLM and GAM are suitable for numeric variables because of their regression nature. MARS tends to be better than CART for numeric variables because hinges are more appropriate for numeric variables than the piecewise constant segmentation used by recursive partitioning. MARS usually do not give as good a fit as GBM, but can be built much more quickly and are more interpretable (i.e. effect of each predictor is clear). Complex models can deal well with both numeric and categorical environmental variables.

For noisy, nonlinear and high-dimensional data, classification tree-based machine learning methods are often suggested. For predicting the distribution of a bird species, random forest is best, CART is second best and ANN is the worst (Kampichler *et al.* 2010). Random forest is preferable for data including categorical predictive variables with different numbers of levels, yet it is biased in favor of those attributes with more levels. Thus, the variable importance scores from random forest are not reliable for this type of data (Deng *et al*. 2011). Predictive accuracy of models was also tested by artificially adding spatial deviances to the occurrences, and the quality of the predictions made with GBM diminished significantly when degraded data were used, while there was no evidence of decline in performance of regression based techniques (GLM, GAM and MARS) and Maxent (Graham *et al.* 2008). Geographical attributes (prevalence, latitudinal range and clumping [spatial autocorrelation]) would also influence predictive accuracy (Marmion *et al.* 2009). GAM, GLM and MDA are highly influenced by the 3 attributes, whereas random forest, ANN and GBM are only moderately influenced, and MARS and CART are only slightly affected (Marmion *et al.* 2009).

Based on a simulated species that provides known species–environment relationships and spatial distribution, Elith and Graham (2009) evaluate differences among methods in relation to the truth. The results suggest that BRT models the overall shape of the response most accurately, followed in order by Maxent, random forest, GLM and GARP. Although the general form of the fitted function was correct for BRT, it was overfitted to the sample; random forest showed even more overfitting (Elith & Graham 2009). This is different from what Breiman (2001b) concludes: that random forest has no overfitting problem. Maxent is able to fit complex functions between response and predictor variables, and can include interaction terms, but to a more limited extent than BRT (Elith *et al.* 2006).

MODEL APPLICATIONS

When applying models, some key steps, such as verification, calibration, validation (evaluation), credibility and qualification, have been well addressed in other reviews (e.g. Rykiel 1996; Guisan & Zimmermann 2000; Guisan & Thuiller 2005). These steps are very important for good modelling practice. Most model systems now provide relevant tools for these steps. However, the greatest challenge for newcomers is determining how to select the most relevant model for their data types and modelling objectives, and how to use them appropriately.

At present, researchers have many options for selecting from a variety of models to deal with different situations relating to, for example, various spatial extent, habitat heterogeneity, types of environmental variables and species–environment relationships. Here, we suggest new processes that are important for model development: model selection, model formulation and parameter estimation.

Model selection

Researchers need to select models on the basis of their research targets (data types: e.g. noisy, nonlinear and high-dimensional, lots of categorical variables) and objectives (e.g. checking variable weight, explaining species–environmental relationships, predicting distribution and estimating future range shifts). Although complex models have better prediction accuracy, simple models have their own unique advantages. GLMs are transparent; that is, all coefficients of explanatory variables (including their quadratic terms and interaction terms) can be clearly shown and well interpreted. On the contrary, other models (i.e. GAM, MARS, MDA, GBM, CART and complex models) have too many parameters, so that meaningful ecological interpretation is not possible. The complex models, such as random forest, are also called ‘black boxes’: they are easy to use yet hard to explain (Breiman 2001b).

Researchers should be aware of the strength of each model. GAM is good for multimodal continuous variables; MARS is suitable for situations with high first-order interaction effects; GBM and CART are preferred for many categorical variables with outlier observations; ANN performs well for very complicated species–environmental relationships. Random forest is ideal for cases with many explanatory variables and high interaction effects. Maxent and GARP are easy to use, because available software can import GIS layers of environmental variables directly, with presence only species occurrence data. Hierarchical modelling is preferable for situations when different regressions (for a few independent processes [e.g. species occurrence and detection]) can be combined together.

Model formulation

Based on Guisan and Zimmermann’s definition, statistical model formulation means: choosing a suited algorithm for predicting the species distributions, defining a particular type of response variable and estimating the model coefficients, and selecting an optimal statistical approach with regard to the modelling context (Guisan & Zimmermann 2000). Model formulation is also referred to as ‘verification’ by some authors (e.g. Rykiel 1996).

In this paper, we narrow down the conceptual extent of model formulation, focusing on the determination of the model structures. Here, model formulation includes the selection of explanatory variables, and the selection of interaction terms and polynomial terms of the variables. Model formulation also includes the determination of certain model parameters, such as the number of splines/knots in GAM, the number of trees in GBM, and the number of units in the hidden layer in ANN. Different options of the model parameters would cause significant difference in model performance (Austin 2007).

A common mistake in using GLMs is to ignore the interaction terms and polynomial terms of the variables (Guisan *et al.* 2002). Presence of the interaction effect is common for species–environmental relationships. For example, a species might prefer shrubs rather than grass at low elevation, but prefer grass to shrubs at high elevation. Under such a situation, vegetation type and elevation have an interaction effect on the species. The quadratic term of explanatory variables is more commonly used for species–environmental relationships, as it represents a favorite value of a variable for a species (e.g. a favorite elevation of 1000 m asl) (note that the suitability of elevation is usually a bell-shaped curve).

Parameter estimation

Estimation of model parameters is usually conducted automatically by the statistical software. It is the key process for model development. Coefficients of variables can be estimated using least squares, maximum likelihood, Markov chain Monte Carlo, Kalman filter, bootstrap, and many other algorithms in machine learning techniques. At this stage, variable selection is also carried out, as the variables for which the coefficients have no significant difference from zero will be removed. Variable selection is performed based on the contribution of the variables to species occurrences as well, measured by information criteria (e.g. Akaike and Bayesian information criteria and BIC) to balance variable contribution and model parsimony. SDMs usually provide some evaluation statistics, such as area under the ROC (receiver operating characteristic) curve (AUC), Cohen's kappa statistic and the true skill statistic. These evaluation statistics represent the overall model predictive accuracy.

Currently, most SDMs provide sufficient tools for coefficient estimation and model evaluation. This most important model development step turns out to be the easiest step in model applications.

DISCUSSION

In this review, we compared 11 species distribution models to clarify the characteristics of the models and suitable situations for their use (in terms of data type and species–environment relationships) and provided guidelines for model application. We did not address the issue of preparation of data (e.g. dealing with pseudoreplication or autocorrelation, data transformation and null values) nor model evaluation. We did not justify the role of SDMs in studies of ecology and climate change, as a few review papers have covered this issue (e.g. Guisan & Zimmermann 2000; Guisan & Thuiller 2005). We aimed to provide a technique-focused picture of current models that are used in predicting species distributions. Among the models we compared, GLM, GAM, ANN, CART, random forest, Maxent and GARP are frequently used in the field of climate change (Table 1). MARS, MDA, GBM and hierarchical modelling were developed in the 1990s, and they are gradually gaining more attention from ecologists. Maxent and GARP, in the form of integrated, user-friendly software, are mostly used for predicting species distributions, whereas others are applied in various other fields, such as social science and economics.

Species–environmental relationships can be diverse, depending on, for example, target species, spatial and time scales and study areas (Marmion *et al.* 2009). We suggest that researchers check their data carefully first before using any models. For example, scatter plots can be used to check the association of species abundance with an explanatory variable, detecting general patterns (e.g. linear trend, bell-shaped curve, multimodal, etc.) or outliers. A correlation matrix can be used to identify correlated explanatory variables. An interaction plot can be used to detect the interaction effect of any 2 explanatory variables.

After familiarizing themselves with their data, researchers should select 1 or a few models that are suitable for their data, and, most importantly, take their time with the model formulation. While conducting model formulation, researchers need to select explanatory variables, check the necessity of polynomial terms and interaction terms, and determine the model parameters if needed (e.g. the number of splines/knots in GAM). Even for Maxent and GARP, which are claimed to be robust with their default settings, researchers should be careful with some model parameters, such as the size of quadrates for the background environmental variable layers and the extent of the study area. It is totally up to users to decide how much marginal area (outside the species distribution range) should be included in the study areas. The parameter estimation process is usually easy because most models can do this well. The model evaluation process is also important, yet current models provide enough statistics to check the model performance. Using many SDMs for a certain dataset and research objective can significantly decrease the risk of obtaining biased results.

ACKNOWLEDGMENTS

This study was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA05080701) and the Public Welfare Project (201209027) of the Ministry of Environmental Protection of China.

REFERENCES

Aertsen W, Kint V, Van Orshoven J, Özkan K, Muys B (2010). Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. *Ecological Modelling* **221**, 1119–30.

Aertsen W, Kint V, Van Orshoven J, Muys B (2011). Evaluation of modelling techniques for forest site productivity prediction in contrasting ecoregions using stochastic multicriteria acceptability analysis (SMAA). *Environmental Modelling Software* **26**, 929–37.

Araujo MB, Guisan A (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography* **33**, 1677–88.

Araujo MB, New M (2007). Ensemble forecasting of species distributions. *Trends in Ecology and Evolution* **22**, 42–7.

Araujo MB, Pearson RG, Thuiller W, Erhard M (2005). Validation of species-climate impact models under climate change. *Global Change Biology* **11**, 1504–13.

Austin M (2007). Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling* **200**, 1–19.

Berliner LM (1996). Hierarchical Bayesian time-series models. In: Hanson KM, Silver RN, eds. *Hierarchical Bayesian Time Series Models*. Fundamental Theories of Physics, vol. 79. Proceedings of the 15th International Workshop on Maximum Entropy and Bayesian Methods; 31 Jul–4 Aug 1995, Santa Fe, NM, USA, pp. 15–22.

Bramer M (2002). Using J-pruning to reduce overfitting in classification trees. *Knowledge-Based Systems* **15**, 301–8.

Breiman L (2001a). Random forests. *Machine Learning* **45**, 5–32.

Breiman L (2001b). Statistical modelling: the two cultures. *Statistical Science* **16**, 199–215.

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984). *Classification and Regression Trees*. Chapman and Hall, New York.

Brotons L, Thuiller W, Araujo MB, Hirzel AH (2004). Presence-absence *versus* presence-only modelling methods for predicting bird habitat suitability. *Ecography* **27**, 437–48.

Coetzee BWT, Robertson MP, Erasmus BFN, van Rensburg BJ, Thuiller W (2009). Ensemble models predict important bird areas in southern Africa will become less effective for conserving endemic birds under climate change. *Global Ecology and Biogeography* **18**, 701–10.

Conroy MJ, Runge MC, Nichols JD, Stodola KW, Cooper RJ (2011). Conservation in the face of climate change: the roles of alternative models, monitoring and adaptation in confronting and reducing uncertainty. *Biological Conservation* **144**, 1204–13.

Costa J, Peterson AT, Beard CB (2002). Ecologic niche modelling and differentiation of populations of *Triatoma brasiliensis neiva*, 1911, the most important Chagas' disease vector in northeastern Brazil (hemiptera, reduviidae, triatominae). *American Journal of Tropical Medicine and Hygiene* **67**, 516–20.

Cressie N, Calder CA, Clark JS, Hoef JMV, Wikle CK (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modelling. *Ecological Applications* **19**, 553–70.

De'ath G, Fabricius KE (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* **81**, 3178–92.

del Barrio G, Harrison PA, Berry PM *et al.* (2006). Integrating multiple modelling approaches to predict the potential impacts of climate change on species' distributions in contrasting regions: comparison and implications for policy. *Environmental Science & Policy* **9**, 129–47.

Deng H, Runger G, Tuv E (2011). Bias of importance measures for multi-valued attributes and solutions. Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN); 14–17 Jun 2011, Espoo, Finland. Springer, New York.

Dormann CF, Schweiger O, Arens P *et al.* (2008). Prediction uncertainty of environmental change effects on temperate European biodiversity. *Ecology Letters* **11**, 235–44.

Elith J, Graham CH (2009). Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. *Ecography* **32**, 66–77.

Elith J, Graham CH, Anderson RP *et al*. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**, 129–51.

Fisher RA (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–88.

Friedman J, Hastie T, Tibshirani R (2000). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics* **28**, 337–407.

Friedman JH (1991). Multivariate adaptive regression splines. *The Annals of Statistics* **19**, 1–67.

Friedman JH (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics* **29**, 1189–232.

Fuller T, Morton DP, Sarkar S (2008). Incorporating uncertainty about species' potential distributions under climate change into the selection of conservation areas with a case study from the Arctic Coastal Plain of Alaska. *Biological Conservation* **141**, 1547–59.

Graham CH, Elith J, Hijmans RJ *et al*. (2008). The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology* **45**, 239–47.

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

Guisan A, Zimmermann NE (2000). Predictive habitat distribution models in ecology. *Ecological Modelling* **135**, 147–86.

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

Hastie T, Tibshirani R (1986). Generalized additive models. *Statistical Science* **1**, 297–318.

Hastie T, Tibshirani R (1996). Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 155–76.

Hastie T, Tibshirani R, Buja A (1994). Flexible discriminant-analysis by optimal scoring. *Journal of the American Statistical Association* **89**, 1255–70.

Hopfield JJ (1982). Neural networks and physical systems with emergent collective computational abilities. *PNAS* **79**, 2554–8.

Iverson LR, Prasad AM, Matthews SN, Peters M (2008). Estimating potential habitat for 134 eastern US tree species under six climate scenarios. *Forest Ecology and Management* **254**, 390–406.

Kampichler C, Wieland R, Calmé S, Weissenberger H, Arriaga-Weiss S (2010). Classification in conservation biology: a comparison of five machine-learning methods. *Ecological Informatics* **5**, 441–50.

Korzukhin MD, TerMikaelian MT, Wagner RG (1996). Process *versus* empirical models: which approach for forest ecosystem management? *Canadian Journal of Forest Research* **26**, 879–87.

Levins R (1966). The strategy of model building in population ecology. *American Scientist* **54**, 421–31.

Li RQ, Tian HD, Li XH (2010). Climate change induced range shifts of Galliformes in China. *Integrative Zoology* **5**, 154–63.

Li WK, Guo QH, Elkan C (2011). Can we model the probability of presence of species without absence data? *Ecography* **34**, 1096–105.

Liu J, Jolly RA, Smith AT *et al*. (2011). Predictive Power Estimation Algorithm (PPEA) a new algorithm to reduce overfitting for genomic biomarker discovery. *PLOS ONE* **6**, e24233.

Marmion M, Luoto M, Heikkinen RK, Thuiller W (2009). The performance of state-of-the-art modelling techniques depends on geographical distribution of species. *Ecological Modelling* **220**, 3512–20.

McCullagh P, Nelder JA (1989). *Generalized Linear Models*. Chapman and Hall, London.

McCulloch W, Pitts W (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5**, 115–33.

McRae BH, Schumaker NH, McKane RB, Busing RT, Solomon AM, Burdick CA (2008). A multi-model framework for simulating wildlife population response to land-use and climate change. *Ecological Modelling* **219**, 77–91.

Meynard CN, Quinn JF (2007). Predicting species distributions: a critical comparison of the most common statistical models using artificial species. *Journal of Biogeography* **34**, 1455–69.

Moisen GG, Frescino TS (2002). Comparing five modelling techniques for predicting forest characteristics. *Ecological Modelling* **157**, 209–25.

Morin X, Thuiller W (2009). Comparing niche- and process-based models to reduce prediction uncertainty in species range shifts under climate change. *Ecology* **90**, 1301–13.

Nelder J, Wedderburn R (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A* **135**, 370–84.

Olden JD, Lawler JJ, Poff NL (2008). Machine learning methods without tears: a primer for ecologists. *Quarterly Review of Biology* **83**, 171–93.

Patt A, Klein RJT, de la Vega-Leinert A (2005). Taking the uncertainty in climate-change vulnerability assessment seriously. *Comptes Rendus Geosciences* **337**, 411–24.

Pearson RG, Thuiller W, Araujo MB *et al*. (2006a). Model-based uncertainty in species range prediction. *Journal of Biogeography* **33**, 1704–11.

Pearson RG, Thuiller W, Araujo MB *et al*. (2006b). Model-based uncertainty in species range prediction. *Journal of Biogeography* **33**, 1704–11.

Phillips SJ, Dudik M (2008). Modelling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* **31**, 161–75.

Phillips SJ, Dudik M, Schapire RE (2004). A maximum entropy approach to species distribution modelling. Proceedings of the 21st International Conference on Machine Learning; 4–8 Jul 2004, Banff, Canada.

Phillips SJ, Anderson RP, Schapire RE (2006). Maximum entropy modelling of species geographic distributions. *Ecological Modelling* **190**, 231–59.

Prasad AM, Iverson LR, Liaw A (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* **9**, 181–99.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Randin CF, Dirnboeck T, Dullinger S, Zimmermann NE, Zappa M, Guisan A (2006). Are niche-based species distribution models transferable in space? *Journal of Biogeography* **33**, 1689–703.

Reibnegger G, Weiss G, Wernerfelmayer G, Judmaier G, Wachter H (1991). Neural networks as a tool for utilizing laboratory information–comparison with linear discriminant-analysis and with classification and regression trees. *PNAS* **88**, 11426–30.

Royle J, Dorazio R (2008). *Hierarchical Modelling and Inference in Ecology*: *The Analysis of Data from Populations, Metapopulations and Communities*. Elsevier, Amsterdam.

Rykiel EJ (1996). Testing ecological models: the meaning of validation. *Ecological Modelling* **90**, 229–44.

Segurado P, Araujo MB (2004). An evaluation of methods for modelling species distributions. *Journal of Biogeography* **31**, 1555–68.

Stockwell D, Peters D (1999). The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science* **13**, 143–58.

Stockwell DRB, Peterson AT (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling* **148**, 1–13.

Sun Y, Todorovic S, Li J (2006). Reducing the overfitting of AdaBoost by controlling its data distribution skewness. *International Journal of Pattern Recognition and Artificial Intelligence* **20**, 1093–116.

Thuiller W, Brotons L, Araujo MB, Lavorel S (2004). Effects of restricting environmental range of data to project current and future species distributions. *Ecography* **27**, 165–72.

Thuiller W, Albert C, Araujo MB *et al*. (2008). Predicting global change impacts on plant species' distributions: future challenges. *Perspectives in Plant Ecology, Evolution and Systematics* **9**, 137–52.

Thuiller W, Lafourcade B, Araujo M (2009a). Mod Operating Manual for BIOMOD. Laboratoire d'Écologie Alpine, Université Joseph Fourier, Grenoble, France.

Thuiller W, Lafourcade B, Engler R, Araujo MB (2009b). BIOMOD–a platform for ensemble forecasting of species distributions. *Ecography* **32**, 369–73.

Tsoar A, Allouche O, Steinitz O, Rotem D, Kadmon R (2007). A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions* **13**, 397–405.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*, 4th edn. Springer, Berlin.

Wiens JA, Stralberg D, Jongsomjit D, Howell CA, Snyder MA (2009). Niches, models and climate change: assessing the assumptions and uncertainties. *PNAS* **106**, 19729–36.

Wikle CK (2003). Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology* **84**, 1382–94.

Wikle CK, Berliner LM, Cressie N (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics* **5**, 117–54.

Yee TW, Mitchell ND (1991). Generalized additive-models in plant ecology. *Journal of Vegetation Science* **2**, 587–602.