Modeling Species' Distributions

Carsten F. Dormann

Helmholtz Centre for Environmental Research-UFZ

Department of Computational Landscape Ecology

Permoserstr. 15

04318 Leipzig

Abstract

Species distribution models have become a commonplace exercise over the last 10 years. Still, analyses differ a lot due to different traditions, aims of applications and statistical background knowledge. In this chapter, I lay out what I consider to be the most crucial steps in a species distribution analysis: data pre-processing and visualisation, dimensional reduction (incl. collinearity), model formulation, model simplification, model type, assessment of model performance (incl. spatial autocorrelation) and model interpretation. For each step, the most relevant considerations are discussed, mainly illustrated with Generalised Linear Models and Boosted Regression Trees as the two most contrasting methods. In the shorter second part, I draw attention to the three most challenging problems in species distribution modelling: identifying (and incorporating into the model) the factors that limit a species' range; separating the fundamental, realised and potential niche; and niche evolution.

Introduction

As species of all types undergo rapid human-induced extinction, understanding why species occur where they do is becoming a highly relevant, pressing and potentially life-saving topic. Conservation actions, such as establishing protected site networks, adapting land use, providing stepping-stone habitats all require an idea of how the target species will respond. Also using organisms as "bioassay", i.e. indicators of environmental trends (such as climate change, air pollution, overfishing) demands an intimate knowledge of the organism's niche. Species distribution modelling attempts to identify the probable causes of species whereabouts. We seek to delineate the realized niche of an organism based on its current distribution

with respect to the environment (see Elith and Leathwick 2009b for definitions and concepts; Guisan and Thuiller 2005; Kearney 2006; Soberón 2007).

Why Species Distribution Modelling?

There are several fundamental challenges to this approach (e.g. first and foremost that it is correlative; see Vaughan & Ormerod (2005) and Dormann (2007b) for a recent critique), and before jumping into the analysis, it is worth considering whether SDMs are actually useful and fit for the purpose of *your* specific problem. For example, at very small spatial scales, differences in environmental conditions may be too small to be of predictive value and biotic interactions (competition, predation) may be of crucial importance. At the global scale, in contrast, data become so coarse that we "only" model the climate niche and specific habitat requirements cannot be detected.

On the other hand, SDMs try to extract ecological information from a species occurrence pattern *when and where it matters*. Expert knowledge usually cannot inform us which trait or limitation will be relevant for our problem at hand. We may know that a palm tree does not survive sub-zero temperatures, but the observed distribution will tell you that even 10 x 10 km grid squares with minimum temperatures well below 0°C harbour this species because of microclimatically suitable places. Thus, at the spatial resolution under investigation, the physiological threshold can be misleading even though it may be true. Overall, SDMs are useful for complementing existing approaches in at least these five areas of research:

1. Small-extent, decision-support for conservation biology (such as Biological Action Plans: Zabel et al. 2003, and numerous others);
2. testing specific hypotheses, e.g. on the spatial scale of habitat selection (Graf et al. 2005; Mackey and Lindenmayer 2001), the species-energy hypothesis (Lennon et al. 2000) or range-size effects on diversity pattern (Jetz and Rahbek 2002);
3. generating hypotheses, e.g. on correlation of species traits with environmental variables (Kühn et al. 2006), which can then be tested experimentally;
4. identifying hierarchies of environmental drivers (Bjorholm et al. 2005; Borcard and Legendre 2002; Pearson et al. 2004);
5. prospective design of surveys, e.g. optimizing sampling schemes for rare species (Guisan et al. 2006).

Now, we shall focus on the technical side, and assume that you know what you are doing, ecologically speaking.

Analyzing the geographic distribution of species' occurrence, abundance or diversity is, essentially, a statistical task. As such, the fundamental ideas and principles of good statistics apply (and can be found in the excellent but advanced book

of Hastie et al. 2008). There are three reasons, at least, why methods for describing or modeling these patterns have reached a higher level of sophistication than many other field in ecology. Firstly, biogeographical data sets are nowadays large (both in terms of number of data points and potentially explanatory variables), necessitating the use of new statistical strategies. Secondly, species distribution data typically carry a largish bunch of common intrinsic statistical problems and accordingly several solution were tailored to these problems (presence-only data, low information content of binary data, spatial autocorrelation, multi-collinearity, model unidentifiability). Thirdly, species distribution modeling (SDM) is "sexy". As habitat of many species is constantly lost, as climate changes and as environmental management becomes a matter of human survival, scientists, decision makers and the general public look for information and predictions of possible future scenarios. In consequence, substantial funding (at least for ecological topic) over the last decade has enabled talented scientists to make a career from SDMs.

Aims of this chapter

The recent developments have made the field of SDM somewhat complex, divers and confusing for the newcomer. The aim of this chapter is thus to (1) provide a recipe for SDM; (2) briefly discuss a few selected "hot" topics; and (3) give an outlook to challenges of a more ecological modeling type (dispersal, occupancy, biotic interactions, functional variables, evolution, changing limiting resources). I shall restrict citations to fundamental or specific methodological papers and will therefore have to ignore the vast amount of good ecological papers that "only" did it right. On the other hand, I am not aware of any paper on species distribution modeling that could tick all elements of the recipe below.

A Species Distribution Modeling recipe

A good cook needs no recipe. Alas, we are more trained in ecology than statistics. Moreover, without the right ingredients (a.k.a. data) and tools (software), no dish will be tasty. Also, I should mention other recipes along this line: see Harrell (2001) for a generic statistical recipe, and Pearson (2007) and Elith & Leathwick (2009a; 2009b) for a specific one on SDMs. As for "cooking tools", I highly recommend using code-based software so that each step of the analysis is documented and easily reproducible. The functions mentioned in this chapter are all from the free R environment for statistical programming (R Development Core Team 2009).

The recipe falls into three section: pre-processing, modelling and model interpretation (Fig. 1). These sections are somewhat arbitrary but useful to structure the

whole endeavor. We shall assume that you have your ingredients well prepared: The observed data are as good as we need them, the explanatory variables are ecologically relevant and at the same resolution and your statistical tools are laid out in front of you. A worked example is available online (*Where's the sperm whale?*), which follows the recipe and provides example data and R-code.

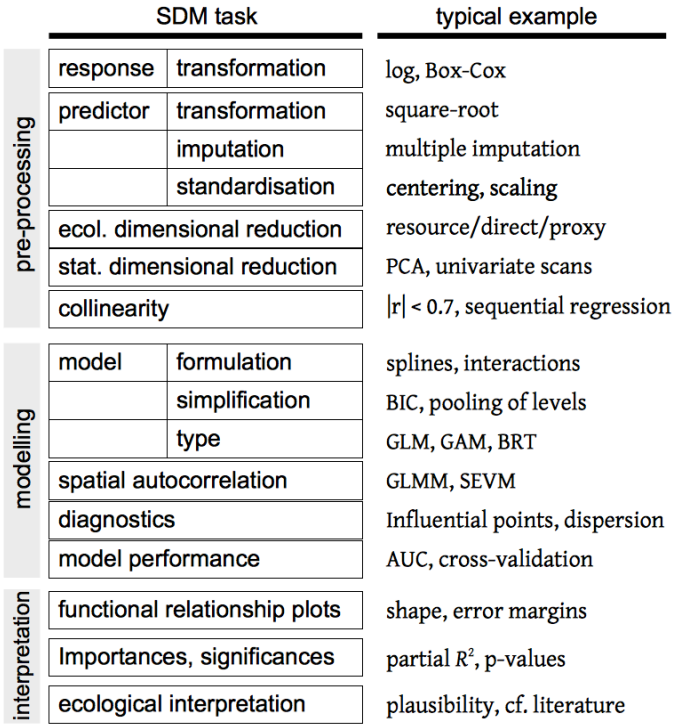

Fig. 1. Overview of the species distribution modelling workflow. The three phases contain various tasks, for which typical examples are given in the right column.

Pre-processing and visualization

The response variable

When the data are presence-absence (i.e. binary) no further preparation is needed. When data are counts or continuous, we have to make sure that assumptions of the

modeling approach are met. For parametric modelling approaches (regressions by means of GLM or GAM), count data are usually assumed to be Poisson distributed but all too often are not. Continuous responses are generally assumed to be normally distributed. These assumptions can be checked only *after modeling*, because we need to look at the residuals or compare log-likelihoods of different distributions. Generally, if too many zeros have been observed, the data are over-dispersed and we have to resort to one of three alternative approaches: a quasi-Poisson distribution (where over-dispersion is explicitly modeled); a negative binomial distribution (where similarly a clumping parameter is fitted); or a separate analysis of zeros and non-zeros (as in zero-inflated or other mixed distribution models: Bolker 2008). Sometimes people log-transform count data (more precisely: y' = log(y+1)), and find the new y' to be normally distributed.

Normally (Gaussian) distributed data show a normal distribution in the model residuals and a straight 1:1 relationship in a QQ-plot of these residuals. Deviations need to be accounted for, e.g. by transforming the data (any good introductory textbook, such as Quinn & Keough (2002), will feature a section on transformations, including useful the Box-Cox[1] transformation).

When we have presence-only data (i.e. only locations where a species occurs but no information where it does not), two alternative approaches are available. We could use purpose-build presence-only methods, or we could use all locations without a presence and call them absences (pseudo-absences). Both approaches have their difficulties (Brotons et al. 2004; Pearce and Boyce 2006). The first suffers from a lack of sound methods (in fact, following e.g. Tsoar et al. (2007) and Elith & Graham (2009), I would currently only recommend MaxEnt[2] in this direction and hope for the approach of Ward et al. (2009) to become publically available). The second approach lacks simulation tests on how to select pseudo-absences and how to weight them (see Phillips et al. (2009) for the cutting edge in this field), although it has been argued that the pseudo-absence approach can be as good or better than the purpose-build presence-only methods (Zuo et al. 2008). In what follows, I only consider presence-(pseudo)absence data.

The explanatory variables

Also explanatory variables may require transforming! Consider a relevant explanatory variables which is highly skewed (e.g. log-normally distributed), as is commonly the case for land-use proportions. Some few high-value data points may completely dominate the regression fitted. To give a more balanced influence to all data points, we want the values of the predictors to be uniformly distributed over their range. This will rarely be achievable, and mostly researchers settle for a

[1] `boxcox` in **MASS** (`typewriter` and **bold** are use to refer to a `function` and its **R-package**)

[2] Phillips et al. (2006b): http://www.cs.princeton.edu/~schapire/maxent/

more or less symmetric distribution of the predictor. Note, however, that ideally we want most data points where they help most. For a linear regression, the mean is always best described, so we would want most data points at the lowest and highest end of the range. For a non-linear function, for example a Michaelis-Menton-like saturation curve, we want most data points in the steep increase, while there is little gained from many points at the high end, once the maximum is reached. As a rule of thumb we need many data points where a curve is changing its slope.

Transformation of explanatory variables is needed particularly for regression-type modeling approaches such as GLM and GAM (see below for explanation). Regression trees (used, e.g., in Boosted Regression Trees, BRT, or randomForest) are far less sensitive, if at all (Hastie et al. 2008). It is a good custom to make a histogram of each explanatory variable before entering it into an analysis! Transformation options are the same as for the response.

Missing data are a (very) special case of transformation. Although generally disliked by many analysts, imputation (replacement of missing data) is often a good idea (see Harrell 2001), particularly if missing data are scattered through the data set (i.e. across several variables!) and we would loose many data points if we simply omitted every data point with missing values. Standard imputation uses the other explanatory variables to interpolate a likely value for the missing one[3]. Replacement by mean is *not* an option!

"Outliers" are (in general) a red herring: If there is no methodological reason why a data point is extremely high (e.g. one data set being recorded in winter, while all other data points are from the summer), then this datum should also be included in the analysis. Otherwise the data set may be poorly sampled, but the "outlier" would still represent a (potentially) valuable datum. It would be good practice to omit it later on and see if the results are robust to this omission. Furthermore, in multi-dimensional data sets (i.e. those with several explanatory variables), a datum might be an "outlier" in one dimension, but a ordinary data point in all others: why delete it?

Finally, all continuous variables should be standardized before the analysis[4]. This reduces collinearity, particularly with interactions (Quinn and Keough 2002). As a convenient side effect, regression coefficients are now directly comparable: the larger their absolute value, the more important is this term in the model (they become standardized regression coefficients).

Collinearity

Collinearity refers to the existence of correlated explanatory variables. Some predictors are only proxies for an underlying, latent variable. For example, consider

[3] `transcan` and `aregImpute` in **Hmisc**

[4] `scale`

temperature and rainfall, which are largely governed by distance to ocean (oceanity), altitude and regional terrain. Collinear predictors can lead to biased models due to inflated variances (Quinn and Keough 2002). There are many cures to this ailment, but no remedy. Logically speaking, if two predictors are tightly linked to an underlying (but elusive) causal variable, there is no way to find out which is the "correct" predictor for our analysis. We may choose precipitation over temperature when modelling plants (or the other way around for insects), but there is no guarantee that this choice allows us a sensible extrapolation of our model. Also using Principal Component Analysis (PCA) or any of the other tailor-made methods for collinearity (Partial Least Squares, penalized regression, latent root regression, sequential regression, and many others) will not solve the ecological problem, only the statistical. These methods will produce either a new data set of uncorrelated variables, or "consider" the correlation when estimating model parameters.

So where is the problem? Imagine an organism whose distribution is governed entirely by its sensitivity to frost. When we combine our climate variables into one or more principal components, model the species' distribution, and then predict to a climate change scenario, the fact that both rainfall and mean summer temperature are correlated with number of frost days will *dilute* its impact in the model. The total effect of "frost" is distributed over all correlated variables. As a consequence, any climate prediction will underestimate the effect of frost and hence yield a "wrong" expected future distribution. If we don't know the true underlying causal mechanism, no statistics can help us here (or at least very little). Any ecological knowledge used in variable pre-selection, however, will lead to a smaller bias in scenario projections!

Dimensional reduction

Often we may have dozens or even hundreds of potential explanatory variables (e.g. from multispectral remote sensing or landscape metrics). We should try to reduce this set too as few as possible for two reasons: (1) The more variables we have, the more they will be correlated. (2) The more variables we have, the more likely one of them will spuriously contribute to our model (type I error). For SDMs, Austin (2002) and Guisan & Thuiller (2005) argue that we should choose "resource" over "direct" and "direct" over "indirect" variables. For example, the abundance of prey (hardly ever available) or nesting opportunities will be a resource variable when analyzing the distribution pattern of a bird of prey. Temperature or human disturbance could be direct variables, impacting on the bird without moderation by other variables. Indirect variables would be altitude or length of road in a grid cell, which are substitutes, surrogates or proxies for other, more directly acting variables. These indirect variables are often not immediately perceivable by the organism (such as altitude by a plant or length of road verges by a rodent). So if we have two (correlated) variables, we should discard the one "further away" from the species' ecology.

If we are unable to reduce the data set sufficiently (i.e. k >> N), we should use dimensional reduction techniques, such as Principal Component Analysis[5] or its more sophisticated variants that also allow categorical variables (nMDS[6]). The scores for the most important axes in this new parameter hyperspace can be used as explanatory variables. Note that interpretation is often extremely impaired by automatic dimensional reductions. It is thus always advisable to use ecological understanding rather than statistical functions at this step!

An alternative is to "filter" the data by importance. We can use a robust and able technique to tell us which variables are important. Next, we use only those five or 12 variables filtered from the initial pool of variables, and continue. Regression-tree based methods are very useful for this, and I recommend randomForest and Boosted Regression Trees. If you plan to model your data with BRT anyway, there is little point in reducing the data before.

Finally, be aware that any model can only find correlations with the variables provided. Of course we know that our hypothetical bird of prey depends on specific prey. Without this information, we may actually be modelling the niche of the prey, not of the predator!

Exploratory data plotting

Can we finally start? No! It is both good practice and highly advisable to look at the data by plotting them in any reasonable combination conceivable (see, e.g., Bolker 2008). Plot thematically related explanatory variables as scatterplot[7] to detect collinearity. Plot each explanatory variable against the response (henceforth called *X* and *y*, respectively) and look for non-linear effects. Plot *y* against two *X*s (Fig. 2) and a hull-polygon around the data to see that 40% or so of the parameter space is not in your data set. This is the area outside the convex hull in fig. 2. The more variables (and hence dimensions) your data set has, the more severe this problem becomes. It is so prominent among statisticians (though not among ecologists) that it is referred to as the "curse of dimensionality" (Bellman 1957; Hastie et al. 2008). Repeat this plotting for any number of variables. Getting a feeling for the data is crucial, and many later errors can be avoided. Every minute invested at this stage saves hours later on.

---

[5] `prcomp`

[6] `isoMDS` in **MASS** or, more conveniently, `metaMDS` in **vegan**

[7] `pairs`

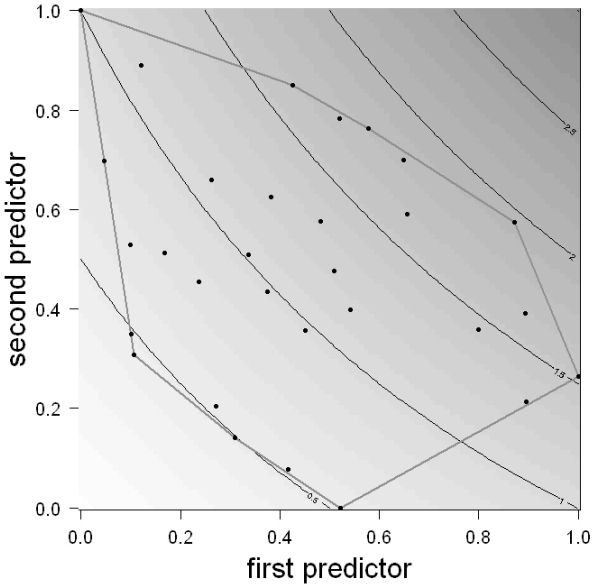

Fig. 2. Visualizing the parameter space supported by data. In this case, the top-right and bottom-left corner of the parameter space of the two predictors has not actually been sampled by data (despite a low correlation of r = –0.26). The actually sampled parameter space, indicated by the convex hull covering 57% or the area, declines dramatically with the number of dimensions ("curse of dimensionality"). In other words: we have few data points to look at interactions of higher order.

Modelling

Here, we arbitrarily divide the process of deriving a "useable" model into two steps. The first, model building, selects the variables to be included, the type of non-linearity and order of interactions considered, and the criteria for selecting the final complexity of a model. The second step, model parameterization, does the final step of using the data to calculate the best estimates for variable effects. It is this model that we want to use for interpolation, hypothesis testing or extrapolation. Note that in some methods these two steps are implicitly taken care of and that there is no two-step process (mainly machine learning, where model selection is done internally through cross-validation in order to prevent models from being "unreasonably" large: e.g. Hastie et al. 2008). For more traditional approaches (and here I am thinking of GLMs), we may want to have these steps functionally separated.

Model formulation

We have reduced our data set to a moderate size of predictors in the step "Dimensional reduction" above. Now we still need to specify in which functional form in which the predictors are allowed to correlate with the response. In early years, both non-linear and interactive model terms were neglected, making many of their findings less trustworthy. Modern methods (such as BRT) will automatically have non-linearity and interactions build-in. It is still important to understand the relevance of non-linearity and interactions, even when using the tree-based methods, because we still have to be able to interpret the results. The information on the importance of a variable often returned by machine-learning algorithms does not allow us to see *how* the variables act. As shown in the case study (see online appendix: *Where's the sperm whale?*), the functional relationship must be plotted to gage its shape. For interactions we need to plot each variable at each level of the other variable, thus visualizing synergistic or compensatory effects of the two variables.

The key idea hind SDM, i.e. the environmental niche of a species, implies a hump-shaped relationship between any environmental predictor and a species' occurrence: there are lower and upper limits. Hence, we *must* allow the model to be non-linear. If we happen to only sample a part of the entire gradient, we also need to consider saturation curves, which are again non-linear. The simplest, and generally sufficient, way to include non-linearity is by generating a new, squared dummy variable for each continuous predictor[8]. This represents the third element of a Taylor series (which can be expanded to represent *any* function). When using GAM or other spline-based approaches, non-linearity is governed by the smoothing function used. Here the issue is not so much *how* to model non-linearity, but rather *how much* non-linearity we allow for. Reducing the "wiggliness" of splines (either by stepwise model selection for the number of knots in each predictor[9] or by shrinkage of spline fits[10]) prevents overfitting and should be the standard approach.

Interactions are similarly relevant. Statistically, an interaction is the product of the participating main effects. Ecologically, it means that we need to know the value of all variables included in the interaction, not only the main effects. Because this is highly relevant and often difficult for the beginner, let me briefly give an example. Assume that global patterns of plant diversity are well-predicted by the predictors "annual precipitation" and "mean annual temperature" – and their interaction. For the main effects, wet or hot means more species, but not necessarily. When a site is hot, it needs to *also* be wet to have high species richness; oth-

---

[8] This can be done either manually (`X1.2 <- X1^2`) or as part of the model formula (`y ~ X1 + I(X1^2)`); higher-order polynomials should be specified using poly (`y ~ poly(X1, degree=3)`), which calculates orthogonal polynomials.

[9] As is proposed for the function `gam` in package **gam**: see `?gam::step.gam`

[10] As proposed for the function `gam` in package **mgcv**: see `?mgcv::step.gam`.

erwise it may well be a barren desert. But when cold, a site will never support many plant species, independent of precipitation. In this example, neither temperature nor rainfall alone is sufficient to predict species richness at any site, but we need to interpret them in concert.

Classification and regression trees (CARTs) embrace non-linearity and interactions in an elegant and natural way. Their boosted (BRT) or bagging (randomForest) extensions hence do not require specification of non-linearity and interactions.

Model simplification

One of the fundamental problems in building statistical models is the trade-off between the variance explained by the model, and the bias it produces when validating it on a new, independent data set (variance-bias-trade-off: Hastie et al. 2008). Smaller models are more robust, i.e. less biased, at the expense of being not very good in explaining variance. The way to derive the "optimal" model size is through cross-validation (CV). For some modelling approaches this is automatically implemented, but the majority of model types require the user to carry out this step[11]. *N*-fold cross-validation encompasses a random assignment of data points to the *N* subset, with *N* usually between 3 and 10. Care should be taken to have equal prevalence in all subsets, e.g. by randomizing 0s and 1s separately (stratified randomization). The model is then fitted to *N*–1 of the *N* subsets and evaluated on (by predicting to) the remaining subset. This is repeated for all *N* subsets and evaluations are averaged. Based on these values, we can select the best modelling strategy (both model complexity and model type). An alternative approach is to bootstrap the entire model building process and use bootstrapped measures of model performance. Since a bootstrap requires several thousand runs, and a CV only a few, CV is far more common.

Information theoretical approaches are based on analytical methods to describe this CV. Hence Akaike's Information Criterion (AIC) or Schwartz'/Bayesian Information Criterion (BIC) are implicitly also based on cross-validation. While it is clear that too large a model will be overfitting, and that too small a model will not capture as much of variation as it should in the data, the "true" model will always remain elusive, and our "optimal" model will only be a caricature of the truth. There is much to be learned from this caricature, however!

Model type

[11] Do not confuse out-of-bag model weights and alike with cross-validation of the entire model. When data are held back during the building of sub-models (e.g. in randomForest or BRT), this does not represent a cross-validation of the entire model (i.e. the average of sub-models).

At this point we have to choose one (or more) method(s) to do our analysis with. The good "traditional" approaches comprise Generalised Linear Models (GLM) and Generalised Additive Models (Guisan and Zimmermann 2000). Discriminant Analysis has been given up on, as have been Neural Networks and CARTs (Guisan and Thuiller 2005). "Modern" approaches are often based on either multi-dimensional extensions of GAMs (such as MARS and SVM) or machine-learning variations of CART (such as BRT and randomForest: Hastie et al. 2008). Anyone using a machine-learning method should familiarize himself with this method. The majority of them are performed on real data set, where the truth is unknown and the performance of a method hence assess by cross-validation. These comparisons show, broadly speaking, that model types differ sometimes dramatically in performance, that each model type can be misused and that both GLM and BRT are reliable methods when used properly.

This is not the place to explain the differences between all of them (see Hastie et al. (2008) for a recent and comprehensive description or Elith & Leathwick (2009a)). It has to suffice to make clear the main difference in the machine-learning approach to "traditional" statistical models. In traditional models (e.g. GLM), we specify the functional relationship between the response and its predictors. So, for example, we decide to include precipitation as a non-linear predictor for plant species richness. This model proposal is then fitted to the data. In machine learning, we propose only the set of predictors, but not the model structure. Here, an algorithm builds a model proposal, fits it to a part of the data set and evaluates its performance on the other part of the data. It then proposes a modification of the original model and so forth. Machine-learning algorithms[12] differ in scope, origin, complexity, speed, but they all share this validation step which is used to steer the algorithm towards a better model formulation. There are plenty of studies comparing different modelling approaches (Guisan et al. 2007; Meynard and Quinn 2007; Pearson et al. 2006; Segurado and Araújo 2004). Rather, we shall continue using GLM and BRT as representatives for the two most common good approaches.

The choice of model type has much to do with availability of software, current fashion and of course with the specific aim of the study. Further complications arise if the design of the survey may require a mixed model approach (e.g. due to repeated measurements or surveys split across observers), if spatial autocorrelation needs to be addressed, if zero-inflated distributions have to be employed, and if corrections for detection probability shall be modeled. The more additional requirements are imposed on the model, the more GLMs become the sole possible method[13]. Alternatively, you may want to go for a Bayesian SDM (see Latimer et al. 2006, for a primer).

---

[12] http://www.machinelearning.org/ is a good place to start exploring this field

[13] Most of these "complications" can be handled by standard extensions of GLMs {see, e.g., \Bolker, 2008 #8207, and various dedicated R-packages). They will, however, make the model less stable, require larger run-times and still rely on getting the distribution right. There is, of

If your data and model require an unusual combination of steps (say a combination of zero-inflated data with nested design and spatial autocorrelation, while predictors are highly correlated and many values missing), and you develop a way to cook this dish, then you should do (at least) two things: Firstly, evaluate your method for its ability to detect an effect that you *know* is there ("power"). Secondly, evaluate your method for its sensitivity to detect effects that you know are *not* there ("type I error"). Both evaluation should be amply replicated, should be based on simulated data (so that you know the truth) and should (finally) confirm that your new methods is reliable!

Spatial Autocorrelation

Spatial autocorrelation (SAC) refers to the phenomenon that data points close to each other in space are more alike than those further apart. For example, species richness in a given site is likely to be similar to a site nearby, but very different from sites far away. This is mainly due to the fact that the environment more similar on short distance. Hence, SAC in the raw data (species richness) is a consequence of SAC in the environment (topography, climate), something Legendre (1993) termed "spatial dependence". In SDMs, we do not care about SAC *per se*, but about SAC in the model's residuals (i.e. unexplained by the environment), because it distorts model coefficients (Bini et al. 2009; Dormann 2007a). To date it is unclear whether this residual SAC is mainly due to model misspecifications (omission of non-linearity and interactions), due to variation in sampling coverage, due to omission of important predictors or due to ecological processes (territoriality, dispersal). Only against some of these problems can *statistical* solution be found. The spatial toolbox is rich in approaches (Beale et al. 2010; Carl et al. 2008; Dormann et al. 2007; Mahecha and Schmidtlein 2008). In any case, SDM residuals should be investigated for spatial autocorrelation, and attempts should be made to correct for it. If spatial models yield similar coefficient estimates (GLM) as non-spatial models, then there seems to be little value in "going spatial": the ranges of the spatial autocorrelation may or may not be related to the ecological scale of movement or behavioral patterns (Betts et al. 2009; Dormann 2009).

Tweaking the model

There are several ways in which to increase the quality of the model (Maggini et al. 2006). One important start is to investigate the model residuals. They indicate

course, the alternative of Bayesian implementations. Since these are also fundamentally maximum likelihood approaches, they are similar to sophisticated GLMs. In any case, there is no Bayesian Boosted Regression Tree (not to speak of a combination with spatial terms and mixed effects). It runs against the Bayesian philosophy to use boosting or bagging, and there is no efficient implementation either.

whether model assumptions were violated (e.g. when residuals are highly skewed or their variance is not the same throughout the range of fitted values) or if some non-linear relationship went unnoticed (residuals may e.g. show a hump-shaped trend against fitted values).

Model diagnostics[14] will also indicate outliers, i.e. data points that have a high influence on the model coefficients. We can use weights to decrease an outlier's impact. Weights are also useful when the balance between presences and absences is very disturbed. Down-weighting the more common category so that model weights sum to the same value for 0s and 1s has been shown to increase the sensitivity of binomial models (Maggini et al. 2006). The same approach is recommended when using pseudo-absences (Elith and Leathwick 2009a).

By including data from other scales or broader geographic coverage, regional or local SDMs can be improved, too. Pearson et al. (2004) used European distribution and climate data to fit a niche model for four plant species. Predicted probabilities of occurrence from this model were then used as input variable alongside land-cover variables in the second-step model for the UK. Thereby the authors obviated the problem that the climate gradient in the UK is much shorter than of the species' global distribution.

Assessing model performance

To quantify how good our model fits the data, we compare model predictions with field data (usually on a hold-out sample, e.g. the subset of a cross-validation). Traditionally, the probability predictions from the model were converted into presences and absences and then a confusion matrix can be used to calculate various parameters of choice (e.g. commission and omission error, kappa, etc: Fielding 2002). The AUC ("area under curve") is currently the most commonly used measure of discriminatory power of a model. Its value (between 0.5 for random and 1 for perfect) quantifies the ability of the model to put the data points into the correct class (i.e. presence or absence), independent of the threshold required by the other measures mentioned. It has recently received justified criticism because its values are not comparable across different prevalences (and the criticism extends to kappa, too: Lobo et al. 2008). Currently, misclassification rates, commission and omission errors are more *en vogue* again, because they can be intuitively interpreted. Furthermore, by assigning different weights to false negatives (omission error) and to false positives (commission error), conservation management can come to more sophisticated and balanced decisions (Rondinini et al. 2006).

Only rarely will a second set of data be available to investigate the quality of our model(s) through external validation. A different recording strategy, another time slice or data from a different geographic location represent really inde-

[14] Diagnostics for GLMs fitted in R are given by plotting the model object.

pendent data, and could thus be considered an external validation. The internal validation (described above as cross-validation) is an optimistic assessment of model quality. When using SDMs to infer underlying mechanisms, external validation is less of an issue than when using them to extrapolate to a future climate or other sites. Because the cross-validated models are optimistic, they give narrower error bands than they should.

Interpretation

Once we have arrived at what we regard as a final model, we should make any effort to understand what it means. A first and most relevant step is to visualize the functional relationships within the model. The plot of how occurrence probability is related to, say, annual precipitation should be accompanied by a confidence band around this line. It may be useful to plot the data as rug into this figure to visualize the support at each point in parameter space (Fig. 3).

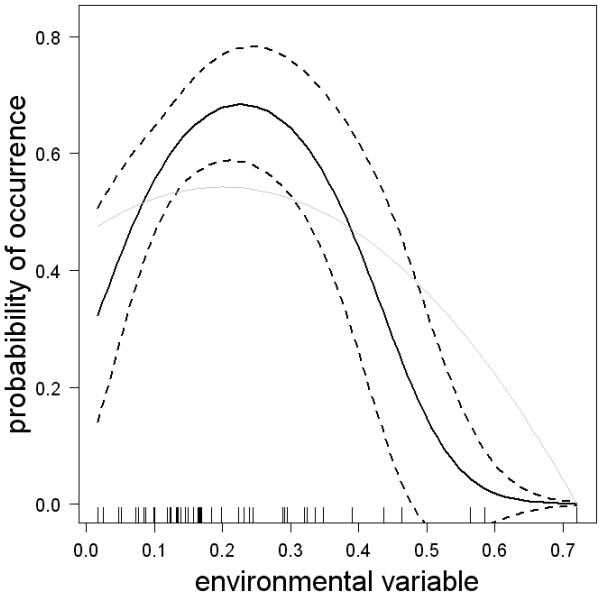

Fig. 3 Functional relationship between an environmental variable and a binary response. Rug (ticks on lower axis) indicate for which x-values data were available. Lines represent a quadratic fit (solid) and its standard deviation (dashed). Thin grey line is the true, underlying, data-generating function. Note the few data points upon which the declining half of the function is based (6 of a total of 50 have a value > 0.35).

For interactions, visualization becomes more difficult. Two-way interactions can still be plotted (e.g. as a 3-D plot or as a contour plot). No confidence bands can be included, though. Here it is again very important to indicate the position of the samples to identify regions of the parameter space that have not been sampled. For higher dimensions, or for model that average across many sub-models, we can do the same plots (called marginal plots for main effects then, because they represent the marginal changes to a predictor, averaging across all other predictors). We can also slice through higher dimensions, i.e. calculate a marginal plot for specific values of other predictors (often their median).

Spending time plotting is again well invested. We will detect errors in the model, scratch our head over inexplicable (and hence overly complex) patterns, and be forced to extract the main conclusions from it. It is this phase where the traditional GLM is superior to the BRT, because variable interpretation is easier. It is, however, also this phase where we may realize that BRTs are superior to GLMs because they can model step-changes and thresholds much better. Personally, I think we should not publish patterns we do not understand. There is, as the previous steps have shown, several decisions that could generate artifacts and their publication cannot be seen as progress.

Beyond recipes: new challenges for species distribution models

The above recipe can be used to derive a static description of environmental correlates with distribution data. But they often leave the analyst unsatisfied. Many assumptions can be suspected to be violated (Dormann 2007b), such as stationarity, unbiased coverage, or equilibrium with environment[15]. Three key challenges are to do with the following problems of current SDMs: (1) A niche model describes the *current* niche, and it is unclear which factors will be limiting elsewhere or in the future. (2) Climate change projections delimit only the *potential* future distribution, and it is unclear whether the species will ever fill this new range, e.g. due to dispersal constraints. And, (3), our understanding of the adaptive potential of a species is currently very poor. This sets, in part, the research agenda for species distribution models. Let us look at these challenges in more detail.

[15] Actually, the term "equilibrium" is a bit misleading. What is meant is that the entire width of its niche is filled. Within this niche, there may well be unoccupied sites, e.g. due to metapopulation dynamics. A problem arises, when a species does not occupy say the dry end of its soil moisture niche for historic reasons. Then the estimate of this end of the niche will be biased.

What limits a species' range?

An organism is constrained in its population dynamics by resources, competitors, predators and diseases, density dependence, reproductive opportunity, mutualists, environmental stochasticity and so forth (e.g. Krebs 2002). The same holds true for its spatial distribution (e.g. Gaston 2009; Holt and Barfield 2009), but additionally spatial constraints come into play (e.g. distance between habitat fragments, minimum territory size, Allee effects due to low population density). With an SDM, we are usually only able to quantify some of these limitations, and, accordingly, SDMs often do not transfer very well to other sites (Schröder & Richter 1999, Randin et al. 2006, Duncan et al. 2009, but see Herborg et al. 2007). In particular biotic interactions are hardly ever quantified explicitly within SDMs (although some of them will also correlate with the environmental data used). But they matter in real data (e.g. Preston et al. 2008; Schweiger et al. 2008), and they impact model performance and predictive ability (as shown, in a simulation study, by Zurell et al. 2009). It is thus a key challenge to incorporate biotic interactions into SDMs, but to date such attempts are far and few between (e.g. Bjornstad et al. 2002). A recent review (Thuiller et al. 2008) indicates some avenues to do so, but all of them are based on the explicit modelling of populations within cells, if not individuals. It is unclear, how to derive a more general dynamic SDM without spending years per species on incorporating detailed ecological knowledge.

Fundamental, potential and realized niches

The reason why many biogeographers refer to SDMs as "Species Distribution Models" and not as "niche models" is because they do not believe that we model the niche of the target species. In fact, as Jiménez-Valverde et al. (2008) argue, because we do not know *why* a species is absent in some sites, we are in the dark about its niche. The discussion of what a niche is, and what we are modeling, has sparked several interesting and not always compatible publications (e.g. Kearney 2006; Soberón 2007). Hence, Araújo & Guisan (2006) have named the "clarification of the niche concept" the first of five challenges for SDMs. While we cannot resolve this issue here, it is important to realize that the "niche" based on the correlation between geographic distributions and environmental conditions is quite a bit more vague than the niche discussed in evolutionary ecology, where resources and other causal drivers are envisaged (see, e.g., Losos 2008).

More to the point in this context is the challenge to quantify how much of the fundamental niche is actually covered by the realized niche as extracted from SDMs. If, on one extreme, the realized niche is pretty much also the fundamental niche (i.e. there are no biotic interactions and alike to constrain the distribution at the scale we are analyzing), then we can merrily predict future distributions of this

species (e.g. under climate or land-use change). At worst, we are overestimating the future, "potential" distribution (if in the future biotic interactions may become limiting or if species do not reach the sites). At the other extreme, if the fundamental niche is considerably wider than what we model, any projection can be fundamentally flawed (Dormann et al. 2010). I am not aware of any study assessing the overlap of realized and fundamental niche for geographic distributions (see also Nogués-Bravo 2009). It could require transplant experiments into areas beyond the current range and the manipulation of biotic interactions there. The few studies going into this direction point at a large discrepancy between fundamental and realized niche. Battisti et al. (2006), for example report on a range shift after a particularly warm summer, which was not reverted afterwards, indicating that it was dispersal limitation that prevented a filling of the niche. Similarly, several studies point at the importance of dispersal limitation (Nekola 1999; Ozinga et al. 2005; Samu et al. 1999; Svenning and Skov 2004), leading to both a bias in the modeled environment-occurrence relationship as well as the width of the niche itself. There is, as yet, no standard way to wed SDMs and dispersal (see Johst et al. 2002; King and With 2002; Lavorel et al. 2000; Lischke et al. 2006; Midgley et al. 2006; Schurr et al. 2007, for attempts; Thuiller 2004).

Niche evolution

Another important and fast developing field related to species distribution modelling is the study of niche evolution. I shall use this term very loosely, as is often done, to also include micro-evolutionary changes, genetic (and ecological) drift within species and genotypic plasticity (Pfenninger et al. 2007). Climate change projections using SDMs rely on the assumption that species are not able to adapt significantly to altering environmental conditions. This assumption is implicit in the extrapolation of the fitted niche: if a species was able to adapt rapidly, then the present niche would not be related to its future niche.

The problem is that we have considerable, if patchy, evidence that niches can rapidly evolve (reviewed in Thompson 1998), change within the fundamental niche (Dormann et al. 2010) or at least that variability within a species is large enough to allow it to shift its niche when confronted with novel environments (e.g. Ackerly et al. 2006; Broennimann et al. 2007; Hajkova et al. 2008; Holt 2003; Holt and Gaines 1992). There is, as yet, no synthesis of niche evolution nor, to my knowledge, any mechanistic approach to incorporate geno- and phenotypic plasticity into SDMs or spatial population models. There is, on the other hand, more than anecdotal evidence that microevolutionary processes are at play and matter ecologically (Hampe and Petit 2005; Phillips et al. 2006a). Hence, this field still awaits being embraced by species distribution models.

Concluding remarks

This chapter tries to strike a balance between guidance for novices to species distribution modeling – by providing a recipe for the most crucial elements of SDMs – and an embedding of SDMs into the currently most relevant statistical challenges. Analyzing a species' distribution can be a very useful starting point for further investigations or process-based modeling attempts. The correlative nature of modelling in general, and species distribution modelling specifically, should always be remembered. Tempting as it may be to incorporate a lot of ecological knowledge into mechanistic or statistical models, only little of this information will actually be relevant *at the focal scale*. The main intellectual challenge remains to not over-interpret one's findings and to seek independent corroboration.

Acknowledgments Over the years, many colleagues helped developing the above recipe. I am particularly grateful to Boris Schröder, Björn Reineking and Jane Elith, as well as the many participants of statistical workshops on this topic. I am also grateful to Fred Jopp, Hauke Reuters and Dietmar Kraft for improving a previous version. Funding by the Helmholtz Association is acknowledged (VH-NG-247).

References

Ackerly DD, Schwilk DW, Webb CO (2006) Niche evolution and adaptive radiation: Testing the order of trait divergence. Ecology 87:S50-S61

Araújo MB, Guisan A (2006) Five (or so) challenges for species distribution modelling. Journal of Biogeography 33:1677-1688

Austin MP (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling 157:101-118

Battisti A, Stastny M, Buffo E, Larsson S (2006) A rapid altitudinal range expansion in the pine processionary moth produced by the 2003 climatic anomaly. Global Change Biology 12:662-671

Beale CM, Lennon JJ, Yearsley JM, Brewer MJ, Elston DA (2010) Regression analysis of spatial data. Ecology Letters 13:246-264

Bellman RE (1957) Dynamic Programming. Princeton University Press, Princeton, NJ

Betts MG et al. (2009) Comment on "Methods to account for spatial autocorrelation in the analysis of species distributional data: a review". Ecography DOI: 10.1111/j.1600-0587.2008.05562.x

Bini LM et al. (2009) Coefficient shifts in geographical ecology: an empirical evaluation of spatial and non-spatial regression. Ecograhpy 32:193-204

Bjorholm S, J.-C. Svenning, F. Skov, Balslev H (2005) Environmental and spatial controls of palm (Arecaceae) species richness across the Americas. Global Ecology & Biogeography 14:423-429

Bjornstad ON, Peltonen M, Liebhold AM, Baltensweiler W (2002) Waves of larch Budmoth Outbreaks in the European Alps. Science 298:1020-1023

Bolker BM (2008) Ecological Models and Data in R. Princeton University Press, Princeton, NJ

Borcard D, Legendre P (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecological Modelling 153:51-68

Broennimann O, Treier UA, Müller-Schärer H, Thuiller W, Peterson AT, Guisan A (2007) Evidence of climatic niche shift during biological invasion. Ecology Letters 10:701-709

Brotons L, Thuiller W, Araújo MB, Hirzel AH (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography 27:437-448

Carl G, Dormann CF, Kühn I (2008) A wavelet-based method to remove spatial autocorrelation in the analysis of species distributional data. Web Ecology 8:22-29

Dormann CF (2007a) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. Global Ecology & Biogeography 16:129-138

Dormann CF (2007b) Promising the future? Global change predictions of species distributions. Basic and Applied Ecology 8:387-397

Dormann CF (2009) Response to Comment on "Methods to account for spatial autocorrelation in the analysis of species distributional data: a review" Ecography 32:379-381

Dormann CF, Gruber B, Winter M, Herrmann D (2010) Evolution of the climate niche in European mammals? Biology Letters 6:229-232

Dormann CF et al. (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30:609-628

Duncan RP, Cassey P, Blackburn TM (2009) Do climate envelope models transfer? A manipulative test using dung beetle introductions. Proceedings of the Royal Society B: Biological Sciences 276:1449-1457

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, Leathwick J (2009a) Conservation prioritisation using species distribution models. In: Moilanen A, Wilson KA, Possingham HP (eds) Spatial Conservation Prioritization: Quantitative Methods and Computational Tools

Elith J, Leathwick JR (2009b) Species distribution models: Ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics 40:677-697

Fielding AH (2002) What are the appropriate characteristics of an accuracy measure? In: Scott JM et al. (eds) Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Washington, pp 271-280

Gaston KJ (2009) Geographic range limits of species. Proceedings of the Royal Society B: Biological Sciences 276:1391-1393

Graf RF, Bollmann K, Suter W, Bugmann H (2005) The importance of spatial scale in habitat models: capercaillie in the Swiss Alps. Landscape Ecology 20:703-717

Guisan A et al. (2006) Using niche-based models to improve the sampling of rare species. Conservation Biology 20:501-511

Guisan A, Graham CH, Elith J, Huettmann F, the NCEAS Species Distribution Modelling Group (2007) Sensitivity of predictive species distribution models to change in grain size. Diversity and Distributions 13:332-340

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

Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. Ecological Modelling 135:147-186

Hajkova P, Hajek M, Apostolova I, Zeleny D, Dite D (2008) Shifts in the ecological behaviour of plant species between two distant regions: evidence from the base richness gradient in mires. Journal Of Biogeography 35:282-294

Hampe A, Petit RJ (2005) Conserving biodiversity under climate change: the rear edge matters. Ecology Letters 8:461-467

Harrell FE, Jr. (2001) Regression Modeling Strategies - with Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer, New York

Hastie T, Tibshirani R, Friedman JH (2008) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, Berlin

Herborg LM, Rudnick DA, Siliang Y, Lodge DM, MacIsaac HJ (2007) Predicting the range of Chinese mitten crabs in Europe. Conservation Biology 21:1316-1323

Holt RD (2003) On the evolutionary ecology of species' ranges. Evolutionary Ecology Research 5:159-178

Holt RD, Barfield M (2009) Trophic interactions and range limits: the diverse roles of predation. Proceedings of the Royal Society B: Biological Sciences 276:1435-1442

Holt RD, Gaines MS (1992) Analysis of adaptation in heterogenous landscapes: implications for the evolution of fundamental niches. Evolutionary Ecology 6:433-447

Jetz W, Rahbek C (2002) Geographic range size and determinants of avian species richness. Science 297:1548-1551

Jiménez-Valverde A, Lobo JM, Hortal J (2008) Not as good as they seem: the importance of concepts in species distribution modelling. Diversity and Distributions 14:885-890

Johst K, Brandl R, Eber S (2002) Metapopulation persistence in dynamic landscapes: the role of dispersal distance. Oikos 98:263-270

Kearney M (2006) Habitat, environment and niche: what are we modelling? Oikos 115:186-191

King AW, With KA (2002) Dispersal success on spatially structured landscapes: when do spatial pattern and dispersal behavior rally matter? Ecological Modelling 147:23-39

Krebs CJ (2002) Ecology: The Experimental Analysis of Distribution and Abundance 5th edn. Benjamin-Cummings, Zug (CH)

Kühn I, Bierman SM, Durka W, Klotz S (2006) Relating geographical variation in pollination types to environmental and spatial factors using novel statistical methods. New Phytologist 72:127-139

Latimer AM, Wu S, Gelfand AE, Silander JA (2006) Building statistical models to analyze species distributions. Ecological Applications 16:33-50

Lavorel S, Davies I, Noble I (2000) LAMOS: a landscape modelling shell. In: Hawkes BC, Flannigan MD (eds) Landscape Fire Modeling Workshop, pp 25-28

Legendre P (1993) Spatial autocorrelation: trouble or new paradigm? Ecology 74:1659-1673

Lennon JJ, Greenwood JJD, Turner JRG (2000) Bird diversity and environmental gradients in Britain: a test of the species-energy hypothesis. Journal of Animal Ecology 69:581-598

Lischke H, Zimmermann NE, Bolliger J, Rickebusch S, Löffler TJ (2006) TreeMig: A forest-landscape model for simulating spatio-temporal patterns from stand to landscape scale. Ecological Modelling 199:409–420

Lobo JM, Jimenez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. Global Ecology And Biogeography 17:145-151

Losos JB (2008) Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. Ecology Letters 11:995-1003

Mackey BG, Lindenmayer DB (2001) Towards a hierarchical framework for modelling the spatial distribution of animals. Journal of Biogeography 28:1147-1166

Maggini R, Lehmann A, Zimmermann NE, Guisan A (2006) Improving generalized regression analysis for the spatial prediction of forest communities. Journal of Biogeography 33:1729-1749

Mahecha MD, Schmidtlein S (2008) Revealing biogeographical patterns by nonlinear ordinations and derived anisotropic spatial filters. Global Ecology And Biogeography 17:284-296

Meynard CN, Quinn JF (2007) Predicting species distributions: a critical comparison of the most common statistical models using artificial species. Journal of Biogeography 34:1455-1469

Midgley GF, Hughes GO, Thuiller W, Rebelo AG (2006) Migration rate limitations on climate change-induced range shifts in Cape Proteaceae. Diversity and Distributions 12:555–562

Nekola JC (1999) Paleorefugia and neorefugia: The influence of colonization history on community pattern and process. Ecology 80:2459-2473

Nogués-Bravo D (2009) Predicting the past distribution of species climatic niche. Global Ecology & Biogeography 18:521-531

Ozinga WA, Schaminée HJ, Bekker RM (2005) Predictability of plant species composition from environmental conditions is constrained by dispersal limitation. Oikos:555-561

Pearce JL, Boyce MS (2006) Modelling distribution and abundance with presence-only data. Journal of Applied Ecology 43:405-412

Pearson RG (2007) Species' Distribution Modeling for Conservation Educators and Practitioners. Synthesis. American Museum of Natural History, http://ncep.amnh.org

Pearson RG, Dawson TP, Liu C (2004) Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. Ecography 27:285-298

Pearson RG et al. (2006) Model-based uncertainty in species range prediction. Journal of Biogeography 33:1704-1711

Pfenninger M, Nowak C, Magnin F (2007) Intraspecific range dynamics and niche evolution in Candidula land snail species. Biological Journal of the Linnean Society 90:303-317

Phillips BL, Brown GP, Webb JK, Shine R (2006a) Runaway toads: an invasive species evolves speed and thus spreads more rapidly through Australia. Nature 439:803

Phillips SJ, Anderson RP, Schapire RE (2006b) Maximum entropy modeling of species geographic distributions. Ecological Modelling 190:231-259

Phillips SJ et al. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19:181-197

Preston KL, Rotenberry JT, Redak RA, Allen MF (2008) Habitat shifts of endangered species under altered climate conditions: importance of biotic interactions. Global Change Biology 14:2501-2515

Quinn GP, Keough MJ (2002) Experimental Design and Data Analysis for Biologists. Cambridge Univ. Press, Cambridge

R Development Core Team (2009) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org

Randin CF, Dirnbock T, Dullinger S, Zimmermann NE, Zappa M, Guisan A (2006) Are niche-based species distribution models transferable in space? Journal of Biogeography 33:1689-1703

Rondinini C, Wilson KA, Boitani L, Grantham H, Possingham HP (2006) Tradeoffs of different types of species occurrence data for use in systematic conservation planning. Ecology Letters 9:1136-1145

Samu F, Sunderland KD, Szinetar C (1999) Scale-dependent dispersal and distribution patterns of spiders in agricultural systems: A review. Journal of Arachnology 27:325-332

Schröder B, Richter O (1999) Are habitat models transferable in space and time? Zeitschrift für Ökologie und Naturschutz 8:195-205

Schurr FM, Midgley GF, Rebelo AG, Reeves G, Poschlod P, Higgins SI (2007) Colonization and persistence ability explain the extent to which plant species fill their potential range. Global Ecology & Biogeography 16:449-459

Schweiger O, Settele J, Kudrna O, Klotz S, Kühn I (2008) Climate change can cause spatial mismatch of trophically interacting species. Ecology 89:3472-3479

Segurado P, Araújo MB (2004) An evaluation of methods for modelling species distributions. Journal of Biogeography 31:1555-1568

Soberón J (2007) Grinnellian and Eltonian niches and geographic distributions of species. Ecology Letters 10:1115-1123

Svenning JC, Skov F (2004) Limited filling of the potential range in European tree species. Ecology Letters 7:565-573

Thompson JN (1998) Rapid evolution as an ecological process. Trends in Ecology & Evolution 13:329-332

Thuiller W (2004) Patterns and uncertainties of species' range shifts under climate change. Global Change Biology 10:2020-2027

Thuiller W et al. (2008) Predicting global change impacts on plant species' distributions: Future challenges. Perspectives in Plant Ecology, Evolution and Systematics 9:137-152

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

Vaughan IP, Ormerod SJ (2005) The continuing challenges of testing species distribution models. Journal of Applied Ecology 42:720-730

Ward G, Hastie T, Barry S, Elith J, Leathwick JR (2009) Presence-only data and the EM algorithm. Biometrics 65:554-563

Zabel CJ, Dunk JR, Stauffer HB, Roberts LM, Mulder BS, Wright A (2003) Northern spotted owl habitat models for research and management application in California (USA). Ecological Applications 13:1027-1040

Zuo W, Lao N, Geng Y, Ma K (2008) GeoSVM: an efficient and effective tool to predict species' potential distribution. Journal of Plant Ecology 1:143-145

Zurell D, Jeltsch F, Dormann CF, Schröder B (2009) Static species distribution models in dynamically changing systems: how good can predictions really be? Ecography 32:733-744