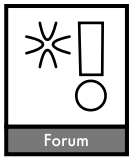

Ecography 31: 272–278, 2008
doi: 10.1111/j.2007.0906-7590.05378.x

Subject Editor: Miguel Araújo. Accepted 13 December 2007

Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson et al. (2007)

Steven J. Phillips

S. J. Phillips (phillips@research.att.com), AT&T Labs Research, 180 Park Avenue, Florham Park, NJ 07932, USA.

The August 2007 issue of Ecography featured the paper ''Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent'' (Peterson et al. 2007), which raised concerns about the ability of the Maxent species distribution modelling method (Phillips et al. 2006) to predict species' distributions in broad unsampled regions and suggested that GARP (Stockwell and Peters 1999) may be the ''winner'' at the ''transferability challenge''. As one of the developers of the Maxent software, I'm pleased to have the opportunity to respond, and I thank the authors for sharing the data from their study. Transferability is an extremely important topic in distribution modelling, and is required by some of the most important applications of modelling. I shall address three issues: first, it is essential to clarify the difference between transferability and the problem of sample selection bias, since the study of Peterson et al. (2007) is better characterised as a study of sample selection bias rather than transferability. Second, when comparing modelling approaches, it is important to include enough species or replicates for the evaluation to have sufficient statistical power to give significant results: despite suggestive patterns in the quantitative results, the three species used by Peterson et al. (2007) were insufficient for achieving statistical significance, so their conclusions were based on subjective visual assessments. Third, since Peterson et al. (2007) focus on the interpretability of Maxent's cumulative output format, I give some clarification of Maxent's output formats, and mention a new logistic output format that may be easier to interpret.

Transferability, climate change and invasive species

''Transferability'' (also called ''generalizability'' or ''generality'') concerns the ability of a model calibrated in one context to be ''transferred'', i.e. to make useful predictions in a different context. There have been a number studies investigating transferability of species distribution models (Araújo et al. 2005, Graf et al. 2006, Pearson et al. 2006, Randin et al. 2006). Here we focus on presence-only modelling methods, which are increasingly being used to predict effects of global climate change on species' potential distribution (Thomas et al. 2004) and for evaluating invasive potential of alien species (Peterson et al. 2003). Peterson et al. (2007) correctly observe that transferability is critical for both applications, and that a recent comprehensive comparison of presence-only methods (Elith et al. 2006) does not shed light on transferability.

Most presence-only models, including those compared by Peterson et al. (2007), are based on distinguishing known occurrence sites for a species from ''background'' data (also known as ''pseudo-absence data'') drawn from the study region. For these methods, transferral involves first creating a model using values of environmental variables at occurrence and background sites in one study region, then transferring the model by applying it to a second set of environmental variables with the same names, but describing a different region or time period (as in Peterson et al. 2003, Thomas et al. 2004). In contrast, Peterson et al. (2007) created models using background data from both the training region and the evaluation region combined. For the climate change application, this would be like mixing environmental data describing both current and future climate conditions while attempting to develop a model of a species' current distribution. This is invalid, as future climate conditions have no bearing on current distributions. Similarly, when predicting an invasive species' potential to invade an as-yet uninvaded region, its niche should be modeled using background data taken only from the native range of the species (as in Peterson et al. 2003), since environmental conditions in regions that the species has not had an opportunity to invade give no clue as to its environmental preferences.

In their experimental design, Peterson et al. (2007) split occurrence data for three North American bird species into two subsets: ''on-diagonal'' (NW and SE quadrants of North America) and ''off-diagonal'' (NE and SW quadrants). They trained models using one subset of the occurrence data, and evaluated the predictions on the other;

however, in all cases background data covers all of North America. There is no transferral involved, as model calibration and model evaluation use exactly the same environmental data. Rather, this is an extreme case of geographic bias in occurrence data. Sample selection bias is an important topic: occurrence data is typically strongly biased towards more accessible areas (Reddy and Dávalos 2003), and such bias can dramatically lower predictive performance of presence-only models (Phillips et al. unpubl.). However, sample selection bias and transferability are different topics, and either one (or both) may be relevant to any particular application of species distribution modelling. It is important to be able to distinguish between the two in order to create the best possible models for any application.

In order for their study to be relevant to the climate change and invasive species applications (and transferability in general), Peterson et al. (2007) should have restricted the background data to the same subset of North America as the presence data. They mention that they did try experimenting with this: ''we also experimented with training Maxent models ... based only on the quadrants used to build models, and then projected them to the entire region ... [but] the general picture of close overfitting to the training region was still observed.'' However, I repeated this comparison, training Maxent models on the same data for one of their species, *Zenaida macroura* (presence data shown in Fig. 1c), using on-diagonal presence data and both on-diagonal background (Fig. 1a) and background covering all of North America (Fig. 1b), and obtain very different results. In particular, the training region is not as readily apparent in (a), and the model transferred to the off-diagonal areas appears to better match the true presences in (a) than in (b). I conclude that the general picture does change with the change in background, but I make no subjective judgement about whether the predictions exhibit ''close overfitting''.

Sample selection bias and discovery of new populations and species

So far we have focused on climate change and invasive species. Peterson et al. (2007) motivate their study with a third application: the use of species distribution models for discovery of new populations and species. However, it isn't clear how transferability is relevant to this application. It could be relevant, for example, if a model developed in one area were projected to another (previously unsampled) area to guide field surveys. However, the more common approach for using species distribution modelling for discovering new populations or species involves creating a model from known occurrences or presence-absence data in a region, then investigating areas in the same region that the model predicts to be suitable (Bourg et al. 2005); when using presence-only data, this means that background data is shared between training and prediction (Raxworthy et al. 2004, Pearson et al. 2007). For brevity, I will refer to the latter approach as simply ''discovery''.

Since the environmental data used to generate the model is the same as the data used in the prediction, discovery does not involve transferral, so it does not require models to be transferable. In contrast, discovery depends heavily on a model's ability to make accurate predictions in those areas that are most affected by sample selection bias. Sample selection bias can impact discovery in multiple ways. First, there may be broad unsampled regions in the study area: this is referred to by Peterson et al. (2007) as the ''challenge of transferability'', but actually succintly defines geographic bias in sampling. This kind of bias is a much greater problem for presence-only models than for presence-absence models, since the bias afflicts presence data but not background data (Phillips et al. unpubl.). A second, more insidious kind of bias primarily impacts species discovery (as opposed to new population discovery), and arises from a disparity between the distribution we seek to model and the data we have available. We seek the distribution of suitable conditions for a species, i.e. the species' potential distribution, while we have only samples from occupied sites, i.e. from its realized distribution. For species discovery, we are most interested in cases where some areas of the potential distribution are occupied not by the species being modelled, but by related species. Sites drawn from the realized distribution thus form a geographically biased sample from the potential distribution. Note that this second form of bias may also impact the climate change and invasive species applications: in both cases, we wish to transfer models of the potential distribution, but models are calibrated using samples from the realized distribution.

Experimenting with sample selection bias

I have argued that the experiment of Peterson et al. (2007) concerns the issue of sample selection bias, rather than transferability, and because of its treatment of background data, it is directly relevant to only one of their motivating applications, new population and species discovery. The experiment has the potential to give useful insight into the effect of sample selection bias, and I would like to suggest four ways in which it could have greater impact.

1) Including a larger number of species in the study would allow for statistically significant comparisons. Model predictivity as measured using AUC scores is lower for GARP in five out of six cases (Table 1 of Peterson et al. 2007); nevertheless, the lack of statistical power made the comparison statistically signicant only for a one-tailed test, not for a two-tailed test, leading Peterson et al. (2007) to conclude that ''GARP and Maxent performed similarly when AUC statistics were examined.''

2) The species were chosen for having broad geographic distributions in the study region (among other factors). This constrains the generality of the study, limiting the conclusions to species with similarly broad distributions. Using a wider selection of species would allow for more general conclusions. In particular, since new population and species discovery is a motivating application, the experiment should include some narrow-range or rare species, i.e. species for which researchers are more likely to want to search for new populations or undiscovered related species.

3) Visual interpretation of predictions is highly subjective. It is therefore essential to have clear criteria for visual evaluation, to avoid being biased by one's prior expectations. For example, when trained on all the data, both

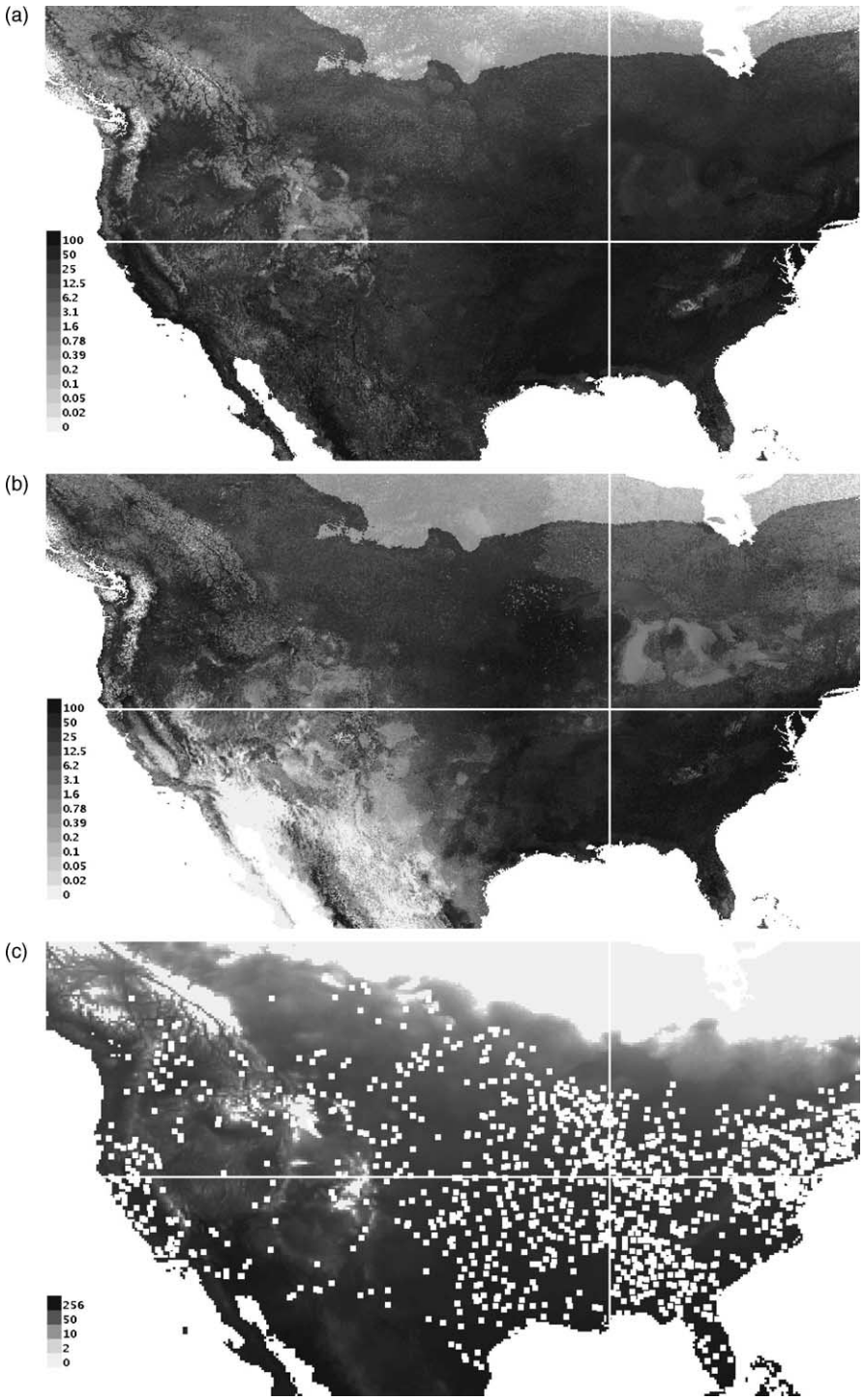

Fig. 1. Transferability versus sample selection bias. Maxent models for *Zenaida macroura* were trained with presence points restricted to NW and SE quadrants, as defined by the white crosshairs. Background data was drawn either from the same quadrants (a) or from the whole map (b). In (a), the model trained on the NW and SE quadrants has been projected onto the whole continent. In (b), the model did not need to be projected; instead, the model has been trained with biased presence data. Presence data is shown as white dots in (c), with the background colored according to annual average temperature (in Celsius, times 10; negative values shown as 0). Presence data derive from the North American Breeding Bird Survey, which covers Canada and the United States, so lack of presences in the north represents likely absence, while presence records are missing in the south because the survey excludes Mexico.

GARP and Maxent produced models for Zenaida macroura that show a dramatic area of over-prediction in the north, extending through Quebec and Ontario up to James Bay and further north-east, hundreds of miles from any recorded presences (see Fig. 2 of Peterson et al. 2007; my Fig. 1a shows the same overprediction). This is true over-prediction, not an artifact of sparse sampling in the north: Fig. 1 of Peterson et al. (2007) shows a band of

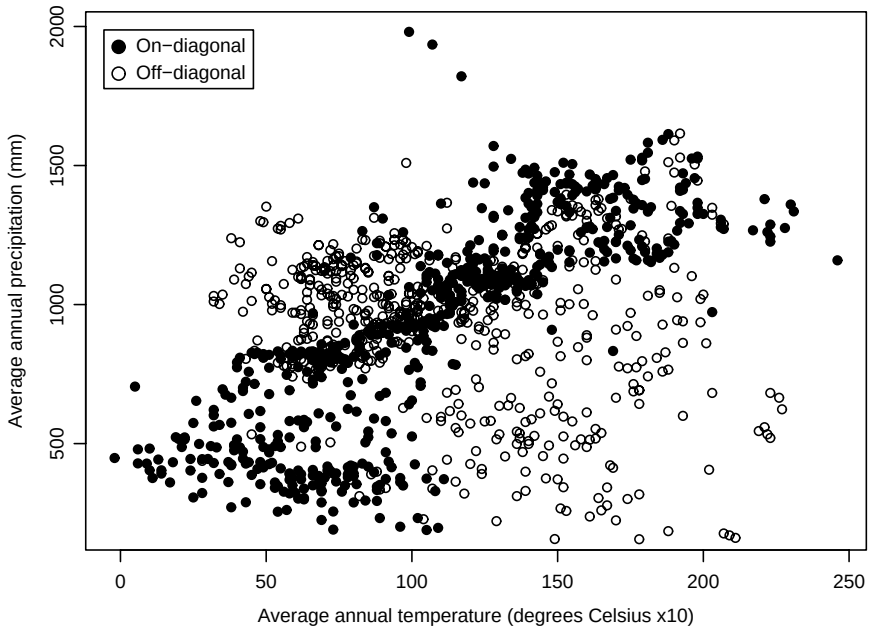

Fig. 2. Environmental space occupied by subsets of occurrence data for *Zenaida macroura.* On-diagonal occurrence points are generally cool and fairly dry (those drawn from the NW quadrant, mostly in the upper midwestern United States) or fairly warm and wet (drawn from the SE quadrant, i.e. the southeastern United States). On-diagonal occurrence data therefore clusters in the bottom left and top right of the pictured environmental space. The off-diagonal occurrence data, in contrast, contain many warm and dry points (drawn from the southwestern United States) and many cool and moist points (drawn from the northeastern United States and southeastern Canada).

absence records above the northernmost recorded presences. For this species, therefore, I would not agree with the visual evaluation of Peterson et al. (2007), that ''both algorithms produced maps that coincided well with the known distribution of the species.'' (Investigating this issue further, we see that the broad area of over-prediction has annual mean temperature values that are well below that of any presence points (Fig. 1c). Maxent models that I created either using bioclimatic variables directly, or just using annual mean temperature, did not suffer from this over-prediction, and GARP models would likely be similarly improved. The over-prediction may be an artefact of Peterson et al. (2007) using variables derived from principal-components analysis, rather than using the bioclimatic and other variables directly. While the effect of variable transformations is an interesting topic, I leave it for future study.)

4) Peterson et al. (2007) suggest that a good model will not be affected by sample selection bias, and that despite biased training data, a model should be preferred if it manages to ''reconstruct much of the species' known distributions''. At first this appears to be a worthy goal: it seems preferable for a modeling algorithm to be robust to sample selection bias. Unfortunately, this may be requiring the algorithm to be preternaturally smart, or not smart at all. In Fig. 2, I have plotted the on-diagonal and off-diagonal occurrences for *Zenaida macroura* in environmental space, defined by average annual precipitation and average annual temperature. I ask the reader, do these two sets of occurrences represent the same ecological niche? Should we seek an algorithm that predicts the same distribution for these two ''species''? The on-diagonal occurrences fill only a small subset of the environmental space occupied by the species, and in particular, they are missing from all warm, dry areas (in the bottom right of the figure). Based on the on-diagonal occurrences, GARP strongly predicts suitable conditions throughout North America (Peterson et al. 2007, Fig. 2). Has it managed to infer that the species tolerates warm, dry conditions, despite being completely missing from such conditions in the training data? Or does it manage to predict into warm, dry conditions simply because it dramatically overpredicts? In this case, GARP has managed to ''fill in'' the environmental space for *Zenaida macroura,* but it is important to note that this overprediction (relative to the training data) would be incorrect for many species.

This fourth point deserves further discussion. All modelling methods that use background data attempt to differentiate the environmental conditions at the presence points from environmental conditions at the background points. If presence points (but not background points) are excluded from a large portion of the study area containing an identifiable subset of environmental space, all such modelling methods will be ''fooled'', as the environmental conditions in that portion of the study area differentiate presence from background. In this way, all presence-background modelling methods (including GARP, Maxent and most other species distribution modelling methods) will necessarily be influenced by sample selection bias. For most of these methods, biased presence data can be used to make accurate estimates of species distributions as long as the bias is known or can be estimated (Phillips et al. unpubl.). If the bias simply restricts presence data to a portion of the study area (as in Peterson et al. 2007, or if only a portion has been surveyed, or when a geographic barrier has prevented a species from dispersing), then the bias can be avoided by simply restricting the study area to that portion; better predictions may be obtained when the

model is then transferred to the rest of the study area (Raza and Anderson pers. comm.).

On the other hand, the sample selection bias may not be known; I see the experiment of Peterson et al. (2007) as modelling this situation. Their results suggest that if there is strong but unknown sample selection bias, then it is unwise to rely on the fine details of a prediction given by a statistical modelling method such as Maxent. This is not controversial: the bias violates the most basic statistical assumption of the method, that the presence data are sampled independently according to the species' distribution in the study area, i.e. without any bias (Phillips et al. 2006). The violated assumption cannot necessarily be fixed by simply adjusting a parameter such as the regularization multiplier, as Peterson et al. (2007) note. However, not using the statistical methods for any applications involving transferability would be like throwing the baby out with the bathwater. Appropriate use and interpretation of these methods can still be beneficial, despite difficulties with data quality. Indeed, Peterson et al. (2007) find that when thresholded appropriately, the Maxent predictions ''closely resembled'' predictions that ''reconstruct much of the species' known distributions''. Rather than concluding that ''these results place some bounds on the applicability of Maxent in ecological niche modeling'', I would suggest that their results argue for using caution when applying any modelling method in situations where the underlying assumptions on the training data are significantly violated. Similar conclusions can be derived from the study of Elith et al. (2006), where the bird data from Ontario is heavily biased, resulting in poor overall predictive performance for many methods (including GARP, which ranked near the bottom for Ontario in Fig. 5 and Table 7 of Elith et al. 2006). The best single-species modelling method for that region was a simple method, BIOCLIM (Busby 1991), that does not use background data, and hence relies on fewer statistical assumptions. However, when bias in the Ontario bird data is properly accounted for, and sampling assumptions are no longer violated, the statistical methods again outperform BIOCLIM (Phillips et al. unpubl.).

Clarification of Maxent output formats

Peterson et al. (2007) base their conclusions on visual representations of Maxent outputs. Their conclusions depend heavily on the particular output format and linear color ramp used to present the outputs, which led them to focus on fine distinctions between the mostly highly-predicted areas. I take this opportunity to clarify certain aspects of Maxent's output. In particular, I outline here why the cumulative output is usually better pictured using a logarithmic scale, as is done by the Maxent software (Phillips et al. 2005), and I mention a new logistic output format, introduced after Peterson et al. (2007), that may be easier to interpret and is naturally displayed with a linear scale.

The primary output of Maxent is an exponential function that describes a probability distribution defined over the set of background points used during model training. The probabilities from that probability distribution (referred to as ''raw'' values) are not intuitive to work with: although they are proportional to modeled conditional probability of presence, it is impossible to determine the constant of proportionality from presence-only data. (Conditional probability of presence is the most intuitive output that we could hope for: it says how likely the species is to be present, given a particular set of environmental conditions.) Raw values are also scale-dependent – using more background points results in smaller raw values. For these reasons, we convert raw values into more easily-used formats. The format used by Peterson et al. (2007) is the cumulative format (Phillips et al. 2006), which was the default format until very recently. It is defined in terms of omission rates: if a point p has cumulative value c, it means that when we set a threshold at p's raw value, c% of the Maxent distribution will have raw values below the threshold. Therefore, if test data are drawn from the Maxent distribution, the predicted omission rate is c%; however, under conditions of strong sample selection bias, we can expect significant deviations from predicted omission rate, as observed by Peterson et al. (2007).

The cumulative format solves the scale-dependence issue, but cumulative values (unlike raw values) are not interpretable in terms of probability of presence. This can be demonstrated by a thought experiment: consider an extreme generalist species whose probability of presence is almost constant across the study area, with miniscule variations ensuring that there are no ties (i.e. that no two sites have exactly the same probability of presence). Cumulative values for the distribution of this species range evenly from 0 to 100 across the study area, since if sites are ranked by probability of presence, the cumulative value for a site is proportional to its rank. Thus, big variations in cumulative value do not necessarily represent big variations in suitability. A side effect of this, noted by Peterson et al. (2007), is that differences between small cumulative values are more important than between large values – for most applications, we would prefer a threshold with low omission, so small values are most important. For this reason, cumulative values should usually be pictured using a logarithmic scale (as in Fig. 1, and by default in output of the Maxent software). Peterson et al. (2007) used a linear scale (as in Phillips et al. 2006), which emphasizes differences between larger values, and is more appropriate when fine distinctions between strongly-predicted areas are relevant (for example, for reserve-design applications). We note also that because small values are important, cumulative output should not be discretized (converted to integers in the range 0–100) as done by Peterson et al. (2007) – this discretizing loses important information at the bottom end of the range, and is the reason why ROC curves for Maxent are truncated at the right-hand end in their Fig. 5.

Given the above concerns about interpreting cumulative values, it is worth noting that ver. 3.0 of the Maxent software has a new default output format, the logistic format, that addresses these concerns: it transforms the exponential function mentioned above into a logistic function representing (with some caveats) probability of presence. It is scale-independent, calibrated so that typical presence points yield a value of 0.5 on a scale of 0 to 1, and pictured by default using a linear scale. More details will be given in a future publication (Phillips and Dudík 2008).

Discussion and conclusions

Peterson et al. (2007) observe that the comparison of Elith et al. (2006) concerns models of species' present geographic distributions, and that there is currently no similarly extensive study to offer guidance for applications requiring transferral. They make the important point that modelling the ecological niche of a species is not necessarily the same as modelling its geographic distribution in a particular study region. Indeed, the relationship between niche and distribution can be quite complex (Pulliam 2000). In general, the geographic distribution of a species may be influenced by such factors as dispersal ability, evolutionary history, and biotic interactions, which are not represented in the definition of its ecological niche. Therefore, a good model of a species' niche may not result in a good model of its current distribution, and vice versa.

The invasive species, climate change and species discovery applications that motivated the study of Peterson et al. (2007) all require models of niches, not geographic distributions. For such applications, we are asking how likely a site is to be suitable for the species, for given environmental conditions. In statistical terms, the response variable being modeled is whether a site is suitable, not whether it is occupied. Therefore, a model of a niche requires a sample of suitable sites as training data (rather than a set of currently occupied sites), and to satisfy the assumptions of statistical modelling methods, that sample should be chosen uniformly at random from the set of all suitable sites. (For simplicity, we ignore the possibility of varying degrees of suitability.) Most often, the available data consists only of currently occupied sites, which, because of the historical and biotic factors mentioned above, may be a biased sample from the set of all suitable sites. When modelling niches, therefore, one must take special care to understand and limit the impact of sample selection bias. When the bias is strong and its details unknown, a simple modelling method may be appropriate (such as BIOCLIM, see above) to avoid making unfounded assumptions about the training data. A better approach is to apply our ecological knowledge of factors such as dispersal limitations, geographic barriers and biotic interactions in order to reduce or account for the bias, so we can successfully use more sophisticated models.

It is important to note that sample selection bias is not the only potential pitfall when modelling niches: the environmental variables must also be chosen with care. Not only are some variable sets less suitable for modelling ecological niches within a region (witness the over-prediction in both GARP and Maxent models of *Zenaida macroura,* perhaps due to the transformation of variables using principle components), but some variables also make a niche model less transferable. Good examples are elevation and aspect, included in the study of Peterson et al. (2007), which are surrogates for variables that have direct biophysical impact on the species being modeled. However, correlations with those variables vary over space and time, making the surrogates unsuitable for climate change and invasive species applications (Peterson 2006). Transferring a model also introduces the potential pitfall of extrapolation, which refers to a model being applied to environmental conditions outside the range on which it was calibrated (Thuiller et al. 2004).

In conclusion, transferability and sample selection bias are exciting and important topics, especially since both are critically important for prediction of species distributional shifts under climate change. I hope that the current discussion leads to improvements in our ability to make species distributions models that are accurate both in the context in which they are trained, and when transferred to contexts that are different from those used during model development.

Acknowledgements – I thank Town Peterson, Rob Anderson, Migúel Araujo, Catherine Graham, Miro Dudík, Jane Elith, Richard Pearson and an anonymous reviewer for helpful comments and suggestions.

References

Araújo, M. B. et al. 2005. Validation of species-climate impact models under climate change. – Global Change Biol. 11: 1504–1513.

Bourg, N. A. et al. 2005. Putting a CART before the search: successful habitat prediction for a rare forest herb. – Ecology 86: 2793–2804.

Busby, J. R. 1991. BIOCLIM – a bioclimate analysis and prediction system. – In: Austin, M. P. and Margules, C. R. (eds), Nature conservation: cost effective biological surveys and data analysis. CSIRO, Melbourne, pp. 64–68.

Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.

Graf, R. F. et al. 2006. On the generality of habitat distribution models: a case study of capercaillie in three Swiss regions. – Ecography 29: 319–328.

Pearson, R. G. et al. 2006. Model-based uncertainty in species range prediction. – J. Biogeogr. 33: 1704–1711.

Pearson, R. G. et al. 2007. Predicting species' distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. – J. Biogeogr. 34: 102–117.

Peterson, A. T. 2006. Uses and requirements of ecological niche models and related distributional models. – Biodiv. Inform. 3: 59–72.

Peterson, A. T. et al. 2003. Predicting the potential invasive distributions of four alien plant species in North America. – Weed Sci. 51: 863–868.

Peterson, A. T. et al. 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. – Ecography 30: 550–560.

Phillips, S. and Dudík, M. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. – Ecography, in press.

Phillips, S. J. et al. 2005. Maxent software for species distribution modeling. – <http://www.cs.princeton.edu/~schapire/maxent/>.

Phillips, S. J. et al. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Model. 190: 231–259.

Pulliam, H. R. 2000. On the relationship between niche and distribution. – Ecol. Lett. 3: 349–361.

Randin, C. F. et al. 2006. Are niche-based species distribution models transferable in space? – J. Biogeogr. 33: 1689–1703.

Raxworthy, C. J. et al. 2004. Predicting distributions of known and unknown reptile species in Madagascar. – Nature 426: 837–841.

Reddy, S. and Dávalos, L. M. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. – J. Biogeogr. 30: 1719–1727.
Stockwell, D. and Peters, D. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. – Int. J. Geogr. Inform. Sci. 13: 143–158.
Thomas, C. D. et al. 2004. Extinction risk from climate change. – Nature 427: 145–148.
Thuiller, W. et al. 2004. Effects of restricting environmental range of data to project current and future species distributions. – Ecography 27: 165–172.