12-2012

Testing the Predictive Performance of Distribution Models

Volker Bahn
Wright State University, volker.bahn@wright.edu

Brian McGill
University of Maine, brian.mcgill@maine.edu

Testing the predictive performance of distribution models

Volker Bahn and Brian J. McGill

V. Bahn (volker.bahn@wright.edu), Dept of Biological Sciences, Wright State Univ., 3640 Colonel Glenn Highway, Dayton, OH 45435, USA.

B. J. McGill, School Biology and Ecology and Sustainability Solutions Initiative, Univ. of Maine, Orono, ME 04469, USA.

Abstract

Distribution models are used to predict the likelihood of occurrence or abundance of a species at locations where census data are not available. An integral part of modelling is the testing of model performance. We compared different schemes and measures for testing model performance using 79 species from the North American Breeding Bird Survey. The four testing schemes we compared featured increasing independence between test and training data: resubstitution, random data hold-out and two spatially segregated data hold-out designs. The different testing measures also addressed different levels of information content in the dependent variable: regression R^2 for absolute abundance, squared correlation coefficient r^2 for relative abundance and AUC/Somer's D for presence/absence. We found that higher levels of independence between test and training data lead to lower assessments of prediction accuracy. Even for data collected independently, spatial autocorrelation leads to dependence between random hold-out test data and training data, and thus to inflated measures of model performance. While there is a general awareness of the importance of autocorrelation to model building and hypothesis testing, its consequences via violation of independence between training and testing data have not been addressed systematically and comprehensively before. Furthermore, increasing information content (from correctly classifying presence/absence, to predicting relative abundance, to predicting absolute abundance) leads to decreasing predictive performance. The current tests for presence/absence distribution models are typically overly optimistic because a) the test and training data are not independent and b) the correct classification of presence/absence has a relatively low information content and thus capability to address ecological and conservation questions compared to a prediction of abundance.

Meaningful evaluation of model performance requires testing on spatially independent data, if the intended application of the model is to predict into new geographic or climatic space, which arguably is the case for most applications of distribution models.

Distribution models are used to predict the occurrence or abundance of a species at locations that have not been censused. Known occurrences or abundances are used as dependent variables to be explained by various environmental variables (e.g. climate, land cover). Unlike information about the distribution of species, we have extensive environmental data that has been interpolated to cover the entire globe. This allows distribution models to make predictions about the occurrence or abundance of a species at uncensused locations. More recently, distribution models built on present day environmental data have been used to predict future species distributions based on projections of future climates from global circulation models (Iverson and Prasad 1998, Skov and Svenning 2004). Distribution modelling has seen an unprecedented amount of attention in recent years (e.g. two special sections in journals in 2006: J. Appl. Ecol. 43 and J. Biogeogr. 33). It is a vital tool in species conservation and land management (Scott and Csuti 1997, Ferrier 2002) and relates to the most fundamental questions of ecology: the abundance and distribution of species (Krebs 1972, Andrewartha and Birch 1984).

Many authors have noted that distribution models omit ecological processes that are known to be important, such as species interactions and dispersal (Davis et al. 1998, Araújo et al. 2005a, Randin et al. 2006). But as Box (1976) noted ‘All models are wrong, but some are useful’, which underscores the crucial importance of an objective, quantitative test of model performance rather than merely noting their many shortcomings. Objective tests of predictions 100 years in the

future are clearly impossible. However, even tests of the predictive success of distribution models in the present day suffer from a confusing array of different approaches and an unclear relationship of performance to complicating factors such as presence/absence (P/A) versus abundance (Ab) modelling and spatial autocorrelation. The goal of our paper is to clarify how the testing of distribution models is influenced by 1) the information content of the dependent variable (in increasing order: presence/absence (P/A) versus, relative abundance (rel.Ab) versus, absolute abundance (Ab)); and 2) the relative independence of the testing data used (in increasing order: resubstitution versus, random hold-out versus, and truly independent (spatially segregated split)).

Several aspects of input data and testing scheme potentially influence the outcome of a model test. For example, the baseline probability for guessing correctly is very different for P/A versus Ab data. If a species has a prevalence of 50% of the cells of a study area then the probability to guess P/A correctly without any further knowledge is 50%. In contrast, the probability of guessing the abundance of a species in a cell correctly would be nearly 0%, if abundance was a truly continuous variable. This ease of prediction is related to the information content of the measure that is to be predicted. Ecologically, knowing the presence or absence of an organism tells us relatively little about the suitability of a location, as a presence could be from a sink population or equally as well a critically important source population (Pulliam 1988, 2000). Relative abundance can at least provide information on the relative suitability of different habitats, while abundance, the most information rich and difficult to predict measure, is linked to survival probability (Dennis et al. 1991).

Different schemes for model testing exist. The most basic scheme, resubstitution, is to judge the goodness-of-fit of a model using the same data the model was fit for (Fielding and Bell 1997). This scheme can inform the researcher about the ability of the chosen model to describe the given data, but it says little about the generality or transferability of the model (Phillips 2008), i.e. if the model can successfully predict onto new, independent data. This scheme is used in basic regression analyses, when a straight line or low-polynomial equation hardly permits overfitting and is often used when scarce data do not allow for a better testing scheme. A more advanced scheme is to randomly hold-out data from the model building process and to test the predictions of the fitted model on this held-out test data (Fig. 1) that was not used in the model building and parameterizing process. This scheme can consist of holding out a fixed percentage of data only once, or holding out a small percentage of the data for testing but rotating the random selection and thus repeating the testing process several times, until each data point was in a test set once, and then averaging over the multiple test results (cross-validation). Splitting the data into training and test data not only gives an impression of the goodness-of-fit of the model to the data but also of its capability to predict onto new data and therefore its generality and transferability. Finally, one can use a hold-out that is spatially segregated. In contrast to the random hold-out, a spatially segregated hold-out prevents spatial intermingling of training and test datasets and thus makes it more likely that the two datasets are truly independent (Peterson et al. 2007).

If the data have no spatial autocorrelation across the modelling extent then there is no difference between the last two approaches (random holdout versus spatially segregated holdout). However, environmental data and distribution data are virtually always autocorrelated in space, which adds

a further complication to the model testing process. Spatial autocorrelation means that, on average, things closer together are more similar than things further apart, resulting in a dependence among locations that decays with distance. Autocorrelation is found in both the independent variables (here the environmental variables) and the dependent variable (the species distribution).

Spatial autocorrelation in dependent and independent variables may not only be a potential violation of many models' assumption that input data and/or error terms be independent (Legendre and Fortin 1989), but may also lead to inflated test measures (Segurado et al. 2006). Consequences of spatial autocorrelation, such as an overestimation of degrees of freedom, a resulting underestimation of variance and overestimation of significance, as well as an influence on variable selection, have been investigated in detail (Legendre 1993, Lennon 2000, Dale and Fortin 2002). However, the consequence of autocorrelation to a test of the predictive power of a model based on data hold-out techniques has received less attention. Several authors have identified this problem (Hampe 2004, Araújo and Guisan 2006, Araújo and Rahbek 2006) and consequently many studies have been conducted testing models on allegedly independent data.

Three categories of 'independent' testing data are typically employed: 1) independently collected data (Elith et al. 2006), 2) temporally independent data (Martinez-Meyer et al. 2004), and 3) spatially independent data (Peterson et al. 2007). We will focus on spatially independent testing data (3), because independently collected data (1) introduce additional variability by potentially using different methods and/or censusing the organism during a different time with different population levels while still not guaranteeing spatial independence. And using temporally

segregated testing data (2) suffers from the same drawbacks as using independently collected data: temporal autocorrelation potentially leads to dependence between training and test data leading to overly optimistic model evaluations, while natural population fluctuations may lead to an overly pessimistic model evaluation.

Searching the literature, we identified 32 studies using spatially independent data for model evaluation. This compilation represents all such studies we could find, but we do not claim it is exhaustive. Given that most authors did not test for spatial independence, we inferred such independence when training and test data were reasonably (dependent on the focal organism) spatially separated. The results of the studies were varied, but more importantly, the interpretation of the studies was varied. For example, Graf et al. (2006) interpreted AUC values between 0.83–0.94 achieved when predicting occurrence in a new area as poor to moderate performance, while (Murray et al. 2011) labelled models challenged with the same task and achieving AUC values of 0.77–0.90 as having excellent discriminative abilities. All but one of the identified studies (Whittingham et al. 2003), used presence/absence or presence only, leaving the effect of autocorrelation on the evaluation of abundance-based models virtually unexplored in the literature. Twenty-three of the 32 studies used fewer than 10 species, with only one based on abundance data (Whittingham et al. 2003), using only one species. The biggest challenge in synthesizing this literature is that studies typically only reported a measure of performance for prediction into a new area. As such, it is hard to say whether the distribution models predicting to the new area performed well or not. For example, is an R^2 of 0.50 or an AUC of 0.7 when predicting into a new area good or bad? Addressing these questions requires a more inferentially systematic way that allows for attributing drops in performance to different contributing factors.

Thus, we believe it is important to implement an evaluation scheme spanning the typically used evaluation methods (resubstitution, random hold-out or CV, or spatial split) in a single study, so that a decrease in the determined performance can be seen relative to the original, goodness-of-fit based estimate of model performance. Equally, we think it is important to run such comparisons across many species and in spatially different areas. None of the investigated studies provides such a complete comparison.

In this paper we systematically and comprehensively investigate the influence of input data (P/A versus Ab), and model testing scheme (resubstitution versus random reserved data design versus spatially segregated data design) on performance tests of distribution models, and show how the testing scheme needs to be matched to the intended purpose of the model to prevent overly optimistic results.

Material and methods

Data sources

We used data from the North American Breeding Bird Survey. This survey was initiated in 1966 and has been conducted yearly since by skilled volunteers under the auspice of the Canadian Wildlife Service and the US Geological Survey's (USGS) Patuxent Wildlife Research Center. Surveys are conducted during the breeding season (mid-May through the first week of July, depending on the latitude of the route) at stops along routes placed on secondary roads. Each of the over 4100 survey routes in the USA and southern Canada is approximately 40 km (exactly

24.5 miles) long and contains fifty regularly spaced (every 0.8 km/0.5 miles) stops. At the stops observers conduct 3 min audio-visual point counts covering a circle with 0.4 km radius. The routes provide a fairly good and random coverage of the study area, albeit with varying density depending on population and road density (Bystrak 1981). Variation among skills of observers introduces noise in the data (Sauer et al. 1994) but there is no indication or reason why this should systematically bias our results. Similarly, the road-side location of the stops and different detectability among species likely introduces more error for some species than for others, but given the large coverage of very different species in our research, there is no reason why this should have generally biased our results.

We averaged counts of 79 selected bird species at 1293 routes which were sampled each of five years (1996–2000) and designated as high quality (good weather and observers) each of those five years. By averaging over several years, we excluded year-to-year population fluctuations for example introduced by winter survival or disease (Sauer et al. 1997) letting us focus on long-term habitat associations rather than dynamics and temporal variation. Fine scale temporal variation and coarse scale temporal trends are not investigated further in this study. The counts were pooled from stop to route level and square root transformed for abundance-based models or turned into binomial presence/absence (non-zero/zero abundance) for P/A models. The resulting coarse spatio-temporal scale aims to exclude much fine scaled variability and fluctuations of bird abundances and environmental variability that do not lend themselves well to modelling with high priority on generality.

The species had to fulfil the following criteria for inclusion in the study: 1) at least 400 occupied locations; 2) land bird; 3) taxonomically stable. We selected the cut-off at a minimum of 400 occupied locations because at this level the positive correlation between model performance and sample size disappeared. All models were restricted to the birds' ranges. Ranges were estimated with the Ripley-Rasson estimator (Ripley and Rasson 1977) based on occupied locations. Consequently, the resulting sample size varied among the 79 bird species and was on average 1041±218 locations (range 492–1365).

We used 27 environmental variables as independent predictors representing land cover (n=11), temperature and precipitation means (n=6), temperature and precipitation extremes (n=2), seasonality in temperature and precipitation (n=4), year to year variation in temperature and precipitation (n=3), and the normalized difference vegetation index (NDVI), which is a measure of vegetation productivity. We used climate data from the CRU CL 1.0 dataset (New et al. 1999) available at < https://crudata.uea.ac.uk/~timm/grid/CRU_CL_1_0.html> and calculated weather variability variables from the United States Historical Climatology Network (HCN) Serial Temperature and Precipitation Data available at <www.ncdc.noaa.gov/ol/climate/research/ushcn/ushcn.html>. Furthermore, we used vegetation land cover data from the USGS Land Cover/Land Use categories available at <http://edcsns17.cr.usgs.gov/glcc/glcc_version1.html#NorthAmerica>. We collapsed the 24 land use categories into 11 and calculated the percentage cover of each category in a 20 km radius circle around the BBS route midpoint. Finally, NDVI came from a NOAA/NASA Pathfinder AVHRR 8 km resolution composite averaged over 1982–1992 for the month June. Given that the average distance between route midpoints in the BBS is 42.4±30.9 km (SD) and that we averaged environmental

variables over 20 km circles to meet the resolution of the routes, the resolution of the environmental data was sufficient.

Statistical techniques

We used random forests (RF) as a very robust and objective method for building the P/A and Ab models (Breiman 2001, Garzon et al. 2006, Prasad et al. 2006). RF are a resampling and subsampling extension of regression trees (RT). We grew 500 trees based on bootstrap samples of the original data and subsamples of the independent variables. RF are not as easily interpreted as RT or multiple regressions, because the final prediction is the model average over 500 individual trees, but this drawback was inconsequential for our study – we were examining the usefulness of predictions from distribution models, not their ability to explain distributions. RF are recognized as one of the best distribution modelling techniques as measured by predictive power (Garzon et al. 2006, Prasad et al. 2006) including specifically in the context of niche modelling (Elith et al. 2006).

Depending on the data, situation, and circumstances, different statistical techniques can perform differently and can come to very different predictions (Pearson et al. 2006, Araújo and New 2007, Thuiller et al. 2009). Therefore, we implemented additional statistical techniques to make sure that our results were not contingent on the use of RF. Our goal was to conduct our comparisons on both, presence/absence and abundance data. Therefore, the included techniques had to be able to handle both types of data, which some popular approaches such as MaxEnt (Phillipset al. 2006) and GARP (Peterson 2001) cannot. In addition to RF we used boosted

regression trees (BRT implemented in the R package gbm), general additive models (GAM implemented in the R package mda), and multivariate adaptive regression splines (MARS implemented in the R package mda). We closely followed the methodology described in Elith et al. (2006) for all techniques. Note that explanatory variables were reduced to eight climatic variables and NDVI for GAM and MARS, because the full set of 27 variables caused convergence problems and deteriorated the performance of GAM and MARS due to intercorrelation among predictors. BRT performed statistically indistinguishably from RF. All other techniques performed worse than RF, especially in the geographically split dataset approaches. Therefore, we will focus results and discussion on RF because statistical model comparisons abound in the literature (Elith et al. 2006) and were not a goal of our work.

For all 79 bird species we built distribution models on P/A and Ab data, using the full dataset (no split), a dataset randomly split in half, a dataset split into quarters along three longitudinal lines, and a dataset split in half along a longitudinal line (Fig. 1). We placed the splits in the longitudinal approaches so that each resulting part contained the same number of occupied locations within the range of a given bird species. While the full dataset approach was tested on the same data that were used for building the models, the split designs built the models using one half of the data and tested it using the other half.

In the random selection we first randomly split all occupied locations and then all unoccupied locations. For splitting the range in half we found the median longitude of all occupied locations and split the dataset along this longitude. We then used one of the halves to build the models and the other to evaluate it and then switched the roles of the halves and averaged over the two sets

of results. The quarter splits (longitudinal) strips worked similarly. We first found the 25th, 50th and 75th percentile longitude of all occupied locations and then split the dataset into four parts along these lines, numbered 1–4 from west to east. We then used part 1 and 3 combined to build the models and part 2 and 4 to evaluated them. Next we switched the roles of the 4 parts and finally averaged over the two sets of results.

The four different approaches represented a progression from no segregation of training and test data (no split) to a more and more spatially segregated split between training and test datasets. Longitudinal splits were chosen because north–south splits would have led to more severe climatic differences between training and test data.

Statistics calculated

P/A and Ab models necessitate different statistics for testing given the difference in variable structure. For P/A models, we reported the widely used area under the curve (AUC) of a receiver operating characteristic curve (ROC) (Fielding and Bell 1997), the square of the point-biserial correlation (r^2_{binom}) calculated simply as a Pearson's product moment correlation between predicted probability of occurrence and observed P/A squared, and Somer's D (also known as Gini coefficient (Engler et al. 2004)), which can be derived from AUC as $D = 2 \times (AUC - 0.5)$, representing a simple standardization of AUC to the more intuitive range of 0 to1. (Note, however, that AUC can be below 0.5 if the model prediction is worse than random chance and thus Somer's D can go below 0.) For the Ab models, we used the familiar coefficient of determination (R^2), based on proportion of variance explained, and the square of the Pearson

correlation coefficient (r^2) between predicted and observed abundances, somewhat analogous to the squared biserial correlation in P/A models. Note that for OLS linear models with only one predictor $R^2 = r^2$ but this does not have to hold in the nonlinear random forest models. R^2 describes the fit between predicted and observed in an absolute way (accuracy), while r^2 describes fit in a relative way (relative abundance, precision). All models were fit in and statistics calculated in R ver. 2.2.1 (R Development Core Team) with the extensions randomForest 4.5-18 and ROCR (1.0-2).

Results

Random forests (RF) led to near perfect discrimination in the presence/absence (P/A) models when tested on the same data they were trained on (Fig. 2): the average AUC scores of the 79 bird species were indistinguishable from 1. Even the squared point-biserial correlation reached very high values (0.95 ± 0.001), showing that while a point-biserial correlation reaching a value of 1 is close to impossible (the continuous variable would have to be distributed perfectly bimodally to match the binary one), it can reach very high values.

When we split the data randomly into halves of equal sample size and trained the RFs on one half and tested predictions on the other (which is spatially interleaved with the training data), the discrimination rate dropped (AUC = 0.90 ± 0.006, $r^2 = 0.43 \pm 0.015$; Fig. 2).

Introducing geographically segregated data splits led to much lower performance measures (Fig. 2). For a complete split into east and west halves, average AUC and r^2 dropped to 0.73 ± 0.012

and 0.12 ± 0.016, respectively. When training and test data were interspersed in four longitudinal strips, the models' tests led to slightly better results (0.79 ± 0.009 and 0.2 ± 0.016, respectively).

Tests of abundance (Ab) models followed similar patterns through the different types of data splits (Fig. 2). The squared correlation coefficients (r^2) were very similar for P/A and Ab models throughout the four testing schemes, indicating that the performances of these two types of models were actually quite similar when tested using a comparable measure. However, the coefficient of determination, R^2, or as colloquially known 'the percentage of variance explained' deviated from the r^2s: it was substantially lower for the schemes with geographically segregated training and test data (Fig. 2). For the quarter split the R^2 remained barely above zero (0.07 ± 0.032), while it dipped below zero when the dataset was split into longitudinal halves (–0.09 ± 0.051). An R^2 below zero indicates that using the average abundance from the training data as prediction over all test locations (analogously to an intercept-only null model in regression models) would have been closer to the truly observed abundances than the model predictions. This means while models retained some capability to predict relative abundance or the relative suitability of locations in spatially segregated test areas, their ability to predict the absolute abundance at new locations was virtually non-existent.

In our tests, boosted regression trees (BRT) performed virtually identical to RF while general additive models (GAM) and multivariate adaptive regression splines (MARS) were clearly outperformed (Fig. 3). Note that these results are for abundance-based modelling only. R^2 values in resubstitution evaluation closer to R^2 in random splits for GAM and MARS versus RF and BRT suggest that MARS and particularly GAM were not as overfit as RF and BRT. Going by

the random split evaluation, we might have concluded that only a small performance gap exists between RF/BRT and GAM/MARS. However, the poor performance of GAM and MARS in the geographic split evaluations (Fig. 3) suggests that these techniques have much more trouble predicting into new areas than do RF and BRT.

Discussion

Our results illustrate how important the selection of a testing scheme is when judging the predictive performance of a distribution model. Resubstitution – i.e. using the same data for model building as for testing – provides unrealistic estimates of performance of modern, flexible models that are prone to overfitting and should be avoided. Doubtlessly, the extremely high measures of performance of RF as judged by resubstitution indicate overfitting. Nobody in the machine learning community would ever suggest judging the fit of a model by resubstitution and RF are typically tested on 'out-of-bag' data that were randomly excluded during bootstrap. This default measure of RFs is analogous to a random-hold out testing scheme and supplies very similar estimates of performance. However, even if RF seem to overfit as judged by resubstitution they have been shown to generalize well (Breiman 2001) and compare very favourably to other methods in prediction on held out data (Garzon et al. 2006, Prasad et al. 2006, Cutler et al. 2007). We verified this in our data by comparing the results of RF to BRT, GAM and MARS models, which we will discuss below in the context of non-analog climate.

The currently most widespread, advanced method in the literature for testing distribution models is either a random hold-out of data, or data collected independently in the same area for testing

purposes (Brotons et al. 2004, Elith et al. 2006, Maggini et al. 2006). This is often described as testing the models on 'independent' data. However, species distribution data typically exhibit spatial autocorrelation. When testing data are randomly held-out, the locations of these data points will be interspersed with the training data locations, in our case leading to average proximity of 42.4±30.9 km (SD) between training and test locations. However, spatial autocorrelation may range much further than this average distance (in our case over several hundred kilometres; Bahn unpubl.), leading to dependence between training and test data (Araújo et al. 2005a). The consequence is that models are already optimised to fit test data during parameterization because of the dependence between training and test data. Therefore, the test data fit the model deceptively well – better than it would if test data were truly independent. This argument is unaltered if the test data were collected by different people at different times by different methods but in the same area (interspersed) with the training data (Edwards et al. 2006, Guisan et al. 2007). Just because collection of the data was independent does not automatically lead to independence in data values.

The overly optimistic testing results of models with randomly held-out but not fully independent test data is illustrated well by the large drop in performance measures we observed when we tested our models on truly independent, spatially segregated data. The models fared slightly better in the interspersed four-split approach than in the halves approach, which could either be an indication of a remaining effect of autocorrelation along the segregation lines (only one segregation line in the halves approach but three in the strips approach) or a reduced problem of models predicting into new climatic and biotic space which may include extrapolation to climates not encountered at the training locations or a violation of the assumption of stationarity

of environmental associations of the species (i.e. that the same functional relationships with the environment govern the abundance of species anywhere in the range, which is expected if a single model is built for the whole range) (Whittingham et al. 2007).

What causes our results? One explanation for our results could be that generic land cover and climate variables at coarse scales have little predictive power for species distributions. If true, it would be a gloomy assessment of our state of understanding and ability to predict the effects of global change at such a scale partly supported by Bahn and McGill (2007).

Two other causes for the drop of predictive power of distribution models when training and test datasets are split geographically include: 1) the effects of extrapolation to non-analog climates (Williams and Jackson 2007), 2) non-stationarity which occurs when the relationship between climate and species presence changes across space (e.g. hot is good in the north but bad in the south) (Whittingham et al. 2007). These three factors are all confounded. To reach truly independent testing data we had to introduce a rather dramatic geographical segregation which effectively broke the dependence via spatial autocorrelation but at the same time presented other problems to distribution models. For distribution models to be successfully applied for prediction into new regions one has to assume stationarity and that the range of combinations of biotic and abiotic factors in the test region were covered in the training regions.

As for the possibility that non-stationarity caused our low success to predict to spatially independent areas, there are a number of threads of evidence for variables and relationships that determine abundance changing across the range of a species. Whittingham et al. (2007) showed

that this was true at the landscape scale within Britain for birds. Similarly, in the few cases where the causes of species range boundaries have been worked out around the entire range (i.e. north, south, east and west) the limiting factors often change. For the Saguaro cactus *Carnegia gigantii*, the eastern and northern limits are set by frost tolerance, the western limit is set by the availability of summer precipitation (the main water source for this shallow rooted plant) and the southern limit is presumably set by being outcompeted by larger columnar cacti (Niering et al. 1963). In another well worked out case along an elevational gradient (Randall 1982), the upper limit of a moth species is set by food availability (the host plant cannot tolerate the colder temperatures) while the lower limit is set by the increasing presence of a particular parasite at warmer temperatures. Finally, Jarema et al. (2009) have shown that climate provides a good predictor of the maximal achievable abundance (90th percentile in quantile regression). However, abundances below that maximum right down to zero were also observed for any given value of any given environmental factor, suggesting that any particular environmental factor is the limiting factor on abundance in a few of locations, while other factors are constraining elsewhere. On one level this non-stationarity makes good sense – ecology has been well known to be a discipline in which many factors are important with their relative importance changing frequently (Quinn and Dunham 1983). Taking non-stationarity into account holds great promise for future improvements in distribution modelling but requires tremendous amounts of training data, as the effects of environmental variables on the distribution of a species will have to be determined for every combination of circumstances individually.

Another possibility is that prediction into non-analog climate caused our low performance measures. Supporting circumstantial evidence for this is the poor performance of GAM and

MARS in split range evaluations. When confronted with climate values out of the range of training data, RF and BRT will apply the prediction from the closest value within the range of training data, also known as 'clamping'. All predictions made by RF and BRT are derived from an average of actual observations and thus will never be completely off the charts. In contrast, extrapolation from training data climatic values can lead to extreme predictions in GAM and MARS, as values are not clamped to the last value contained in the training range but are free to rise or fall strongly from there on. Therefore, a likely cause for the abysmal performance of GAM and MARS in our tests are regular occurring completely 'off-the-charts' predictions, strongly negatively influencing the R^2 values. Despite this circumstantial evidence for climatic extrapolation happening, our evaluation is reasonable from a practical standpoint. Every location is climatically non-analog to any other location if one only looks closely enough (uses enough variables and their interactions). The question thus is not whether prediction into non-analog conditions happened (it certainly has), but how far the models had to extrapolate in climate space and whether our tests are reasonable from a practical stand point of what these models are typically used for.

However, whether spatially segregated holdouts lead to such low predictive power due to the elimination of the effects of spatial autocorrelation on testing (i.e. a more rigorous evaluation) or the unintended effects of non-stationarity and predicting into non-analog climate (or a mixture of all), the conclusion of our research remains similar: coarse-scale environment-based distribution models predict weakly when they are forced to predict upon truly spatially independent (and thus segregated) locations and/or into new climates, which is often the goal of these models. Such predictions simply always carry the risk of predicting into non-analog climates or areas where

climatic effects on a species differ from the training region. Therefore, testing them on data that is interspersed with training data within the range of autocorrelation is misleading when the performance of prediction into a new area or into new conditions (e.g. climate change) is to be judged. Bahn and McGill (2007) showed for North American Breeding Bird Survey data that if new locations to be predicted upon are not truly independent in space (i.e. they are closely interspersed with surveyed locations), simple spatial interpolation from surveyed locations is as powerful for prediction as an environment-based modelling approach. And here we showed that if locations are truly independent, as in our geographically segregated split approaches, the environment-based models are not consistently useful. This conclusion may well be different at finer scales and using more direct resource gradients as explanatory variables (Vanreusel et al. 2007).

Our results require careful interpretation. First of all, the goal of our study needs to be clear. Our intention was to determine the influence of a gradient in dependence between training and test data on the outcome of model tests that are used to determine the models' predictive power (R^2 and other measures). There are many other current and important methodological questions for distribution models which we do not address. We did not investigate the effects of autocorrelation on model building, parameter estimates, estimates of degrees of freedom, variance or hypothesis test statistics. This has all been covered in detail elsewhere (Legendre 1993, Lennon 2000, Dale and Fortin 2002, Dormann 2007). Also not the subject of our study and covered elsewhere are detailed comparisons of different modelling techniques (Elith et al. 2006, Garzon et al. 2006, Prasad et al. 2006, Cutler et al. 2007), influence of ecology/life-history on models (Austin 2007, McPherson and Jetz 2007), scale (Storch and Gaston 2004, Araújo et al.

2005b, Betts et al. 2006), and spatio-temporal variability/population dynamics/equilibrium (Johnson et al.1992, Maurer and Taper 2002, Svenning and Skov 2007), to name the most prominent distribution modelling issues. Instead, we focused on a top performing, robust modelling technique, averaged out spatio-temporal dynamics, and tried to match the scale of dependent and independent variables to the best of their availabilities. Moreover, to gain as much generality as possible, we included as many species as possible and used general environmental variables, doubtlessly sacrificing some explanatory power that could have been achieved by building individual models for each of the 79 bird species based on a detailed review of their ecology (Austin 2002).

P/A models and Ab models performed very similarly in the different testing schemes according to a comparable test statistic: the squared correlation between predicted and observed values. However, why did the Ab models show a much worse performance when the coefficient of determination R^2 was the test measure and the P/A models a seemingly better performance when the test measure was AUC or the equivalent but scaled to 0 to 1 Somer's D (Fig. 2)? First, the difference between the squared correlation coefficient r^2 and R^2 is that the former relies on relative abundance (i.e. areas of higher abundance must be predicted to have a relatively higher abundance than areas of low abundance, but the absolute value of abundance need not be right – it could be strongly biased high or low), while R^2 describes the absolute accuracy and precision of the predictions. Thus, the probability of occurrence generated by P/A models was an equally successful predictor of relative abundance as the abundance estimates generated by the Ab models – a result which concurs with Pearce and Ferrier (2001). However, relative abundance cannot be translated into absolute abundance without additional information. Second, AUC and

Somer's D take us even one step below r^2 in terms of information content of the dependent variable, namely, to the predicted classification into presence and absence. Therefore, they give seemingly better test results. This is further amplified in the case of AUC by being scaled to 0.5 to 1 rather than 0 to 1 as most other statistics used for model testing. Thus, the higher values of AUC and Somer's D than r^2 and of r^2 than R^2 have to be seen in the light of the differences in information content in the dependent variable. From an ecological and conservation point of view, knowledge on the absence or presence of an organism at a location is less useful than an estimate of its relative abundance which is in turn less useful than an estimate of absolute abundance (or density). After all, a presence could stem from an extinction prone sink population just as well as from a very high density population of core importance to the species (Pulliam 1988, 2000).

We reached these conclusions using a dataset that is of outstanding quality and quantity, and a selection of species with high quality and quantity of data. Our test of P/A models on independent data was well in line with other studies (Manel et al. 1999, Betts et al. 2006, Elith et al. 2006). In addition, using the longitudinal split, we tried to avoid extrapolating in environmental space (e.g. predicting for a high-temperature region based on a model derived from a low-temperature region), although some applications of these models, such as global warming scenarios, try to do exactly that. Given that the BBS data has a good spatial coverage and that we had true absences and did not have to generate absences from background conditions, we also had a low danger of sample selection bias (Phillips 2008).

Our results corroborate and extend the results of previous work. Several studies tested predictions from distribution models on spatially segregated data (Fielding and Haworth 1995, Peterson 2003, Randin et al. 2006, Segurado et al.2006, Peterson et al. 2007, Phillips 2008) or temporarily segregated data (Martinez-Meyer et al. 2004, Araújo et al. 2005a). All but one (Whittingham et al. 2003) of these studies exclusively dealt with presence–absence or presence only data and most were only based on a few species. Although there was some variation in results and the difference in methods and criteria makes a rigorous comparison difficult, we concluded that within this segment of our study (presence–absence tested on segregated data) our results were similar (AUCs in the range of 0.7–0.8) to other studies. Interpretation of these results varied wildly, though, with some researchers concluding that species distributions are rather complex and unpredictable and others enthusiastically declaring an excellent predictive capability. For example Graf et al. (2006) predicted distributions of *Tetrao urogallus* based on models using training data from one region in other, spatially segregated regions in Switzerland, achieving AUC (area under the curve of a receiver operating curve) values of 0.83–0.94. However, they concluded that '[t]he regional models performed well in the region where they had been calibrated, but poorly to moderately well in the other regions'. In contrast, Murray et al. (2011) interpreted a discriminative ability of 0.77–0.90 across a few different models used for predicting *Petrogale penicillatato* occurrence to adjacent areas as 'excellent'. We hope that our comprehensive study puts these results into a better perspective, on the one hand putting them into the full gradient of independence from resubstitution to spatial segregation, and on the other hand comparing them to the ecologically more interesting and statistically easier to interpret results from abundance models all well replicated on a large number of species covering different areas.

In general, we encourage a greater distinction between interpolation and extrapolation (see also Peterson et al. 2007). Interpolation uses distribution models to fill in holes within the geographic and environmental space of the original data (unsampled sites surrounded by sites which were sampled) which actually benefit from the autocorrelation in the data (Bahn and McGill 2007). Extrapolation uses distribution models to make predictions about a time or place geographically or environmentally distinct from where the measurements (training data) were taken. This is commonly done both for predicting the new ranges of invasive species (Higgins et al. 1999, Thuiller et al. 2005) and for predicting the ranges of species in the future under global warming (Peterson et al. 2002, Oberhauser and Peterson 2003, Thomas et al. 2004). When extrapolation is the intended goal for a model, model testing on resubstituted or randomly held-out data will make the model appear more successful at prediction than it will be under the new conditions.

How widespread are these extrapolation types of applications of SDM's for which our results are relevant? Araújo and Peterson (2012) recently reviewed applications of bioclimatic envelope models and 'suggest that criticism has often been misplaced, resulting from confusion between what the models actually deliver and what users wish that they would express'. A first differentiation they make is between models that aim to explain relationships between environmental conditions and an organism and models that aim at predicting the potential distribution of an organism. However, a failure to evaluate a model rigorously on independent data can lead to overfitting. An overfit model may give misleading conclusions on the importance of explanatory variables and thus fail to achieve the goal of understanding relationships between an organism and the environment. Further, Araújo and Peterson (2012) list

six common applications for climate envelope models: 1) discovery of new populations or species; 2) reserve selection and design; 3) restoration, translocation, or reintroductions; 4) evaluating risk of species invasions and disease transmission; 5) climate change impacts on biodiversity; and 6) niche evolution. Four out of six of these common applications explicitly have prediction into new geographic or climatic space as a goal (1, 3, 4, 5), while 2 and 6 could at least include such cases. We believe that it is fair to say that most applications of SDM's or bioclimatic envelope models are used to predict into new geographic and/or environmental space and thus are subject to the findings we present in this paper.

Conclusion

We showed that the currently most widely used and accepted method for testing presence/absence distribution models, namely randomly holding out test data, led to estimates of performance (average AUC >0.9 in our models) that were inflated in comparison to a more rigorous test that accounted for two additional factors not commonly considered. Incorporating these two factors drastically reduced the apparent performance of the model (relative abundance mean r^2 <0.15 and absolute abundance mean R^2<0). These two factors were: 1) the presence of strong autocorrelation in the data, which prevents random hold-out data from being truly independent (even if it was collected independently) and creates a false sense of predictive power in model tests, and 2) the assessment of prediction of presence/absence classification (as measured by AUC) rather than prediction of relative or absolute abundance – the more ecologically meaningful information (as measured by r^2 and R^2, respectively).

The absence of these more rigorous methods in currently widely used model evaluation practices has far reaching consequences. Evaluation is used in two critical ways: model selection and judging our confidence in the model. In the former, using a testing scheme that leads to overly optimistic evaluations, as is the case with any test on not fully independent test data, will lead to the selection of overfit models that can both lead to false insights in which factors are important to the distribution of a species and also lead to false predictions of which conditions are generally suitable to a species. The second problem, a misjudgement of our confidence in the models, seems to have less severe consequences, unless an overly optimistic model evaluation suggests that we are able to predict potential species occurrences or abundances with moderate accuracy when in reality, we are doing no better than random (i.e. AUC$\leq$0.5 or $R^2 \leq 0$).

For many urgent questions, distribution modelling is currently the only tool available and we do not suggest discarding it. However, our results suggest that current opinion about how well distribution models perform may be overly optimistic when extrapolating into new areas or new climate regimes for either prediction or understanding and when testing is done with interspersed (spatially autocorrelated) test data. Distribution modellers should exercise caution when using such models in a predictive fashion, especially under radically changed conditions such as exploring the effects of future climate change. The testing scheme used to judge the usefulness of a model needs to match the intended purpose. A model that is intended to predict into new areas or conditions needs to be tested using truly independent, spatially segregated data.

Acknowledgements

We are indebted to thousands of volunteers and their coordinators for making the North American Breeding Bird Survey data publicly available. Funding was provided by NSERC. We thank Deanna Newsom for assistance.

References

Andrewartha, H. G. and Birch, L. C. 1984. The ecological web: more on the distribution and abundance of animals. – Univ. of Chicago Press.

Araújo, M. B. and Guisan, A. 2006. Five (or so) challenges for species distribution modelling. – J. Biogeogr. 33: 1677–1688.

Araújo, M. B. and Rahbek, C. 2006. How does climate change affect biodiversity? – Science 313: 1396–1397.

Araújo, M. B. and New, M. 2007. Ensemble forecasting of species distributions. – Trends Ecol. Evol. 22: 42–47.

Araújo, M. B. and Peterson, A. T. 2012. Uses and misuses of bioclimatic envelope modeling. – Ecology 93: 1527–1539.

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

Araújo, M. B. et al. 2005b. Downscaling European species atlas distributions to a finer resolution: implications for conservation planning. – Global Ecol. Biogeogr. 14: 17–30.

Austin, M. P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. – Ecol. Modell. 157: 101–118.

Austin, M. 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches .– Ecol. Modell. 200: 1–19.

Bahn, V. and McGill, B. J. 2007. Can niche-based distribution models outperform spatial interpolation? – Global Ecol. Biogeogr. 16: 733–742.

Betts, M. G. et al. 2006. The importance of spatial autocorrelation, extent and resolution in predicting forest bird occurrence. – Ecol. Modell. 191: 197–224.

Box, G. E. P. 1976. Science and statistics. – J. Am. Stat. Ass. 71: 791–799.

Breiman, L. 2001. Random forests. – Machine Learning 45: 5.

Brotons, L. et al. 2004. Presence–absence versus presence-only modelling methods for predicting bird habitat suitability. – Ecography 27: 437- 448.

Bystrak, D. 1981. The North American breeding bird survey.– In: Ralph, C. J. and Scott, J. M. (eds), Estimating numbers of terrestrial birds. Cooper Ornithol. Soc., pp. 34–41.

Cutler, D. R. et al. 2007. Random forests for classification in ecology. – Ecology 88: 2783–2792.

Dale, M. R. T. and Fortin, M.-J. 2002. Spatial autocorrelation and statistical tests in ecology. – Ecoscience 9: 162–167.

Davis, A. J. et al. 1998. Making mistakes when predicting shifts in species range in response to global warming. – Nature 391: 783–786.

Dennis, B. et al. 1991. Estimation of growth and extinction parameters for endangered species. – Ecol. Monogr. 61: 115–144.

Dormann, C. F. 2007. Effects of incorporating spatial autocorre- lation into the analysis of species distribution data. – Global Ecol. Biogeogr. 16: 129–138.

Edwards, J. T. C. et al. 2006. Effects of sample survey design on the accuracy of classification tree models in species distribution models. – Ecol. Modell. 199: 132–141.

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

Engler, R. et al. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. – J. Appl. Ecol. 41: 263–274.

Ferrier, S. 2002. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? – Syst. Biol. 51: 331–363.

Fielding, A. H. and Haworth, P. F. 1995. Testing the generality of bird-habitat models. – Conserv. Biol. 9: 1466–1481.

Fielding, A. H. and Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/ absence models. – Environ. Conserv. 24: 38–49.

Garzon, M. B. et al. 2006. Predicting habitat suitability with machine learning models: the potential area of Pinus sylvestris L. in the Iberian Peninsula. – Ecol. Modell. 197: 383–393.

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.

Guisan, A. et al. 2007. What matters for predicting the occurrences of trees: techniques, data or species' characteristics? – Ecol. Monogr. 77: 615–630.

Hampe, A. 2004. Bioclimate envelope models: what they detect and what they hide. – Global Ecol. Biogeogr. 13: 469–471.

Higgins, S. I. et al. 1999. Predicting the landscape-scale distribution of alien plants and their threat to plant diversity. – Conserv. Biol. 13: 303–313.

Iverson, L. R. and Prasad, A. M. 1998. Predicting abundance of 80 tree species following climate change in the eastern United States. – Ecol. Monogr. 68: 465–485.

Jarema, S. I. et al. 2009. Variation in abundance across a species' range predicts climate change responses in the range interior will exceed those at the edge: a case study with North American beaver. – Global Change Biol. 15: 508–522.

Johnson, A. R. et al. 1992. Animal movements and population dynamics in heterogeneous landscapes. – Landscape Ecol. 7: 63–75.

Krebs, C. J. 1972. Ecology: the experimental analysis of distribution and abundance. – Harper and Row.

Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? – Ecology 74: 1659–1673.

Legendre, P. and Fortin, M.-J. 1989. Spatial pattern and ecological analysis. – Vegetatio 80: 107–138.

Lennon, J. J. 2000. Red-shifts and red herrings in geographical ecology. – Ecography 23: 101–113.

Maggini, R. et al. 2006. Improving generalized regression analysis for the spatial prediction of forest communities. – J. Biogeogr. 33: 1729–1749.

Manel, S. et al. 1999. Alternative methods for predicting species distribution: an illustration with Himalayan river birds.– J. Appl. Ecol. 36: 734–747.

Martinez-Meyer, E. et al. 2004. Ecological niches as stable distributional constraints on mammal species, with implications for Pleistocene extinctions and climate change projections for biodiversity. – Global Ecol. Biogeogr. 13: 305–314.

Maurer, B. A. and Taper, M. L. 2002. Connecting geographical distributions with population processes. – Ecol. Lett. 5: 223–231.

McPherson, J. M. and Jetz, W. 2007. Effects of species' ecology on the accuracy of distribution models. – Ecography 30: 135–151.

Murray, J. V. et al. 2011. Evaluating model transferability for a threatened species to adjacent areas: implications for rock-wallaby conservation. – Austral. Ecol. 36: 76–89.

New, M. et al. 1999. Representing twentieth-century space-time climate variability. Part I. Development of a 1961–90 mean monthly terrestrial climatology. – J. Climate 12: 829–856.

Niering, W. A. et al. 1963. The saguaro: a population in relation to environment. – Science 142: 15–23.

Oberhauser, K. and Peterson, A. T. 2003. Modeling current and future potential wintering distributions of eastern North American monarch butterflies. – Proc. Natl Acad. Sci. USA 100: 14063–14068.

Pearce, J. and Ferrier, S. 2001. The practical value of modelling relative abundance of species for regional conservation plan- ning: a case study. – Biol. Conserv. 98: 33–43.

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

Peterson, A. T. 2001. Predicting species' geographic distributions based on ecological niche modeling. – Condor 103: 599–605.

Peterson, A. T. 2003. Predicting the geography of species' invasions via ecological niche modeling. – Q. Rev. Biol. 78: 419–433.

Peterson, A. T. et al. 2002. Future projections for Mexican faunas under global climate change scenarios. – Nature 416: 626–629.

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

Phillips, S. J. 2008. Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson et al. (2007). – Ecography 31: 272–278.

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

Prasad, A. et al. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction.– Ecosystems 9: 181–199.

Pulliam, H. R. 1988. Sources sinks and population regulation.– Am. Nat. 132: 652–661.

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

Quinn, J. F. and Dunham, A. E. 1983. On hypothesis testing in ecology and evolution. – Am. Nat. 122: 602–617.

Randall, M. G. M. 1982. The dynamics of an insect population throughout its altitudinal distribution: Coleophora alticolella (Lepidoptera) in Northern England. – J. Anim. Ecol. 51: 993–1016.

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

Ripley, B. D. and Rasson, J. P. 1977. Finding the edge of a poisson forest. – J. Appl. Probabil. 14: 483–491.

Sauer, J. R. et al. 1994. Observer differences in the North American breeding bird survey. – Auk 111: 50–62.

Sauer, J. R. et al. 1997. The North American Breeding Bird Survey results and analysis. Ver. 96.4. –Patuxent Wildlife Res. Center.

Scott, J. M. and Csuti, B. 1997. Gap analysis for biodiversity survey and maintenance. – In: Reaka-Kudla, M. L. et al. (eds), Biodiversity, II. Understanding and protecting our biological resources. Joseph Henry Press, pp. 321–340.

Segurado, P. et al. 2006. Consequences of spatial autocorrelation for niche-based models. – J. Appl. Ecol. 43: 433–444.

Skov, F. and Svenning, J.-C. 2004. Potential impact of climatic change on the distribution of forest herbs in Europe.– Ecography 27: 366–380.

Storch, D. and Gaston, K. J. 2004. Untangling ecological complexity on different scales of space and time. – Basic Appl. Ecol. 5: 389–400.

Svenning, J.-C. and Skov, F. 2007. Ice age legacies in the geographical distribution of tree species richness in Europe.– Global Ecol. Biogeogr. 16: 234–245.

Thomas, C. D. et al. 2004. Extinction risk from climate change.– Nature 427: 145–148.

Thuiller, W. et al. 2005. Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. – Global Change Biol. 11: 2234–2250.

Thuiller, W. et al. 2009. BIOMOD - a platform for ensemble forecasting of species distributions. – Ecography 32: 369–373.

Vanreusel, W. et al. 2007. Transferability of species distribution models: a functional habitat approach for two regionally threatened butterflies. – Conserv. Biol. 21: 201–212.

Whittingham, M. J. et al. 2003. Do habitat association models have any generality? Predicting skylark Alauda arvensis abundance in different regions of southern England. – Ecography 26: 521–531.

Whittingham, M. J. et al. 2007. Should conservation strategies consider spatial generality? Farmland birds show regional not national patterns of habitat association. – Ecol. Lett. 10: 25–35.

Williams, J. W. and Jackson, S. T. 2007. Novel climates, no-analog communities and ecological surprises. – Front. Ecol. Environ.5: 475–482.

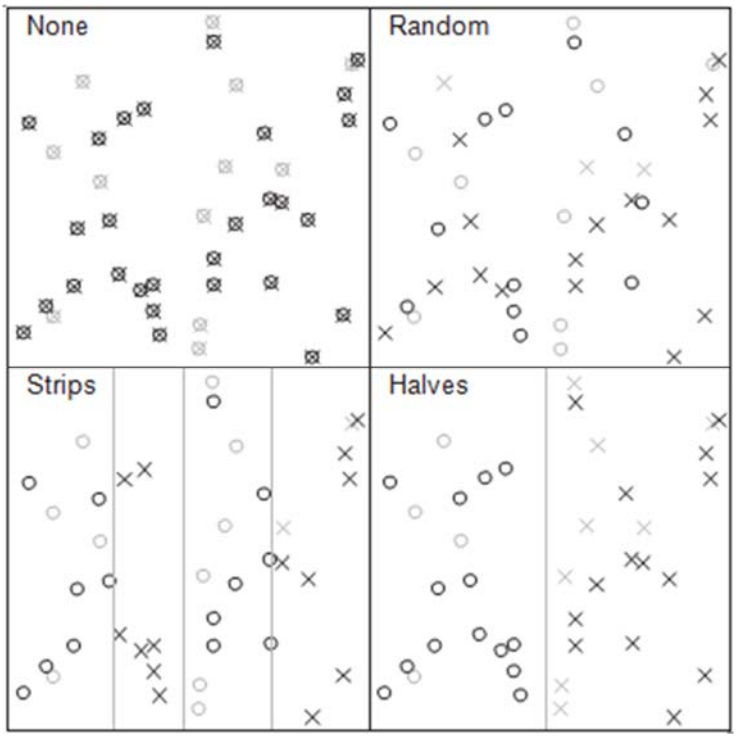

Figure 1. Schematic representation of data splitting approaches. Black symbols indicate locations at which the species is present, grey symbols indicate locations at which it is absent, circles indicate locations used in the training dataset, and ×'s locations that were used for testing of the model. Splits for strips and halves were selected so that an equal number of occupied (black) locations fell into each part.

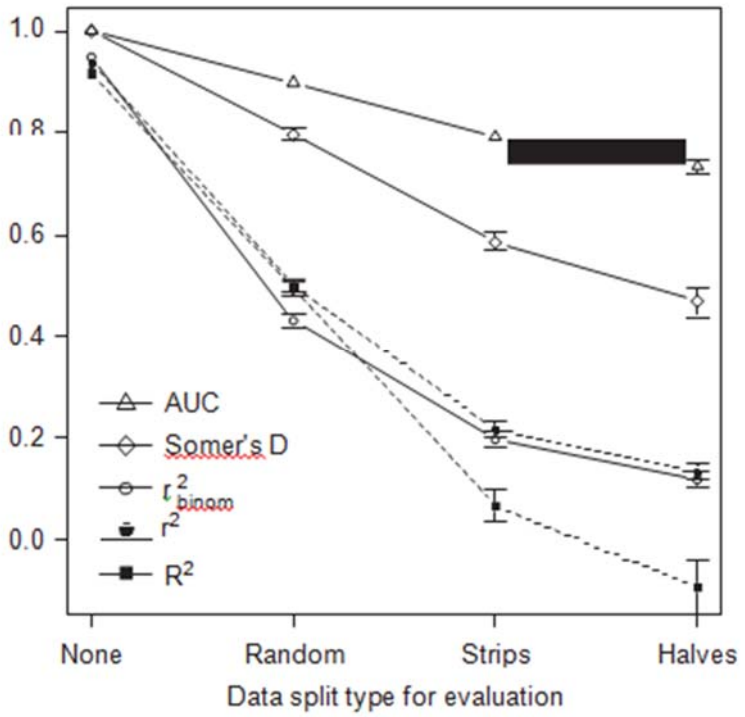

Figure 2. Influence of model testing scheme and choice of performance measure on perception of performance of distribution models. The average over models for 79 bird species is shown. Dependent variables were either bird presence/absence (P/A; solid lines and open symbols) or abundance (Ab; dashed lines and closed symbols). The statistics for the P/A models were area under the curve (AUC) of a receiver operating characteristic curve, Somer's D (2(AUC-0.5)), and squared point-biserial correlation (r^2_{binom}). The statistics for the Ab models were squared Pearson's correlation coefficient (r^2) and the coefficient of determination (R^2). Models were built on training data and tested on progressively more independent test data implemented by different splitting schemes: none (training data=test data), random (dataset split in half randomly), strips (dataset quartered into longitudinal strips, interspersed as training and test data), halves (dataset split in longitudinal halves). Standard errors >0.011 are shown as error bars.

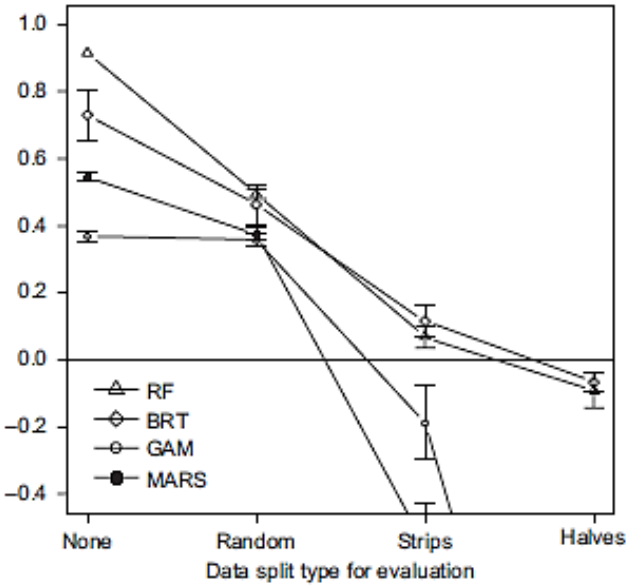

Figure 3. Influence of modelling technique on performance of distribution models. The median coefficient of determination (R^2) over models for 79 bird species is shown for four different modelling techniques: random forests (RF), boosted regression trees (BRT), general additive models (GAM) and multivariate adaptive regression splines (MARS). Models were built on training data and tested on progressively more independent test data implemented by different splitting schemes: none (training data=test data), random (dataset split in half randomly), strips (dataset quartered into longitudinal strips, interspersed as training and test data), halves (dataset split in longitudinal halves). Standard errors >0.011 are shown as error bars. Note that R^2 is cut off at -0.4 in the graph because a negative R^2 clearly indicates failed models and displaying even more negative values would have made the graph less appealing while not adding any valuable information.