RESEARCH PAPER

Can niche-based distribution models outperform spatial interpolation?

Volker Bahn* and Brian J. McGill

Department of Biology, McGill University, Montreal, Quebec, H1A 1B1, Canada. E-mails: volker.bahn@mcgill.ca; brian.mcgill@mcgill.ca

ABSTRACT

Aim Distribution modelling relates sparse data on species occurrence or abundance to environmental information to predict the population of a species at any point in space. Recently, the importance of spatial autocorrelation in distributions has been recognized. Spatial autocorrelation can be categorized as exogenous (stemming from autocorrelation in the underlying variables) or endogenous (stemming from activities of the organism itself, such as dispersal). Typically, one asks whether spatial models explain additional variability (endogenous) in comparison to a fully specified habitat model. We turned this question around and asked: can habitat models explain additional variation when spatial structure is accounted for in a fully specified spatially explicit model? The aim was to find out to what degree habitat models may be inadvertently capturing spatial structure rather than true explanatory mechanisms.

Location We used data from 190 species of the North American Breeding Bird Survey covering the conterminous United States and southern Canada.

Methods We built 13 different models on 190 bird species using regression trees. Our habitat-based models used climate and landcover variables as independent variables. We also used random variables and simulated ranges to validate our results. The two spatially explicit models included only geographical coordinates or a contagion term as independent variables. As another angle on the question of mechanism vs. spatial structure we pitted a model using related bird species as predictors against a model using randomly selected bird species.

Results The spatially explicit models outperformed the traditional habitat models and the random predictor species outperformed the related predictor species. In addition, environmental variables produced a substantial R^2 in predicting artificial ranges.

Main conclusions We conclude that many explanatory variables with suitable spatial structure can work well in species distribution models. The predictive power of environmental variables is not necessarily mechanistic, and spatial interpolation can outperform environmental explanatory variables.

*Correspondence: Volker Bahn, Department of Biology, McGill University, Stewart Biology Building, 1205 avenue Docteur Penfield, Montreal, QC, H3A 1B1, Canada.
E-mail: volker.bahn@mcgill.ca

Keywords
Birds, distribution modelling, habitat, macroecology, model evaluation, neighbourhood, niche, spatial autocorrelation, spatial interpolation, species distributions.

INTRODUCTION

Explaining and predicting the abundance and distribution of species is one of the fundamental tasks of ecology (Andrewartha & Birch, 1954; Guisan & Thuiller, 2005). The ability to model species distributions has proved to be vital to species conservation, land management and protected area design (Scott & Csuti, 1997; Ferrier, 2002) because complete censuses of threatened species are typically impractical.

Accordingly, the field of distribution modelling has developed rapidly in the past two decades, spawning a rich literature and methodology (for overviews of the field see Guisan & Zimmermann, 2000; Scott *et al.*, 2002; Guisan & Thuiller, 2005). At the same time, the development of remote sensing and high-resolution climate

DOI: 10.1111/j.1466-8238.2007.00331.x
www.blackwellpublishing.com/geb

interpolations has given distribution modelling a strong impetus to use explanatory variables from one category: the physical habitat. Records of species occurrences have increased but not kept up with the dramatic increase in information about the physical environment; data on species abundances covering large extents are still very rare. With the help of the statistical models, sparsely sampled species censuses are related to the physical environment, allowing the prediction of occurrences at all points where physical environment variables are available, which is nearly everywhere. Although this approach is typically correlative, it is assumed to be capturing the essential mechanisms underlying species distributions, albeit in a phenomenological fashion.

Unfortunately, most large-spatial-scale censuses of species record only presence/absence information rather than abundance. The ready availability of data about the physical environment and presence/absence data has led to the status quo of distribution modelling becoming presence/absence models predicted using habitat characteristics only (Scott *et al.*, 1993). The sheer number of tools developed in this specific application of the field (see the list in Guisan & Thuiller, 2005) is an indicator of the dominance of this approach, attributable at least in part to its practicality. While different modelling strategies have been explored [e.g. spatially explicit presence/absence models (Pereira & Itami, 1991; Augustin *et al.*, 1996), spatially explicit abundance models (Lichstein *et al.*, 2002) and community-based models (Elith *et al.*, 2006)], presence/absence-based habitat models still dominate current practice. Habitat modelling has even been expanded in temporal scope: it is now used for predicting species occurrences 100 years into the future, with the goal of predicting the consequences of human-induced climate change (e.g. Iverson & Prasad, 1998; Bakkenes *et al.*, 2002; Skov & Svenning, 2004; Thomas *et al.*, 2004; Hijmans & Graham, 2006).

Many modellers have noted spatial structure in the residuals of species distribution models and have found that spatially explicit models outperform spatially implicit models (Lennon, 2000; Keitt *et al.*, 2002; Lichstein *et al.*, 2002; Bahn *et al.*, 2006; Segurado *et al.*, 2006). In species distributions, such spatial patterns can result from either the spatial patterns in underlying conditions and resources (*exogenous* to the dependent variable), such as soil or temperature, or from population processes that directly generate spatial patterns in the species abundances (*endogenous* to the dependent variable), such as dispersal (Bahn *et al.*, 2006). The dominant thinking is that the niche part of the model captures the effects of local conditions on species in a mechanistic way, and also captures the spatial structure of the local conditions themselves (exogenous), as long as all important predictors are included (Diniz-Filho *et al.*, 2003). The question remains: can all spatial structure in distributions be explained by the spatial structure in the niche variables (exogenous autocorrelation), or is there residual spatial structure caused, for example, by the movement of organisms (endogenous autocorrelation)? Currently, it seems that the majority of studies (Lichstein *et al.*, 2002; Hawkins *et al.*, 2003; Bahn *et al.*, 2006) support the notion that endogenous sources of spatial structure are important in species distributions (but see Diniz-Filho *et al.*, 2003).

However, this line of thinking assumes that all spatial structure implicitly captured by environmental variables is legitimately and mechanistically explained by environmental variables. Under this assumption, endogenous spatial effects gain credibility only if there is spatial structure left in the residuals of a niche-based model. The 'new kid on the block' — endogenous spatial effects — is left with the burden of proof that it has something additional to offer. We question this view and ask: what if environmental variables capture spatial structure in a haphazard way by randomly matching the type and scale of spatial structure present in a species distribution? Why is an endogenous spatial effect required to demonstrate that it cannot be explained away by environmental variables, while environmental variables do not have to prove that they can explain variability in species distributions above and beyond what can be modelled by pure spatial interpolation? Such a question would seem perfectly normal to a mining engineer, who is used to modelling ore deposits with only spatial interpolation (kriging) between bore holes and needs to be convinced that introducing covariates is of any predictive value.

In the present study, we turn the standard question around and ask: how much can niche models explain that is not attributable to coinciding spatial structure? We approach this question by contrasting niche models with pure spatial interpolation models. We also model artificial ranges with empirical environmental variables, to gauge the feasibility of environmental variables explaining distributions haphazardly by randomly capturing spatial structure. Finally, we ask whether all variables with a suitable spatial structure could be decent predictors, and investigate this question by contrasting the use of related bird species as predictors with the use of random bird species.

METHODS

We used data from the North American Breeding Bird Survey (BBS) for our empirical investigation. The data were averaged at the route level at 1368 locations over 5 years (1996–2000). The locations covered the conterminous USA and southern Canada. We excluded locations in Alaska because they were too disjunct. We used all routes that had the highest reliability rating from the administrators of the BBS. We used all 190 land bird species that had at least 200 occupied locations (i.e. that were recorded at least once in the 5 years in at least 200 locations) and that were taxonomically stable over the time period covered (i.e. species that were joined or split were excluded). For each species we defined the range boundary using the Ripley–Rasson estimate (Ripley & Rasson, 1977) based on all occupied locations (with a non-zero averaged abundance). Only locations within this range were used in the individual species models leading to an average of 797 ± 26 (SE) locations per species.

The restriction of modelled sites to the range of the species was necessary to avoid excessive numbers of unoccupied locations in the data. High numbers of locations with zero abundance lead to strongly skewed distributions and models with inflated R^2. The latter is due to the ease with which independent variables can segregate out habitat far outside of the range: for example, a model that includes tundra for a sub-tropical bird will be very

successful at predicting low abundances in the tundra and thus will have a high R^2, but will not be correspondingly interesting since we already know that the bird does not occur anywhere close to the tundra. Including sites known to be unsuitable and unoccupied leads to higher R^2 but less useful models. A moderate to high number of unoccupied locations was present within the range of all bird species. These unoccupied locations characterize unsuitable habitat better than unoccupied locations outside of the range, where it is unknown whether birds are absent due to unsuitable conditions or due to their failure to colonize the location. To support this choice, we also ran the core models without restricting locations to the range of species for comparison.

We find that using abundance in distribution modelling rather than presence/absence has advantages. First, the measures of goodness-of-fit are more intuitive, better known and possibly more robust than the measures used in presence/absence models. The measures currently most accepted in binary models — AUC (area under the curve) of ROC (receiver operating characteristic) curves and Cohen's Kappa (Manel *et al.*, 2001; McPherson *et al.*, 2004) — are not as widely and as intuitively understood as the R^2 of an ordinary linear regression. R^2 surrogate measures are seldom used in conjunction with logistic regressions because they are typically low (for reasons why the coefficient of determination is typically very low with binary responses see Cox & Wermuth, 1992). Second, presence/absence modelling obscures the abundance structure of species distributions and, with it, the potential additional information contained in the orders of magnitude variation in abundance across regions, which is treated as identical by presence/absence information. We square root transformed the abundance (count) data to approach a normal distribution (Zar, 1996). Regression tree models are invariant to monotonic transformations in the independent variables but not in the dependent variable.

We used regression trees (RTs) to build statistical models that explain and predict bird abundances (Breiman, 1984). We calculated the models using the package RPART (Therneau & Atkinson, 1997) for R (R Development Core Team, 2005) — both in version 2.2.0. This technique recursively partitions the data into two groups at an optimal value of an independent variable. This optimal splitting value is found by testing all values of all independent variables in the data set as potential locations for a split, and picking the one split that best homogenizes the resulting groups with respect to the sums of squares in the dependent variable. For example, this technique might determine that a split of all locations by average January temperature > –5 °C leads to an optimal grouping into low and high abundance. Then the algorithm would continue searching for the next split within each of the newly formed groups, until given stopping criteria are reached. In our case these stopping criteria were: at least 15 cases had to be left in a node to allow a search for a new split, at least 5 cases had to end up in each new group after a split and each split had to increase the R^2 by at least 0.001. The resulting regression tree is a hierarchical accumulation of splitting rules, leading to $n + 1$ 'terminal groups of cases' or 'endnodes' in a tree with n splits. Any combination of values of independent variables unambiguously leads into one endnode, which gives a prediction for the dependent variable as the average of the dependent variables of the cases in this endnode from the original training set.

Typically, the above stopping criteria lead to over-fitted trees that do very well at explaining the training data but perform poorly on reserved or new data. Algorithms exist to scale back or prune a tree so that there is an optimal trade-off between explaining the learning data set and predicting upon a reserved data set. We pruned the RTs using a 10-fold cross validation (CV) and the one standard error rule (Breiman, 1984). In 10-fold cross validation, 90% of the data are used to create a model, which is then tested against the reserved 10%. This process is repeated 10 times until all data have been used as a reserve once. This method is more robust than traditional methods, with a constant split in the training and validation data sets, while still testing the model against different data than that for which the model was fit. This approach is thought to err on the side of under-fitting rather than over-fitting (Breiman, 1984), giving us confidence that neither the habitat model nor the spatial interpolation model performance estimates were biased from over-fitted models. The R^2-like measures (derivatives of reductions in deviance) we present here were also calculated during CV. Therefore, the models were tested on reserved data not included in building the RTs in the given step of CV, and the R^2-like measures not only signify goodness-of-fit of the models but also their predictive abilities (Therneau & Atkinson, 1997).

The benefits of RTs over ordinary least squares regression are manifold in our situation. Had we used ordinary regressions or general linear models (GLMs), any automated method of variable selection (e.g. stepwise) would have been subject to Freedman's paradox (Freedman, 1983) and any attempt to select variables for 190 species and 11 models by hand and expert opinion would have been daunting and subjective. RTs gave us an imperfect but solid and objective way of variable selection allowing for an objective comparison among the different model types. In this context it is important to note that RTs not only take care of variable selection but also model nonlinearities and interactions — another major potential source of error and subjective decisions in ordinary multiple regressions. In addition, RTs do not have distributional assumptions for the model errors. Had we used ordinary regressions, we would have had to check all models for Gaussian distribution of the errors and would have had to apply subjective remedial measures in the many cases of violations that would have had arisen. Furthermore, in RTs the deviance explained at every split and thus by every variable is known, so that a partitioning of variance method such as introduced by Borcard *et al.* (1992) was unnecessary to determine which variables (or groups of variables) explained which part of the variance. And finally, the ease of interpretation is typically greater in regression trees than in multiple regressions. A regression tree gives a clear path of the 'decisions' taken to get to a certain predicted abundance in an endnode, which is simply the average abundance of all cases from the training data set found in that endnode. For example, such a path may state that locations with July temperature < 25 °C and January temperature > –5 °C and coniferous forest > 70% can on average expect 2.2 bird detections of a certain species at a typical BBS point count. The effect of a discrete value of a variable is immediately obvious in the model,

while in a multiple regression, for example, slope coefficients of several variables have to be viewed in conjunction with each other and an intercept to interpret a model.

We used a range of models to investigate the performance of habitat models and how it relates to spatial patterns. First, we built classical habitat models (HAB) based on a large variety of environmental data and rooted in niche theory (Guisan & Zimmermann, 2000).

The environmental climate data came from the CRU CL 1.0 data set (New *et al.*, 1999) available at http://www.cru.uea.ac.uk/~timm/grid/CRU_CL_1_0_text.html on a 0.5° resolution grid. 0.5° corresponds to roughly 55 km and the average distance between nearest neighbours in the BBS data set as we used it is 42.4 ± 30.9 km (SD). Thus the resolution of the climate data was appropriate. The weather variability variables were calculated from the United States Historical Climatology Network (HCN) Serial Temperature and Precipitation Data available at http://www.ncdc.noaa.gov/ol/climate/research/ushcn/ushcn.html. The vegetation landcover data came from the USGS Land Cover/Land Use categories available at http://edcsns17.cr.usgs.gov/glcc/glcc_version1.html#NorthAmerica in a Lambert azimuthal projection where the USGS 24 land use categories were collapsed into 11. For each of the 11 land cover classes, the percentage cover of a circle 20 km in radius around the BBS route mid-point was calculated and used as an independent variable. Finally, the normalized difference vegetation index (NDVI) came from a NOAA/NASA Pathfinder AVHRR 8-km resolution composite averaged over 1982–92 for the month of June. We included 27 environmental variables representing landcover (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 NDVI, which is a measure of productivity. The variety of environmental data we included meets or exceeds all habitat modelling studies we know of. In particular, some authors have shown that just two to three climatic variables are sufficient (e.g. Bartlein *et al.*, 1986; Austin *et al.*, 1990); our model included these typical two or three variables as well as 23 more.

We used two spatial models not containing environmental variables to estimate how much variation in the distributions could be explained through endogenous and exogenous spatial patterns. The spatial models, which are a form of spatial interpolation, can recover both the endogenous spatial patterns caused by environmental variables and the exogenous spatial patterns caused by population dynamics and movement. However, they can only recover variation caused by exogenous (environmental) sources insofar as these sources vary in a spatially smooth and autocorrelated way. Even if all environmental variables were autocorrelated and spatially smooth at a suitable scale, the recovery of variation produced by environmental effects through purely spatial models should still be imperfect, because while autocorrelation means more similarity of neighbours than expected by random chance, it does not mean that neighbours need be identical or even very similar.

We used two different spatial interpolation techniques. The first, which includes geographical coordinates as independent variables (COORD), takes advantage of large-scale spatial trends (Legendre & Legendre, 1998) such as the hypothesized bell-curve structure of abundance across a species' range (Brown, 1995). The second, which uses the average of a neighbourhood (contagion; CONT) as an independent variable, takes advantage of fine scale spatial autocorrelation (Augustin *et al.*, 1996; Araújo & Williams, 2000).

The COORD models using regression trees can dissect the species' range into geographical regions somewhat like a polynomial trend surface, but in a more flexible way. Thus they can use coarse spatial patterns for non-causal interpolation/prediction. We used coordinates projected in Lambert equal-area azimuthal, central meridian-100 and reference latitude 45.

The second spatial null model, CONT, was aimed at fine-scale spatial interpolation. We wanted to determine if a simple distance weighted average of species abundances at neighbouring locations would make for a good predictor in a distribution model. We selected 200 km as a suitable maximum neighbourhood, given that birds are very mobile and that we had found autocorrelation with similar ranges in residuals of ordinary linear regression models of bird abundances. Within this neighbourhood, the average number of neighbours was 23.7 ± 16.8 (SE) (range 0–87), the average distance to neighbours was 130.2 ± 47.4 km and the average distance to the closest neighbour was 42.3 ± 25.7 km. We distance weighted the average abundance of neighbours using a spherical model (Legendre & Legendre, 1998):

$$\text{weight}_{ij} = \frac{1 - 1.5\,\text{distance}_{ij}/200\,\text{km} + 0.5(\text{distance}_{ij}/200\,\text{km})^3}{\sum \text{weights}_j}$$

where weight_{ij} is the weight assigned to the neighbour j of the location i, and distance_{ij} is the distance of neighbour j to location i. We standardized the weights so that their sum over all neighbours of a location i was 1, and then used them for a weighted average of the abundances at the neighbouring locations, giving the one contagion value per location i.

In a similar spirit to the contrasting of niche models and pure spatial models, we contrasted models that included related species as predictors with models containing random species as predictors. Species interactions (especially competition within a guild) are often cited as influencing species distributions (MacArthur, 1972). With birds, family membership is a reasonable approximation for a guild, as members of a family typically eat similar food in a similar fashion and thus should be more likely to share a similar distribution. Therefore, we hypothesized that using the abundances of species from the same family as predictors for a species' distribution would be more likely to lead to useful predictions than using a random selection of predictor bird species. An alternative hypothesis was that other bird species would work as predictors by randomly having a suitable spatial structure. In this case, family membership should not matter.

For the first type of model, we included the abundance of predictor species from the same family aiming at modelling positive or negative species interactions (RESP, for RElated SPecies). This approach led to a varying number of predictors for each species, ranging between 0 and 27 with a mean of 10.3 ± 0.62 (SE). The second type of model used a random selection of predictor species,

which may or may not be from the same family, rather than related species (RASP, for RAndom SPecies). It used exactly the same number of predictor species as the first type of model (RESP).

We used simulated data to determine the susceptibility of RTs to random effects, to validate our conclusions on the partitioning of explanatory power, and to gain further insights into the explanatory power of random spatial structure matches. We generated two types of simulated data with 190 variables each across the 1368 locations: complete spatial randomness (CSR) and simulated ranges that were constructed around randomly placed centres (SIM).

Including many independent variables in multiple regressions can lead to high R^2 values even if the variables are not connected to the dependent variable (Freedman, 1983). We used a completely spatially random data set (CSR) to investigate the susceptibility of the RT models to the selection of random predictors from a large data set. In addition, this complete random data set is most consistent with the classical meaning of a null hypothesis in the traditional statistical sense of H_0 or no effect, which in regression models would be the inclusion of an intercept only. The data were generated from a lognormal distribution, following closely the distribution of real abundances. In addition, we added similar numbers of zeros to the simulated data as were found in the real bird data. However, the random data and the zeros were randomly distributed across all locations, so that no spatial patterns or coherent ranges resulted.

The simulated ranges (SIM), in contrast, were closely modelled on the characteristics of real ranges. The idea was to create ranges that mimicked real ranges in characteristics and scale, but were not causally related to environmental conditions in any way. Trying to predict these artificial ranges with real environmental data could then tell us how much variability environmental predictors might be explaining through a random spatial match of structure.

We started out with random locations as centres. We then generated an artificial range with abundance decaying with distance from the centre following a Gaussian distribution (Brown *et al.*, 1995). We used a scaling factor [(sum of real range + a)/b] to model the sum of the abundance at all locations after the real bird ranges. The sigma of the Gaussian distribution was scaled to result in a similar number of occupied locations in the real and the simulated distributions [(number of occupancies + c)/d]. The simulated abundances were thus based directly on characteristics from the real bird ranges (number of occupied locations and total sum of abundance) and calculated according to the following formula:

$$\text{abundance of species } j \text{ at location } i = \frac{\text{sum.det}_j + a}{b} \frac{1}{\sigma\sqrt{2\pi}} \exp(-x_i^2/2\sigma^2)$$

with

$$\sigma = \frac{(\text{no.occ}_j + c)}{d}$$

where a, b, c and d are scaling parameters derived from empirical trials; x_i is the Euclidean distance between location i and the centre, π is the ratio of the circumference of a circle to its diameter, sum.det_j is the sum of all abundances for real bird species j, and no. occ_j is the number of occupied locations for real bird species j. The empirically derived values for a, b, c and d were 50, 0.0001, 220.6777 and 0.001505665, respectively.

In addition, we added random noise to the abundances and set all abundances < 0.2 to 0, to create a range boundary. With this method we created 190 random ranges that resembled the 190 original ranges very closely in the distribution of: (1) total sum of abundances; (2) standard deviation in abundances; (3) number of occupied sites; and (4) maximum abundance. The simulated abundances were square root transformed analogously to the real bird abundances.

RESULTS

We ran 13 models. Descriptions of the models and summarized results are in Table 1. We found that spatial interpolation (COORD and CONT) led to better predictive models than habitat-based models (HAB) and selecting species at random as predictor variables led to better models than species interaction models (RESP) based on related birds. While the 190 habitat models (HAB) based on climate and landcover had an average R^2 of 0.32 ± 0.015 (SE), the models that used coordinates as the only independent variables (COORD) resulted in an average R^2 of 0.36 ± 0.016, and the model that used the contagion term as the only independent variable (CONT) led to an average R^2 of 0.43 ± 0.014.

Combining the HAB and COORD models led to an intermediate R^2 of 0.34 ± 0.015. In multiple regressions, introducing additional independent variables cannot decrease the R^2. In regression trees this can happen because they are not forward-looking and do not select splits with consideration for future splits. Thus, if many additional variables are included that are decent predictors, such as the 27 environmental variables included here, even if slightly inferior to the original few predictors — here the two coordinates — they can by random chance capture splits by being the slightly better predictor in a specific situation but subsequently lead to slightly worse trees. If the additional variables were terrible predictors, however, RT would practically never pick them by random chance (see test results below). In the combined HAB and COORD model, environmental variables captured an R^2 of 0.25 compared with 0.09 captured by the coordinates. This result also indicates that the information content in environmental variables and coordinates with regard to bird distributions is very similar. Otherwise, one would expect that the addition of 27 environmental variables would add to the R^2 of the COORD model, by explaining different parts of the variability in the bird distributions. It appears that the two kinds of variables compete to explain the same part of the variability rather than complementing each other.

However, when the environmental variables were added to the CONT model, they were only able to capture an R^2 of 0.02 and the overall R^2 of this combined model was 0.42 ± 0.014 compared to 0.43 ± 0.014 in the CONT model alone. Despite the 27:1 ratio of environmental variables to contagion, environmental variables were not able to capture a substantial part of the explained variability or add to the predictive power of the CONT models.

Table 1 Results from the 13 regression tree models. SE is one standard error of the mean R^2 of 190 species models. The average number of variables included in the regression tree models is k, while p is the number of variables that were available to the models. The average number of locations included for the 190 species was 797 ± 26 (SE).

Dependent variable	Independent variable(s)	R^2	SE	k	p
Bird abundance	Environment	0.32	0.015	2.96	27
Bird abundance	Coordinates	0.36	0.016	1.56	2
Bird abundance	Contagion	0.43	0.014	0.95	1
Bird abundance	Coordinates + environment	0.34	0.015	3.05	29
Bird abundance	Contagion + environment	0.42	0.014	1.13	28
Bird abundance	Related birds	0.24	0.015	2.18	10.3
Bird abundance	Random birds	0.33	0.016	2.98	10.3
Bird abundance	Environment + related birds	0.38	0.015	3.59	37.3
Bird abundance	Coordinates + contagion	0.42	0.014	1.02	3
Bird abundance	Simulated ranges	0.09	0.009	1.29	10.3
Simulated ranges	Environment	0.24	0.011	2.03	27
Simulated ranges	Coordinates	0.42	0.008	1.93	2
Bird abundance	189 random variables	0.00	0.002	0.29	189

The abundances of related birds as independent variables (RESP) led to models with an average R^2 of 0.24 ± 0.015. Despite the fact that these abundances are rarely included in traditional habitat models, this R^2, when viewed alone, was large enough to lend support to the idea that species interactions might play an important role in driving species ranges. However, using the same number of species as independent variables per predicted bird, but picking these species randomly (RASP), rather than using family members, led to an increase in average R^2 to 0.33 ± 0.016.

Combining our two predictive models by adding the 27 environmental variables and the related bird species as independent variables in a single model led to an average R^2 of 0.38 ± 0.015, only slightly higher than the COORD null model and lower than the CONT null model. In contrast, adding coordinates to the contagion model did not improve the R^2 (0.42 ± 0.014, vs. 0.43 ± 0.014 for the contagion model only), indicating that the coordinates did not contain information important beyond what the contagion (averaged neighbourhood) could explain. This last fact suggests that the relevant spatial processes might be on the scale of a few hundred kilometres rather than a continental scale.

The simulated ranges, which were constructed spatially completely independently of real ranges and environmental data, gave an impression of how much variability could be explained by random coincidence of spatial structure alone without any functional connection. Using the same selection of the number of independent variables as in the related bird species approach, only this time selected from simulated ranges, the average R^2 was 0.09 ± 0.009. Conversely, using the simulated ranges as dependent variables and the unrelated environmental variables as independent variables led to an average R^2 of 0.24 ± 0.011. Using coordinates in models on simulated ranges led to an average R^2 of 0.42 ± 0.008. The increase in R^2 from models on real ranges using coordinates as predictors (0.36 ± 0.016) to models on simulated ranges is likely explained by the better spatial coherence and more regular shape of the simulated ranges. Conversely, the higher R^2 when real bird ranges were modelled by related birds, rather than simulated ranges (0.24 ± 0.015 vs. 0.09 ± 0.009), is likely due in part to species interaction and in part to natural ranges tracing bioregions, barriers and other natural structures that species may share without necessarily having direct interaction.

The RT models were not subject to Freedman's paradox (Freedman, 1983). Including 189 completely spatially random (CSR) variables as independent variables led to an average R^2 of 0.00 ± 0.002. The number of variables included did not in and of itself seem to have an effect on the average R^2. After all, including 190 random variables still led to an R^2 of 0.00 ± 0.002, and the contagion model with only one independent variable outperformed a combined habitat and related species model with an average number of roughly 37 independent variables. We included the number of available variables and the average number of variables included per model in Table 1.

Running the models without a range restriction on included locations increased the R^2, as expected, to 0.41 ± 0.015 for the HAB model, 0.47 ± 0.015 for the COORD model, 0.52 ± 0.013 for the CONT model and 0.43 ± 0.009 for the SIM model with environmental variables as predictors. Thus, the relationships between the models were virtually unchanged except for a dramatic increase in the predictive power of environmental variables on simulated ranges. However, as explained in the methods section, we do not find these values of R^2, based on a combination of modelling the range itself and modelling abundances within the range, as informative or interesting as the ability of the different models to predict within the range.

DISCUSSION

In this study we asked whether environmental variables could add predictive power to pure spatial interpolation at a coarse scale. We found no indication that environmental variables were able to add any predictive power above and beyond what spatial interpolation could provide. Spatial models had on average a higher R^2 than habitat-based models. While habitat variables

were somewhat competitive when compared with and added into a model with coordinates, suggesting that they convey very similar information in relation to species distributions, they were clearly outperformed by the fine-scale contagion variable. In a classic partitioning of variance approach (Borcard *et al.*, 1992), the 'environment' partition would have been zero, because the models combining spatial and environmental variables ended up having lower R^2 than the pure spatial models. (The reduction in R^2 from adding environmental variables is arguably an artefact of RTs, but the point remains that the environmental variables had no additional capacity to explain variability in the species distribution than what had already been captured in the spatial variables.)

Why did spatial interpolation outperform habitat models?

We cannot be certain why spatial interpolation outperformed habitat models. Possible explanations fall somewhere between the extremes of: (a) environmental variables had no explanatory power and were but poor surrogates for spatial variables, capturing spatial structure haphazardly; and (b) spatial variables were merely good surrogates for environmental variables that explained nothing but outperformed the included habitat variables because the measurement, scale and/or inclusion of environmental variables was poor. We believe that the explanation lies closer to (a) than (b) for the following reasons: (1) environmental variables performed well at predicting artificial ranges; (2) our set of included environmental variables met or exceeded the standards of other current distribution models at a coarse scale; and (3) the success of using randomly selected species as predictors for a target species' distribution also points to the importance of spatial structure to making good predictors.

Regarding point (1), environmental variables generated a substantial R^2 in predicting artificial ranges. This R^2 was about 75% of the R^2 generated by environmental variables explaining real bird ranges. The explanatory power of environmental variables on simulated ranges indicates that the spatial structure of environmental variables alone could account for three-quarters of the explanatory power of environmental variables on real ranges, independent of the actual biological implications of the environment. Not much explained variance would then be left for the alleged mechanistic function of environment on species distributions. In addition, climate follows natural structures such as mountain ranges and coasts, which species also follow, possibly at least in part independently of climate effects. This means that climate has an even higher potential to spatially predict species ranges without a causal connection in real ranges than in the simulated ranges. The 75% figure may thus be too low rather than too high.

On point (2), the comparatively low R^2 of the environmental models could be partly caused by an inadequate set of included variables. We have no defence against such a claim other than that we included all the types of variables typically included in habitat models and more. In addition to land cover, which accounted for only 0.4% of the explained variability in the environmental models, we included all the usual measures of climate (precipitation, temperatures and their extremes), which accounted for 73.4% of the explained variability. We also included within-year and between-year variability in climate, which accounted for 25.0% of the explained variability, and the NDVI index of productivity, which accounted for 1.2% of the explained variability. The latter two groups of variables are an addition to typical habitat models that performed well (e.g. Bartlein *et al.*, 1986; Austin *et al.*, 1990). Thus, while we cannot exclude the possibility that spatial models performed better than environmental models because the environmental variables were of poor quality and/or poorly selected, we can at least claim that that would then also be the case for most, if not all, other published habitat-based distribution models at such a coarse scale.

On point (3), we note that the predictive abilities of other species as independent variables may have little to do with 'species interactions'. Indeed, this is made likely by the fact that randomly chosen species outperformed closely related species in prediction. Alternative explanations could be that other species are good integrators over environmental and other abiotic conditions and thus make for good predictors, or that other species offer a wide range of spatial structures that lend themselves to modelling other species distributions in a haphazard way (enhanced by being realized on the same physical landscape with barriers, eco regions and other structural–spatial features). If the former were true, we would have expected that related species work better as predictors than random species, because they should have more similar environmental preferences. Therefore, we conclude that the latter hypothesis has some merit. Also, if other species were simply good integrators over the environment, we would have expected no increase in R^2 when environmental variables were included as predictors in a model that already contains random predictor species (observed increase from 0.33 to 0.42).

Regression tree models

We have great confidence in the regression tree models we used for automated variable selection and creation of the 11 models for 190 species. Even when offered a high number of random variables, the models did not construct spurious relationships or R^2 distinguishable from zero. The objective variable selection of the RTs ensured that we did not influence the performance of different kinds of models through our own knowledge and perceptions of ecological phenomena. The 10-fold cross-validation ensured that the pruned RTs were stable and that the listed R^2-like measures were based on data not included in the creation of the models. For our comparative purposes this modelling strategy was ideal.

A criticism of RT models is that they do not look ahead, meaning that the first split is selected purely by its performance in reducing the variance but does not take into account how well the following splits perform. In practice, this could mean that the initial best split sets up the two groups of data so that only poor splits or none can follow and the resulting tree could be inferior in total explained variability to a tree started with the second best or worse split than the first split but followed by a strong tree.

Because the possibilities of permutation grow exponentially with tree size, building 'ahead-looking' trees that take into consideration more than the best split are computationally too costly.

The reduction in R^2 from the spatial models to models including spatially explicit variables and environmental variables has to be understood in the light of the inability of the regression trees to look ahead. While the environmental variables performed slightly worse than the spatial variables, they were more numerous and thus had a better chance of randomly having a slight edge over the spatial ones for one particular split. However, they would then set the tree up for an overall loss in R^2, compared to a scenario in which it is run with only the spatial variables.

An additional caveat in the use of RT in the context of distribution modelling is that the variance calculation underlying the split algorithm does not take spatial autocorrelation into account. Thus, the variable and split selection could be subject to red-shift (Lennon, 2000). In previous unpublished work, V.B. has observed that autocorrelation in the residuals of RT models is lower than in the residuals of multiple linear regression models based on the same variables. The reason is likely to be that the data partitioning method for the RTs allows them to geographically subdivide the data based on spatially implicit information contained in environmental variables even if no spatially explicit variables are included. In addition, a red-shift would rather favour environmental variables (Lennon, 2000) than bias against them, so that our results on the predictive power of habitat models should, if anything, be biased in their favour.

Caveats

An important aspect of our work is scale. Scale has been shown to influence virtually all aspects of ecology, and species distributions are no exception. It is rightly thought that distributions at a fine scale are determined by the most immediate requirements of species survival and reproduction: nutrients, space, shelter, etc. At the coarser scale at which our study was set such direct requirements give way to more indirect, large-scale variables such as climate (Thuiller *et al.*, 2003). Our study is valid only in the context of such a coarse scale and the alleged working of coarse-scaled variables.

Similarly, our study targeted only common species (> 200 locations of occurrence) for statistical reasons. Rarer species may have more narrow niches and thus lend themselves better to environmental modelling than to spatial interpolation due to their sparse distribution. Thus, an extrapolation of our results to rare species, different scales, different geographical areas or other organisms has to be conducted with the scrutiny and reservation that any form of extrapolation deserves.

Last but not least, any form of spatial interpolation for prediction is based on pre-existing local information on the abundance (or presence) of the dependent variable, here the abundance of the bird species. For example, the range of the neighbourhood we selected for the contagion variable was 200 km, based on the authors' experience with the data set. The contagion model has zero predictive capability beyond these 200 km, while the environmental model presumably retains predictive capabilities beyond this distance. In particular, an extrapolation in geographical space soon renders the spatial interpretation models useless, while the environmental model may perform decently if the geographical extrapolation remains within the range of environmental conditions covered by the model. This effect is continuous: with everything else being equal, the sparser the existing coverage of sample locations for the dependent variable, the worse the spatial interpolation will perform.

Spatial patterns

Given the success of the spatial interpolation models, it is worth looking more closely at the biological mechanisms enabling them. Some researchers have hypothesized that spatial autocorrelation, or the influence of neighbours, is caused by dispersal *sensu latu*, leading to a connection between neighbouring populations (Selmi & Boulinier, 2001; Trenham *et al.*, 2001; Bahn *et al.*, 2006). However, such a functional connection is neither widely accepted by ecologists nor currently widely used in distribution models. Quite a number of studies use spatial models, including either coordinates or local neighbourhood for species distribution models (e.g. Augustin *et al.*, 1996; Araújo & Williams, 2000; Lichstein *et al.*, 2002). However, the most cited reason for using spatially explicit models is to address the statistical problems stemming from the lack of independence in residuals and less often to recover an ecological mechanism. In short, spatial autocorrelation is often treated as a nuisance rather than a source of additional predictive power.

We concur with colleagues who use spatially explicit models and thus include spatial autocorrelation as an additional strong predictor, an approach long adopted by mining engineers who use kriging as a predictive tool (Costanza & Ruth, 2001). Ignoring spatial autocorrelation or only acknowledging its effect on the estimation of degrees of freedom seems to squander a very important source of information, as we demonstrated in this research. Specifically, our work puts a challenge out to ecologists and distribution modellers: can ecological mechanisms be used to create better models than merely copying the abundance or presence/absence value of a neighbouring site 50–100 km away? The success of coordinates and contagion relative to environment in modelling species distributions suggests that we are missing fundamental ecological elements underlying distributions when we model them using environmental conditions only. Moreover, our results suggest that conservation planners could do better by simply sampling the species very sparsely (as is already done for input to habitat models) and creating a smoothed surface from the results. If resources allow, a combination of both approaches in spatially explicit models, such as general linear models with a neighbourhood-based covariance matrix in the errors, or conditional autoregressive models (CARs), would be ideal. The success, however, would depend on the careful selection of variables. In our work here, such a combined model did not perform better than the contagion model alone.

Our research cannot explain the phenomenon of strong spatial autocorrelation, which is so useful in prediction. However, several plausible processes leading to such strong spatial structures beyond what can be explained by the environment are

known or can be hypothesized. For example, it is possible that population processes such as dispersal give species more spatial structure than the environment. At the same time, environmental dynamics and lack of equilibrium in the distribution with the environment could further weaken the niche approach. In other words, some of the same dynamics (e.g. cycles of good and bad years, dispersal and dispersal limitation, lags in population growth and recolonization) that decrease the predictive power of environmental conditions may increase autocorrelation in distributions and perpetuate a strong peak and tail distribution pattern at the range scale. Instead of ignoring these patterns or treating them as a nuisance that invalidates statistical analyses (by introducing dependence in the residuals), we suggest that they should be viewed as an opportunity to create better distribution models and to make progress in the as yet underappreciated field of population dynamics in species distributions.

Conclusion

Our study is not to be misunderstood as a condemnation of habitat-based distribution modelling. Initiatives such as the GAP programme (Scott *et al.*, 1993) are valuable contributions to conservation efforts, working in an emergency triage environment where any result is better than none. Most often the information available for species is much poorer than the data we worked with; information on abundances covering large extents and with good spatial coverage as used in our research is a rare exception. Often only a few museum-collection data points are available (Graham *et al.*, 2004). Therefore, rather than condemning habitat-based distribution modelling, our point was to deepen our understanding of the successes and failures of habitat models and the relative importance of the mechanisms driving species distributions.

In conclusion, the predictive power of habitat-based models at a coarse scale may be in substantial part due to coincidence in spatial structure between habitat variables and species distributions rather than a functional relationship. Other spatially structured predictors, such as a random selection of other birds as predictor variables, reach similar levels of predictive power as environmental variables, and variables capturing spatial structure directly outperform habitat variables. The future lies in better understanding why spatial patterning in species distributions is such a strong predictor, which processes lead to the spatial patterning, and how these processes can be modelled mechanistically rather than phenomenologically.

ACKNOWLEDGEMENTS

We thank Deanna Newsom and five anonymous referees for reviewing and improving the manuscript and the many thousands of volunteer observers and organizers who contributed to the BBS data under the auspice of the US Geological Survey's (USGS) Patuxent Wildlife Research Center and the Canadian Wildlife Service's National Wildlife Research Center. Both authors thank NSERC and McGill University for supporting this research.

REFERENCES

Andrewartha, H.G. & Birch, L.C. (1954) *The distribution and abundance of animals*. University of Chicago Press, Chicago.

Araújo, M.B. & Williams, P.H. (2000) Selecting areas for species persistence using occurrence data. *Biological Conservation*, **96**, 331–345.

Augustin, N.H., Muggleston, M.A. & Buckland, S.T. (1996) An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology*, **33**, 339–347.

Austin, M.P., Nicholls, A.O. & Margules, C.R. (1990) Measurement of the realized qualitative niche: environmental niches of five eucalyptus species. *Ecological Monographs*, **60**, 161–177.

Bahn, V., O'Connor, R.J. & Krohn, W.B. (2006) Importance of spatial autocorrelation in modeling bird distributions at a continental scale. *Ecography*, **29**, 835–844.

Bakkenes, M., Alkemade, J.R.M., Ihle, F., Leemans, R. & Latour, J.B. (2002) Assessing effects of forecasted climate change on the diversity and distribution of European higher plants for 2050. *Global Change Biology*, **8**, 390–407.

Bartlein, P.J., Prentice, I.C. & Webb, T., III (1986) Climatic response surfaces from pollen data for some eastern North American taxa. *Journal of Biogeography*, **13**, 35–57.

Borcard, D., Legendre, P. & Drapeau, P. (1992) Partialling out the spatial component of ecological variation. *Ecology*, **73**, 1045–1055.

Breiman, L. (1984) *Classification and regression trees*. Wadsworth International Group, Belmont, CA.

Brown, J.H. (1995) *Macroecology*. University of Chicago Press, Chicago.

Brown, J.H., Mehlman, D.W. & Stevens, G.C. (1995) Spatial variation in abundance. *Ecology*, **76**, 2028–2043.

Costanza, R. & Ruth, M. (2001) Dynamic systems modeling. *Institutions, ecosystems, and sustainability* (ed. by R. Costanza, B.S. Low, E. Ostrom and J. Wilson), pp. 21–29. Lewis Publishers, Boca Raton, FL.

Cox, D.R. & Wermuth, N. (1992) A comment on the coefficient of determination for binary responses. *American Statistician*, **46**, 1–4

Diniz-Filho, J.A.F., Bini, L.M. & Hawkins, B.A. (2003) Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology and Biogeography*, **12**, 53–64.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, T.A., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Ferrier, S. (2002) Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology*, **51**, 331.

Freedman, D.A. (1983) A note on screening regression equations. *American Statistician*, **37**, 152–155.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A.T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, **19**, 497.

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

Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

Hawkins, B.A., Porter, E.E. & Diniz-Filho, J.A.F. (2003) Productivity and history as predictors of the latitudinal diversity gradient of terrestrial birds. *Ecology*, **84**, 1608–1623.

Hijmans, R.J. & Graham, C.H. (2006) The ability of climate envelope models to predict the effect of climate change on species distributions. *Global Change Biology*, **12**, 2272–2281.

Iverson, L.R. & Prasad, A.M. (1998) Predicting abundance of 80 tree species following climate change in the eastern United States. *Ecological Monographs*, **68**, 465–485.

Keitt, T.H., Bjornstad, O.N., Dixon, P.M. & Citron-Pousty, S. (2002) Accounting for spatial pattern when modeling organism–environment interactions. *Ecography*, **25**, 616–625.

Legendre, P. & Legendre, L. (1998) *Numerical ecology*, 2nd English edn. Elsevier, Amsterdam.

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

Lichstein, J.W., Simons, T.R., Shriner, S.A. & Franzreb, K.E. (2002) Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs*, **72**, 445–463.

MacArthur, R.H. (1972) *Geographical ecology: patterns in the distribution of species*. Harper & Row, New York.

Manel, S., Williams, H.C. & Ormerod, S.J. (2001) Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, **38**, 921–931.

McPherson, J.M., Jetz, W. & Rogers, D.J. (2004) The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, **41**, 811–823.

New, M., Hulme, M. & Jones, P. (1999) Representing twentieth-century space-time climate variability. Part I: Development of a 1961–90 mean monthly terrestrial climatology. *Journal of Climate*, **12**, 829–856.

Pereira, J.M.C. & Itami, R.M. (1991) GIS-based habitat modeling using logistic multiple-regression: A study of the Mt Graham red squirrel. *Photogrammetric Engineering and Remote Sensing*, **57**, 1475–1486.

R Development Core Team (2005) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna (http://www.r-project.org/).

Ripley, B.D. & Rasson, J.P. (1977) Finding the edge of a Poisson forest. *Journal of Applied Probability*, **14**, 483–491.

Scott, J.M. & Csuti, B. (1997) Gap analysis for biodiversity survey and maintenance. *Biodiversity, II. Understanding and protecting our biological resources* (ed. by M.L. Reaka-Kudla, D.E. Wilson and E.O. Wilson), pp. 321–340. Joseph Henry Press, Washington, DC.

Scott, J.M., Anderson, H., Davis, F., Caicoo, S., Csuti, B., Edwards, T.C., Jr, Noss, R., Ulliman, J., Groves, C. & Wright, R.G. (1993) Gap analysis: a geographic approach to protection of biological diversity. *Wildlife Monographs*, **123**, 1–41.

Scott, M.J., Heglund, P.J., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A. & Samson, F.B. (2002) *Predicting species occurrences: issues of accuracy and scale*. Island Press, Washington, DC.

Segurado, P., Araújo, M.B. & Kunin, W.E. (2006) Consequences of spatial autocorrelation for niche-based models. *Journal of Applied Ecology*, **43**, 433–444.

Selmi, S. & Boulinier, T. (2001) Ecological biogeography of Southern Ocean islands: the importance of considering spatial issues. *The American Naturalist*, **158**, 426–437.

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

Therneau, T.M. & Atkinson, E.J. (1997) *An introduction to recursive partitioning using the rpart routines*. Department of Health Science Research, Mayo Clinic, Rochester, MN. (Available at http://mayoresearch.mayo.edu/mayo/research/biostat/splus-functions.cfm; accessed 21 February 2007.)

Thomas, C.D., Cameron, A., Green, R.E., Bakkenes, M., Beaumont, L.J., Collingham, Y.C., Erasmus, B.F.N., de Siqueira, M.F., Grainger, A., Hannah, L., Hughes, L., Huntley, B., van Jaarsveld, A.S., Midgley, G.F., Miles, L., Ortega-Huerta, M.A., Peterson, A.T., Phillips, O.L. & Williams, S.E. (2004) Extinction risk from climate change. *Nature*, **427**, 145.

Thuiller, W., Araújo, M.B. & Lavorel, S. (2003) Generalized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science*, **14**, 669–680.

Trenham, P.C., Koenig, W.D. & Shaffer, H.B. (2001) Spatially autocorrelated demography and interpond dispersal in the salamander *Ambystoma californiense*. *Ecology*, **82**, 3519–3530.

Zar, J.H. (1996) *Biostatistical analysis*, 3rd edn. Prentice Hall, Upper Saddle River, NJ.

BIOSKETCHES

Volker Bahn is interested in macroecological patterns and their causes. In particular, the processes underlying the distribution of species and the statistical techniques for elucidating these processes are a focus of his research. He aspires to make his work relevant to the fields of ecology, conservation biology and conservation management.

Brian McGill is interested in a variety of ecological topics on large spatial, time or taxonomic scales. Recently he has explored the causes and implications of patterns in the geographical ranges of species, traditional community ecology questions such as the causes of the species abundance and topics in macroevolution/palaeoecology.

Editor: Jack Lennon