ISEIS

Journal of Environmental Informatics 17(2) 91-101 (2011)

Journal of Environmental Informatics

www.iseis.org/jei

A Robust Test of Spatial Predictive Models: Geographic Cross-Validation

D. J. Lieske[1,2*] and D. J. Bender[1]

[1]Department of Geography, University of Calgary, 2500 University Dr. NW, Calgary, Alberta T2N 4N1, Canada
[2]Department of Geography and Environment, Mount Allison University, 144 Main Street, Sackville, New Brunswick E4L 1A7, Canada

Received 25 October 2009; revised 27 April 2011; accepted 30 May 2011; published online 15 June 2011

ABSTRACT. Predictive modeling is an important tool for identifying areas for conservation prioritization. But the reliability of any model depends on how well its predictions can be generalized beyond the area surveyed. Recent work points to the potential for enhancing predictive power by incorporating such spatial processes as autocorrelation or the influence of location, so this study addressed two questions: (1) what affect does model complexity, spatial autocorrelation and spatial location have on model accuracy? (2) how generalizable are different methods when applied to new geographic test regions? On average, predictive power declined 22.7% ± 2.7% SE when models were used to predict occurrences in "unsampled" geographic test regions. Overall variability in performance depended on the method used. AUTO and GAM models tended to be amongst the least variable, but results depended upon species. Our results suggest that models with complex functional relationships between the response and predictor variables (such as GAMs fit with up to 5 knots) tended to either improve accuracy, or perform more consistently across species, but not both at the same time. In general, it is very difficult to accurately extrapolate model predictions into unsampled geographic areas. However, we found that habitat specialists such as the Sedge Wren were consistently well predicted, regardless of method, and that autocorrelated regression (using a Gibbs sampler and simulation of presence/absence) could be more reliably generalized for species showing strong social structure (e.g., patchiness). GWR was especially sensitive to the plots used to train the model.

Keywords: geographic cross-validation, predictive models, generalizability, accuracy

1. Introduction

An important consideration in assessing the accuracy and reliability of any predictive model is the generalizability of its predictions across a range of novel conditions. One way to assess this is to use a completely new set of independent data (Fielding and Bell, 1997; Justice et al., 1999), but as pointed out by Vaughan and Ormerod (2005), such test data may be logistically unfeasible to gather, or impossible in the case of retrospective analyses (Araujo and Guisan, 2006). Plus, proper test data must be representative and of sufficient sample size (Vaughan and Ormerod, 2005). In these situations one solution is to use resampling methods (Verbyla and Litvaitis, 1989; Vaughan and Ormerod, 2005). Regardless of the origin of the test data (newly gathered vs. resampled from the same data set) we must bear in mind that all accuracy assessments are provisional tests of natural systems which are likely to be in an open and non-equilibrial state (Oreskes et al., 1994; Justice et al., 1999).

Generalizability has often been assessed through the use of training data cross-validated by a randomly-selected set of test points (Randin et al., 2006). But as pointed out by Fielding and Haworth (1995), true predictive generality stems from models which are capable of extrapolating beyond the geographic range used to train the model, something that is rarely ever assessed (Guisan et al., 2006; Randin et al., 2006). This is a much stricter and potentially more informative test of generalizability than other kinds of validation, as biases in the model may only surface when attempts are made to apply the model beyond the area or years used to train the model (Vaughan and Ormerod, 2003).

Spatial predictive modelling is a key component of many research and applied conservation programs, where the output is often a series of maps showing preferred habitat or priority areas for species of interest. Given the geographic nature of the problem ('where is the species most abundant?') it should be unsurprising that a number of recent papers have drawn attention to the way that complicated spatial processes can interact to influence species distribution (Lieske and Bender, 2009; Kissling and Carl, 2008; Diniz-Filho, 2008). For example, in a survey by Dormann (2007) the importance of the presence (or abundance) of neighbouring locations in determining presence at any given location was highlighted, illustrating the widespread importance of spatial autocorrelation (Legendre, 1993). In addition to autocorrelation, model relationships do not necessarily remain constant over the entire region, a phenomenon known as non-stationarity (Foody, 2004; Fortin and Dale, 2005; Jetz et al., 2005). When location is important, we expect global

[*] Corresponding author. Tel.: +1 506 3642315; fax: +1 506 3642625.
E-mail address: dlieske@mta.ca (D.J. Lieske).

ISSN: 1726-2135 print/1684-8799 online
doi:10.3808/jei.201100191

models to mask potentially important and informative local variations in response (Fotheringham et al., 2002).

Our study presents the novel application of a new model validation method, which we call ‘geographic cross validation’, to perform a comparative assessment of the predictive accuracy and generalizability of a suite of important species distribution modeling methods. This is the first time these methods have been evaluated and compared in this way. Through the use of a set of covariates for land cover, climate, and elevation – as well as information about spatial structure – we produced a series of predictive models. We postulated that spacing behaviour (spatial autocorrelation) might exert a measurable effect at the finest scales of this study and expected that incorporating this additional information – in the form of a spatial autologistic regression model (AUTO) – might improve predictive accuracy. Furthermore, given the large geographical area of this study it is possible that local-scale adaptation could bring about geographically-distinct species responses. Under these circumstances we would expect geographically-weighted regression (GWR) potentially well-suited to capture this variability and improve predictive accuracy. More generally, and given the concern that complex models are more vulnerable to overfitting the data at hand (Harrell et al., 1996; Justice et al., 1999; Fielding, 2002; Randin et al., 2006), we also assessed the overall impact of model complexity.

2. Methods

2.1. Geographic Cross Validation

This method requires the geographic partitioning of the study area into sub regions, which we suggest is most relevant for large study areas likely to encompass sufficient geographic variation to warrant this approach. In this study, eight geographic sub-regions were used, each of which was approximately evenly sampled (Figure 1). We then performed an 8-fold cross-validation, with each single sub-unit being successively reserved to test the predictive accuracy of models generated using the data from the remaining seven regions (the training set). As all model predictions were probabilities of occurrence at individual locations, this introduced a threshold dependency in deciding the cutoff point for determining whether test points should be expected to contain an occurrence for that species (Fielding and Bell, 1997). Predictive accuracy, therefore, was assessed using the area under the receiver operating characteristic curve (ROC curve, Zweig and Campbell, 1993; Fielding and Bell, 1997), which avoids the threshold problem by integrating across all combinations of possible thresholds.

The area under the curve (AUC) of the ROC curve represents the proportion of cases in which the model predictions are consistent with the observed test points (where model predictions are higher for presence points than absence points), with a value of 0.50 indicating a model no more capable of predicting occurrence/absence than random chance. As pointed out by Elith et al. (2006), values less than 0.50 indicate models which actually perform poorer than random prediction. The AUC of the ROC curve is especially suitable when the main goal (as in our study) is to assess the ability of model predicttions to be used to rank the relative importance of landscape units (Pearce and Ferrier, 2000; Pearce et al., 2001).

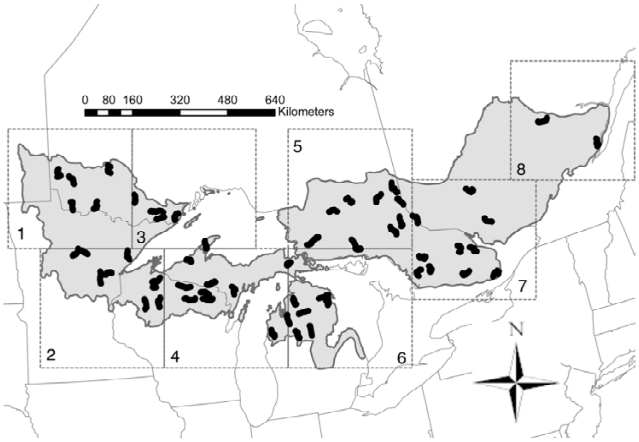

Figure 1. Delineation of the boreal-hardwood transition zone (shaded area), which encompasses portions of Southern Manitoba, Ontario and Quebec as well as Northern Minnesota, Wisconsin, and Michigan (approx. 618,000 km^2). Also indicated are the 56 Breeding Bird Survey routes used in the study (black dots), as well as the geographic test regions (dotted lines, numbered one through eight).

2.2. Study Area and Land Cover Information

The 618,000 km^2 study region (Figure 1) constitutes a transition zone between mixed hardwood and boreal forest, and is heavily influenced by the presence of the Great Lakes (Ontario Partners, 2006). The forest communities of this region represent a heterogeneous mix of oaks, maples, birch and pines in the southern portions of the region, shifting to coniferous species in the more northern, boreal portions (Ontario Partners, 2006).

The Moderate Resolution Imaging Spectroradiometer (MODIS) of the NASA Earth Orbiting System (Friedl et al., 2002; Huete et al., 2002) provided an index of vegetation greenness (the Enhanced Vegetation Index, or EVI) as well as a supervised land cover classification image. Preliminary analysis suggested that an average EVI based on a 3 km × 3 km neighbourhood (EVIMEAN) was a better predictor than the local EVI value (Lieske and Bender, 2009). An additional advantage to using this neighbourhood-based approach was that it allowed us to calculate an index of coarse-scale habitat heterogeneity, approximated by the standard deviation of EVI values in the 3×3 window (EVISD). Due to the rarity of many land cover classes, categories were regrouped, and only the major ones retained for model building (at 1-km resolution): conifer-dominated forest (CONIFER), cropland/vegetation mosaic (CROPVEG), deciduous-dominated forest (DECID), and mixed (conifer-deciduous) forest (MIXEDF). These four land cover classes were combined with a default class (OTHER) to produce a dummy-coded omnibus variable (LANDCOV). Climatic measurements were obtained from the global climate data of Mitchell and Jones (2005), for the years 1997 to 2002. Considerations of multicollinearity forced us to retain only three of the variables in that dataset (Lieske and Bender, 2009): mean monthly diurnal temperature range (DTR, in □); total annual precipitation (PRECIP, in mm), and mean monthly temperature (TEMP, in □). Finally, we used a 1-km resolution elevation dataset obtained from the GTopo30 global digital elevation model of the U.S. Geological Survey's EROS Data Center in Sioux Falls, South Dakota (U.S. Geological Survey, 1996). The grid is approximately 1-km resolution, and resulted in the elevation variable ELEV (in m).

2.3. Species Occurrence Data

Based on Partners in Flight (PIF) ranking of conservation priority, we chose four species of sufficient importance to place them on the PIF Watch List (Rich et al., 2004): Blackburnian Warbler *Dendroica fusca*, Canada Warbler *Wilsonia canadensis*, Purple Finch *Carpodacus purpureus*, and Sedge Wren *Cistothorus platensis*. Additionally, we included the American Crow *Corvus brachyrhynchos,* a breeding species which is common and widespread throughout the study area. Species occurrence data was obtained from the North American Breeding Bird Survey (BBS), a monitoring project initiated in 1966 (Robbins et al., 1986). While primarily intended to detect long-term trends in species abundance, individual volunteer surveys consist of 50 3-min. stop point observations (0.8 km apart) along a defined route and hence, contain valuable spatial information. For this study, species count data (at the level of the individual stop point) was reclassified as "used" when non-zero counts were noted across any of 7 years (from 1997 to 2003). Precisely georeferenced stop points were available for only 7 routes in the study area, so we were forced to employ a linear referencing operation in ArcGIS (Environmental Systems Research Institute, 2002) to subdivide individual routes to obtain a larger sample of stop locations. This resulted in *n* = 56 routes (2799 stop points).

2.4. Modelling Methods: Generalized Linear and Generalized Additive Models

Global models (estimated using all locations simultaneously) were estimated as either generalized linear models (GLMs; McCullagh and Nelder, 1999) or generalized additive models (GAMs; Hastie and Tibshirani, 1990), and were implemented within the freely available *R* Statistical Package (Ihaka and Gentleman, 1996). GLM models were of the following general form:

$$\log it(Y) = \log\left(\frac{P(Y)}{1-P(Y)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \tag{1}$$

where the log-link of species occurrence was modeled as a combination of a y-intercept term (β_0) and $\beta_1 \ldots \beta_p$ globally estimated coefficients for $X_1 \ldots X_p$ covariates (see Section 2.2). We used two different GLM model formulations: GLM1, in which all relationships between species occurrences and candidate predictor variables were assumed to be simple linear trends (1 degree of freedom), and GLM2 where unimodal relationships were modeled as quadratic polynomials (2 degrees of freedom).

"Simple" GAM models (GAM1) closely mirrored the GLM2 models in terms of the degrees of freedom allocated to model relationships; while "complex" GAMs involved spline smoothing with up to 5 knots in order to accommodate strongly non-linear relationships (GAM2):

$$\log it(Y) = \log\left(\frac{P(Y)}{1-P(Y)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j (X_j) \tag{2}$$

where the log-link of species occurrence was modeled as a combination of a y-intercept term (β_0) and $\beta_1(X_1) \ldots \beta_p(X_p)$ globally estimated smoothing functions of the $X_1 \ldots X_p$. covariates (see Section 2.2). We used two different GAM model formulations: GAM1, in which all relationships between species occurrences and candidate predictor variables were modeled with the same degrees of freedom as used in GLM2, and GAM2 where 5 degrees of freedom were allocated to all smoothing functions.

As the aim of this study was to compare the accuracy and performance of a number of predictive models, our goal was to not to exhaustively explore alternative model structures but to objectively select a reasonable base specification for comparing each of the methods. We used an all-combinations procedure to identify models with the lowest Akaike Information Criterion (AIC). This approach avoids hypothesis testing and makes use of information theory to identify plausible models while guarding against the tendency for models to retain variables which provide only marginal improvements in information content. We found the all-combinations approach to be a practical, robust, and objective way to produce this reduced set of variables, provided that we confined variable selection to a computationally simpler algorithm (GLM2). To identify this baseline model structure we fit GLM2 regressions using all combinations of predictor variables. Another important advantage of this approach was that it allowed us to avoid the potential vagaries of stepwise model selection. For instance, an important criticism of stepwise model selection is that important combinations of variables can escape consideration due to the premature discard of key variables in earlier model selection steps. Interactions were not tested.

We used the simple (but objective) rule of choosing the final set of predictor variables that resulted in the GLM2 regression with the lowest AIC value, recognizing that: (1) it is the relative difference in AIC values that is important, not the absolute values, and (2) some alternative specifications of predictor variables resulted in models that were virtually indistinguishable in terms of relative AIC differences. The final set of predictor variables defined the base specification that was used to build all subsequent distribution models. In this way we were able to eliminate variability attributable to differences in model selection procedures, and to focus on the head-to-head performance of each of the methods. We advise practitioners who are applying one of the modelling methods in isolation, and who have recourse to sufficient time and computational resources, to consider the use of bootstrapping to assess the relative importance of predictor variables (Harrell, 2001) or calculate model-averaged estimates for each parameter (Diniz-Filho et al., 2008; Burnham and Anderson, 2002).

2.5. Modelling Methods: Autologistic and Geographically-weighted Regression.

Proximity (autocorrelation) was incorporated by extending the GLM2 model estimated above, through inclusion of a spatially-lagged autocovariate term (AUTO):

$$\log it(Y) = \log\left(\frac{P(Y)}{1-P(Y)}\right) = \beta_0 + \gamma AUTO + \sum_{j=1}^{p} \beta_j X_j \qquad (3)$$

where the log-link of species occurrence was modeled as a combination of a y-intercept term (β_0), a globally estimated autocorrelation term (γ), and $\beta_1 \ldots \beta_p$ globally estimated coefficients for $X_1 \ldots X_p$ covariates (see Section 2.2). The AUTO covariate was the product of an $n \times n$ weights matrix (W) and an $n \times 1$ binary vector (y, a 1/0 dummy variable indicating presence or absence at neighbouring points). A simple weighting function was applied to all points:

$w_{ij} = 1/8$, with each neighbouring presence point weighted equally

= 0, otherwise

We adopted the Markov Chain Monte Carlo Gibbs Sampler (as described below) of Augustin et al. (1996, 1998) to enable predictions to be made throughout the entire surface and not just at sampled locations ($T = 11$ iterations). Initial “presences” (the prior knowledge) were simulated based on the starting predictions obtained from the GLM2 model (Section 2.4, Lieske and Bender, 2009) and the autocovariate estimated in subsequent iterations:

1. GLM2 (without an autocovariate) was used to produce the initial set of probabilities of occurrence;

2. A random number generator, drawing from a Bernoulli distribution, was used to simulate presences for unsampled grid locations;

3. The logistic regression was recomputed, this time with the autocovariate term included;

4. The random number generator was re-applied to simulate presences for unsampled grid locations;

5. A Gibbs Sampler was applied, with unsampled points chosen at random (one at a time), the autocovariate re-calculated, the conditional probability of occurrence recomputed, and a new random number generator applied to that point. With each iteration of the Gibbs Sampler, the probabilities of occurrence at any given point were progressively updated (given the conditional dependence on the neighbours), ultimately resulting in a model of the joint distribution of all grid points (Augustin et al., 1998).

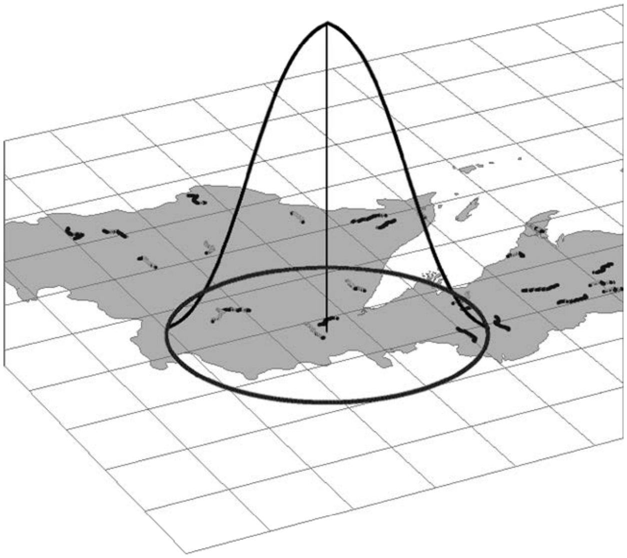

Figure 2. Illustration of the geographically-weighted regression (GWR) framework used in this study. At each location the response variable was modelled as a function of a limited set of observations (defined by the kernel radius, which was ¼ of the study area or about 500 km), each of which was differentially weighted as a continuously decaying function of distance from the kernel centre.

Location (non-stationarity) was modelled using the binomial GWR framework of Fotheringham et al. (2002). Using this approach, parameters were estimated at each sample location using the local neighbourhood of observations, each of which was differentially weighted as a continuously decaying function of distance from the center (Fotheringham et al., 2002; Figure 2):

$$\log it(Y) = \log\left(\frac{E(Y)}{1-E(Y)}\right) = \beta_0(x,y) + \sum_{j=1}^{p} \beta_j(x,y) X_j \qquad (4)$$

where $\beta_0(x, y)$ is a locally estimated y-intercept term, and $\beta_1(x, y) \ldots \beta_p(x, y)$ are locally estimated coefficients for $X_1 \ldots X_p$ covariates (Section 2.2). In general, for n samples there will be n parameter estimates, each a function of location (Cartesian x and y coordinates). Due to computational demands associated with the use of the adaptive kernel, we used a fixed Gaussian radius of 1/4 of the width of the study area (about 500 km). We estimated these parameters using the code originally implemented in the R statistical language by C. Brunsdon.

3. Results

3.1. Variable Selection

The results of the variable-selection process (based on all-combinations variable selection) are shown in Table 1. Mo-

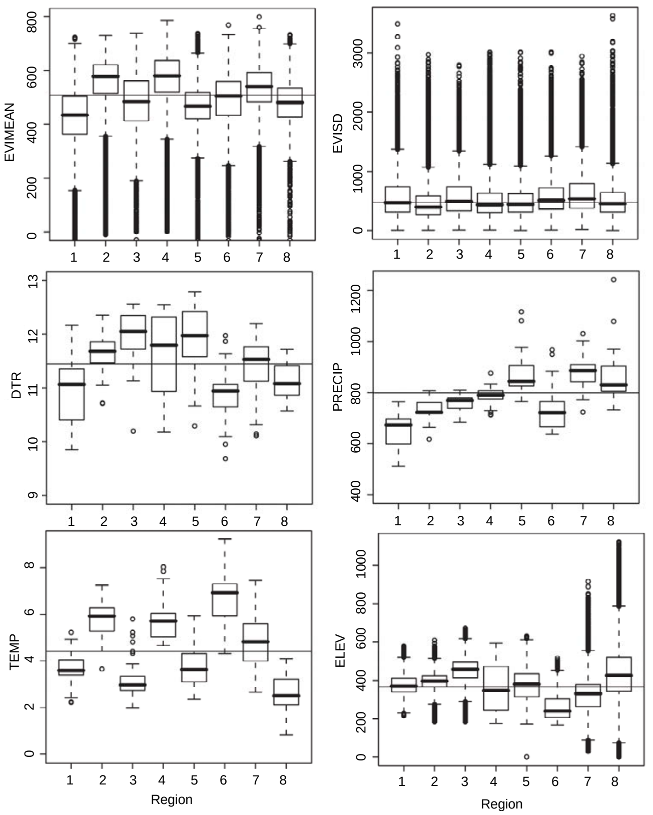

Figure 3. Environmental characteristics of the eight sub-regions used to perform the geographic cross-validation. Six continuous environmental predictor variables were examined: mean enhanced vegetation index for a 3 km × 3 km spatial neighbourhood (EVIMEAN); standard deviation of enhanced vegetation indices for the same spatial neighbourhood (EVISD); average diurnal temperature range (DTR, in □); annual precipitation (PRECIP, in mm); mean monthly temperature (TEMP, in □); and elevation (ELEV, in m). See Figure 1 for the locations of each of the sub-regions.

st of the variables were retained for the American Crow, Blackburnian Warbler and Sedge Wren. In the case of the Canada Warbler, LANDCOV and EVI-based variables (e.g. EVIMEAN, EVISD) were dropped from the final model, and simpler linear relationships to PRECIP and ELEV were favoured over higher order polynomials (PRECIP2 and ELEV2). Similarly for the Purple Finch, EVI-based variables were dropped, as was the PRECIP and higher order ELEV2 variables. LANDCOV was retained in the final model.

Predictions from the species distribution models were used to produce predictive occurrence maps, which we illustrate for each species (Figures A1 to A3). In all cases, predicttions were binned into five categories (quintiles), which improved the readability of the map and simplified comparison of the distribution of areas most likely to be used by each species. Some of the impressions conveyed by the distribution maps included: patchiness in peak probability of occurrence for the Blackburnian and Canada Warblers (Figure A1, f-j and Figure A2, a-e), and southern and westerly peaks in probability of occurrence for the American Crow (Figure A1, a-e) and Sedge Wren (Figure A3, a). In general, patterns in peak probability of occurrence were qualitatively similar acro- ss the gradient of model complexity, although the relative importance of some geographic regions appeared to decline for some species (e.g. the eastern half of the study area for Sedge Wren, obvious from comparisons of GLM1 with GLM2, and GAM1 with GAM2).

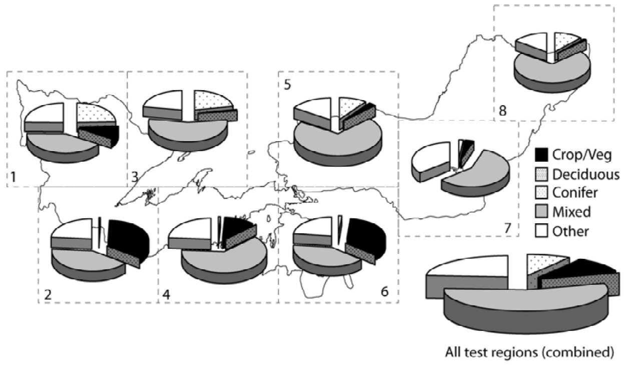

Figure 4. Land cover characteristics of the eight sub-regions used to perform the geographic cross validation. Five land cover classes were defined: conifer-dominated forest (CONIFER), cropland/vegetation mosaic (CROPVEG), deciduous-dominated forest (DECID), and mixed (conifer-deciduous) forest (MIXEDF). These four land cover classes were combined with a default class (OTHER) to produce a dummy-coded omnibus variable (LANDCOV). Solid bars indicate the distribution of land cover classes for the entire study area, while grey bars indicate the distributions for individual regions.

Table 1. Results of the All-Combinations Model Selection Procedure for Each of the Five Species[*]

Species[**]	Model	n.p.	AIC	AUC
AMCR	EVIMEAN + EVIMEAN^2 + EVISD + EVISD^2 + DTR + DTR^2 + PRECIP + PRECIP^2 + TEMP + TEMP^2 + ELEV + LANDCOV	16	3452.4	0.71
BLBW	EVIMEAN + EVIMEAN^2 + DTR + DTR^2 + PRECIP + PRECIP^2	15	1620.6	0.77
CAWA	DTR + DTR^2 + PRECIP + TEMP + TEMP^2 + ELEV	7	860.2	0.72
PUFI	DTR + DTR^2 + TEMP + TEMP^2 + ELEV+ LANDCOV	10	1478.1	0.70
SEWR	EVIMEAN + EVIMEAN^2 + EVISD + DTR + DTR^2 + PRECIP + PRECIP^2 + TEMP + TEMP^2 + ELEV + ELEV^2 + LANDCOV	16	652.2	0.92

[*] Presented in the table are: the "best" model (model with the lowest AIC value); the numbers of parameters (n.p.); the Akaike Information Criterion (AIC); and the apparent predictive accuracy (AUC), which was estimated using the same data used to build the model.

[**] AMCR = American Crow, BLBW = Blackburnian Warbler, CAWA = Canada Warbler, PUFI = Purple Finch, and SEWR = Sedge Wren.

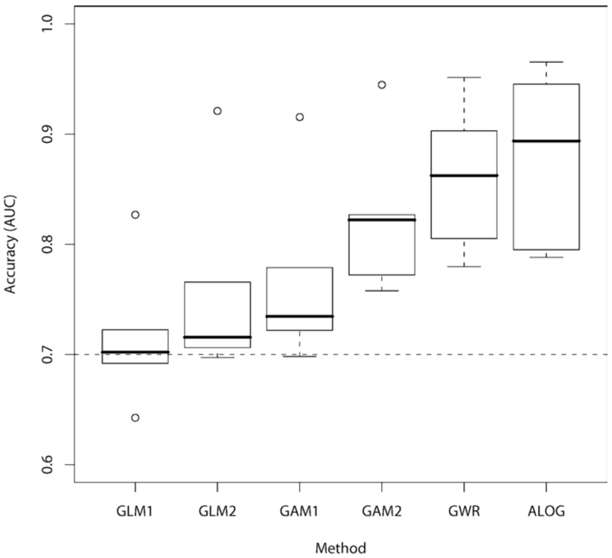

Figure 5. Accuracy assessment based on the same observations used to train the models, with boxplots representing the median and variation in AUC values (across five species). The models included: non-polynomial (GLM1) and polynomial (GLM2) logistic regression; generalized additive modelling (GAM1 and GAM2); geographically-weighted regression (GWR); and autologistic regression (ALOG).

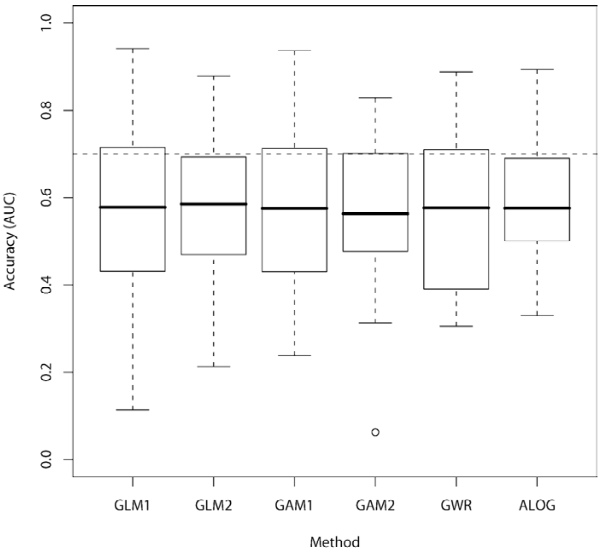

Figure 6. Predictive accuracy (AUC, based on the area under the receiver operating characteristic curve) for: non-polynomial (GLM1) and polynomial (GLM2) logistic regression; generalized additive modelling (GAM1 and GAM2); geographically-weighted regression (GWR); and autologistic regression (ALOG). Boxplots represent the median and distribution of AUC values for 8 iterations of geographic cross validation across five different species.

3.2. Characterization of the Environmental Envelope

Environmental conditions for the continuous predictor variables are indicated in Figure 3. The average value for each variable, for the entire study area, is indicated by a solid horizontal line. Odd numbered sub-regions (#1, 3, 5, 7+8) were at more northern latitudes, while most even-numbered sub-regions (#2, 4, 6) were more southerly. As the sub-regions were num- bered from west to east, lower sub-region numbers refer to locations in the west side of the study area and larger numbers to regions on the east side.

With regards to EVIMEAN values, northern sub-regions (#1, 3, 5, 8) tended to exhibit lower than average greenness values compared to southern regions (#2, 4, 6). There were no perceptible west-east patterns. Variation in EVISD suggested that region # 2 (within the State of Minnesota and in the western side of the study area) was lower (more homogeneous) than average while region # 7 (within the Ottawa region) was higher than average. Diurnal temperature range (DTR) appeared most extreme for the central regions of the study area (#3, 4, 5), particularly those in the vicinity of Lake Superior and Lake Michigan. Average annual precipitation (PRECIP) showed a strong west-east gradient, with eastern portions of the study area receiving substantially more precipitation than those in the west. Not unexpectedly, average monthly temperature (TEMP) exhibited a strong latitudinal gradient, with southern-most regions (#2, 4, 6) experiencing higher than average temperatures for the entire study area. The northern-most region (#8) was subjected to the lowest average monthly temperatures of all. With respect to elevation, differences in median elevation were less dramatic than variation in the range of values, which was greatest for regions in the eastern portion of the study area (#7 and 8). Clearly, regions #7 and 8 has more variable topography than the others.

Based on the overall distribution of land cover classes (Figure 4), the region was best described as predominantly mixed wood forest. Pixels classified as pure deciduous forest were rare. The relative differences in abundance of each class varied depending upon geographic sub region. The southern regions (#2, 4, and 6) consisted of large quantities of cropland, compared to the two north-western regions (#1 and 3) which had more conifer forest.

3.3. Assessment of Model Accuracy

Testing accuracy on the basis of the same points used to train the model conveys an optimistic picture of model predictive power (Figure 5). The use of more complex (non-linear) functional responses (GLM2 and GAM2) resulted in higher median predictive accuracies than simple linear ones (GLM1 and GAM1). Incorporating the effects of location (GWR) and proximity (ALOG) resulted in even greater improvements. Median accuracy for GWR and AUTO (across species) exceeded 0.85. Clearly, the results of this analysis show that spatially-explicit models have the potential to be substantially more accurate than non-spatially explicit models. But this apparent gain in predictive power came at a cost: while more accurate overall, the interquartile ranges of the spatially-explicit models was considerably wider, suggesting that there was greater variability in model performance. Occurrences for some species were predicted more accurately than for others.

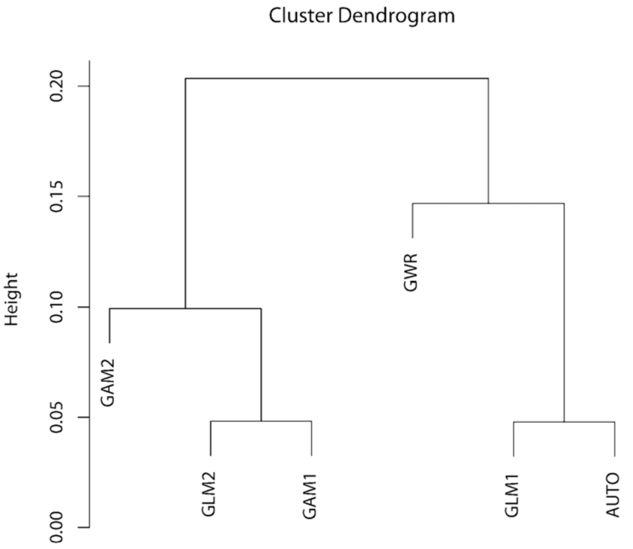

Figure 7. Dendrogram (based on cluster analysis) grouping methods on the basis of their performance under geographic cross-validation.

Of note was the lone outlier defining the uppermost limit of accuracy for the non-spatially explicit models; this point was the Sedge Wren, a species predicted consistently well (> 0.80) regardless of the method used.

Geographic cross-validation provided a very rigorous test of predictive power and contrasted with the previous results (Figure 6). Immediately noteworthy was a substantial decline in predictive accuracy for all methods (indicated by the medians of the accuracy boxplots falling below the horizontal 0.70 cutoff value). Model performance was compared using a hierarchical cluster analysis, and the resulting dendrogram (Figure 7) indicated that the spatially-explicit models formed a separate branch distinct from the non-spatially explicit GAMs as well as GLM2. Curiously, AUTO performed more similarly to the simple linear regression (GLM1) than to any of the other methods. GWR and GAM2 formed solitary leaves. GLM2 and GAM1 formed a distinct group. Overall variation in accuracy values was much greater when geographical cross validation was used (indicated by the wider interquartile ranges). For instance, interquartile range for GLM1 was 3.1% based on the optimistic assessment, but 28.4% for the geographic cross-validation. Not only were predictions from all methods less accurate when extrapolated geographically, but they varied to a large degree depending on sub-region.

In terms of general performance, region-specific comparisons (Figure 8) indicated that sub-region #1 tended to confound all models. Presences in sub-region #8 tended to be predicted more accurately than the others. In terms of model-specific performance, GLM1 models showed a linear trend in predictive accuracy that paralleled the east-west precipitation gradient of Figure 3. GLM2 and GAM1 models showed very similar patterns of accuracy across sub-regions, re-iterating the clustering results of the dendrogram (Figure 7). Clearly there were no substantive differences between GLM and GAM models when the structure (set of predictor variables) and complexity (i.e., degrees of the freedom) of the models were the same. GAM2 models were different from the others, showing fairly consistent performance excepting the very poor fit for one species in sub-region #8. This was indicated by the prominently asymmetric distribution and a very low lower whisker. GWR models showed promise for sub-regions 1 through 4, exhibiting relatively consistent performance across species, but predictions became increasingly unreliable for sub-regions 5 and 6. The GWR model was unable to produce predictions for sub-regions 7 and 8. Finally, AUTO models were the most consistent performer across sub-regions, as indicated by the narrower range in confidence intervals. But AUTO models struggled at either of the east-west extremes of the study area, producing less accurate predictions for the sub-regions and an overall “bowed” shape in performance.

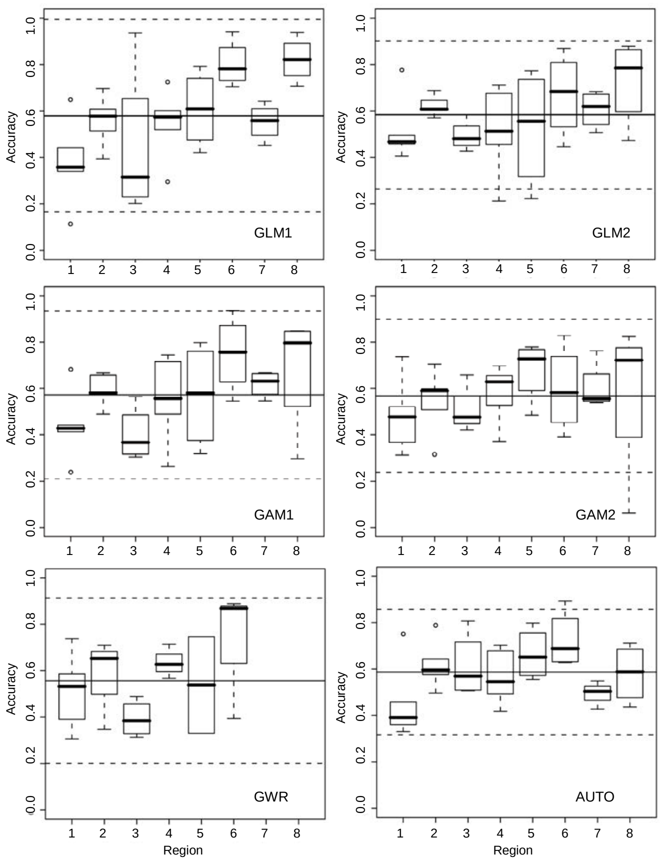

Figure 8. Geographically cross-validated accuracies for each of the models used in this study. For each model, average predictive accuracy is indicated by a solid horizontal line, while the upper and lower 95% percentiles are represented by hatched horizontal lines.

3.4. Species-Specific Results

To better understand the differences in model performance under the regime of geographic cross-validation we teased apart performance by species (Figure 9). For the American Crow overall accuracy was highest for method AUTO. Region-by-region variability was also considerably lower for ALOG

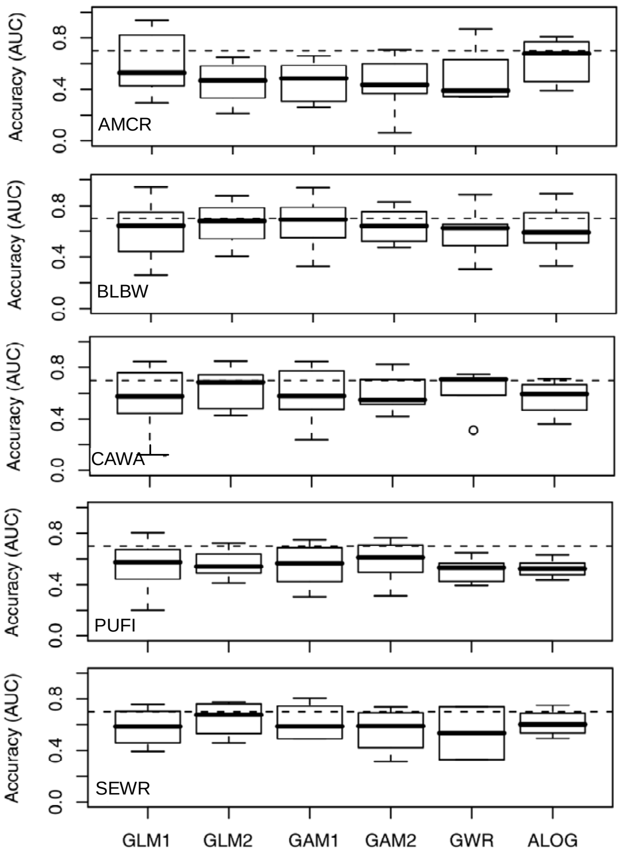

Figure 9. Results of the species-specific geographic cross-validations, summarized by method. Species included: American Crow (AMCR); Blackburnian Warbler (BLBW); Canada Warbler (CAWA); Purple Finch (PUFI); and Sedge Wren (SEWR). Modelling methods included: non-polynomial logistic regression (LOG); polynomial logistic regression (LOGPOLY); generalized additive modelling (GAM); geographically-weighted regression (GWR); and autologistic regression (ALOG). Boxplots represent the median and distribution of AUC values for 8 iterations of geographic cross validation.

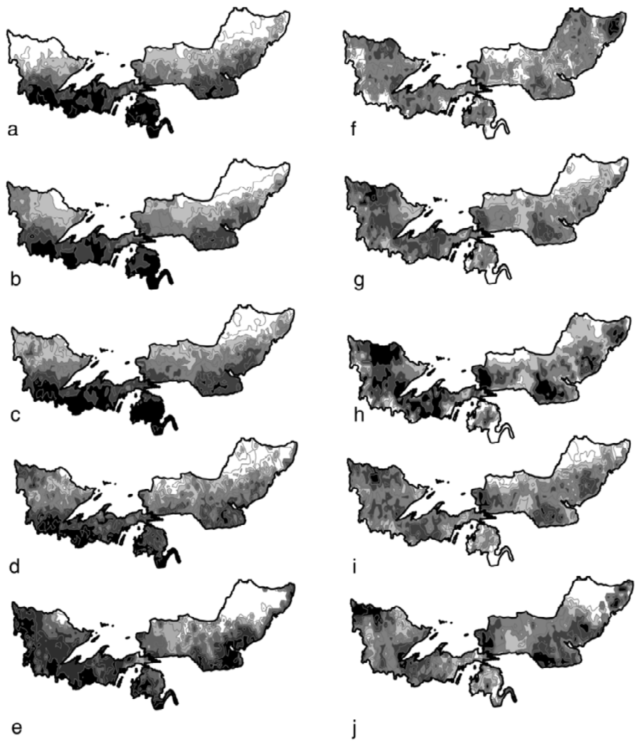

Figure A1. Predictive species occurrence maps for the American Crow (left column, a-e) and the Blackburnian Warbler (right column, f-j). Maps in the first row (a and f) were derived from non-polynomial GLM1 models. The second row (b and g) was derived from polynomial GLM2 models, the third row (c and h) from GAM1 models, the fourth row (d and i) from GAM2 models, and the fifth row (e and j) from GWR models. To assist the visualisation, the prediction surfaces were obtained by applying an inverse-distance weighting smoother to quintile (1 through 5) rankings of the raw probabilities of occurrence.

models of the Purple Finch and Sedge Wren. GWR didn't appear to enhance predictive accuracy for any particular species. GLM2 and GAM1 performed similarly regardless of species. The simplest model of all, GLM1 – which assumed only linear relationships – lead to higher median accuracies for some species, but with much less consistency in performance.

4. Discussion

4.1. The Effect of Model Complexity, Spatial Autocorrelation and Spatial Location

The results of the geographic cross-validation demonstrated that in some cases the linear GLM1 model could produce median levels of accuracy comparable to more complex models. But this came at the price of greater inconsistency across sub-regions. On the basis of optimistic accuracy assessments (Figure 5) predictive accuracy is potentially higher for GAM and spatially-explicit (GWR and ALOG) models, but this depends on the use to which those predictions are to be put. When sampling is uniform and comprehensive for the region of interest, and predictions are desired for points within that sampling frame, GAMs, GWR and ALOG methods can be expected to yield improved accuracy. However, when the goal is to extrapolate into poorly- or non-sampled sub regions, overall accuracy could actually be lower and more variable.

4.2. Geographic Generalizability

When exposed to a regime of geographic cross-validation there was an overall average decline in predictive accuracy of 22.7% ± 2.7% SE (compare Figures 5 and 6), which was similar to the findings of Menke et al. (2009) who applied a geographic test to Californian occurrences of the Argentine ant. Of all the methods we examined GWR appeared to be the most sensitive to the geographic locations used to calibrate the model, as it resulted in the lowest overall accuracy and gave the widest range in performance. Curvilinear models (GAM2,

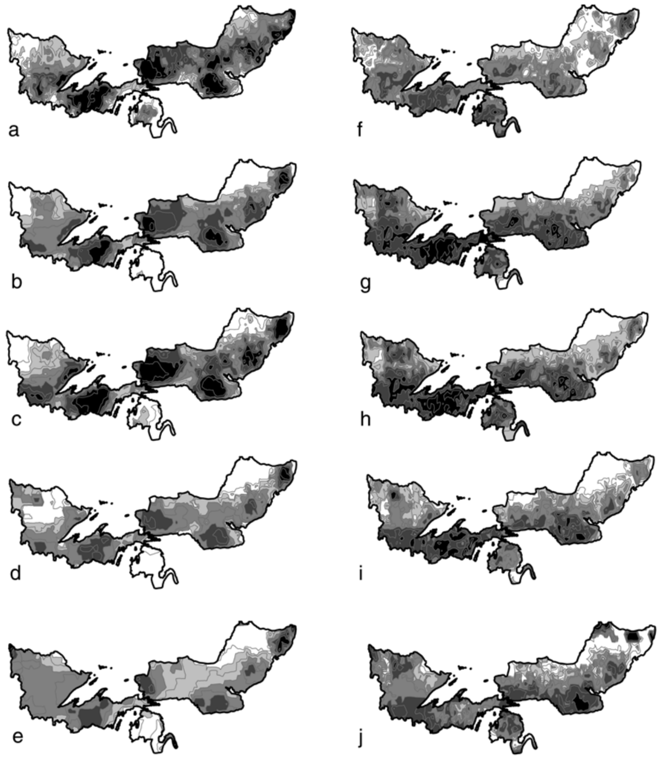

Figure A2. Predictive species occurrence maps for the Canada Warbler (left column, a-e) and the Purple Finch (right column, f-j). Maps in the first row (a and f) were derived from non-polynomial GLM1 models. The second row (b and g) was derived from polynomial GLM2 models, the third row (c and h) from GAM1 models, the fourth row (d and i) from GAM2 models, and the fifth row (e and j) from GWR models. To assist the visualisation, the prediction surfaces were obtained by applying an inverse-distance weighting smoother to quintile (1 through 5) rankings of the raw probabilities of occurrence.

GAM1) seemed an improvement over simple linear ones (GLM1), in terms of having higher overall accuracy and exhibiting less variability across species, but this didn't seem to apply to GAM2 models. In this case, our results suggested that having more complex functional relationships between the response and predictor variables (GAMs with up to 5 knots) tended to be either more accurate, or to perform more consistently across species, but not both at the same time.

5. Conclusions

In general, it is very difficult to accurately extrapolate model predictions into unsampled geographic regions. All methods examined in this study experienced substantial declines in predictive power when this was attempted. But our results suggest some strategies that may help improve the generalizability of predictive models:

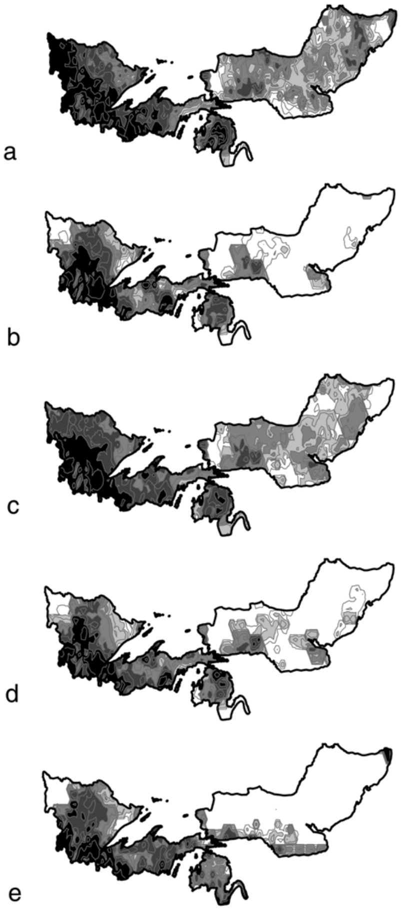

Figure A3. Predictive species occurrence maps for the Sedge Wren. The maps in rows a through e were derived from the following models: non-polynomial GLM1, polynomial GLM2, GAM1, GAM2, and GWR. To assist the visualisation, the prediction surfaces were obtained by applying an inverse-distance weighting smoother to quintile (1 through 5) rankings of the raw probabilities of occurrence.

It is better to concentrate on modeling the dominant predictor variables, especially those where the species response “signal” is strong and relatively simple in form (linear or curvilinear). The use of more complex functional relationships, such as with GAM2 models (fit with up to 5 knots) tended to either improve accuracy or perform more consistently across species, but not both at the same time. As with Pearce et al. (2001) and Segurado and Araujo (2004), we found our habitat specialist (Sedge Wren) tended to produce accurate models regardless of method.

Spatial autoregressive approaches are particularly effective for species where spacing behaviour operates strongly at the scale of interest. Incorporating spacing behaviour (through the use of an autocorrelated predictor variable) appears to provide extra “contextual” information and was especially beneficial for the American Crow, a highly social species with a tendency to be patchily distributed.

GWR has the potential to improve model predictive accuracy, but not when the predictions are extrapolated as done here. This method appears quite sensitive to the data used to train the model.

Acknowledgments. The authors acknowledge the commitment and effort of the many volunteer BBS participants as well as the scientists and administrators of the Canadian Wildlife Service and the Patuxent Wildlife Research Center of the USGS for supporting and overseeing the BBS. DJL acknowledges the Department of Geography and the Faculty of Graduate Studies of the University of Calgary for personal support in the form of a Graduate Research Scholarship and a Graduate Teaching Fellowship. The authors thank C. Brunsdon for kindly providing R code for implementing the GWR models.

References

Augustin, N.H., Mugglestone, M.A., and Buckland, S.T. (1996). An autologistic model for the spatial distribution of wildlife, *J. Appl. Ecol.*, 33, 339-347. doi:10.2307/2404755

Augustin, N.H., Mugglestone, M.A., and Buckland, S.T. (1998). The role of simulation in modelling spatially correlated data, *Environmetrics*, 9, 175-196. doi:10.1002/(SICI)1099-095X(19 9803/04)9: 2<175:: AID-ENV294>3.0.CO;2-2

Burnham, K.P., and Anderson, D.R. (2002). *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. Springer, New York, pp. 488.

Diniz-Filho, J.A.F., Rangel, T.F.L.V.B., and Bini, L.M. (2008). Model selection and information theory in geographical ecology, *Global Ecol. Biogeogr.*, 17, 479-488. doi:10.1111/j.1466-8238.2008.0039 5. x

Dormann, C.F. (2007). Effects of incorporating spatial autocorrelation into the analysis of species distribution data, *Global Ecol. Biogeogr.*, 16, 129-138. doi:10.1111/j.1466-8238. 2006.00279.x

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., McC. Overton, J., Peterson, A.T., and Phillips, S.J. (2006). Novel methods improve prediction of species' distributions from occurrence data, *Ecography*, 29, 129-151. doi:10.1111/j.2006.0906-7590.04596.x

Environmental Systems Research Institute. (2002). *ArcGIS version 8. 3. Environmental Systems Research Institute*, Redlands, California.

Fielding, A.H., and Haworth, P.F. (1995). Testing the generality of bird-habitat models. *Conserv. Biol.*, 9, 1466-1481. doi:10.1046/j. 1523-1739.1995.09061466.x

Fielding, A.H. (2002). Appropriate characteristics of an accuracy measure. Predicting species occurrences: issues of accuracy and scale (ed.by J.M. Scott, J.M., P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. Wall, W.A. and F.B. Samson), Island Press, pp. 271-280.

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. doi:10.1017/S0376892997 000088

Foody, G.M. (2004). Spatial nonstationarity and scale dependency in the relationship between species richness and environmental determinants for the sub-Saharan endemic avifauna, *Global Ecol. Biogeogr.*, 13, 315-320. doi:10.1111/j.1466-822X.2004.00097.x

Fortin, M.J., and Dale, M.R.T. (2005). *Spatial analysis: a guide for ecologists*, Cambridge University Press, New York.

Fotheringham, A.S., Brunsdon, C., and Charlton, M. (2002). *Geographically weighted regression*, Wiley & Sons, Ltd., Chichester, England.

Friedl, M.A., McIver, D.K., Hodges, J.C.F., Zhang, X.Y., Muchoney, D., Strahler, A.H., Woodcock, C.E., Gopal, S., Schneider, A., Cooper, A., Baccini, A., Gao, F., and Schaaf, C. (2002). Global land cover mapping from MODIS: algorithms and early results, *Remote Sens. Environ.*, 83, 287-302. doi:10.1016/S0034-4257(02)00078-0

Guisan, A., Lehmann, A., Ferrier, S., Austin, M., Overton, J. MC. C. Aspinall R.,and Hastie T. (2006). Making better biogeographical predictions of species' distributions. *J. Appl. Ecol.*, 43, 386–392. doi: 10.1111/j.1365-2664.2006.01164.x

Harrell Jr., F.E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer, New York, pp. 568.

Harrell Jr., F.E., Lee, K.L., and Mark, D.B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, *Stat. Med.*, 15, 361-387. doi:10.1002/(SICI)1097-0258 (19960229) 15:4 <361::AID-SIM168>3.0.CO;2-4

Hastie, T.J., and Tibshirani, R.J. (1990). *Generalized Additive Models*, Chapman & Hall/CRC, New York.

Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., and Ferreira, L.G. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices, *Remote Sens. Environ.*, 83,195-213. doi:10.1016/S0034-4257(02)000 96-2

Ihaka, R., and Gentleman, R. (1996). R: a language for data analysis and graphics, *J. Comput. Graph. Statistics*, 5, 299-314.doi:10.230 7/1390807

Jetz, W., Rahbek, C., and Lichstein, J.W. (2005). Local and global approaches to spatial data analysis in ecology, *Global Ecol. Biogeogr.*, 14, 97-98. doi:10.1111/j.1466-822X.2004.00129.x

Justice, A.C., Covinsky, K.E., and Berlin, J.A. (1999) Assessing the generalizability of prognostic information, *Annals of Internal Medicine*, 130, 515-524.

Kissling, W.D., and Carl, G. (2008). Spatial autocorrelation and the selection of simultaneous autoregressive models, *Global Ecol. Biogeogr.*, 17, 59-71.

Legendre, P. (1993) Spatial autocorrelation: trouble or new paradigm? *Ecology*, 74, 1659-1673. doi:10.2307/1939924

Lieske, D.J., and Bender, D.J. (2009). Accounting for the influence of geographic location and spatial autocorrelation in environmental models: a comparative analysis using North American songbirds, *J. Env. Inform.*, 13, 12-32. doi:10.3808/jei.200900137

McCullagh, P., and Nelder, J.A. (1999). *Generalized linear models*, CRC Press, New York.

Menke, S.B., Holway, D.A., Fisher, R.N., and Jetz, W. (2009). Characterizing and predicting species distributions across environments and scales: Argentine ant occurrences in the eye of the beholder, *Global Ecol. Biogeogr.*, 18, 50-63. doi:10.1111/j.1466-8238.2008.00420.x

Mitchell, T.D., and Jones, P.D. (2005). An improved method of constructing a database of monthly climate observations and associated high-resolution grids, *Int. J. Climatol.*, 25, 693-712. doi: 10.1002/joc.1181

Ontario Partners in Flight. (2006) *Ontario Landbird Conservation Plan: Boreal Hardwood Transition (North American Bird Conservation Region 12): priorities, objectives and recommended actions*. Version 1.0. EC/MNR.

Oreskes, N., Shrader-Frechette, K., and Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences,*Science*, 263, 641-646. doi:10.1126/science.263.5147.641

Pearce, J., and Ferrier, S. (2000). Evaluating the predictive performance of habitat models developed using logistic regression, *Ecol. Model.*, 133, 225-245. doi:10.1016/S0304-3800(00)00322-7

Pearce, J., Ferrier, S., and Scotts, D. (2001). An evaluation of the predictive performance of distributional models for flora and fauna in north-east New South Wales, *J. Environ. Manage.*, 62, 171-184. doi:10.1006/jema.2001.0425

Randin, C.F., Dirnbock, C., Dullinger, S., Zimmermann, N.E., Zappa, M., and Guisan, A. (2006). Are niche-based species distribution models transferable in space? *J. Biogeogr.*, 33, 1689-1703. doi:10. 1111/j.1365-2699.2006.01466.x

Rich, T.D., Beardmore, C.J., Berlanga, H., Blancher, P.J., Bradstreet, M.S.W., Butcher, S., Demarest, D.W., Dunn, E.H., Hunter, W.C., Inigo-Elias, E.E., Kennedy, J.A., Martell, M.A., Panjabi, A.O., Pashley, D.N., Rosenberg, K.V., Rustay, C.M., Wendt, J.S., and Will, C.T. (2004). *Partners in Flight North American Landbird Conservation Plan*. Cornell Lab of Ornithology. Ithaca, NY.

Robbins, C.S., Bystrak, D., and Geissler, P.H. (1986). The *breeding bird survey: its first fifteen years*, 1965-1979. U.S. Fish & Wildlife Service Publication, pp.157.

Segurado, P., and Araujo, M.B. (2004). An evaluation of methods for modelling species distributions. *J. Biogeogr.*, 31, 1555-1568. doi: 10.1111/j.1365-2699.2004.01076.x

U.S. Geological Survey Survey. (1996). GTopo30. http://edcdaac.usgs. gov/ gtopo30 /w10090.asp, Sioux Falls, South Dakota.

Vaughan, I.P., and Ormerod, S.J. (2003). Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data*.* *Conserv. Biol.*, 17, 1601-1611. doi:10.1111/j.1523-1739.2003.00359.x

Vaughan, I.P., and Ormerod, S.J. (2005). The continuing challenges of testing species distribution models, *J. Appl. Ecol.*, 42, 720-730. doi:10.1111/j.1365-2664.2005.01052.x

Verbyla D.L. and Litvaitis J.A. (1989). Resampling methods for evaluating classification accuracy of wildlife habitat Models. *Environ. Manage.*, 13, 783-787. doi: 10.1007/BF01868317

Zweig, M.H., and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine, *Clin. Chem.*, 39, 561-571.