ELSEVIER

Ecological Modelling 157 (2002) 281–300

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Evaluating resource selection functions

Mark S. Boyce [a,*], Pierre R. Vernier [b], Scott E. Nielsen [a], Fiona K.A. Schmiegelow [c]

[a] *Department of Biological Sciences, University of Alberta, Edmonton, Alta., Canada T6G 2E9*
[b] *Department of Forest Sciences, University of British Columbia, Vancouver, BC, Canada V6T 1Z4*
[c] *Department of Renewable Resources, University of Alberta, Edmonton, Alta., Canada T6G 2H1*

Abstract

A resource selection function (RSF) is any model that yields values proportional to the probability of use of a resource unit. RSF models often are fitted using generalized linear models (GLMs) although a variety of statistical models might be used. Information criteria such as the Akaike Information Criteria (AIC) or Bayesian Information Criteria (BIC) are tools that can be useful for selecting a model from a set of biologically plausible candidates. Statistical inference procedures, such as the likelihood-ratio test, can be used to assess whether models deviate from random null models. But for most applications of RSF models, usefulness is evaluated by how well the model predicts the location of organisms on a landscape. Predictions from RSF models constructed using presence/absence (used/unused) data can be evaluated using procedures developed for logistic regression, such as confusion matrices, Kappa statistics, and Receiver Operating Characteristic (ROC) curves. However, RSF models estimated from presence/available data create unique problems for evaluating model predictions. For presence/available models we propose a form of k-fold cross validation for evaluating prediction success. This involves calculating the correlation between RSF ranks and area-adjusted frequencies for a withheld sub-sample of data. A similar approach can be applied to evaluate predictive success for out-of-sample data. Not all RSF models are robust for application in different times or different places due to ecological and behavioral variation of the target organisms.

Keywords: Habitat selection; Logistic regression; Model selection; Prediction; RSF; Resource selection functions; Validation

1. Introduction

The idea of a resource selection function (RSF) has its origins with the theory of natural selection (Manly, 1985), but with the intent to characterize selection of resources by animals. A RSF is defined as any function that is proportional to the probability of use by an organism (Manly et al., 1993). The units being selected by animals (e.g. pixels of land) are conceived as resources and predictor variables associated with these resource units may be 'resource' variables or covariates of the resources, e.g. elevation or human-disturbance. RSF models overlap substantially with methods that have been developed for mapping distributions of organisms using species–environment

* Corresponding author. Tel.: +1-780-492-0081; fax: +1-780-492-9234
E-mail address: boyce@ualberta.ca (M.S. Boyce).

0304-3800/02/$ - see front matter
PII: S0304-3800(02)00200-4

patterns (Guisan and Zimmermann, 2000); indeed in some instances the methods are identical. Such predictive geographic modeling has been particularly well developed by plant ecologists (e.g. Austin et al., 1990, 1994).

A RSF model is a form of habitat suitability index (HSI; US Fish and Wildlife Service, 1981) but with statistical rigour. Some HSI models are created using expert opinion and other approaches not directly tied to statistical estimation, whereas RSF models are always estimated directly from data. A RSF usually is estimated from observations of (1) presence/absence (used-vs.-unused), or (2) presence/available (used-vs.-available) resource units. For both of these sampling designs the prevailing statistical model is a binomial generalized linear model (GLM), usually logistic regression, although in the case of presence/available sampling designs, logistic regression is used as an estimating function and not for statistical inference. When linked to a geographical information system (GIS), RSF models can be powerful tools in natural resource management, with applications for cumulative effects assessment, land-management planning, and population viability analysis (Boyce et al., 1994; Boyce and McDonald, 1999; Boyce and Waller, 2000).

As many applications of RSF models have conservation and management implications, we believe that the most important consideration for evaluating RSF models is prediction. If a model reliably predicts the locations of organisms, it is a good model. Other measures for model evaluation relate to the model selection process, such as how well the model fits the data. In some circumstances we need to know if the model does better than a random null model, i.e. statistical inference. Still other evaluation criteria must be considered relative to model robustness, e.g. how well a habitat model can predict distribution and abundance in other places and times.

1.1. Sampling designs

If we have access to presence/absence (used-vs.-unused) resource units, logistic regression can be used to generate RSF models. Here we assign a 0 to sites where the organism is absent and a 1 to sites where present; and a set of predictor variables is selected a priori using hypothesized habitat relations based on the ecology of the organism (Burnham and Anderson, 2001). With this sort of data, logistic regression yields the probability of use of a resource unit, e.g. a pixel or polygon. Models that can predict the probability of use have been termed resource selection probability functions (RSPF; Manly et al., 1993).

A concern with the used-versus-unused approach to fitting RSF models is that it may be difficult to demonstrate non-use, especially for mobile and cryptic animals. Also, non-use can depend on sampling intensity, so that a more extensive search might result in unused sites being reclassified as used sites. This means that the errors are imbalanced because we can be assured that used sites indeed have been used, but we are less certain about unused sites. Depending on the interpretation of the results, this may not be a serious problem if a representative sample of used and unused sites is based on an unbiased sample of use. Even though some sites might be used eventually, the fact that a representative sample did not show presence is still an unbiased sample of use and non-use. Building a model using such a sample is acceptable if one can be satisfied with model predictions that are proportional to the probability of use (i.e. RSF), but imbalanced errors can result in biased estimates of the actual probability of use (RSPF) because the true probabilities of use are higher than predicted by the model.

In some sampling situations we cannot estimate a sample of unused sites. For example, radiotelemetry data on animals can identify points used by the animals, but there is an infinite number of points on the landscape which may have been used, so we cannot define ‘unused’ points. With plants we are often confronted with herbarium records or other ‘presence-only’ data where no field effort has attempted to characterize sites where the plant did not occur (Zaniewski et al., 2002 this volume). In these situations, the appropriate design structure is one of presence versus availability, i.e. characterizing a sample of sites where the organism is present from a sample of what is available on the landscape. What is

'available' can be based on a complete census of the available habitats, or more plausibly, we can take a random sample of landscape locations. Hence, the domain of available sites is a critical consideration in framing the scale of the study (Johnson, 1980).

We note that a variety of sampling designs is possible and additional alternatives for estimating RSF models are reviewed by Manly et al. (1993).

1.2. Statistical inference

In practice, simply showing that a habitat model provides a significant fit to data is not very revealing. In fact, in nearly all cases, testing whether or not the use of habitats is non-random is a trivial exercise, because organisms are always differentially distributed on the landscape (Cherry, 1998). Generally, statistical inference is not a very useful concept in habitat modeling.

Perhaps the most common approach to evaluating RSF models, albeit misguided (Yoccoz, 1991; Cherry, 1998; Johnson, 1999; Anderson et al., 2000; Boyce, 2001), is to use probability or significance levels. A small P-value is presumed to give some measure of confidence that the model is not due to chance alone. Most of us learned to use statistics for hypothesis testing consistent with the Popperian approach to science (Popper, 1934). Yet, there are many long-appreciated difficulties with the hypothesis-testing focus in biostatistics (Platt, 1964). Thresholds for significance are entirely arbitrary or based on convention, e.g. $\alpha = 0.05$. The focus is on achieving statistical significance (Yoccoz, 1991). Null hypotheses commonly are framed in ways that are obviously false or uninformative (Cherry, 1998). Often little or no information is reported about effect sizes, precision, or sample sizes, such that we cannot properly evaluate the meaning of P-values (Anderson et al., 2000).

Most habitat data are burdened with spatial and/or temporal autocorrelation (Otis and White, 1999)—a particular problem with the new GPS-telemetry technology that permits frequent observations of individuals. This lack of independence means that we are likely to commit a Type I error because we underestimate variances associated with model coefficients (Lennon, 1999, 2000). A variety of approaches have been suggested, such as rarefying (e.g. using only every i-th observation for analysis) data (Swihart and Slade, 1985a,b). Applications of such techniques, however, have proven to be highly conservative, with a reduction in data of greater than 90% for one study (McNay et al., 1994). Most ecologists are loath to discard such hard-earned data. Moreover, rarefaction procedures tend to remove highly selected habitats that can be quite important to an animal's fitness (e.g. the re-use of important feeding patches or bedding sites). A more satisfactory approach is to use variance inflators to obtain robust standard errors (Nielsen et al., 2002; Vernier et al. 2002; also see Section 2.4.2). These do not alter the model coefficients but can have quite marked consequences for P-values, almost always making them larger when accounting for spatial-temporal autocorrelation.

1.3. Model selection

In many instances, a key step in analysis is to ask which of several alternative models best explains the data (Burnham and Anderson, 1998). Clearly, P-values are an insufficient measure of the appropriateness of alternative models (Burnham and Anderson, 1998; Anderson et al., 2000), and a much better alternative is to use some form of information criteria, such as the Akaike Information Criteria (AIC) or the Bayesian Information Criteria (BIC). These are based on the principle of parsimony, helping to identify the model that accounts for the most variation with the fewest variables. AIC tends to select models with too many parameters when sample sizes are large, whereas BIC may be too conservative, yielding overly simplistic models because of the heavy penalty on the addition of additional parameters (Hastie et al., 2001).

Application of information criteria in model selection does not preclude the utility of conventional hypothesis testing, but biologists need to understand when one or the other is appropriate. Information-criteria approaches to model selection have the potential to increase our biological understanding, through application of data to

process-based (mechanistic) models. Herein, lies a new opportunity and challenge in biostatistics. We should be building models based on ecological principles that can be evaluated with data. This requires a fundamental understanding of relevant theory to propose models that are most likely to direct our understanding of the system. We must strive to ensure model adequacy if we want to ensure biological relevancy (Boyce, 2001).

Since sampling design is crucial to the methods available for evaluating predictions from RSF models, we have organized this paper according to the two primary designs: presence/absence (used-vs.-unused) and presence/available (used-vs.-available) resource units.

2. Presence/absence (used-vs.-unused) designs

With this design, each resource unit is classified as being occupied or not. For example, pixels in a GIS image are identified as being locations where an animal was present (used) or where the animal did not appear (unused). In plant ecology, these are typically presence/absence data for individual species, such as would be accumulated in quadrat data.

2.1. Model selection

Information criteria such as AIC and BIC are the most powerful approaches for model selection from a set of alternative plausible models (Burnham and Anderson, 1998). Other methods exist for evaluating how well the data are explained by the model, but these may not be penalized appropriately for the number of variables in the model.

The process of building a logistic regression model is similar to that for multiple linear regression analysis (Guisan and Zimmermann, 2000), and because most users of logistic regression software are likely to be familiar with linear regression, there is a demand for an easy-to-interpret measure of how well the model fits the data. The coefficient of determination, r^2, is just such a linear regression statistic included with most statistical software packages offering logistic-regression analogues. Several variations exist including the Cox and Snell r^2 and the Nagelkerke r^2. For GLMs one typically compares -2 log likelihood (-2LL) for a model with the -2LL for a null model and calculate the percent deviance explained (StataCorp., 2001). Another common method used for assessing model fit is the use of goodness-of-fit tests. Long (1997) provides a description of several scalar measures developed to summarize the overall goodness-of-fit for regression models of continuous, count, or categorical dependent variables. Likelihood ratio or Pearson χ^2 goodness-of-fit tests are questionable, however, when the number of covariate patterns is close to the number of observations. In such circumstances, we might regroup the data into nearly equal-size groups based on ordered model probabilities, followed by a χ^2 goodness-of-fit test (Hosmer and Lemeshow, 1989).

2.2. Prediction

2.2.1. *Confusion matrices and classification tables*

Standard output from statistical software for logistic regression usually includes a classification table or 'confusion matrix' of predicted and observed values based on the fitted model, typically with a χ^2 statistic. The observed response variable is (0,1) so this does not present a problem. Predicted values vary as a probability ranging from a possible low value of 0 to a possible high value of 1. Typically, a mid-point cut-off level of 0.5 is the default for creating a classification table. Even with presence/absence data, the cut-off is seldom equal to 0.5 and one must select a threshold cut-off level (Guisan et al., 1998). Classification accuracy is also sensitive to the relative frequency (prevalence) of observations of the species within the sample as threshold cut-off levels are varied. For instance, high classification accuracy for a rare species may be obtained by simply shifting the probability cut-off level to 1, hence, never predicting the species to occur (Pearce and Ferrier, 2000).

2.2.2. *Kappa statistic*

An additional method for evaluating classification effectiveness of a model is with the Kappa statistic. The basic premise of Kappa is that a

certain level of random or chance agreement is going to occur within a classification table and, therefore, one must adjust classification rates accordingly. Kappa adjusts for such bias by measuring the actual agreement minus the agreement expected by chance (Cohen, 1960). Important advantages of Kappa are the use of agreement, instead of association (traditional classification tables). This is particularly important when the prevalence of the species is low (Fielding and Bell, 1997)—a frequent phenomenon in wildlife RSF modeling because models often are developed for rare, threatened, or endangered species. Under such circumstances, agreement (vs. classification tables) becomes much more valuable, because it distinguishes and incorporates both false positives and false negatives. Kappa is also less sensitive to zero values within the matrix (Manel et al., 2001). Remote-sensing classifications frequently use Kappa as a measure of accuracy assessment (Congalton, 1991; Lillesand and Kiefer, 1994). As in a confusion matrix, a threshold value for prediction must be identified, as illustrated in the Guisan et al. (1998) model for the distribution of an alpine plant. Under situations where one might consider some errors in the classification table to be less or more important, one can use a weighted Kappa, controlling the seriousness of each possible disagreement (Cohen, 1968). This weighted Kappa technique has proven to be useful in assessing thematic maps where the categories are ordered or classified according to a nominal or ordinal scale (Naesset, 1996; Guisan and Harrell, 2000). Such a measure might be used for ordered classes of RSF scores, but we suggest caution if the method is used for presence/available data, for reasons explained below. Alternatively, one could compare different maps/models developed for the same area using a Kappa statistic (Monserud and Leemans, 1992).

2.2.3. *ROC*

A problem with all threshold dependent measures of classification is the potential for distortions or bias when dichotomizing an inherently continuous variable (Altman et al., 1994). Recognizing this limitation, a classification approach called the Receiver Operating Characteristic (ROC) has been developed that is independent to probability cut-off levels. Although the medical literature is rich in the use of ROC to evaluate models, its use in ecological studies has only recently occurred (e.g. Murtaugh, 1996; Fielding and Bell, 1997; Cumming, 2000; Pearce and Ferrier, 2000). An advantage of the ROC approach over traditional classification tables is the ability of ROC analyses to evaluate the proportion of correctly and incorrectly classified predictions over a continuous range of threshold probability cut-off levels (Pearce and Ferrier, 2000). To calculate a ROC curve, sensitivity and specificity are evaluated at different probability cut-off levels within the data to produce pairs of sensitivity/specificity values (Metz, 1978). Sensitivity is defined as the probability that a model yields a positive prediction where an animal actually occurs (i.e. 1), whereas, specificity is the probability that a low score is predicted where no animal is observed (i.e. 0). As the cut-off threshold is varied, different proportions of positive and negative cells are included. Plotting sensitivity as a function of 1−specificity for each threshold yields a ROC curve. From this ROC curve, we can integrate the area under the curve (AUC) as an assessment of model performance or predictive power (Cumming, 2000). A model with no predictive power would have an AUC of 0.5 (e.g. a 45° line), while a perfect model would correspond to an AUC of 1.0.

2.2.4. *k-Fold cross validation*

Researchers are frequently limited in both time and money and are, therefore, unlikely to have independent data available for prospective sample evaluations. An alternative is to withhold a fraction of the data using a k-fold partitioning of the original samples (Fielding and Bell, 1997), where k represents the number of partitions ranging from 2 to N−1 (number of observations minus one). This method works particularly well for studies having a single intensive period of data collection (e.g. GPS radio telemetry data) across only one region (Nielsen et al., 2002). A variety of model-testing strategies then can be used to evaluate the reliability of the model using the partitioned data set(s), including most of the in-sample resubstitu-

tion techniques (e.g. ROC, goodness-of-fit, etc.) described in the previous sections.

k-Fold cross validation methods also can be used to evaluate spatially explicit RSF model predictions (i.e. GIS maps) by partitioning *k* random subsets from the original data. Following model development, the study area can be classified for the probability of occurrence using the RSF model in a GIS, and tallied (binned) into an arbitrary number of categories of RSF scores (Boyce and Waller, 2000). For each withheld observation, an RSF value can be calculated from the model constructed with the training set. Then, one can plot the frequency of observations of RSF scores, adjusted for area, within that particular RSF-score category. Adjusted frequencies should be highly correlated with the RSF scores if the model is a good one, i.e. indicating that the RSF model was indeed predicting the relative probability of occurrence of the organisms on the landscape. To make this evaluation, we recommend use of the Spearman-rank correlation. As we develop below, this *k*-fold cross validation method is particularly important for presence/availability designs.

2.3. Robustness

There is an element of circularity in evaluating a model based on the data that were used to estimate model coefficients in the first place. Evaluations made from the data used to calibrate the model often result in optimistic measures of classification success (Fielding and Bell, 1997). RSF models are presumed to apply in the area for which they were developed (Manly et al., 1993). Applying the models to a new area or different time period (prospective sampling) results in changes in the availability of various habitats and this will usually result in a change in the model coefficients and apparent selection. Nevertheless, one measure of the robustness of a habitat model is when it can be applied in other areas. Some models, e.g. those for the Northern Spotted Owl (*Strix occidentalis caurina*), apply over vast areas partly because the models are dominated by old growth forest, a good predictor of Northern Spotted Owl habitat throughout their range (Meyer et al., 1998).

Seeing how well one can predict an animal's distribution in space and time can be an important measure of the robustness of a model, and for some applications may be essential. However, there are a number of biological problems that interfere with model evaluation based on prospective sampling:

1) Habitat selection changes seasonally due to changes in resources. This is dramatically illustrated by grizzly bear habitat-use patterns (*Ursus arctos*; Boyce and Waller, 2000). Early in the season after emergence from their dens, bears often are found at lower elevations where they search for food. Higher elevation areas still are covered with snow so they are restricted from effectively foraging in those areas. Later, the bears move to higher elevations where they forage on vegetation, and may move to berry-producing areas if berries are abundant. Bears are opportunistic and may feed extensively on fish, ungulates, moths or even garbage when these resources become available.
2) Habitat selection can change among years, due to fluctuating resources, or to shifts in local distribution that result from changes in abundance of territorial species. For example, in years of relatively high abundance, individuals of territorial species may occupy a wider range of habitats than in years where abundance is lower, and, thus, selection would appear to vary. A failure to take into account such variations can lead to poor model fit or inappropriate inferences (Schooley, 1994). There are many examples of this (Kneeland et al., 2002), e.g. Case Study 1 below.
3) Habitat selection can vary spatially depending on the spectrum of habitats available in an area (Osborne and Suárez-Seoane, 2002 this volume). Often a habitat is used more as it becomes more available. This can occur because the species in question learns how to use the habitat more effectively, i.e. they develop a search image for the habitat or resource type. In the context of predator–prey theory, this change in selection associated with availability is sometimes called a functional response

(Mysterud and Ims, 1998). Some investigators have taken the perspective that the existence of such functional responses amount to a fundamental flaw in RSF models (Garshelis, 2000). We take an opposing view that such functional responses can be modeled explicitly, so, if given information about what resources are available, we can estimate RSF coefficients for a particular locality (Kneeland et al., 2002). One might consider such a model to be a 'localized' model in contrast to a 'robust' model that applies over a broad array of localities.

Even if a model performs similarly across space and time, we cannot assume that the model has a high predictive value. The model may simply be predicting equally poorly between sample periods or regions, but maintaining a high degree of predictive correlation due to similar patterns of selection. As an example, Manly et al. (1993) describe the predictive correlation between two habitat models for the galaxiid fish *Galaxias vulgaris*, along the Shag River in New Zealand (McIntosh et al., 1992). Models were fitted for 'trout' (brown trout [*Salmo trutta*]) and 'non-trout' sites to predict the density (log-linear models) of galaxiids based on five principal components obtained from 14 measured environmental variables. The effect of introduced trout on habitat selection, however, was not evident because parameter estimates for models of 'trout' and 'non-trout' streams were quite similar. In fact, the correlation between model predictions was strong ($r = 0.992$, $P < 0.001$), indicating that the distribution of galaxiids under the 'trout' model predicted the 'non-trout' data well and vice versa. Although this consistency in RSF scores between streams suggests similar patterns of habitat selection, we still cannot be confident that the models predicted well. Furthermore, within dynamic systems, model consistency from present and past data is no guarantee of future predictive performance (Oreskes et al., 1994). Particular care should be taken in interpreting model results beyond their original domain, a difficult task given the fact that ecologists are often charged with evaluating future scenarios (global warming, cumulative effects, population viability, etc.).

2.4. Case study 1: songbird–habitat relationships

We developed RSF models using presence/absence data for 5 boreal forest songbird species: Black-throated Green Warbler (*Dendroica virens*), Red-breasted Nuthatch (*Sitta canadensis*), White-throated Sparrow (*Zonotrichia albicolis*), Yellow-rumped Warbler (*D. coronata*), and Yellow Warbler (*D. petechia*). This is the same suite of species for which we recently developed abundance models (Vernier et al., 2002). The study area encompasses ≈ 140 km^2 of boreal mixed-wood forest near Calling Lake, in north-central Alberta, Canada (55°N, 113°W). Trembling aspen (*Populus tremuloides*), balsam poplar (*P. balsamifera*), and white spruce (*Picea glauca*) were the most abundant upland tree species, often occurring together in old, mixed stands, whereas black spruce (*P. mariana*) characterized hydric sites.

Our objectives with this case study were to (1) assess the temporal variability in songbird-habitat relationships (direction, strength, and significance of estimated coefficients), (2) evaluate the temporal variability in the predictive performance of models, and (3) determine if the predictive performance of the models increased with the number of years of data used to fit the model.

2.4.1. Songbird and habitat variables

We used bird abundance data collected by point-count surveys conducted between 1993 and 1999, as part of the Calling Lake Fragmentation Experiment and related studies (e.g. Schmiegelow et al., 1997). A total of 406 permanent sampling stations were located within 65 sites, which we define as contiguous areas of similar forest type and age. Site types included areas clearcut in 1993 as part of the experimental design, young and old deciduous forests, mature coniferous forests, and mixed-wood forests. There was at least 200 m between each sampling station. Details of the sampling protocol and study area can be found in Schmiegelow et al. (1997).

We measured habitat patterns around each bird sampling station using 1:20 000 digital Alberta

Vegetation Inventory (AVI) maps. We used original and derived map layers to measure habitat characteristics around each bird sampling station at two spatial scales: the local-scale, which matched the size and shape of the circular bird sampling stations (inner buffer of 100 m radius, or 3.14 ha), and the neighborhood scale, which extended from 100 to 500 m beyond the sampling stations (outer buffer, 75.4 ha). Habitat variables we selected (Table 1) either had been used in the literature previously, or were hypothesized correlates of species abundance based on the ecology of the species. The process of generating the variables as well as their evaluation for inclusion in statistical models is described in Vernier et al. (2002).

2.4.2. Statistical analysis

All presence/absence models (GLM, logistic regression) were developed using the same general approach we recently used to develop abundance models (Vernier et al., 2002; GLM, Poisson regression), where the set of variables included for each species' model was selected from among five alternative habitat model formulations using AIC). We used STATA's cluster option to calculate variance estimates that are robust to influential observations, within site correlations, and undetected over-dispersion (StataCorp., 2001). We evaluated models using ROC curves. The AUC was used as a measure of model performance for both in-sample/resubstitution (ROC_{in}) and out-of-sample/prospective sampling (ROC_{out}) data. Following Swets (1988) and Manel et al. (2001), we considered models with a ROC value ranging between 0.7 and 0.9 as 'useful applications' and those with values greater than 0.9 as being of 'high accuracy'.

To assess the temporal variability in songbird–habitat relationships (objective 1), we developed separate models for each year (1993–1999) and recorded the direction, strength, and significance of the estimated coefficients. To evaluate the temporal variability in the predictive performance of the models (objective 2), we used the models developed in objective 1 to calculate ROC_{in} for each year (e.g. 1995) and calculated ROC_{out} using data from the following year (e.g. 1996), excluding, the last year of sampling. To determine if the predictive performance of the models increased with the number of years used to fit the model (objective 3), we developed models using 1 year of data, 2 years of data, 3 years of data, and so on, and validated each of these models using data from the following year. For example, a model

Table 1
Definitions for predictor variables in the songbird-habitat relationships case study

Variable	Variable type	Range of values	Description
Local			
L_CCUT	Binary	0 or 1	Survey station located in recent clearcut (< 15 years)
L_YDEC	Binary	0 or 1	Survey station located in young deciduous stand (< = 90 years)
L_ODEC	Binary	0 or 1	Survey station located in old deciduous stand (> 90 years)
L_PINE	Binary	0 or 1	Survey station located in pine stand
L_SIZE	Numeric	0.5–703.4 ha	Area of stand in which survey station is located
L_CROWN	Numeric	0-85.5%	Mean crown closure among forested polygons
L_DEC	Numeric	0–1.0	Mean deciduous proportion of forested polygons
L_HT	Numeric	0–31.0 m	Mean stand height of forested polygons
Neighborhood			
N_CUT	Numeric	0–0.66	Proportion of neighborhood in a clearcut
N_MID	Numeric	0–0.99	Proportion of neighborhood in mid seral forest (15–90 years)
N_LATE	Numeric	0–1.00	Proportion of neighborhood in late seral forest (> 90 years)
N_DEC	Numeric	0–1.00	Proportion of neighborhood in deciduous forest
N_SW	Binary	0 or 1	Presence of white spruce forest
N_SIMP	Numeric	0–0.83	Habitat patch diversity (Simpson's index)
N_EDGEN	Numeric	0-85.3 m/ha	Natural edge density

using 5 years of data would be developed three times (i.e. 1993–1997, 1994–1998, 1995–1999) and tested three times using ROC_{in} (i.e. 1993–1997, 1994–1998, 1995–1999) and two times using ROC_{out} (i.e. 1998 and 1999). Out-of-sample tests for models that included 1999 data were not possible. We summarized our results graphically using the average ROC_{in} and ROC_{out} value for each 'number of years' group (e.g. the mean of the three models developed with 5 years of data). We made no attempt to interpret inter-model variability because the number of possible models decreased linearly as the number of years included in the model increased. For instance, a model based on 1 year of data could be developed seven times, while one based on 7 years of data could only be developed once.

2.4.3. Results

There was moderate variability among years in the strength and significance of songbird–habitat model coefficients, and the overall model (all years) was generally not a good indicator for individual-year models (Table 2). Exploration of changes in abundance between years could provide further insight into why this might occur. Conversely, the direction (sign) of the coefficients was largely consistent across years (16/21 over all species). For each species except Yellow-rumped Warbler, only one variable was consistently significant across years, and only for Red-breasted Nuthatch and Yellow Warbler was this variable also significant for the overall (all years) model.

The predictive performance of songbird models was more variable across years when assessed using out-of-sample data (ROC_{out}) than when using in-sample data (ROC_{in}) (Table 3 and Fig. 1). Generally, all bird species had good model accuracy with ROC_{in} and ROC_{out} values >0.7 for all years; the exceptions being Yellow-rumped Warbler in 1993 and Red-breasted Nuthatch in 1993–1995. For three of the bird species (Black-throated Green Warbler, White-throated Sparrow, and Yellow Warbler), the values of ROC_{in} and ROC_{out} are very similar and show little variability over time (Fig. 1). In contrast, but only for the first 3 years, the other two species (Red-breasted Nuthatch and Yellow-rumped Warbler) have very different values and exhibit high variability. Differences in patterns for the first 3 years may be accounted for by landscape-level adjustments to forest harvesting in the area (Schmiegelow and Hannon, 1999; Norton et al., 2000). Nevertheless, although the strength and significance of model coefficients are quite variable, the models themselves are consistently in the 'useful applications' and 'high accuracy' categories, with the exceptions noted above. In other words, when prediction is the objective, the models appear to be robust, even though their interpretation may vary across years.

The relationship between mean model performance (i.e. the average of the models with the same number of years of data) and the number of years used to develop the model is summarized in Fig. 2. With the exception of Black-throated Green Warbler, out-of-sample tests (ROC_{out}) were more variable than in-sample tests (ROC_{in}). In fact, in-sample tests appeared to be little affected by the number of years used to develop the models. Out-of-sample evaluations were more variable, but only in the case of Red-breasted Nuthatch did performance increase consistently with number of years. This result is not surprising, given that among those species we analyze here, the Red-breasted Nuthatch exhibits the highest spatial and temporal variance in distribution and abundance (Carlson and Schmiegelow, 2002). For two species, White-throated Sparrow and Yellow-rumped Warbler, out-of-sample tests actually indicated a loss in predictive performance, albeit minor, with number of years. We make no attempt to assign significance, as differences in the number of possible models as a function of the number of years included in the model made interpretation of variance problematic. Nevertheless, both in-sample and out-of-sample model performance was always greater than 0.7 indicating 'useful applications' and 'high accuracy' models.

3. Presence/available (use-vs.-availability) designs

For some types of investigation, e.g. using radiotelemetry, we obtain attribute data from the used sites, and from a random sample of what is available. Resource units (sites) where the organ-

Table 2
Estimated coefficients and standard errors for species presence/absence models for the years 1993–1999

Species/variable	All years	1993	1994	1995	1996	1997	1998	1999
Black-throated Green Warbler								
l_ht	−0.018 (0.045)	0.251 (0.057)***	0.279 (0.097)***	0.272 (0.060)***	0.195 (0.060)***	0.136 (0.023)***	0.179 (0.032)***	0.122 (0.046)***
n_cut	2.138 (1.737)	4.609 (4.063)	3.569 (1.265)***	3.514 (1.261)***	0.023 (1.34)	0.164 (0.963)	−0.813 (1.229)	2.704 (1.722)
n_late	5.032 (1.194)***	0.929 (1.167)	2.671 (1.236)**	4.409 (1.484)***	2.233 (1.034)**	2.133 (0.890)**	2.929 (1.120)***	2.523 (0.901)***
n_dec	5.575 (0.760)***	4.999 (0.992)***	3.098 (1.134)***	1.953 (0.864)**	1.142 (0.819)	3.184 (0.960)***	1.975 (0.725)***	4.875 (0.550)***
n_sw	0.566 (0.363)	1.624 (0.668)**	0.679 (0.484)	1.239 (0.658)*	0.479 (0.504)	0.891 (0.342)***	0.127 (0.279)	0.997 (0.358)***
n_simp	7.942 (1.648)***	2.577 (1.87)	4.447 (1.914)**	4.174 (1.431)***	4.122 (1.315)***	3.789 (1.534)**	4.865 (1.607)***	4.834 (1.258)***
Constant	−10.185 (1.307)***	−11.634 (1.747)***	−13.083 (3.000)***	−13.847 (2.198)***	−9.549 (2.311)***	−8.804 (1.898)***	−9.945 (1.709)***	−10.234 (1.091)***
Red-breasted Nuthatch								
l_ht	0.127 (0.058)**	0.164 (0.028)***	0.149 (0.033)***	0.19 (0.062)***	0.16 (0.023)***	0.232 (0.034)***	0.295 (0.038)***	0.144 (0.032)***
l_dec	−4.053 (1.101)***	0.023 (0.613)	−4.088 (0.830)***	−1.477 (1.38)	−3.602 (0.586)***	−2.986 (0.719)***	−2.994 (0.823)***	−1.463 (0.594)**
Constant	−1.473 (1.694)	−4.314 (0.814)***	−1.719 (0.610)***	−2.31 (0.672)***	−2.538 (0.381)***	−2.031 (0.399)***	−4.807 (0.778)***	−3.093 (0.965)***
White-throated Sparrow								
n_dec	7.317 (1.823)***	5.659 (1.368)***	7.083 (1.916)***	0.587 (2.093)	3.309 (1.461)**	4.234 (1.045)***	3.114 (1.188)***	6.006 (1.139)***
n_simp	4.233 (2.322)*	5.009 (3.51)	6.993 (2.053)***	−0.013 (3.542)	2.812 (2.42)	5.227(1.886)***	4.515 (1.905)**	4.832 (2.170)**
l_dec	2.184 (1.725)	2.839 (1.141)**	2.207 (0.885)**	4.921 (0.737)***	4.247 (0.840)***	3.129 (0.619)***	2.442 (1.071)**	2.657 (0.817)***
l_ccut	19.297 (0)	−0.933 (1.009)	1.533 (0.681)**	22.229 (0)	5.753 (1.197)***	20.489 (0)	3.967 (0.918)***	−0.389 (0.828)
n_mid	−2.85 (1.773)	−1.646 (0.892)*	−5.632 (1.348)***	−3.856 (1.545)**	−1.99 (1.308)	−3.618 (0.887)***	−4.532 (0.805)***	−2.168 (0.787)***
l_pine	1.082 (1.248)	0.446 (0.938)	0.44 (1.006)	−3.728 (1.829)**	−1.439 (0.955)	−0.946 (0.859)	0.307 (0.761)	0.736 (0.88)
n_edgen	−0.014 (0.028)	−0.085 (0.031)***	−0.029 (0.016)*	−0.03 (0.023)	−0.036 (0.026)	−0.036 (0.017)**	−0.024 (0.02)	−0.06 (0.021)***
Constant	−4.246 (1.645)***	−2.222 (1.745)	−3.804 (1.119)***	0.293 (3.206)	−3.022 (1.365)**	−2.45 (1.017)**	−3.146 (1.341)**	−2.935 (1.063)***
Yellow-rumped Warbler								
n_late	5.48 (1.103)***	3.197 (1.989)	3.14 (0.899)***	0.505 (1.173)	1.864 (0.818)**	1.712 (0.906)*	2.361 (0.925)**	4.72 (1.117)***
l_odec	−0.362 (1.293)	−1.959 (1.087)*	−3.203 (1.034)***	−16.922 (0.797)***	−0.403 (0.746)	−0.46 (0.653)	0.325 (0.532)	−1.092 (1.02)
l_ccut	17.191 (0)	−6.52 (1.505)***	−4.28 (1.356)***	−20.862 (0.893)***	−25.478 (0)	−4.466 (0.827)***	−3.831 (0.731)***	−3.491 (1.205)***
n_mid	3.044 (1.327)**	4.32 (1.347)***	3.275 (1.562)**	0.908 (1.367)	7.121 (1.653)***	3.141 (1.179)***	2.58 (1.058)**	2.974 (1.113)***
l_ydec	0.939 (1.223)	−3.122 (1.542)**	−2.128 (1.321)	−17.01 (0)	−3.037 (0.900)***	−1.774 (0.978)*	−0.94 (0.634)	−0.577 (1.149)
l_size	−0.006 (0.001)***	−0.005 (0.001)***	−0.004 (0.001)***	−0.001 (0.001)	−0.003 (0.001)***	−0.003 (0.001)***	−0.001 (0.001)	−0.005 (0.001)***
Constant	−0.358 (0.846)	2.447 (1.973)	2.372 (1.131)**	18.232 (1.054)***	1.121 (0.909)	1.425 (0.979)	−0.005 (0.79)	0.736 (1.077)
Yellow Warbler								
l_odec	2.458 (0.784)***	2.635 (0.447)***	3.333 (0.433)***	2.188 (0.517)***	2.82 (0.370)***	1.496 (0.499)***	2.441 (0.388)***	2.603 (0.457)***
l_crown	−0.026 (0.013)**	−0.005 (0.011)	−0.014 (0.008)*	−0.01 (0.006)	−0.038 (0.009)***	−0.038 (0.008)***	−0.039 (0.010)***	−0.012 (0.01)
Constant	−1.419 (0.89)	−2.599 (0.614)***	−1.945 (0.539)***	−1.107 (0.404)***	−0.604 (0.39)	−0.253 (0.337)	−0.749 (0.468)	−2.226 (0.570)***

Robust standard errors in parentheses. * Significant at 10%; ** significant at 5%; *** significant at 1%.

Table 3
Area under the ROC curve (AUC_{in}), and 1-year ahead predicted area under the ROC curve (AUC_{out}) for the years 1993–1999

Species	Year	Prevalence	N_{in}	AUC_{in}	N_{out}	AUC_{out}
Black-throated Green Warbler	93	0.477	174	0.881	235	0.770
	94	0.396	235	0.867	311	0.820
	95	0.447	311	0.860	204	0.842
	96	0.412	204	0.866	337	0.855
	97	0.282	337	0.861	355	0.831
	98	0.352	355	0.855	355	0.861
	99	0.324	355	0.875		
Red-breasted Nuthatch	93	0.224	174	0.866	235	0.653
	94	0.353	235	0.733	311	0.697
	95	0.209	311	0.806	204	0.534
	96	0.618	204	0.694	337	0.742
	97	0.157	337	0.777	355	0.805
	98	0.530	355	0.851	355	0.870
	99	0.324	355	0.869		
White-throated Sparrow	93	0.747	174	0.934	235	0.836
	94	0.626	235	0.932	311	0.879
	95	0.823	311	0.912	204	0.854
	96	0.902	204	0.898	337	0.936
	97	0.798	337	0.950	355	0.935
	98	0.859	355	0.941	355	0.878
	99	0.710	355	0.901		
Yellow-rumped Warbler	93	0.810	174	0.939	235	0.591
	94	0.749	235	0.958	311	0.819
	95	0.759	311	0.860	204	0.788
	96	0.765	204	0.786	337	0.934
	97	0.656	337	0.951	355	0.873
	98	0.685	355	0.884	355	0.834
	99	0.656	355	0.857		
Yellow Warbler	93	0.161	174	0.825	235	0.778
	94	0.145	235	0.778	311	0.870
	95	0.270	311	0.870	204	0.785
	96	0.407	204	0.785	337	0.828
	97	0.279	337	0.828	355	0.731
	98	0.231	355	0.759	355	0.776
	99	0.217	355	0.795		

N_{in} and N_{out} refer to the number of observations in the model building and model testing sets, respectively. Prevalence indicates the frequency of occurrence of the species over all stations for a given year.

ism is present will be scored 1 whereas all possible units (or a sample of available units) will be scored 0. If used sites are points, there are infinitely many available points, thus, the presence/absence approach is not appropriate. Observed locations are presumed to have been drawn from a sample of available sites, thus, observed values must be a subset of what is available. We assume a particular function between the relative probability of use, $w(\cdot)$, and a vector of n predictor variables, $x = x_1$, x_2, x_3, …, x_n, often using the log-linear form:

$$w(x) = \exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n) \quad (1)$$

Due to the mathematical relationship between the Poisson and binomial distributions, we can estimate β coefficients in the model at Eq. (1) using coefficients from a logistic regression. So we will define an estimating function,

$$\tau(x) = \frac{\exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}{[1 + \exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)]} \quad (2)$$

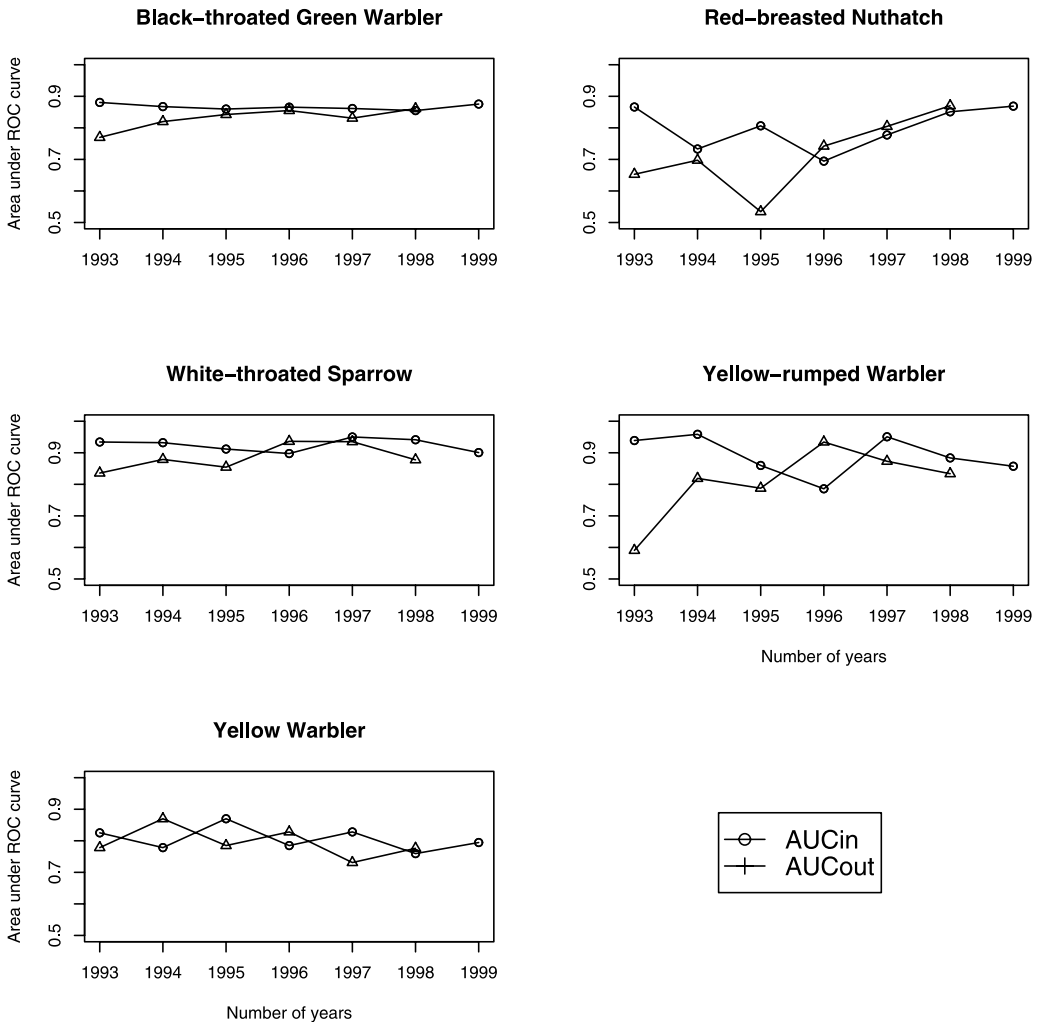

Fig. 1. Temporal variability in ROC (AUC_{in} and AUC_{out}) for five species of boreal songbirds in Alberta 1993–99. All models were developed using 1 year of data and validated using the following year of data. No validation was possible for the 1999 models.

with occupied resource units assuming a 1 and random landscape locations a 0. The predicted RSF values, $w(x)$, are commonly scaled so that they are bounded by 0 and 1, say by dividing by the maximum RSF value, but this is not necessary. Sampling protocols for estimating RSF models have been developed in detail (Manly et al., 1993). Although evaluation procedures for RSF models based on presence/absence data using logistic regression have received attention (Hosmer and Lemeshow, 1989; Fielding and Bell, 1997), evaluation of RSF models for use/availability data are little studied.

3.1. Model selection

We see no reason to question the appropriateness of AIC or BIC for use/availability data when selecting the best RSF model from a set of alternative models. Likelihood methods should not be affected by the structure of presence/available data and we would follow Burnham

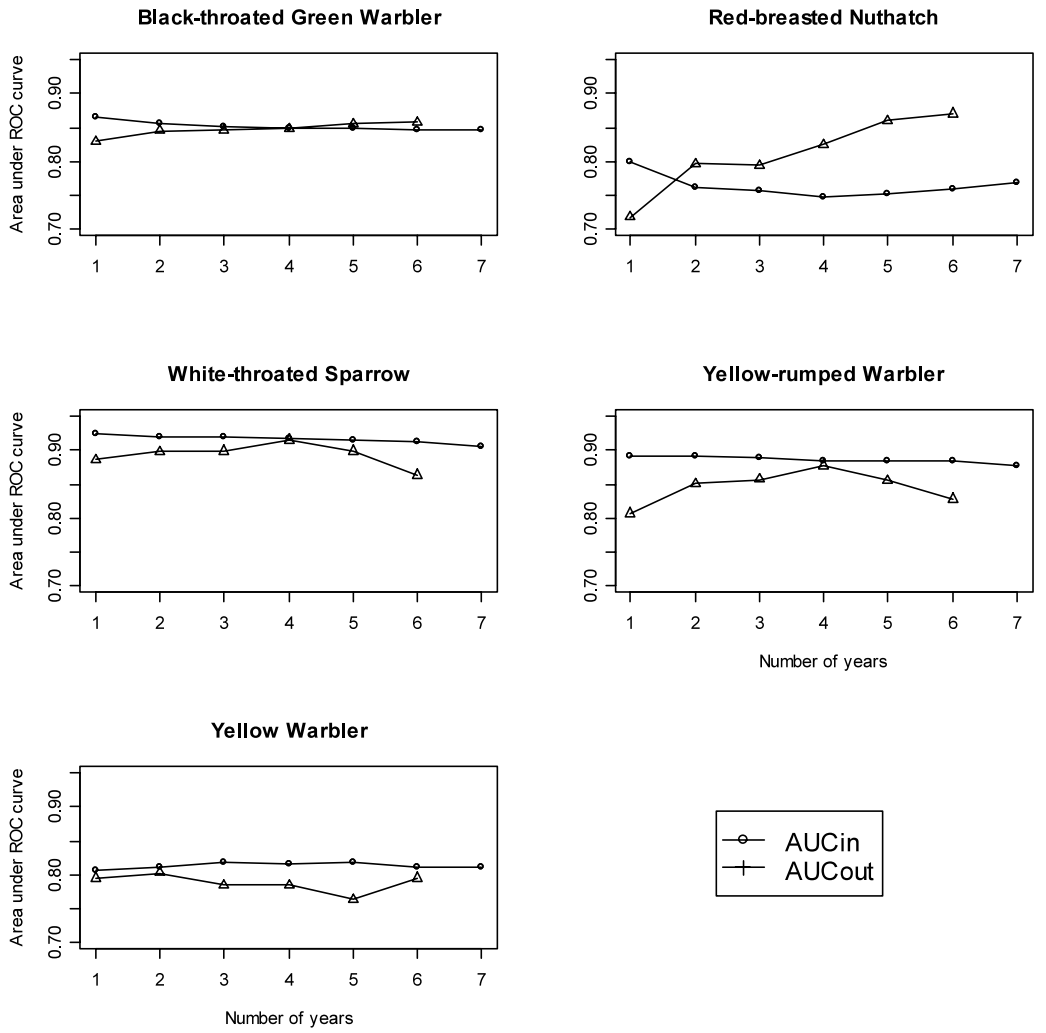

Fig. 2. Relationship between mean model performance (AUC_{in} and AUC_{out}) and the number of years used to build the models. Data were validated using data from the following year. Note that out-of-sample predictions were not possible for models developed using all 7 years of data.

and Anderson (1998) in advocating the use of AIC and related methods for model selection.

3.2. Prediction

Many methods for evaluating logistic regression model predictions are inappropriate for presence/available data. For presence/available data, the distribution of used sites is drawn directly from the distribution of available sites, so these are not exclusive categories as in usual applications of logistic regression. Used sites are expected to have 1's but we also expect that these same sites will be included amongst the set of available sites with 0's. The categories of predicted and observed values are not unique and as a consequence, classification success may be low. All methods related to the confusion matrix, including Kappa and ROC, are flawed when data come from a sampling scheme involving presence/available data. This often leads to confusion with researchers believing that the model is poor because of low classification success.

But even a 'perfect' RSF model might not predict a value greater than some low value.

Of the methods that we reviewed under Section 2 for presence/absence data, none of the standard statistics for logistic regression are appropriate. We propose a method that evaluates prediction success from RSF models built with presence/available data using a form of k-fold cross validation.

3.3. Robustness

The same issues that limit the robustness of models constructed from used/unused data plague models constructed from use/available designs. However, for presence/available data we cannot use out-of-sample binary classification success as was possible for used/unused designs.

3.4. Case study 2: grizzly bear habitat selection

We evaluated within-home-range resource selection for grizzly bears (*U. arctos*) in wilderness areas of the Greater Yellowstone Ecosystem (GYE) of Wyoming, Montana, and Idaho, USA. We stratified bear presence data into two ecological seasons following Mattson et al. (1998) (1) summer or early hyperphagia (15 July–31 August); and (2) fall or late hyperphagia (1 September to denning). Here we used presence/availability data to fit a RSF model of the form of Eq. (1). Our objective for this case study was to evaluate model prediction using a RSF model based on a presence/available design.

3.4.1. Bear and habitat variables

We used 1072 VHF radiotelemetry locations from 92 grizzly bears gathered between 1989 and 1997 (after the fires of 1988). Availability of resources was sampled using 10 318 randomly generated points in a GIS. Environmental predictor variables included elevation from a digital elevation model (DEM), the square (Gaussian transformation) of elevation, greenness derived from a tasseled-cap transformation of spectral reflectance from a Landsat image, and habitat cover type following aggregate functional habitats outlined by Mattson et al. (1998). In total, 17 separate habitat classes were delineated, with dry Douglas fir forests representing the reference category for comparison, because categorical dummy variable coding required one habitat to be removed. An indicator contrast was used where estimated coefficients were based on their comparison with this reference category.

3.4.2. Statistical analysis

We divided the data, by season (early hyperphagia and late hyperphagia), into cross-validation groups following a k-fold partitioning design (Fielding and Bell, 1997; Hastie et al., 2001). Huberty's (1994) rule of thumb was used to determine the model training-to-testing ratio. Based on this rule, a testing ratio of 20% was determined and a k-fold partition of five groups considered. Using cross-validation procedures, we trained our model iteratively on four of the five data sets using logistic regression. Validation was based on the remaining testing set. We estimated all 19 parameters of interest (greenness, 16 habitats, elevation, and elevation^2) in full models. Since habitat was a categorical variable, we used dry Douglas fir forests as the reference category. All habitat estimates are in comparison with this reference.

To examine model performance, we investigated the pattern of predicted RSF scores for partitioned testing data (presence-only) against categories of RSF scores (bins). A Spearman-rank correlation between area-adjusted frequency of cross-validation points within individual bins and the bin rank was calculated for each cross-validated model. Area-adjusted frequencies in this case were simply the frequency of cross-validated use locations with a bin adjusted (divided) by the area of that range of RSF scores available across the landscape. An area-adjusted frequency of 1.0 would indicate that cross-validated use locations occurred at rates expected by chance. A model with good predictive performance would be expected to be one with a strong positive correlation, as more use locations (area-adjusted) would continually be falling within higher RSF bins. To determine bin size and number, we divided predictions into 20 equal-interval bins scaled between the minimum and maximum scores. We further simplified the 20 bins

into 10 more-or-less equal sample sized bins, since rare RSF bins were occurring at the tail distributions of scores. These bins were often without validation points or reliable samples of available scores and, thus, warranted merging. A number of different approaches can be employed to stratify bins, including histogram-equalized stretches that base bin levels on frequency of occurrence (Lillesand and Kiefer, 1994).

3.4.3. Results

During summer (early hyperphagia), grizzly bears selected for areas of high greenness, low elevation, and mesic forest-opening habitats. In comparison, during the fall (late hyperphagia) season, bears selected for high elevation sites along with fen, grouse-whortleberry, and whitebark pine communities. Resource switching, following the phenology of foods, was the likely cause of seasonal differences. Bears were likely using green herbaceous foods during early hyperphagia, while keying in on berries and whitebark pine nuts during fall (late hyperphagia).

Area-adjusted frequencies displayed significant positive rank values (Spearman-rank correlation) across RSF bins for summer and fall seasons (Table 4). Although the summer model was slightly more significant overall, there was greater variation within k-folded sets, particularly at high RSF values (Fig. 3). All individual summer model sets, however, demonstrated significant Spearman-rank correlations, indicating little evidence for poor model performance. In fall (late hyperphagia) we also found significance for all model sets, although individual sets appeared to be more consistent. Frequencies for fall models, however, displayed stronger differentiation between low and high RSF bins (Fig. 4). By comparison, a number of the summer models were not stable at high RSF bins as evidenced by decreasing frequency values. Although both seasons appear to predict bear occurrence well, fall models are considered more consistent across all RSF bins, with both low frequency in lower bins and higher frequency in high bins.

Table 4
Cross-validated Spearman-rank correlations (r_s) between RSF bin ranks and area-adjusted frequencies for individual and average model sets

Set	Summer		Fall	
	r_s	P	r_s	P
1	0.842	<0.01	0.936	<0.001
2	0.915	<0.001	0.929	<0.001
3	0.939	<0.001	0.897	<0.001
4	0.782	<0.01	0.951	<0.001
5	0.867	<0.002	0.874	<0.002
Average	0.988	<0.001	0.972	<0.001

Results are presented by seasons: summer (early-hyperphagia) and fall (late-hyperphagia).

4. Discussion and conclusions

Some form of bootstrapping would appear to be a useful approach for evaluating predictive success, but bootstrapped errors are likely to be too small because of overlap between the training dataset and the test sample. An area for development would be to devise a bootstrap method similar to that proposed by Hastie et al. (2001) (pp. 217–220) that could be applied to evaluate prediction success for RSF models.

A number of in-sample and out-of-sample model evaluation techniques are available for presence/absence modeling (Fielding and Bell, 1997). Although in-sample resubstitution techniques are frequently used, they have a tendency to produce over-fitted models, optimistic estimates of model performance and loss of generality (Chatfield, 1995). In a number of situations (museum specimens, herbarium records and animal telemetry data), the data used to model the occurrence of the species follows a presence/available design, not a presence/absence design (Manly et al., 1993). There is an important distinction between the two, because group membership is known in the former but uncertain in the latter. Additionally, the presence/available sampling design entails obtaining two samples requiring different statistical treatment (Manly et al., 1993). Problems can arise when a confusion matrix (classification table, ROC, etc.) is used on presence/available data. Confusion matrices and related methods of Kappa and ROC are not recommended for presence/

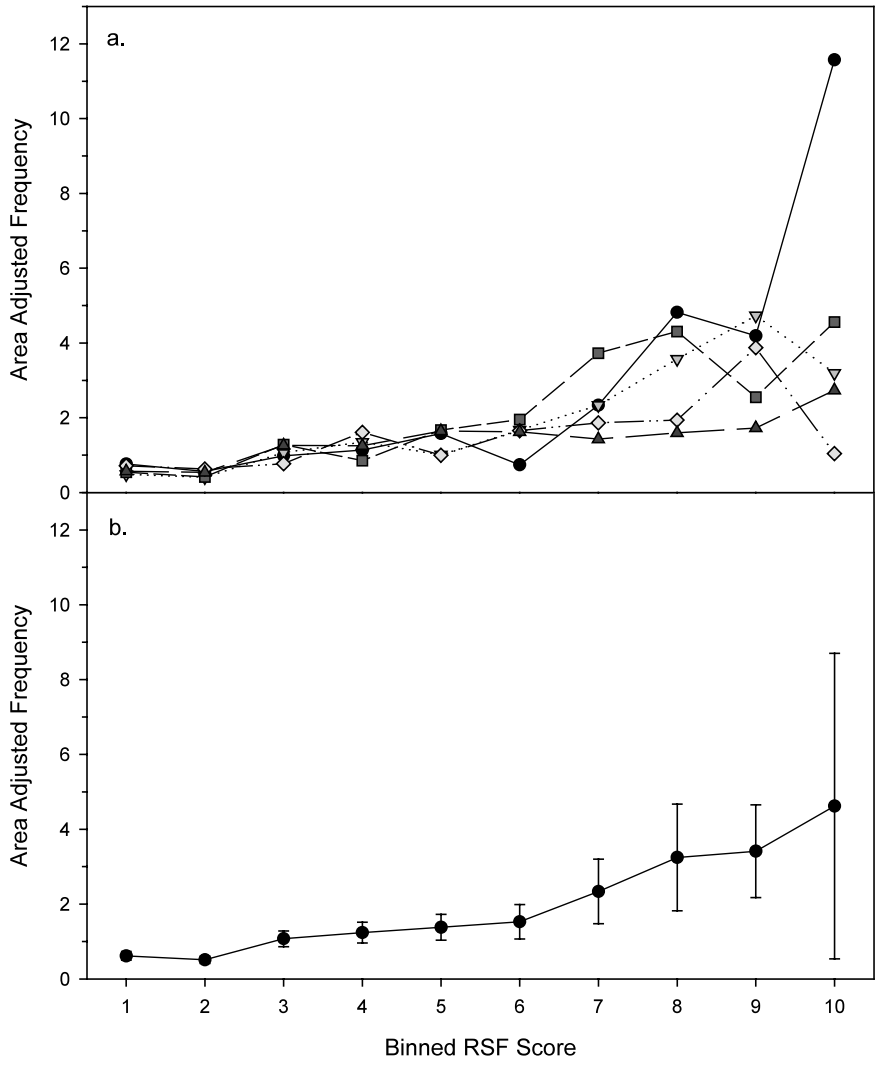

Fig. 3. Area-adjusted frequency of categories (bins) of RSF scores for withheld locations of grizzly bears for summer (early-hyperphagia) RSF models in the Greater Yellowstone Ecosystem, USA. Frequency values for individual cross-validation sets ($n = 5$) are depicted with unique symbols (graph a). Mean ($\pm$S.D.) frequency values by RSF-score bin are illustrated in graph b. A Spearman-rank correlation for mean frequency values by bins ($r_s = 0.988$, $P < 0.001$) indicates that the model predicted cross-validated use locations well.

available data. Instead, the k-fold cross validation method that we developed here, or out-of-sample evaluation techniques are preferred.

In some circumstances, the development of robust RSF models may be quite straightforward (e.g. habitat specialists like Northern Spotted Owls). However, given the spatially and temporally dynamic nature of habitat selection common to many species, robust RSF models are not necessarily expected. If models are not robust, we believe that there are opportunities to strive for general models by understanding functional responses (Mysterud and Ims, 1998; Boyce et al., 1999) and the influence of environmental variation on the availability or quality of habitat resources. A mechanistic approach and analysis of RSF model evaluation is an important next step that traditionally has been lacking in most RSF ana-

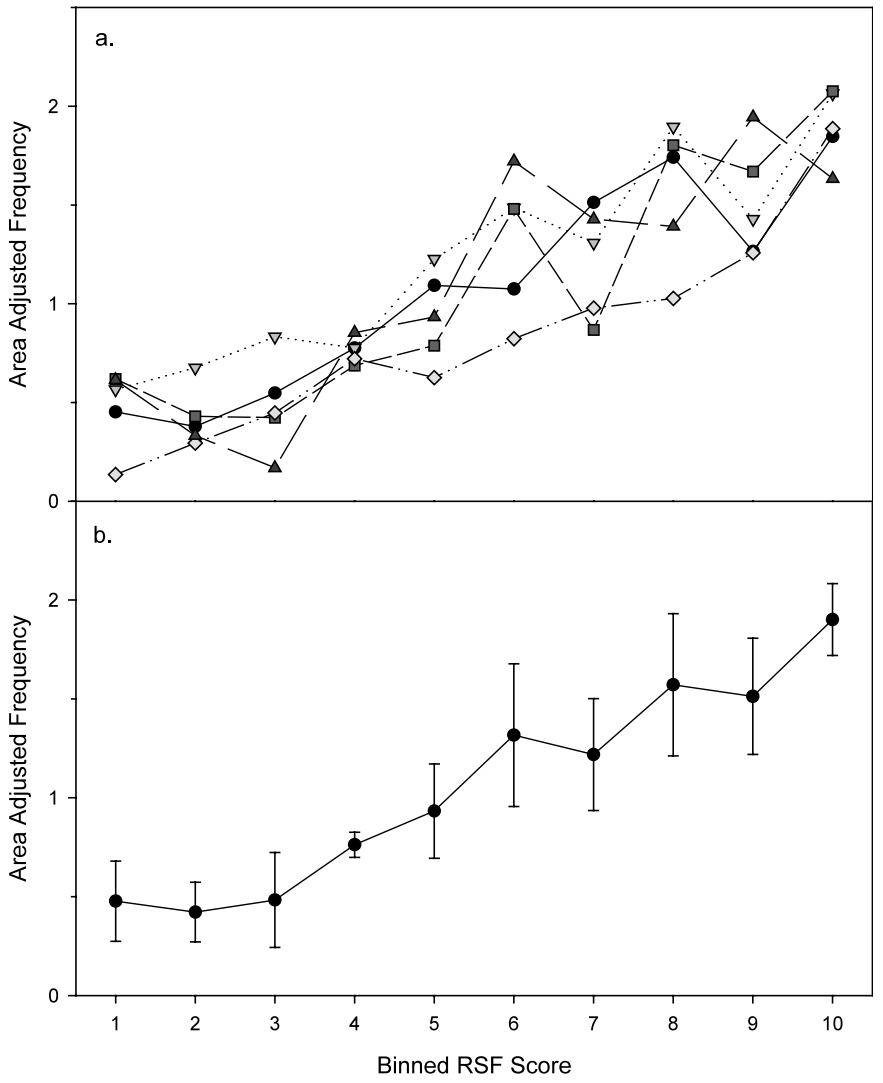

Fig. 4. Area-adjusted frequency of binned cross-validated use locations for fall (late-hyperphagia) RSF models in the Greater Yellowstone Ecosystem, USA. Frequency values for individual cross-validation sets ($n = 5$) are depicted with unique symbols (graph a). Mean ($\pm$S.D.) frequency values by bin are illustrated in graph b. A Spearman-rank correlation for mean frequency values by bins ($r_s = 0.972$, $P < 0.001$) indicates that the model predicted cross-validated use locations well.

lyses (Garshelis, 2000). Understanding such relationships is of crucial importance in natural resource management and conservation, because managers and conservationists are asked to provide habitat-based models describing the influence of changing land-use activities on sensitive or rare species (cumulative effects assessments, population viability analyses, climate change models, etc.). A less satisfying but possibly necessary approach is simply to develop different RSF models for different seasons, years or localities (Jaberg and Guisan, 2001). Generalist species, like grizzly bears, likely will require such an approach, because substantial differences in selection are apparent between seasons (Boyce and Waller, 2000; Nielsen et al., 2002) and over regional scales. Our first case study points to such problems, where estimated resource selection coefficients were quite

variable between years, making the generation of general models (multi-year response) difficult for most species. Such adaptive behavior by some bird species encumbers possible application of such models in natural resource management and conservation planning.

In general, RSF models based on presence/available data are not well evaluated using usual metrics of classification success. Better model evaluation is achieved by withholding data (k-fold partitioning) for testing model predictions or by comparing RSF predictions using models developed at other times and places (prospective sampling). Without such evaluations it is difficult to interpret the predictive ability of the RSF model, and therefore, the reliability of these models as resource management tools.

Acknowledgements

Thanks to the Natural Sciences and Engineering Research Council of Canada, the National Science Foundation, the Alberta Conservation Association, the Canadian Foundation for Innovation (M.S. Boyce), and the Sustainable Forest Management Network NCE (PRV & FKAS) for support. We benefited from discussions and correspondence with A. Guisan, S. Cherry, J. Elith, T. Hastie, C. Johnson, S. Lele, B.F.J. Manly, L.L. McDonald, E.H. Merrill, D. Ouren and C. Schwartz.

References

Altman, D.G., Lausen, B., Sauerbrei, W., Schumacher, M., 1994. Dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. J. Natl. Cancer Inst. 86, 829–835.

Anderson, D.R., Burnham, K.P., Thompson, W.L., 2000. Null hypothesis testing: problems, prevalence, and an alternative. J. Wildl. Manage. 64, 912–923.

Austin, M.P., Nicholls, A.O., Margules, C.R., 1990. Measurement of the realized qualitative niche: environmental niches of five *Eucalyptus* species. Ecol. Monogr. 60, 161–177.

Austin, M.P., Nicholls, A.O., Dherty, M.D., Meyers, J.A., 1994. Determining species response functions to an environmental gradient by means of a β-function. J. Veg. Sci. 5, 215–228.

Boyce, M.S., 2001. Statistics as viewed by biologists. J. Agric. Biol. Environ. Stat. 7, 51–57.

Boyce, M.S., McDonald, L.L., 1999. Relating populations to habitats using resource selection functions. Trends Ecol. Evol. 14, 268–272.

Boyce, M.S., Waller, J., 2000. The application of resource selection functions analysis to estimate the number of grizzly bears that could be supported by habitats in the Bitterroot ecosystem. In: Servheen, C. (Ed.), Grizzly Bear Recovery in the Bitterroot Ecosystem. Final Environmental Impact Statement. Appendix 21B.US. Fish & Wildlife Service, Missoula, Montana, USA, pp. 6-231–6-241.

Boyce, M.S., Meyer, J.S., Irwin, L.L., 1994. Habitat-based PVA for the northern spotted owl. In: Fletcher, D.J., Manly, B.F.J. (Eds.), Statistics in Ecology and Environmental Monitoring, Otago Conference Series No. 2. University Otago Press, Dunedin, New Zealand, pp. 63–85.

Boyce, M.S., McDonald, L.L., Manly, B.F.J., 1999. Reply to Mysterud and Ims. Trends Ecol. Evol. 14, 490.

Burnham, K.P., Anderson, D.R., 1998. Model Selection and Inference: A Practical Information—Theoretic Approach. Springer, New York, p. 353.

Burnham, K.P., Anderson, D.R., 2001. Kullback–Leibler information as a basis for strong inference in ecological studies. Wildl. Biol. 28, 111–119.

Carlson, M.J., Schmiegelow, F.K.A., 2002. Cost-effective sampling design for broad-scale avian monitoring. Conserv. Ecol., submitted for publication.

Chatfield, C., 1995. Model uncertainty, data mining and statistical inference. J. R. Stat. Soc., Series A 158, 419–466.

Cherry, S., 1998. Statistical tests in publications of the Wildlife Society. Wildl. Soc. Bull. 26, 947–953.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46.

Cohen, J., 1968. Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol. Bull. 70, 213–220.

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens. Environ. 37, 35–46.

Cumming, G.S., 2000. Using between-model comparisons to fine-tune linear models of species ranges. J. Biogeogr. 27, 441–455.

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

Garshelis, D.L., 2000. Delusions in habitat evaluation: measuring use, selection, and importance. In: Boitani, L., Fuller, T.K. (Eds.), Research Techniques in Animal Ecology: Controversies and Consequences. Columbia University Press, New York, pp. 111–164.

Guisan, A., Harrell, F.E., 2000. Ordinal response regression models in ecology. J. Veg. Sci. 11, 617–626.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186.
Guisan, A., Theurillat, J.P., Kienast, F., 1998. Predicting the potential distribution of plant species in an Alpine environment. J. Veg. Sci. 9, 65–74.
Hastie, T., Tibshirani, T., Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, p. 533.
Hosmer, D.W., Jr, Lemeshow, S., 1989. Applied Logistic Regression. Wiley, New York, p. 307.
Huberty, C.J., 1994. Applied Discriminant Analysis. Wiley Interscience, New York.
Jaberg, C., Guisan, A., 2001. Modelling the distribution of bats in relation to landscape structure in a temperate mountain environment. J. Appl. Ecol. 38, 1169–1181.
Johnson, D.H., 1980. The comparison of usage and availability measurements for evaluating resource preference. Ecology 61, 65–71.
Johnson, D.H., 1999. The insignificance of statistical significance testing. J. Wildl. Manage. 63, 763–772.
Kneeland, M.C., Boyce, M.S. Kubisiak, J.F., 2002. Habitat ecology of a cyclic ruffed grouse population. Can. J. Zool., under revision.
Lennon, J.J., 1999. Resource selection functions: taking space seriously. Trends Ecol. Evol. 14, 399–400.
Lennon, J.J., 2000. Red-shifts and red herrings in geographical ecology. Ecography 23, 101–113.
Lillesand, T.M., Kiefer, R.W., 1994. Remote Sensing and Image Interpretation, third ed.. Wiley, New York, p. 750.
Long, J.S., 1997. Regression models for categorical and limited dependent variables. In: Advanced Quantitative Techniques in the Social Sciences 7. SAGE Publications, London.
Manel, S., Ceri Williams, H., Ormerod, S.J., 2001. Evaluating presence–absence models in ecology: the need to account for prevalence. J. Appl. Ecol. 38, 921–931.
Manly, B.F.J., 1985. The Statistics of Natural Selection. Chapman & Hall, London, p. 484.
Manly, B.F.J., McDonald, L.L., Thomas, D.L., 1993. Resource Selection by Animals: Statistical Design and Analysis for Field Studies. Chapman & Hall, London, p. 177.
Mattson, D.J., Barber, K., Maw, R. and Renkin, R., 1998. Coefficients of productivity for Yellowstone's grizzly bear habitat. Report of the U.S. Geological Survey Forest and Rangeland Ecosystem Science Centre, Moscow, Idaho, p. 74.
McIntosh, A.R., Townsend, C.R., Crowl, T.A., 1992. Competition for space between introduced brown trout (*Salmo trutta* L.) and a native galaxiid (*Galaxias vulgaris* Stokell) in a New Zealand stream. J. Fish Biol. 41, 63–81.
McNay, R.S., Morgan, J.A., Brunnell, F.L., 1994. Characterizing independence of observations in movements of Columbian black-tailed deer. J. Wildl. Manage. 58, 422–429.
Metz, C.E., 1978. Basic principles of ROC analysis. Sem. Nucl. Med. 8, 283–298.
Meyer, J.S., Irwin, L.L., Boyce, M.S., 1998. Influence of habitat abundance and fragmentation on spotted owls in western Oregon. Wildl. Monogr., No. 139 1–51.
Monserud, R.A., Leemans, R., 1992. Comparing global vegetation maps with the Kappa statistic. Ecol. Model. 62, 275–293.
Murtaugh, P.A., 1996. The statistical evaluation of ecological indicators. Ecol. Appl. 6, 132–139.
Mysterud, A., Ims, R.A., 1998. Functional responses in habitat use: availability influences relative use in trade-off situations. Ecology 79, 1435–1441.
Naesset, E., 1996. Use of the weighted Kappa coefficient in classification error assessment of thematic maps. Int. J. Geogr. Inf. Syst. 10, 591–604.
Nielsen, S.E., Boyce, M.S., Stenhouse, G.B., Munro, R.H.M., 2002. Modeling grizzly bear habitats in the Yellowhead ecosystem of Alberta: taking autocorrelation seriously. Ursus. (in press).
Norton, M.R., Hannon, S.J., Schmiegelow, F.K.A., 2000. Fragments are not islands: patch vs. landscape perspectives on songbird presence and abundance in a harvested boreal forest. Ecography 23, 209–223.
Oreskes, N., Shrader-Frechette, K., Belitz, K., 1994. Verification, validation, and confirmation of numerical models in the earth sciences. Science 263, 641–646.
Osborne, P.E., Suárez-Seoane, S., 2002. Should data be partitioned spatially before building large-scale distribution models? Ecol. Model., (in press).
Otis, D.L., White, G.C., 1999. Autocorrelation of location estimates and the analysis of radiotracking data. J. Wildl. Manage. 63, 1039–1044.
Pearce, J., Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Model. 133, 225–245.
Platt, J.R., 1964. Strong inference. Science 146, 347–353.
Popper, K.R., 1934. Logik der Forschung. Julius Springer Verlag, Vienna, Austria, p. 480.
Schmiegelow, F.K.A., Hannon, S.J., 1999. Forest-level effects of fragmentation on boreal songbirds: the Calling Lake Fragmentation Studies. In: Rochelle, J.A., Lehmann, L.A., Wisniewski, J. (Eds.), Forest Fragmentation: Wildlife and Management Implications. Brill, Leiden, pp. 201–221.
Schmiegelow, F.K.A., Machtans, C.S., Hannon, S.J., 1997. Are boreal birds resilient to forest fragmentation? An experimental study of short-term community responses. Ecology 78, 1914–1932.
Schooley, R.L., 1994. Annual variation in habitat selection: patterns concealed by pooled data. J. Wild. Manage. 58, 367–374.
StataCorp., 2001. STATA STATISTICAL Software, Release 7.0. Stata Corporation, College Station, Texas.
Swets, J.A., 1988. Measuring the accuracy of diagnostic systems. Science 240, 1285–1293.
Swihart, R.K., Slade, N.A., 1985a. Influence of sampling interval on estimates of home-range size. J. Wildl. Manage. 49, 1019–1025.
Swihart, R.K., Slade, N.A., 1985b. Testing for independence of observations in animal movements. Ecology 66, 1176–1184.
U.S. Fish and Wildlife Service, 1981. Standards for the development of suitability index models. Ecology Service

Manual 103, U.S. Fish and Wildlife Service, Division of Ecological Sciences, Washington, DC, USA.

Vernier, P.R., Schmiegelow, F.K.A., Cumming, S.G., 2002. Modeling bird abundance from forest inventory data in the boreal mixed-wood forest of Canada. p 559–571. In: Scott, J.M., Heglund, P.J., Morrison, M., Raphael, M., Haufler, J., Wall, B. (Eds.), Predicting Species Occurrences: Issues of Scale and Accuracy. Island Press, Covello, California (draft version available at http://www.rr.ualberta.ca/research/best/birdmodels.pdf).

Yoccoz, N.G., 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. Bull. Ecol. Soc. Am. 72, 106–111.

Zaniewski, A.E., Lehmann, A., Overton, J.McC., 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecol. Model. (in press).