Journal of Applied Ecology 2006
43, 405–412

REVIEW

Modelling distribution and abundance with presence-only data

JENNIE L. PEARCE* and MARK S. BOYCE†

**Canadian Forest Service, Great Lakes Forestry Centre, Landscape Analysis and Application Section, 1219 Queen St E, Sault St Marie, ON P6A 2E5 Canada; and †Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E9 Canada*

Summary

1. Presence-only data, for which there is no information on locations where the species is absent, are common in both animal and plant studies. In many situations, these may be the only data available on a species. We need effective ways to use these data to explore species distribution or species use of habitat.

2. Many analytical approaches have been used to model presence-only data, some inappropriately. We provide a synthesis and critique of statistical methods currently in use to both estimate and evaluate these models, and discuss the critical importance of study design in models where only presence can be identified

3. Profile or envelope methods exist to characterize environmental covariates that describe the locations where organisms are found. Predictions from profile approaches are generally coarse, but may be useful when species records, environmental predictors and biological understanding are scarce.

4. Alternatively, one can build models to contrast environmental attributes associated with known locations with a sample of random landscape locations, termed either 'pseudo-absences' or 'available'. Great care needs to be taken when selecting random landscape locations, because the way in which they are selected determines the modelling techniques that can be applied.

5. Regression-based models can provide predictions of the relative likelihood of occurrence, and in some situations predictions of the probability of occurrence. The logistic model is frequently applied, but can rarely be used directly to estimate these models; instead, case–control or logistic discrimination should be used depending on the sample design.

6. Cross-validation can be used to evaluate model performance and to assess how effectively the model reflects a quantity proportional to the probability of occurrence. However, more research is needed to develop a single measure or statistic that summarizes model performance for presence-only data.

7. *Synthesis and applications.* A number of statistical procedures are available to explore patterns in presence-only data; the choice among them depends on the quality of the presence-only data. Presence-only records can provide insight into the vulnerability, historical distribution and conservation status of species. Models developed using these data can inform management. Our caveat is that researchers must be mindful of study design and the biases inherent in presence data, and be cautious in the interpretation of model predictions.

Key-words: case–control, distribution, habitats, logistic discrimination, logistic regression, presence-only studies, pseudo-absences, resource selection functions, RSF, sampling

Journal of Applied Ecology (2006) **43**, 405–412
doi: 10.1111/j.1365-2664.2005.01112.x

Correspondence: Jennie Pearce, 1405 Third Line East, Sault Ste Marie, ON P6A 6J8, Canada (e-mail: jlpearce@shaw.ca).

Introduction

To manage a species effectively, conservation projects may require a description of a species' geographical distribution or use of habitats. Examples include reserve design (Araújo & Williams 2000), population viability analysis (Boyce *et al*. 1994; Akçakaya *et al*. 2004) and species or resource management (Johnson *et al*. 2004). Rarely are survey data available to describe species presence at every location on the landscape. Thus models are used to interpolate, or extrapolate beyond the locations where species presence is known, by relating species presence to environmental variables. This has been facilitated by remotely sensed data, allowing assessment of the distribution of resources over large, and even inaccessible, areas.

Many approaches have been used to model 'presence–absence' or 'used–unused' data (see Guisan & Zimmermann 2000 for a review). However, there is growing interest in making use of 'presence-only' data, consisting only of observations of the organism but with no reliable data on where the species was not found. Sources for these data include atlases, museum and herbarium records, species lists, incidental observation databases and radio-tracking studies.

Developing models of species distribution for presence-only data is challenging (Graham *et al*. 2004). Several approaches have been used; however, the choice among them is not clear. Terminology also differs between studies. For example, some studies refer to the 'presence' of a species, whereas faunal studies on wide-ranging species often refer to 'used' locations, rather than species presence. Here, we refer to the presence of a species for consistency. We review the various steps in modelling the distribution of species when we know some of the locations where they occur on the landscape, but have no information on where they do not occur.

We provide a synthesis and critique of statistical methods currently in use to both estimate and evaluate these models, and discuss the critical importance of study design in models where only presence can be identified. Our objective is to provide ecologists and managers with a wide range of approaches to explore patterns in presence-only data, and to identify analytical aspects that require further development.

Statistical model formulation

We review four approaches taken to describe the presence of a species in relation to environmental predictors when only presence is known. These are:

1. Describing the distribution of the presence-only records.

2. Contrasting the distribution of presence records with that of pseudo-absences.

3. Contrasting the distributions of presence records and available sites.

4. Modelling abundance when abundance given presence is known.

The modelling approaches for (2) and (3) derive from different sampling motivations. In (2), biologists wish to contrast used or consumed resource units such as plots of land, denning or nesting sites, prey or food items, with characteristics of resource units that have not been used or where use has not been recorded. Plants provide the clearest example of this view, where individuals are either present or truly absent at any given point on the landscape, within a given time-frame. Models provide predictions of the relative probability of a resource unit being used, given its characteristics. This differs from the motivation behind (3), where all resource units within the sampling domain are assumed to be available to be used, but some are used more frequently than others. Radiotelemetry studies of species such as grizzly bears *Ursus arctos* provide an example of this view, where bears might potentially be recorded at any point within their home range, but some locations are used more frequently than others. The difference between these sampling motivations is subtle, but explains the historical development of different approaches for similar problems.

DESCRIBING THE DISTRIBUTION OF THE PRESENCE-ONLY RECORDS

This first group of modelling techniques, termed profile techniques, seeks to characterize environmental conditions associated with the presence records without reference to other data points. Environmental envelope techniques are the most widely applied (e.g. Busby 1986; Caughley *et al*. 1987; Lindenmayer *et al*. 1991; Law 1994; Pearce & Lindenmayer 1998; Walther, Wisz & Rahbek 2004). Chief among these techniques have been BIOCLIM (Busby 1986, 1991) and HABITAT (Walker & Cocks 1991). Environmental envelopes enclose presence records into a multidimensional envelope within environmental space. The various techniques use different classification algorithms, but often provide similar results. Predictions are summarized typically as the degree of classification within subenvelopes.

A recent variation on this approach has been the development of support vector machines (SVM) for one-class problems (e.g. Guo *et al*. 2005). SVMS seek to identify an environmental envelope or hyperspace containing the data points, in which the envelope is optimized with respect to the number of points in the envelope and to the number of outliers. The distance between the point and the centre of hyperspace determines membership of the hyperspace. The advantage of this approach over BIOCLIM, for example, is that the SVM hyperspace can be any shape, whereas BIOCLIM uses hyperboxes to enclose the presence data (Guo *et al*. 2005). HABITAT also is more flexible than BIOCLIM, defining the environmental envelope using a convex hull and the relative density of observations within environmental space. SVM, therefore, may be considered a refinement of the HABITAT approach.

Multivariate association methods such as DOMAIN (Carpenter, Gillison & Winter 1993) also require only presence data. DOMAIN defines the degree of similarity among presence sites in terms of environmental conditions. The method can be used to determine either environmental envelopes or a continuous map of similarity.

At a finer scale, utilization distributions (UD) can be used to characterize the distribution of animals. The UD is a probability density function that quantifies an individual's or group's relative use of space (van Winkle 1975). Marzluff *et al*. (2004) have extended this approach by modelling the intensity of use relative to environmental covariates.

Profile techniques summarize environmental characteristics at presence locations, and typically each record has equal weight within the model. Because of this, these techniques are highly dependent on biases in the presence records. Some approaches, such as BIOCLIM, can be highly sensitive to the inclusion of outliers. Elith & Burgman (2002) provide a discussion of the pros and cons of geographical and climatic envelope-based techniques. Predictions from presence-only approaches are generally coarse, but may be useful at meso-scales to describe poorly understood species when species records, environmental predictors, and biological understanding are scarce.

CONTRASTING THE DISTRIBUTION OF PRESENCE VS. PSEUDO-ABSENCE

Many studies have sought to apply presence–absence techniques to presence-only data by generating pseudo-absence data from background areas from which species data are missing. These sites may be selected without replacement from within the study region either randomly (Stockwell & Peterson 2002), randomly with case-weighting to reduce the effective sample size of pseudo-absences (Ferrier & Watson 1996; Ferrier *et al*. 2002), or by using environmentally weighted random sampling (Zaniewski, Lehmann & Overton 2002). Pseudo-absences are assumed to represent true absences, although because sites were not searched some pseudo-absences might represent presence locations (Graham *et al*. 2004). Generalized linear models and generalized additive models have been the most widely applied statistical methods (e.g. Ferrier *et al*. 2002). However, other approaches such as tree-based methods (e.g. Ferrier & Watson 1996) and genetic algorithms (e.g. GARP; Stockwell & Peters 1999) also have been considered.

Regression models have generally performed better than tree-based methods or genetic algorithms in predicting species presence (Ferrier & Watson 1996). Tree-based methods are expected to be highly sensitive to biases within the sample data (Hastie *et al*. 2001), and the underlying model used to make predictions in GARP is largely inaccessible and difficult to interpret (Elith & Burgman 2002).

When using presence-only data it is generally not possible to calculate probabilities of presence; instead we aim to predict the relative likelihood of presence. There are two reasons for this: (a) separate samples of presence and pseudo-absence data have been selected where sampling fractions are not known, and (b) the pseudo-absence data contains an unknown number of presences, and is thus a contaminated sample of absences. To understand this we examine the logistic function and its assumptions. The logistic regression model assumes that a sample is selected, and that this sample contains observations of either the presence ($y = 1$) or the absence ($y = 0$) of a species. For each observation there is a set of habitat measurements $\mathbf{x}$. From this the probability of occurrence $[P(y = 1|)]$ can be estimated:

$$P(y = 1|\mathbf{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)} \qquad \text{eqn 1}$$

This assumes that presence and absence observations were recorded from a sample of resource units in which the presence of the species at a resource unit was not known prior to sampling. Thus the sample contains presence and absence sites in approximate proportion to their occurrence on the landscape. In the absence of habitat information, the probability of occurrence then can be estimated directly from the proportion of observations in the sample at which the species was present. For example, if in a sample of 100 observations, 20 contain the species, the probability of occurrence is 0·2 [= 20/(20 + 80)]. However, with presence-only data, we sample the presence locations independently and then select a sample of pseudo-absence locations, and so the proportion of presences within the sample does not represent the true prevalence of the species in the population, but rather the relative proportion chosen by the researcher. For example, we have a sample of 20 presence records and we select independently a set of 80 'pseudo-absence' records. In this case the probability of occurrence is also 0·2 [= 20/(20 + 80)]. However, if we select 200 pseudo-absence locations, then the probability of occurrence is 0·09 [= 20/(20 + 200)].

When samples for $y = 1$ and $y = 0$ are selected in advance, we need to modify the logistic model to account for the probability that a location has been sampled to obtain probabilities of occurrence. We do this by correcting the model using P_1 and P_0, the proportion of occupied and unoccupied locations, respectively, selected from the total number of occupied and unoccupied locations in the landscape. This also is known as a case–control design.

$$P(y = 1|\mathbf{x}, \text{sampled}) = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \ln\left(\frac{P_1}{P_0}\right)\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \ln\left(\frac{P_1}{P_0}\right)\right)} \qquad \text{eqn 2}$$

In practice we rarely know what proportion of the used and unused locations we have selected in our samples, and so P_0 and P_1 are unknown. Model predictions using the uncorrected logistic function are therefore only relative predictions. Alternatively, we can interpret model coefficients in terms of odds ratios, where the odds that a species will be present given covariate pattern **x**, is compared to a reference habitat, usually one in which the values for x_1 to x_p are set to zero (Keating & Cherry 2004). Thus:

$$\frac{\dfrac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})}}{\dfrac{P(y=1|\mathbf{x}_{\text{reference}})}{P(y=0|\mathbf{x}_{\text{reference}})}} = \exp(\beta_1 x_1 + \ldots + \beta_p x_p) \qquad \text{eqn 3}$$

A further complication of this sampling scheme is that the process of generating pseudo-absences randomly from the landscape of interest means that these locations are actually an unknown mixture of presence and absence locations, unless the species is very rare on the landscape. Keating & Cherry (2004) discuss the difficulties of deriving probabilities of occurrence in case–control designs under these circumstances. However, unless the level of contamination (proportion of presences within the absence sample) is very high, the model may provide acceptable predictions of the relative likelihood of occurrence, or odds-ratios. Based on simulations, Lancaster & Imbens (1996) obtained unbiased estimates of β_is with contamination rates less than 20%. Also, they provide an algorithm for dealing with situations where greater contamination rates exist. This approach seeks to calculate the predicted probability of species presence where presence locations are contrasted with control sites, which are an unknown mixture of occupied and unoccupied locations. The implementation of this approach is complex, not available in standard statistical packages, and frequently fails to converge to a unique solution (Keating & Cherry 2004). Barry, Elith & Pearce (unpublished data) provide a worked example of this approach for habitat studies.

CONTRASTING THE DISTRIBUTION OF PRESENCE SITES WITH AVAILABLE SITES

A slightly different approach has been applied in studies of wide-ranging animals. These studies do not refer to the presence or absence of a species, but rather to how well a habitat is 'used', usually determined through radiotelemetry studies (Frair *et al*. 2004). In these studies, the landscape is considered to be available to the species of interest and potentially used to some extent, but some habitats are occupied more frequently than others within a given time period. These models describe the relative probability of use for different resource units (e.g. a pixel) over the study area, as described by habitat characteristics. The distinction between this approach and the pseudo-absence approach is subtle, because in practice the sampling schemes are similar. However, the underlying conceptual difference between contrasting unoccupied-vs.-occupied locations, and used-vs.-available locations has resulted in the development of a wide range of alternative modelling approaches.

Four approaches have been used to model presence-availability. The first of these, ecological niche factor analysis (ENFA) implemented in the BIOMAPPER package (Hirzel, Hausser & Perrin 2004) is similar to profile techniques. ENFA uses factor analysis to quantify the environmental conditions of the presence sites by comparing them to the environmental conditions of the entire region of interest, and predictions are provided as a habitat suitability index (Hirzel *et al*. 2002; Dettki, Löfstrand & Edenius 2003; Reutter *et al*. 2003; Brotons *et al*. 2004; Chefaoui, Hortal & Lobo 2005). ENFA considers the density of points within subenvelopes of data and is therefore an improvement on presence-only approaches. This technique is generally optimistic regarding species distribution, which may be an advantage when a species does not occupy all suitable habitats on the landscape (Hirzel, Helfer & Metral 2001; Brotons *et al*. 2004). The two-class SVD model uses a similar approach to ENFA, except that it does not assume a particular probability distribution for the data (Guo *et al*. 2005).

A second approach to modelling presence-availability involves using case–control logistic regression where used resource units are contrasted with random locations within an activity area available to individuals. There are different sampling designs available to conduct this, where cases may be matched or unmatched with controls (Collett 1991; Arthur *et al*. 1996; Manly *et al*. 2002). Examples of this approach include contrasting wood turtle *Clemmys insculpta* locations with paired random locations (Compton, Rhymer & McCollough 2002) and contrasting superb parrot *Polytelis swainsonii* nest trees with paired random trees (Manning, Lindenmayer & Barry 2004). Models estimated using case–control logistic regression are based on the contrasts between used and control resource units and can be interpreted as odds ratios or relative likelihoods of occurrence (Keating & Cherry 2004). The discussion in the previous section about the contamination of controls also applies here.

A third approach proposed by Manly *et al*. (2002) uses logistic regression to estimate relative likelihoods using an exponential model:

$$P(y=1|\mathbf{x}) = \exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p) \qquad \text{eqn 4}$$

This model has been used widely in resource selection studies (e.g. Campos *et al*. 1997; Johnson *et al*. 2002; Nielsen *et al*. 2002; Boyce *et al*. 2003), rather than the logistic function, because it avoids the problem of different denominators encountered in the logistic model. However, as Manly *et al*. (2002: 101–102) point out, this approach assumes a particular sampling scheme. In particular, this approach requires that one sample of presence locations and one sample of available locations be taken, and that any single location selected that occurs in both the presence and the available samples be included only in the available sample. McDonald (2003) shows that the duplicate records can be removed from the available sample rather than the observed sample unless the number of duplicates is high. Manly *et al.* (2002) show how, with known sampling frequencies of presence and available samples, probabilities of occurrence can be calculated, although Keating & Cherry (2004) question this model. However, in practice sampling probabilities are unknown, and irrespective of the validity of the model formulation, model predictions provide relative likelihoods of occurrence (i.e. the RSF). When interpreted as relative likelihoods, it is not necessary that the predictions are constrained to lie below 1, a concern raised by Keating & Cherry (2004).

A fourth approach is to use the logistic regression algorithm to approximate a logistic discrimination model. Here we use the logistic model to estimate a function that discriminates between two distributions of habitat covariates, one set associated with locations where the species is present $f_{y=1}(\mathbf{x})$ and another set associated with random (available) locations $f_{y=0}(\mathbf{x})$ (Keating & Cherry 2004). We sample independently from each distribution, with probability π_1 of a sampled observation (from the joint distribution of presence and available sites) being a presence record, and π_2 of it being an available record. We can assume (Seber 1984: 308) that the probability of a species being present at a location with covariates **x**, given that it was sampled is:

$$\log\left(\frac{f_{y=1}(\mathbf{x})}{f_{y=0}(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \left(\frac{\pi_1}{\pi_2}\right) \quad \text{eqn 5}$$

We can combine the sampling constant $\log(\pi_1/\pi_2)$ with the intercept term β_0. Because we have no information on the sampling proportions we can calculate the relative probability of occurrence (dropping the intercept term). This approach is suitable for discriminating between random sites and sites at which the species has been observed. Naturally the discriminant function cannot discriminate between sites at which a species was present and sites at which it was absent (from a contaminated sample of occupied and unoccupied locations) (Keating & Cherry 2004). Again, predictions need not be constrained to lie below 1, because predictions are relative likelihoods rather than probabilities of occurrence.

The logistic discrimination model is very similar to the exponential model suggested by Manly *et al.* (2002), and in practice its application differs only because resource units that appear in the used sample also can appear in the sample of available units (Johnson *et al.* 2006). The logistic discrimination model does not require as many assumptions as the exponential model: assumptions that Keating & Cherry (2004) suggest might sometimes be violated. Seber (1984: 309) suggests that the logistic discrimination model may be relatively robust to observations occurring in both the presence and the available sample.

MODELLING SPECIES ABUNDANCE WHEN ONLY ABUNDANCE-GIVEN-PRESENCE IS KNOWN

Often, estimates of relative abundance are made at locations where the species has been detected. Examples are counts of individuals, indices of abundance such as the Braun–Blanquet scale for plants (Kent & Coker 1992) and density measurements. Few studies have tried to model data of this type, even when data were acquired through systematic surveys. Regression approaches modelling abundance given only presence are possible using a truncated Poisson or negative binomial distribution. Alternatively, it may be possible to modify zero-inflated Poisson or negative binomial (ZIP or ZINB) regression models (Welsh *et al.* 1996; Barry & Welsh 2002; Dirnböck & Dullinger 2004; Nielsen *et al.* 2005) to model abundance given availability, where available locations are assigned a value of zero. This requires further investigation. We are unaware of any application explicitly modelling abundance given presence only.

Sampling issues

Knowledge of only the presence of a species presents a number of data-quality issues. Central among these are difficulties presented by choice of scale. Models of distribution or abundance can be highly sensitive to the scale of resolution (grain) as well as the extent (domain) (Soberón & Peterson 2005). There are no obvious guidelines about which choice of scale is appropriate, because such choice will depend on the ecology of the organism at hand and the objectives of the investigation (Boyce *et al.* 2003). If the intent is to model the global distribution of a species, obviously one should be using a very different scale than if the objective were to model use of habitats within a species' home range (Johnson 1980).

However, selection of extent can be a difficult question when using presence-only data. Implicit in that selection is an understanding of the sampling design by which the presence records were obtained. In studies where the data were obtained by survey, such as in the study of *Phytophthora ramorum* (Guo *et al.* 2005) or caribou *Rangifer tarandus* (Johnson *et al.* 2004), then the geographical, temporal and environmental boundaries of the study are known. However, presence-only data might be 'found' data – collated from multiple sources such as herbarium or museum records and for which there is no information on survey effort. Not knowing the sampling extent prevents us from defining available habitat adequately. For example, many herbarium databases are biased towards roads. A model of sampling effort would identify that only locations close to roads would have a high probability of being sampled, therefore only sites near roads should be included in the available sample. Not accounting for these biases may complicate model interpretation because the resulting model might describe sampling effort more than resource selection.

Once sampling scale and extent have been identified, the question then arises as to how to choose random locations from a potentially large area to contrast with the presence records. Little guidance exists in the literature; however, as Manly *et al*. (2002) argue, it is most important to minimize sampling errors, selecting data in such a way as to be fully representative of the study area. This implies that a large number of locations be selected randomly from the landscape to contrast with presence locations. McDonald (2003) suggests that several orders of magnitude more available units than used units be employed when applying the exponential model. Using GIS databases, such high sampling intensity for random landscape locations is feasible.

Modern biotelemetry systems such as GPS radiotelemetry (Frair *et al*. 2004) permit the collection of huge data sets of animal locations, with short time intervals between locations. Similarly, atlas data are usually obtained using grid coverage of the entire region, and so adjacent sampling squares are not independent (Augustin, Mugglestone & Buckland 1996). Such data are inherently plagued with both temporal and spatial autocorrelation because such frequent locations are not independent in time or space (Nielsen *et al*. 2002). To avoid committing a Type I error, adjustments for autocorrelation can be achieved using *post-hoc* methods of variance inflation (Nielsen *et al*. 2002) such as the Newey–West method (Newey & West 1987), or autocorrelation can be modelled more explicitly using mixed models (Laidre *et al*. 2004).

Validating presence-only models

Models based on presence-only data can be validated with data composed of presences and absences using existing evaluation statistics for presence–absence data [such as the area under the receiver operator characteristic (ROC) curve, or the kappa statistic]. However, when validation data consist only of presence data, model evaluation is more difficult because of the absence of a truly binary statistic. These issues are discussed by Boyce *et al*. (2002), who present an approach based on use-availability data to explore model performance; this approach has been developed further by Hirzel (unpublished data).

In this method *k*-fold cross-validation is used to correlate prediction ranks with area-adjusted frequencies of predicted values. Prediction ranks are obtained by breaking the range of predicted values into 10 (or some arbitrary number) evenly spaced bins. Area-adjusted frequencies of predicted values are then obtained by counting the number of occupied sites within the predicted value bins, and dividing these values by the area of the study area assigned the predicted values associated with than bin. This graphical approach holds great promise as a method to visualize predictive performance and to assign thresholds of prediction. However, as yet there is no suitable single measure of performance (or statistic) available to compare and contrast models.

An important feature of this approach is being able to examine how well model predictions are related to the probability of occurrence. A good model is one in which model predictions are proportional to the probability of occurrence (Manly *et al*. 2002). In the *k*-fold cross-validation graph, this would imply linear correspondence between the test-case area-adjusted frequencies and model predictions. There is no guarantee that any of the models described above will capture the true shape of the selection function, and thus might not be proportional to the probability of occurrence. Standard transformations of model predictions, e.g. logarithmic, square root, etc., might be necessary to scale the resource selection function appropriately. Proportionality is important because it allows model predictions to be used explicitly, such as when linking habitats to populations (Boyce & McDonald 1999; McDonald & McDonald 2001).

Conclusion

A number of statistical procedures are available for exploring patterns in presence-only data; the choice among them depends on the quality of the presence-only data. Profile techniques are most useful when species records, environmental predictors and biological understanding are scarce. However, when data quality is higher, regression-based techniques have generally proved more informative than profile techniques. The choice among regression modelling strategies depends on the sampling scheme for the 'absence' or 'control' records. The logistic regression model should not be used directly in most instances. Instead, either a case–control or discrimination approach should be adopted to contrast presence records with available resource units.

All techniques (profile and regression-based) may effectively rank habitats. Regression-based approaches might also provide predictions describing the relative likelihood of occurrence. If information on the relative proportions of presence and 'available' locations that were sampled is known, then predictions of the probabilities of occurrence are possible. *k*-Fold cross-validation can be used to examine model performance and proportionality.

Many conservation projects require a complete description of a species' geographical distribution or use of habitats to manage the species or environment effectively. However, for rare and endangered species, newly introduced species, or species requiring large geographical areas to meet all their life requirements, presence–absence data can be difficult or impossible to collect. Presence-only records can provide insight into the vulnerability, historical distribution and conservation status of species; models developed using these data can inform management. Our caveat is that researchers must be mindful of study design and the biases inherent in the presence data and be cautious in the interpretation of model predictions.

Acknowledgements

We thank Bryan Manly, Lisa Venier and colleagues at the Species Modelling workshop in Reideralp, Switzerland for invaluable discussion on the use of presence-only data to model species distribution, and Lisa Venier, Dan McKenney, Bryan Manly and Simon Ferrier for their comments on drafts of this manuscript. MSB acknowledges support from the Natural Sciences and Engineering Research Council of Canada (NSERC), the National Science Foundation (grant no. DEB-0078130) and the Alberta Conservation Association.

References

Akçakaya, H.R., Burgman, M.A., Kindvall, O., Wood, C.C., Sjögren-Gulve, P., Hatfield, J.S. & McCarthy, M.A. (2004) *Species Conservation and Management*. Oxford University Press, Oxford, UK.

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

Arthur, S.M., Manly, B.F.J., McDonald, L.L. & Garner, G.W. (1996) Assessing habitat selection when availability changes. *Ecology*, **77**, 215–227.

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

Barry, S.C. & Welsh, A.H. (2002) Generalised additive modelling and zero inflated count data. *Ecological Modelling*, **157**, 179–188.

Boyce, M.S., Mao, J.S., Merrill, E.H., Fortin, D., Turner, M.G., Fryxell, J. & Turchin, P. (2003) Scale and heterogeneity in habitat selection by elk in Yellowstone National Park. *Ecoscience*, **10**, 321–332.

Boyce, M.S. & McDonald, L.L. (1999) Relating populations to habitats using resource selection functions. *Trends in Ecology and Evolution*, **14**, 268–272.

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

Boyce, M.S., Vernier, P.R., Nielsen, S.E. & Schmiegelow, F.K.A. (2002) Evaluating resource selection functions. *Ecological Modelling*, **157**, 281–300.

Brotons, L., Thuiller, W., Araújo, M.B. & Hirtzel, A.H. (2004) Presence–absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, **27**, 437–448.

Busby, J.R. (1986) A biogeoclimatic analysis of *Nothofagus cunninghamii* (Hook.) Oerst. in southeastern Australia. *Australian Journal of Ecology*, **11**, 1–7.

Busby, J.R. (1991) BIOCLIM – a bioclimatic analysis and prediction system. *Nature Conservation: Cost Effective Biology Survey and Data Analysis* (eds C.R. Margules & M.P. Austin), pp. 64–68. CSIRO, Australia.

Campos, D., Kaur, A., Patil, G.P., Ripple, W.J. & Taillie, C. (1997) Resource selection by animals: the statistical analysis of binary response. *Coenoses*, **12**, 1–21.

Carpenter, G., Gillison, A.N. & Winter, J. (1993) DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation*, **2**, 667–680.

Caughley, G., Short, J., Grigg, G.C. & Nix, H. (1987) Kangaroos and climate: an analysis of distribution. *Journal of Animal Ecology*, **56**, 751–761.

Chefaoui, R.M., Hortal, J. & Lobo, J.M. (2005) Potential distribution modelling, niche characterisation and conservation status assessment using GIS tools: a case study of Iberian *Copris* species. *Biological Conservation*, **122**, 327–338.

Collett, D. (1991) *Modelling Binary Data*. Chapman & Hall, London, UK.

Compton, B.W., Rhymer, J.M. & McCollough, M. (2002) Habitat selection by wood turtles (*Clemmys insculpta*): an application of paired logistic regression. *Ecology*, **83**, 833–843.

Dettki, H., Löfstrand, R. & Edenius, L. (2003) Modelling habitat suitability for moose in coastal northern Sweden: empirical vs. process-oriented approaches. *Ambio*, **32**, 549–556.

Dirnböck, T. & Dullinger, S. (2004) Habitat distribution models, spatial autocorrelation, functional traits and dispersal capacity of alpine plant species. *Journal of Vegetation Science*, **15**, 77–84.

Elith, J. & Burgman, M.A. (2002) Habitat models for PVA. *Population Viability in Plants* (eds C.A. Brigham & M.W. Schwartz), Springer-Verlag, New York, NY.

Ferrier, S. & Watson, G. (1996) *An Evaluation of the Effectiveness of Environmental Surrogates and Modelling Techniques in Predicting the Distribution of Biological Diversity*. Consultancy report prepared by the New South Wales National Parks and Wildlife Service for the Department of Environment, Sport and Territories.

Ferrier, S., Watson, G., Pearce, J. & Drielsma, M. (2002) Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. *Biodiversity and Conservation*, **11**, 2275–2307.

Frair, J.L., Nielsen, S.E., Merrill, E.H., Lele, S., Boyce, M.S., Munro, R.H.M., Stenhouse, G.B. & Beyer, H.L. (2004) Removing habitat-induced, GPS-collar bias from inferences of habitat selection. *Journal of Applied Ecology*, **41**, 201–212.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A.T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, **19**, 497–503.

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

Guo, Q., Kelly, M. & Graham, C.H. (2005) Support vector machines for predicting distribution of Sudden Oak Death in California. *Ecological Modelling*, **182**, 75–90.

Hastie, T., Tibshirani, R. & Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Verlag, New York.

Hirzel, A.H., Hausser, J., Chessel, D. & Perrin, N. (2002) Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology*, **83**, 2027–2036.

Hirzel, A.H., Hausser, J. & Perrin, N. (2004) *Biomapper 3·0, User's Manual* [online]. Available at: http://www.unil.ch/biomapper [accessed September 2004].

Hirzel, A.H., Helfer, V. & Metral, F. (2001) Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, **145**, 111–121.

Johnson, C.J., Nielsen, S.E., Merrill, E.H., McDonald, T.L. & Boyce, M.S. (2006) Resource selection functions based on use-availability data: theoretical motivation and evaluation methods. *Journal of Wildlife Management*, in press.

Johnson, C.J., Parker, K.L., Heard, D.C. & Gillingham, M.P. (2002) A multiscale behavioural approach to understanding the movements of woodland caribou. *Ecological Applications*, **12**, 1840–1860.

Johnson, C.J., Seip, D.R. & Boyce, M.S. (2004) A quantitative approach to conservation planning: using resource selection functions to map the distribution of mountain caribou at multiple spatial scales. *Journal of Applied Ecology*, **41**, 238–251.

Johnson, D.H. (1980) The comparison of usage and availability measurements for evaluating resource preference. *Ecology*, **61**, 65–71.

Keating, K.A. & Cherry, S. (2004) Use and interpretation of logistic regression in habitat selection studies. *Journal of Wildlife Management*, **68**, 774–789.

Kent, M. & Coker, P. (1992) *Vegetation Description and Analysis: a Practical Approach*. John Wiley and Sons, Chichester, West Sussex, UK.

Laidre, K.L., Heide-Jorgensen, M.P., Logdson, M.L., Hobbs, R.C., Heagerty, P., Dietz, R., Jorgensen, O.A. & Treble, M.A. (2004) Seasonal narwhal habitat associations in the high Arctic. *Marine Biology*, **145**, 821–831.

Lancaster, T. & Imbens, G. (1996) Case–control studies with contaminated controls. *Journal of Econometrics*, **71**, 145–160.

Law, B.S. (1994) Climatic limitation of the southern distribution of the common blossom bat *Syconycteris autralis*, New South Wales. *Australian Journal of Ecology*, **19**, 366–374.

Lindenmayer, D.B., Nix, H.A., McMahon, J.P., Hutchinson, M.F. & Tanton, M.T. (1991) The conservation of leadbeater's possum, *Gymnobelideus leadbeateri* (McCoy): a case study of the use of bioclimatic modelling. *Journal of Biogeography*, **18**, 371–383.

Manly, B.F.J., McDonald, L.L., Thomas, D.L., McDonald, T.L. & Erickson, W.P. (2002) *Resource Selection by Animals*, 2nd edn. Kluwer Academic Publishers, Dordrecht, the Netherlands.

Manning, A.D., Lindenmayer, D.B. & Barry, S.C. (2004) The conservation implications of bird reproduction in the agricultural 'matrix': a case study of the vulnerable superb parrot of south-eastern Australia. *Biological Conservation*, **120**, 363–374.

Marzluff, J.M., Millspaugh, J.J., Hurvitz, P. & Handcock, M.S. (2004) Relating resources to a probabilistic measure of space use: forest fragments and Stellar's jays. *Ecology*, **85**, 1411–1427.

McDonald, T.L. (2003) Estimation of resource selection functions when used and available samples overlap. *Resource Selection Methods and Applications* (ed. S.V. Huzurbazar), pp. 35–39. Omnipress, Laramie, WY.

McDonald, T.E. & McDonald, L.L. (2001) A new ecological risk assessment procedure using resource selection models and geographical information systems. *Wildlife Society Bulletin*, **30**, 1015–1021.

Newey, W.K. & West, K.D. (1987) A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, **55**, 703–708.

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*, **13**, 45–56.

Nielsen, S.E., Johnson, C.J., Heard, D.C. & Boyce, M.S. (2005) Can models of presence–absence be used to scale abundance? Two case studies considering extremes in life history. *Ecography*, **28**, 1–12.

Pearce, J. & Lindenmayer, D. (1998) Bioclimatic analysis to enhance reintroduction biology of the endangered Helmeted Honeyeater (*Lichenostomus melanops cassidix*) in southeastern Australia. *Restoration Ecology*, **6**, 238–243.

Reutter, B.A., Helfer, V., Hirzel, A.H. & Vogel, P. (2003) Modelling habitat-suitability using museum collections: an example with three sympatric *Apodemus* species from the Alps. *Journal of Biogeography*, **30**, 581–590.

Seber, G.A.F. (1984) *Multivariate Observations*. Wiley, New York, NY.

Soberón, J. & Peterson, A.T. (2005) Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics*, **2**, 1–10.

Stockwell, D. & Peters, D. (1999) The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographyraphic Information Science*, **13**, 143–158.

Stockwell, D.R.B. & Peterson, A.T. (2002) Controlling bias in biodiversity data. *Predicting Species Occurrences: Issues of Accuracy and Scale* (eds J.M. Scott, P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. Wall & F.B. Samson), pp. 537–546. Island Press, Washington, DC.

Walker, P.A. & Cocks, K.D. (1991) HABITAT: a procedure for modelling a disjoint environmental envelope for a plant or animal species. *Global Ecology and Biogeography Letters*, **1**, 108–118.

Walther, B., Wisz, M. & Rahbek, C. (2004) Known and predicted African winter distributions and habitat use of the endangered Basra reed warbler (*Acrocephalus griseldis*) and the near-threatened cinereous bunting (*Emberiza cineracea*). *Journal of Ornithology*, **145**, 287–299.

Welsh, A.H., Cunningham, R.B., Donnelly, C.F. & Lindenmayer, D.B. (1996) Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling*, **88**, 297–308.

van Winkle, W. (1975) Comparison of several probabilistic home-range models. *Journal of Wildlife Management*, **39**, 118–123.

Zaniewski, A.E., Lehmann, A. & Overton, J.M. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261–280.

Received 12 May 2005; final copy received 1 June 2005
Editor: Rob Freckleton