Journal of
Plant Ecology

VOLUME 5, NUMBER 1,
PAGES 22–31

MARCH 2012

doi: 10.1093/jpe/rtr042

available online at
www.jpe.oxfordjournals.org

Dealing with detection error in site occupancy surveys: what can we do with a single survey?

Subhash R. Lele[1,]*, *Monica Moreno*[1] *and Erin Bayne*[2]

[1] *Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1*
[2] *Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1*
*Correspondence address. Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G2G1. Tel: +1-780-492-4290; Fax: +1-780-492-6826; E-mail: slele@ualberta.ca

Abstract

Aim

Site occupancy probabilities of target species are commonly used in various ecological studies, e.g. to monitor current status and trends in biodiversity. Detection error introduces bias in the estimators of site occupancy. Existing methods for estimating occupancy probability in the presence of detection error use replicate surveys. These methods assume population closure, i.e. the site occupancy status remains constant across surveys, and independence between surveys. We present an approach for estimating site occupancy probability in the presence of detection error that requires only a single survey and does not require assumption of population closure or independence. In place of the closure assumption, this method requires covariates that affect detection and occupancy.

Methods

Penalized maximum-likelihood method was used to estimate the parameters. Estimability of the parameters was checked using data cloning. Parametric boostrapping method was used for computing confidence intervals.

Important Findings

The single-survey approach facilitates analysis of historical datasets where replicate surveys are unavailable, situations where replicate surveys are expensive to conduct and when the assumptions of closure or independence are not met. This method saves significant amounts of time, energy and money in ecological surveys without sacrificing statistical validity. Further, we show that occupancy and habitat suitability are not synonymous and suggest a method to estimate habitat suitability using single-survey data.

Keywords: abundance estimation • biodiversity • BBS • closed population • data cloning • penalized likelihood • species occurrence

Received: 30 July 2011 Revised: 3 October 2011 Accepted: 17 October 2011

INTRODUCTION

The ability to accurately measure the distribution and abundance of species is at the core of ecological research and monitoring (Krebs 1999). Understanding and quantifying species distributions are essential to predict the effects of global climate change, colonization by exotic species or changes in land use. Many sampling methods and statistical analyses have been developed to estimate species abundance. Most methods are designed to measure density, i.e. the number of individuals per unit area. While density is a desirable state variable to report, there are numerous practical problems in estimating density (Krebs 1999). Even when it is possible to measure density accurately, the economics of doing so can be prohibitive for large-scale applications. This has led many research programs and monitoring initiatives to rely on rates of occurrence to get coarse measures of species abundance.

Collecting presence–absence data at a series of locations has become a preferred method of evaluating ecological status and trends because of the simplicity of data collection. Analysis of presence–absence data is often done to estimate the relationship between site occupancy probability and site characteristics and then use these models to predict the number of sites in a larger landscape that are occupied by that species. Most applications of species presence–absence data use logistic regression or contingency table analysis to associate species with temporal, habitat or spatial covariates. The assumption inherent in this approach is that the proportion of sites where a species is detected is equivalent to the actual occupancy rate. Occupancy rate is the true proportion of sites where a species occurs. It can only be estimated when the detection probability is equal to 1. The statistical analysis of presence–absence data and the ability to draw conclusions from such data has been questioned (MacKenzie *et al.* 2002) because detection is seldom perfect; hence, true occupancy rates are generally underestimated.

Detection probability is the probability that a species is observed, when it was present, at a particular site during the survey period. Detection probability can be less than one for a variety of reasons. Observers may be unable to detect the presence of the species due to survey-specific conditions such as rain, temperature or lack of visibility at the time of the survey. Detection rate can vary between habitats because the structure of the habitat may alter the ability of an observer to detect a species. Imperfect detection, if not taken into account, leads to a biased estimator of occupancy probability (MacKenzie *et al.* 2002). Recently, several researchers (Gu and Swihart 2004; MacKenzie *et al.* 2002; Martin *et al.* 2005; Stauffer *et al.* 2004; Tyre *et al.* 2003) have developed methods for estimating occupancy probability when the probability of detection is less than 1.

A common requirement for current methods to estimate occupancy probability when detection probability is less than 1 is that sites must be sampled repeatedly to estimate detection error rate. Repeated sampling can take the form of visiting multiple locations within a larger area of interest or by visiting the same location at different times. Using a repeated-visit approach, the target species is recorded as being detected or not detected at each visit. At locations where the species is present, detection error will occasionally result in species not being detected even though it is present at the site. Assuming true occupancy status does not change over the duration of repeated visits, called the closed-population assumption, changes in the detection and non-detection at a particular site can be attributed to detection error. This allows estimation of detection probability, which can be used to correct biases in the naïve estimator of occupancy probability.

Repeatedly visiting sites to estimate detection error is not always feasible. Returning to a site more than once multiplies the cost of most monitoring and research programs. For example, if an observer has to return on different days, which may be required to ensure the additional assumption of independence of surveys is met, the cost of travel is effectively doubled. Requirement of multiple visits reduces the number of sites that can be visited within a given sampling season for the same cost, reducing the generality of the results to broader areas. Repeated visits to study sites can also have an adverse effect on the survival of the individuals under observation (Rotella *et al.* 2000) making multiple visits undesirable from a conservation perspective. Analysis of the vast historical datasets (i.e. monitoring data) that did not conduct multiple visits is also not possible in this context (Hirzel *et al.* 2002). Validity of the repeated visit methodology depends crucially on the assumption that the population is closed, i.e. site occupancy status remains the same throughout the study period. To ensure closure, many researchers have used very short-time intervals between revisits. However, the closure assumption can be violated, even on very short-time scales, because of within territory movement. Replicate visits also require statistical independence. For plant ecologists, this will typically require different observers to return to sites to ensure that observers do not 'remember' where they found rare species. Bayne, Lele and Solymos (unpublished manuscript) show that within-territory movement by birds can introduce severe biases in occupancy estimates. Replicate visit methodology also assumes statistical independence between surveys, which is likely to be violated with shorter durations between revisits (Kendall and White 2009).

There is a growing belief among ecologists that repeat surveys are essential if predictive species-habitat models are to be useful. For example, Bolker (2008, p. 333) claims: 'there is no way to identify catchability—the probability that you will observe an individual—from a single observational sample; you simply don't have the information to estimate how many animals or plants you failed to count'. We whole-heartedly agree that ecologists need to be more rigorous in reporting accurate occupancy rates so that rigorous comparisons can be made across studies. However, does this mean that single-visit data and the resulting inferences are without value or is there something that can be done with single-visit data that allows us to account for detection error?

Fortunately, multiple surveys are not essential for estimating site occupancy parameters in the presence of detection error. In general, site occupancy and detection probability parameters can be estimated using a single survey provided (i) probability of occupancy and probability of detection depend on covariates and (ii) the set of covariates that affect occupancy and the set of covariates that affect detection differ by at least one variable. We surveyed nearly 100 papers on site occupancy and found that, when covariates were used for modeling occupancy and detection, most of them satisfied these conditions. Clearly, the conditions needed for the single survey-based estimation are not esoteric and are often satisfied in practice. In the following, we describe a methodology for dealing with detection error in single survey, present simulation results that show how the methodology works and present an analysis of a common type of data for which only single-survey data have been available for over 50 years (Breeding Bird Survey (BBS)).

STATISTICAL MODEL AND ESTIMATION PROCEDURE

Let us assume that there are N sites that will be surveyed in the study area. Let Y_i=1 if the ith site is occupied and Y_i=0 if the ith site is unoccupied. These are the true states that are unobserved. Let W_i=1 if the ith site is 'observed to be occupied' and W_i=0 if the ith site is 'observed to be unoccupied'. The probability of occupancy is denoted by $P(Y_i{=}1){=}\psi_i$ and the probability of detection by $P(W_i{=}1|Y_i{=}1){=}p_i$. We assume that if the species is not present, it will not be misidentified and hence $P(W_i{=}0|Y_i{=}0){=}1$. Simple probability calculations show that $P(W_i{=}0){=}1-p_i\psi_i$ and $P(W_i{=}1){=}P(W_i{=}1|Y_i{=}1)P(Y_i{=}1){=}p_i\psi_i$. These probabilities can depend on the habitat and other covariates. Let $\underline{X}$ denote the set of covariates that affect occupancy, and $\underline{Z}$ denote the set of covariates that affect detection. Some covariates may affect only detection, some covariates may affect only occupancy and some covariates may affect both detection and occupancy. For example, type of forest cover may affect both occupancy and detection, whereas time of the day or weather conditions may affect only detection. Thus, some of the covariates in the sets $\underline{X}$ and $\underline{Z}$ might be the same. With this notation, $\psi_i{=}\psi(\underline{X}_i,\beta)$ and $p_i{=}p(\underline{Z}_i,\theta)$. These functions should be such that $0\leqslant\psi(\underline{X}_i,\beta)\leqslant1$ and $0\leqslant p(\underline{Z}_i,\theta)\leqslant1$.

The necessary conditions under which the parameters (β,θ) are identifiable using single-survey data $\underline{W}$ are (i) there should exist at least one numeric (not categorical) covariate that affects probability of detection and probability of occupancy and (ii) the set of covariates $\underline{X}$ and $\underline{Z}$ should be such that there is at least one covariate that is in one set but not the other.

From a survey of previous applications of site occupancy models, it seems that most practical situations satisfy these conditions. In fact, for 94 out of 100 cases, the covariates that affect detection and covariates that affect occupancy were disjoint. A limitation of our methodology is that the case of constant probability of occupancy and constant probability of detection cannot be estimated. However, it appears that this restriction is not important in practice as in most papers we have reviewed the probability of occupancy and detection both were seldom constant. It is not possible to provide general result about identifiability conditions under every possible model. However, the data cloning method (Lele *et al.* 2007, 2010) can be used for both estimation and detection of possible non-estimability. The parameters are estimable if and only if the posterior variance converges to zero as the number of clones increases. This test is built into our software for the analysis of single-survey site occupancy data (Solymos and Moreno 2010).

The goal of the statistical analysis is to estimate (β,θ) given the observations $W{=}\{W_1,W_2,\ldots,W_N\}$. The likelihood function for these data is:

$$L\left(\beta,\theta;\underline{W}\right)=\prod_{i=1}^{N}\left(\psi(X_i,\beta)p(Z_i,\theta)\right)^{W_i}\left(1-\psi(X_i,\beta)p(Z_i,\theta)\right)^{1-W_i}.$$

Maximum likelihood estimators (MLEs) are obtained by maximizing this function with respect to (β,θ). If the sample size is large, one can use any numerical optimization technique to obtain the MLE. However, for small samples, this likelihood function is not a well-behaved, concave function. The problem of ill-behaved likelihood function is not unique to single-survey situation. In Moreno and Lele (2010), it is shown that even in the multiple-survey approach used by Mackenzie *et al.* (2002) that when the number of sites and/or number of surveys is small the likelihood function is ill-behaved and will not reach the proper solution. Moreno and Lele (2010) use penalized likelihood to obtain estimators from a multiple visit approach that are somewhat biased in small samples but have substantially smaller mean-squared error than the MLE. They also show that the confidence intervals based on the penalized likelihood estimators are substantially shorter and have close to nominal coverage than using the standard likelihood. The concept of maximum penalized likelihood estimator (MPLE), its intuitive justification and simulation results comparing its properties with the usual maximum-likelihood estimator under various sample sizes and different levels of occupancy and detection error are presented in Moreno and Lele (2010). Similar improvements over the standard maximum-likelihood estimator are obtained using the penalization idea in the single-survey situation as well. We emphasize here that the difference between MPLE and MLE vanishes as the sample size increases. The penalization simply stabilizes the likelihood function for small sample sizes. The penalized likelihood estimators in the single-visit case are obtained using the following steps:

Step 1: Obtain the MLE for(β,θ) by maximizing the likelihood function in Equation (1). Let us denote these by (β_M,θ_M).

Step 2: Obtain the naïve estimator of β by maximizing:

$$L\left(\beta;\underline{W}\right)=\prod_{i=1}^{N}\psi(X_i,\beta)^{W_i}\left(1-\psi(X_i,\beta)\right)^{1-W_i}.$$

This estimator, denoted by $\hat{\beta}_{\mathrm{naive}}$, is based on the assumption that there is no detection error. This is stable but biased with the magnitude of bias depending on how large the detection error is.

Step 3: Obtain the naïve estimator of θ by maximizing:

$$L\left(\theta;\underline{W}\right)=\prod_{i=1}^{N}p(Z_i,\theta)^{W_i}\left(1-p(Z_i,\theta)\right)^{1-W_i}.$$

This estimator, denoted by $\hat{\theta}_{\mathrm{naive}}$, is based on the assumption that all sites are occupied. This estimator is stable but biased.

Step 4: Maximize the penalized likelihood function with respect to (β,θ):

$$\log\mathrm{PL}(\beta,\theta;\underline{W})=\log L(\beta,\theta;\underline{W})-\lambda_1|\beta-\hat{\beta}_{\mathrm{naive}}|-\lambda_2|\theta-\hat{\theta}_{\mathrm{naive}}|,$$

where $\lambda_1{=}\left(1-\hat{\psi}_{\mathrm{naive}}\right)\hat{p}_M\sqrt{\mathrm{tr}(\mathrm{Var}(\hat{\theta}_M))}$ and $\lambda_2{=}\left(1-\hat{p}_{\mathrm{naive}}\right)\hat{\psi}\times{}_M\sqrt{\mathrm{tr}(\mathrm{Var}(\hat{\beta}_M))}$ and $(\hat{p}_{\mathrm{naive}},\hat{\psi}_{\mathrm{naive}})$ and $(\hat{p}_M,\hat{\psi}_M)$ denote the average occupancy and detection probabilities under the naïve method of estimation and MLE, respectively. Justification for this penalty function is along the same lines as described in Moreno and Lele (2010). The confidence intervals for MPLE can be based on the bootstrap technique and are shown to have good coverage (Moreno and Lele 2010). A computer program written in R to implement this method is available in Solymos and Moreno (2010).

SIMULATION RESULTS

These simulations have two goals. The first goal is to support the claim of estimability of the parameters using a single survey. If the parameters are consistently estimable then, as we increase the sample size, the estimates should converge to the true values. The second goal is to show that these estimators give reasonable inferences in practical situations.

To achieve this goal, and for the purpose of considering a variety of scenarios commonly found in this type of analysis, a total of 54 cases were simulated. These cases were defined by considering different levels for factors such as the sample size, the type of link function, the probability of occupancy and detection and the configuration of the covariates, i.e. whether or not there was a common covariate for both occupancy and detection.

For each case, 100 datasets were generated using two covariates for occupancy and two covariates for detection. For the case with no common covariates, the probability of occupancy was calculated using the logistic link $\psi_i=\frac{\exp(\beta_0+\beta_1X_{i1}+\beta_2X_{i2})}{1+\exp(\beta_0+\beta_1X_{i1}+\beta_2X_{i2})}$ where covariate values were generated using$X_{1i}\sim\text{Normal}(0,1)$ and $X_{2i}\sim\text{Bernoulli}(0.55)$. Similarly, the probability of detection was calculated using either the logistic link $p_i=\frac{\exp(\theta_0+\theta_1Z_{i1}+\theta_2Z_{i2})}{1+\exp(\theta_0+\theta_1Z_{i1}+\theta_2Z_{i2})}$ or the log–log link $p_i=\exp(-\exp(\theta_0+\theta_1Z_{i1}+\theta_2Z_{i2}))$ where covariate values were generated using $Z_{1i}\sim\text{Normal}(0,1)$ and $Z_{2i}\sim\text{Bernoulli}(0.65)$.

For the common covariate cases, the covariate for the occupancy model was taken as the common one for both. For instance, if the common covariate is a continuous one and the link for both occupancy and detection is the logistic link, the probability of occupancy for the site i is $\psi_i=\frac{\exp(\beta_0+\beta_1X_{i1}+\beta_2X_{i2})}{1+\exp(\beta_0+\beta_1X_{i1}+\beta_2X_{i2})}$ and the probability of detection is $p_i=\frac{\exp(\theta_0+\theta_1X_{i1}+\theta_2Z_{i2})}{1+\exp(\theta_0+\theta_1X_{i1}+\theta_2Z_{i2})}$.

The set of parameters was selected to obtain the desired level of occupancy and detection required according to the case and the estimates were obtained by using the MPLE described in the Section 2 of this paper. Figures 1 and 2 present the results from two representative cases obtained from the simulations.

Figure 1 shows the box plots of the estimated parameters when the mean probability of occupancy is 0.27, the mean probability of detection 0.27 and the covariates for occupancy and detection are separable. Clearly as the sample size increases, the distributions of the parameters become more symmetric and their centers are closer to the true value. It is also observed that as the sample size increases, the spread of the distributions decreases. Figure 2 presents the results obtained for the case in which there is a discrete covariate that is common to both occupancy and detection. Again, in this case, the mean probability of occupancy and detection are low (0.27 and 0.31, respectively). Similar to the separable covariates case, as the sample size increases, the centers of the density functions get closer to the true value, and also, that the variance decreases as the sample size increases.

For most of the situations considered in our simulations the mean occupancy and mean detection probability can be estimated reasonably well at sample sizes of 100–200, whereas a good estimation of regression coefficients occurs at sample sizes of 300 or larger. See Fig. 3 for an example. If the main goal of an analysis is estimation of mean occupancy rate, one does not have to worry as much about sample size as when accurate estimation of the regression coefficient *per se* is the objective.

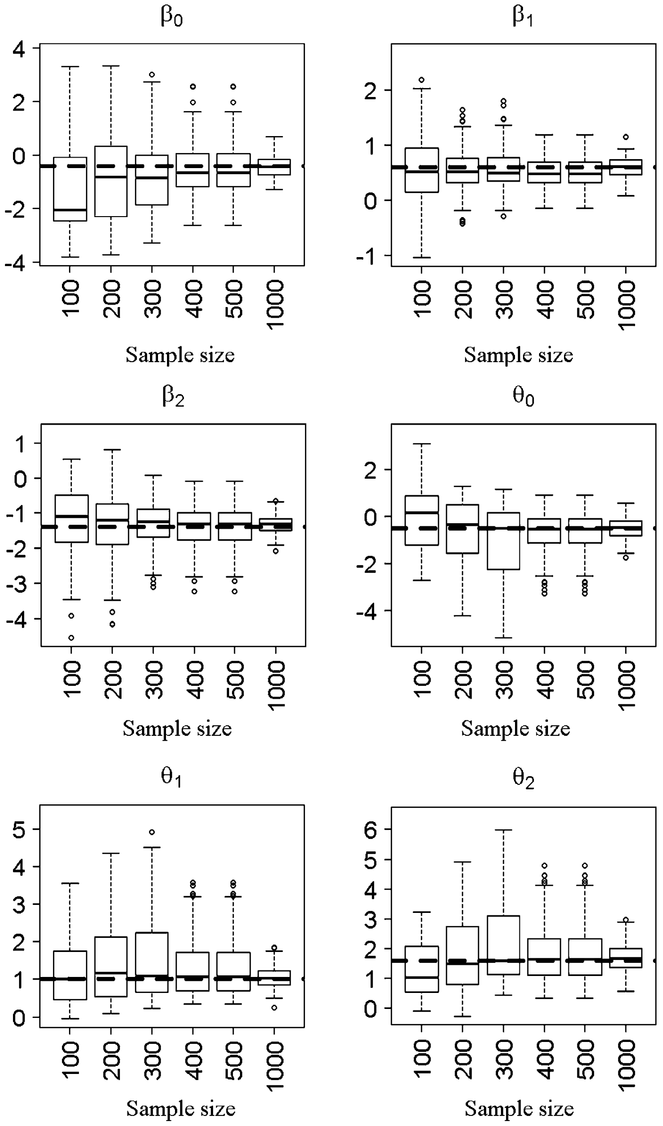

Figure 1: simulations showing estimability of the parameters using a single survey when the covariates that affect occupancy and detection are distinct. The parameters $(\beta_{0,}\beta_{1,}\beta_2)$ correspond to the occupancy and $(\theta_{0,}\theta_{1,}\theta_2)$ correspond to the detection models. As the sample size increases, the estimates converge to the true value.

These results along with the results in the Appendix show that the occupancy and detection parameters are identifiable using single survey data. This holds even when the set of covariates for occupancy and detection have some overlap.

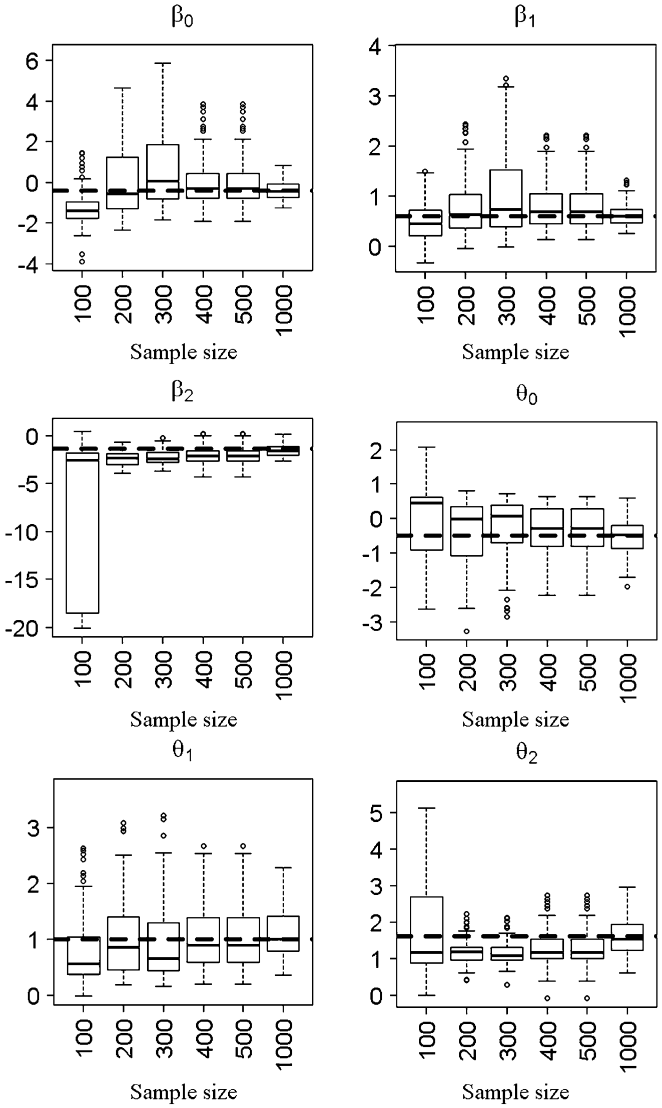

Figure 2: Simulations showing estimability of the parameters using a single survey when there is a categorical common covariate that affect occupancy and detection. The parameters $(\beta_0, \beta_1, \beta_2)$ correspond to the occupancy and $(\theta_0, \theta_1, \theta_2)$ correspond to the detection models. As the sample size increases, the estimates converge to the true value.

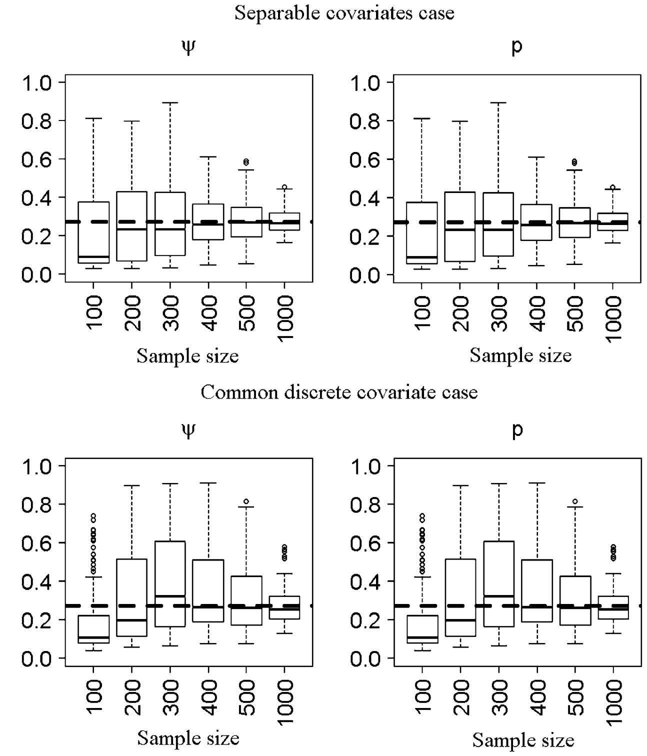

Figure 3: Simulations showing estimability of the mean occupancy and detection for both cases: using separable covariates, and using a categorical common covariate that affects occupancy and detection.

AN ILLUSTRATION OF THE METHOD

To illustrate the estimation of the parameters for an occupancy model using a single survey, we consider detected–not detected data for Ovenbirds (*Seiurus aurocapilla*). Data were collected in 1999 using BBS Protocols (Downes and Collins 2003) in the boreal plains eco-region of Saskatchewan. The goal of the study was to determine whether the occupancy of this species was influenced by the amount of forest around each survey point. Data were collected along 36 BBS routes each consisting of 50 survey locations with survey locations separated by 800 m. To increase independence of observations, we used every second survey point along each route (thus each point was 1.6 km apart) in our analysis (n = 900 survey locations). Attributes about the forest type and amount of forest remaining with a 400-m radius were estimated from the Saskatchewan Digital Land Cover Project (MacTavish 1995).

The habitat requirements of the Ovenbird are well-understood (Hobson and Bayne 2002) and we expected that the probability a location was occupied by the Ovenbird would be positively influenced by the amount of deciduous forest remaining (forest deciduous proportion). We also included longitude as the study covered an east–west gradient over 1000 km in length although *a priori* we were not sure what effect this would have on occupancy.

We expected four factors to influence detection probability: observer, time of day, time of year and amount of forest. Observers differ in their ability to hear birds in part not only because of skill but also because of fundamental differences in the distance over which they hear things. In general, male songbirds sing very regularly early in the breeding season making it easy to detect individuals that are present. As the breeding season progresses, however, the males spend less time singing as they focus on other activities. This often results in lower detectability later in the breeding season. We included Julian date as a variable influencing detection error. Male songbirds also have a tendency to sing earlier in the day, shortly after sunrise and then later in the morning focus on guarding the mate or foraging. To account for this, we included time of the day as a factor influencing detectability. Detectability can also be influenced by habitat attributes. In more open environments, it is plausible that birds can be heard from long distances increasing the chance an individual is detected (Schieck 1997). Alternatively, in areas with more forest, the chance of multiple males singing may be higher, increasing detection probability relative to areas with less forest where only one individual may exist.

We considered several different models and used Akaike information criterion (AIC) to select the final model. We also used the receiver operating curve (ROC) and the area under the ROC (AUC) to heuristically compare the fit of the full model, detection and occupancy together. Table 1 gives the details on the various models that were considered and the corresponding AIC and AUC values. The final model has AUC of 0.82 indicating a fairly good predictive capacity for the full model, detection and occupancy together. It also has a smaller AIC value relative to other candidate models. Table 2 presents the estimated parameters, the 90% confidence intervals and the estimated standard errors for occupancy and detectability. Figure 4 depicts graphically how the probability of occupancy and detection vary with the covariates. The confidence intervals and the standard errors were estimated using 200 bootstrap samples. As expected, the proportion of deciduous forest has a positive effect on the probability of occupancy. This relationship was best fit using a log transformation of forest deciduous proportion. Longitude was not statistically significant but it suggested that Ovenbird occupancy rate increased as surveys were done further west.

The amount of forest cover was the strongest predictor of detection probability. Detection probability was highest in areas with higher forest cover. This suggests that increased numbers of birds in areas with more forest increase the probability of detecting the species while in areas, with low forest cover the reduced numbers of birds means the chances of detecting the species given they are present is considerably lower. Although not strictly statistically relevant according to AIC (and, hence not included in the final model), detection probability did differ among observers and was affected by the time of the day and Julian date in a sensible fashion (Fig. 5). One observer (S.V.Wilgenburg) was much more likely to detect birds in areas with less forest than the others. Previous experience in other projects has demonstrated that this individual is able to hear birds over far greater distances than other people so this result was not unexpected. Detection probability was negatively related to Julian date indicating decreased singing activity later in the season was reducing observer ability to detect birds given they were present. Time of day had a positive relationship with detection probability. This was somewhat unexpected. However, time of day was the least significant effect and surveys were done in a very narrow time window (4:00 AM to 9:00 PM local time). The estimated mean probability of occupancy for all the sites, based on the final model, was 0.523, with a mean probability of detection of 0.466. The mean probability of occurrence without correcting for detection error, on the other hand, was 0.297.

OCCUPANCY AND HABITAT SUITABILITY

An occupancy survey answers the question: At the time of the survey, was the site occupied? Such information is useful for monitoring studies where one wants to know the current status of the study area. However, researchers should be aware that occupancy status of a location depends not only on its habitat suitability but also on the overall population density. For example, if the population density is low, even in a highly suitable habitat only a few of the sites will be occupied. In such

Table 1: models for the ovenbird data sorted from smallest to largest AIC

Model	Occupancy model covariates	Detection model covariates	AUC	AIC	BSIC
1	Log (proportion of deciduous forest)	Proportion of forest	0.823	825.500	844.656
2	Log (proportion of deciduous forest)	Proportion of forest, Julian date, time of day	0.826	826.550	855.283
3	Log (proportion of deciduous forest); log (non-deciduous forest)	Proportion of forest	0.823	827.129	851.074
4	Proportion of deciduous forest, longitude	Proportion of forest, Julian date, time of day, observer	0.828	827.907	875.797
5	Log (proportion of deciduous forest), longitude	Proportion of forest, Julian date, time of day	0.826	828.490	862.013
6	Log (proportion of deciduous forest), longitude	Proportion of forest, Julian date, time of day, observers	0.827	830.775	878.664
7	Log (proportion of agricultural area), longitude	Proportion of forest, Julian date, time of day, observers	0.818	841.286	889.175
8	Proportion of agricultural area, longitude	Proportion of forest, Julian date, time of day, observers	0.820	843.457	891.347

We also provide the Schwarz Information Criterion (BIC) and the AUC. Smaller the AIC or BIC, better is the model fit; larger the AUC, better is the fit.

Table 2: estimated parameters, confidence intervals and standard errors for the occupancy and detection model for Ovenbird occupancy survey data

Model	Covariates	Point estimate (90% confidence interval)
Occupancy model	Intercept	−0.255 (−0.843, 0.566)
	Log (proportion of deciduous forest)	1.546 (0.836, 3.561)
Detection model	Intercept	0.476 (0.047, 1.127)
	Proportion of forest area	0.951 (0.588, 1.763)
Average occupancy probability		0.496 (0.405, 0.643)
Average detection probability		0.489 (0.377, 0.608)
Naïve estimate of average occupancy		0.297 (0.278, 0.318)

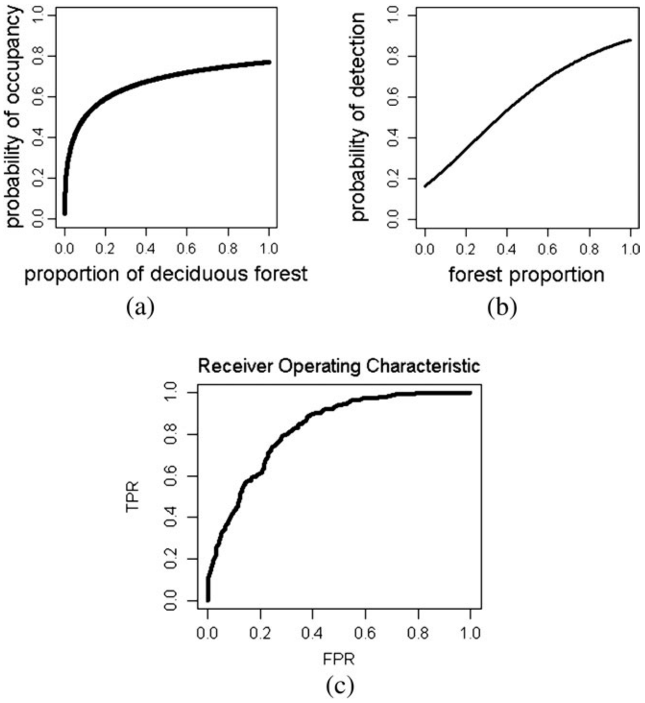

Figure 4: (**a**) Probability of occupancy, (**b**) probability of detection and (**c**) receiver operating curve for the Ovenbird data.

a situation, unoccupied sites are unoccupied not because they have unsuitable habitats but because there are not enough individuals to occupy them. Qualitatively such unoccupied sites are different than the sites that are unoccupied because they are truly unsuitable. Presence–absence data cannot differentiate between these two kinds of unoccupied sites. Similarly, when the population size is high, even sites with low-quality habitats are occupied. In the extreme, if all sites are occupied, although occupancy can be estimated, it is mathematically impossible to distinguish between suitability of different sites. Lele, Merrill, Keim and Boyce (submitted manuscript) provide a detailed discussion of the mathematical relationship between selection, habitat suitability and occupancy. Researchers using occupancy surveys should be aware that a naïve interpretation of the results of the occupancy model as 'habitat suitability' is incorrect and can be misleading.

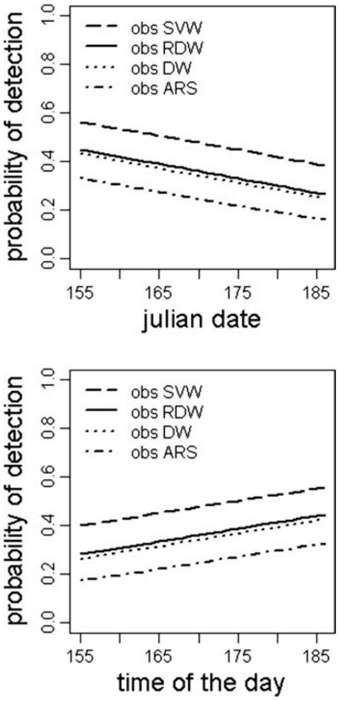

Figure 5: Estimates of the effects of the observers, Julian date and time of the day over the probability of detection for the Ovenbird data.

Although presence–absence data are uninformative about habitat suitability, count data (e.g. number of birds), abundance data (e.g. plant biomass) or ordinal data (e.g. low, medium or high level of invasion) are useful to determine habitat suitability. For such data, the extra zeros arising out of low population size can be modeled using zero-inflated distributions. On the other hand, in the presence of high population size, the relationships between values 1 and beyond provide information about habitat suitability. For example, for count data, one can model extra zeros arising because of low population sizes using zero inflated Poisson distribution or zero-inflated negative binomial distribution. For ordinal data, one can model these extra zeros using zero-inflated multinomial models. In the Appendix, we show that one can correct for zero inflation along with detection zeros by using the method of conditional likelihood. We discuss the mathematical details for computing conditional likelihood for the general case in the Appendix. P. Solymos *et al.* (2012, in press) study the statistical performance of this method. They use zero-inflated Poisson distribution for analyzing the ovenbird count data and compare the results with the results in this paper. A detailed application to multinomial data to study the spread of invasive plants is discussed elsewhere.

Habitat suitability is also commonly studied using the 'presence-only data' (Lele and Keim 2006; Phillips *et al.* 2006). Although vastly popular, presence-only data is uninformative about habitat suitability. Mathematically speaking, analyzing presence-only data is an ill-conditioned (non-identifiable) problem. The only way to make inference about habitat suitability using presence only data is to assume complete knowledge of the available distribution (Lele and Keim 2006; Manly *et al.* 2002). The choice of available distribution is arbitrary and its validity cannot be tested. Different choices lead to vastly different conclusions about habitat suitability with no recourse to knowing which one is better supported by the data. The mathematical details and practical implications of this non-identifiability issue are discussed in a separate publication.

DISCUSSION

Ecologists have long relied on single surveys to estimate the relative abundance of organisms between habitats. The key assumption underlying this approach is that by randomizing sources of detection error across the variables of primary interest that the correct relative pattern of habitat suitability is revealed even though the absolute abundance or occupancy rate is underestimated. However, multiple survey techniques have demonstrated that patterns in habitat selection can be strongly influenced by detection error when habitat conditions influence the ability to detect an organism. In addition, multiple survey methods have been useful in shifting ecologists away from relative measures to absolute estimates of abundance, facilitating informed decision making and better ecological inference. When the crucial assumptions of population closure and independence of surveys are satisfied and costs are not a major issue, then multiple survey methods will generally provide statistically more efficient estimators than a single survey-based approach. However, multiple survey methods have assumptions that, if not met, will result in biased estimates of occupancy rate.

The concept of a closed population is clear in some situations. For stationary organisms, the assumption of population closure is met quite naturally because detection error cannot be due to movement of the organism out of the sampling area. Detection error in such situations is clearly due to sampling conditions such as weather, date, time of day, observer effects or the ability to see a species in one habitat over another. For mobile organisms, on the other hand, a closed population is difficult to define. For example, many wildlife ecologists record species presence using point surveys whereby individuals heard or seen within a given radius of the observer's location are counted. For species that are not territorial and move widely across the landscape, multiple visit methodologies could fail to provide accurate estimates of occupancy rate. Even for territorial species, the issue of spatial scale of sampling relative to spatial scale of movement of the animal creates problems for multiple-visit surveys. For example, when only part of an individual animal's territory is within the sampling area, an observer may detect the animal imperfectly over multiple visits simply because the animal is in a part of its territory that is outside the sampling area during one visit and inside the sampling area during another visit. E. Bayne *et al.* (unpublished manuscript) show that within-territory movement can introduce biases in occupancy estimates that dramatically overestimate density. See also Rota *et al.* (2009) for other factors that result in the assumption of closure being violated.

Financial and logistical costs are pivotal in the design of effective monitoring and scientific studies to track biodiversity. Proponents of multiple visit methods often suggest the increased costs of multiple visits to be negligible. Admittedly, if the repeated visits occur on the same day, the travel costs to a site will be relatively small. However, whether such an approach will achieve independence of visits is not well-established. In addition, monitoring programs such as the BBS already maximize the number of stops that an observer can do during the ideal period of observations for birds. Requiring multiple visits would force the BBS and other monitoring programs to return to the same locations on a different day. Returning to a site on a different day increases travel costs and personnel time, which will typically reduce the total number of sites that can be visited during a survey season in direct proportion to the number of sites that have to be revisited. We estimate that a shift from a single visit to four visits for a bird monitoring program we are involved with in northern Alberta, where many sites are visited by helicopter, would increase the annual costs of monitoring 720 stations per year from ~$180,000 to ~$700,000 per year.

The development of the single-survey approach provides an additional tool to ecologists that allows for correction of detection error, does not have the critical assumption of population closure and has the logistical flexibility of conventional single-survey designs.

FUNDING

Natural Sciences and Engineering Research Council of Canada; Alberta Biodiversity Monitoring Initiative; Environment Canada.

ACKNOWLEDGEMENTS

Conflict of interest statement. None declared.

APPENDIX

COMPUTATIONAL ISSUES FOR PENALIZED LIKELIHOOD ESTIMATION

As described in Moreno and Lele (2010), because $\mathrm{tr}(\mathrm{Var}(\theta_M)) \to 0$ and $\mathrm{tr}(\mathrm{Var}(\beta_M)) \to 0$ as the sample size increases, the penalized likelihood function approaches the likelihood function if the number of sites is large. Penalization simply stabilizes the likelihood function for small sample sizes. If the MLE of average detection probability is high, naïve estimates of the occupancy are reasonable. In this case, the first component of the penalty function is large, thus shrinking the occupancy parameters toward their naïve estimates. Similarly, if the MLE of the average occupancy is high, naïve estimates of the detection parameters are reasonable. In this case, the second component of the penalty function is large, thus shrinking the detection parameters toward their naïve estimates.

In Step 1 of the penalized likelihood estimation algorithm, we need to compute the MLE and its variance. If the number of sites is smaller than 100, using a gradient-based optimization technique to find the location of the maximum tends to be tricky as it tends to lead to nearly singular Hessian matrices (Moreno and Lele 2010). Hence, we cannot use the inverse of the Hessian matrix to approximate the variance of the MLE. Instead of using a local, gradient-based technique to find the MLE and its variance in Step 1, we use a global stochastic search method, a variant of the well-known simulated annealing method, called data cloning (Lele *et al.* 2007, 2010). In data cloning, as in simulated annealing, the MLE is obtained as the mean of the posterior distribution. This avoids the task of numerically differentiating a non-smooth function. To obtain the variance of the MLE, one can either use the bootstrap (Efron and Tibshirani 1993) or it can also be approximated by the variance of the posterior distribution (Lele *et al.* 2007, 2010). This avoids the problem of inverting a nearly singular Hessian matrix to approximate the variance of the MLE. The penalized likelihood function (Step 4) is maximized using the standard numerical optimization techniques.

ZERO-INFLATED DISTRIBUTIONS, DETECTION ERROR AND CONDITIONAL LIKELIHOOD

This Appendix describes the mathematics behind the conditional likelihood approach for dealing with zero-inflated distributions in the presence of detection error. A detailed study of the performance of the conditional likelihood estimators for the zero-inflated Poisson model is provided in Solymos *et al.* (2012, in press). Throughout this description, as in the paper, we assume availability of covariates for detection and occupancy models. In the following, for pedagogical convenience, we assume that the response variable is a discrete random variable. These results can be generalized to continuous response variables.

Let Y^* denote a random variable that takes values in the set $S=\{0,1,2,\ldots,K\}$ where $K\geqslant 2$ and which can potentially be infinity. For example, Y^*can be a Poisson, negative binomial or a multinomial random variable. Let Xdenote the covariates and $\boldsymbol{\beta}$ denote the regression coefficients. Let $P(Y=y|X=x)=\boldsymbol{\psi}(y;x,\boldsymbol{\beta})$ for $y\in\{0,1,2,\ldots,K\}$ denote the probability mass function (p.m.f.). The regression coefficients $\boldsymbol{\beta}$ in this model inform us about the habitat suitability of various covariates. Let us denote the zero-inflated version of the random variable Y^* by$Y$. Its p.m.f. is given by $P(Y=y)=(1-\boldsymbol{\phi})\boldsymbol{\psi}(y;x,\boldsymbol{\beta})+\boldsymbol{\phi}I_{(y=0)}$. The parameter $\boldsymbol{\phi}$ corresponds to the proportion of additional unoccupied sites because of low population size in the study area.

Because of detection error, the true status of the site cannot be observed. For example, observed counts might be different than the true counts. In the case of the multinomial response, the observed class might be different than the true status, e.g. a medium abundance site may be observed as low abundance because one may miss some of the invasive plants. It is also possible that true low abundance be classified as 'medium abundance' if there is a big patch of invasive plants near the observer location. Let W denote the observed status. This variable also takes values in the set $S=\{0,1,2,\ldots,K\}$. We assume that if species is absent, observer does not imagine its existence. That is, if $Y=0$ then $W=0$ with probability 1.

The p.m.f. of W depends on the true status of the site (Y) as well as the conditions at the time of observation and other habitat characteristics of the site (Z). We denote the p.m.f. of W by $P(W=w|Y=y;Z=z,\boldsymbol{\delta})=p(w;y,z,\boldsymbol{\delta})$. Basic probability laws show that for $w\neq 0$, $P(W=w|X=x,Z=z)=\sum_{y\in S}P(W=w|Y=y;z)\ P(Y=y)=\left(1-\phi\right)\sum_{y\in S}p(w;y,z,\delta)\psi(y;x,\beta)$ and $P\left(W=w|W>0\right)=\frac{\sum_{y\in S}p(w;y,z,\delta)\psi(y;x,\beta)}{\sum_{w=1}^{K}\sum_{y\in S}p(w;y,z,\delta)\psi(y;x,\beta)}$.

Notice that this conditional probability does not depend on the zero-inflation parameter $\boldsymbol{\phi}$. The habitat suitability parameters $\boldsymbol{\beta}$ and detection parameters $\boldsymbol{\theta}$ can be estimated using only the data from the non-zero sites by maximizing the conditional likelihood function: $\mathrm{CL}(\beta,\theta;\underline{w})=\prod_{i=1}^{n}\{P(W_i=w_i|W>0)\}^{I_{(w_{i\neq 0})}}$.

REFERENCES

Bolker B (2008) Ecological Models and Data in R. Princeton, NJ: Princeton University Press.

Downes CM, Collins BT (2003) The Canadian Breeding Bird Survey, 1967-2000. Canadian Wildlife Service—Progress Notes # 219. Ottawa, ON: National Wildlife Research Centre.

Efron B, Tibshirani RJ (1993) An Introduction to Bootstrap. New York, NY: Chapman and Hall.

Gu W, Swihart RK (2004) Absent or undetected? Effects of non-detection of species occurrence on wildlife-habitat models. *Biol Conserv* **116**:195–203.

Hirzel AH, Hausser J, Chessel D, *et al.* (2002) Ecological niche factor analysis: how to compute habitat suitability maps without absence data. *Ecology* **83**:2027–36.

Hobson KA, Bayne EM (2002) Breeding bird communities in boreal forest of Western Canada: consequences of ''unmixing'' the mixed woods. *Condor* **102**:759–69.

Kendall WL, White GC (2009) A cautionary note on substituting spatial subunits for repeated temporal sampling in studies of site occupancy. *J Appl Ecol* **46**:1182–8.

Krebs CJ (1999) Ecological Methodology (2nd edn). Menlo Park, CA: Pearson Education Press.

Lele SR, Dennis B, Lutscher F (2007) Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecol Lett* **10**:551–63.

Lele SR, Keim JL (2006) Weighted distributions and estimation of resource selection probability functions. *Ecology* **87**:3021–8, 2006.

Lele SR, Nadeem K, Schmuland B (2010) Estimability and likelihood inference for generalized linear mixed models using data cloning. *J Am Stat Assoc* **105**:1617–25.

MacKenzie DI, Nichols JD, Lachman GB, *et al.* (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83**:2248–55.

MacTavish P (1995) Saskatchewan digital landcover mapping project. Report I-4900-15-B-95. Saskatoon, SK: Saskatchewan Research Council.

Manly BFJ, McDonald LL, Thomas DL, *et al.* (2002) Resource Selection by Animals: Statistical Design and Analysis for Field Studies (2nd edn). New York, NY: Kluwer Academic Publishers.

Martin TG, Wintle BA, Rhodes JR, *et al.* (2005) Zero tolerance ecology: improving ecological inference by modeling the source of zero observations. *Ecol Lett* **8**:1235–46.

Moreno M, Lele SR (2010) Improved estimation of site occupancy using penalized likelihood. *Ecology* **91**:341–6.

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecol Model* **190**:231–59.

Rota CT, Fletcher RJ, Dorazio RM, *et al.* (2009) Occupancy estimation and the closure assumption. *J Appl Ecol* **46**:1173–81.

Rotella JJ, Taper ML, Hansen AJ (2000) Correcting nesting success estimates for possible observer effects: maximum likelihood estimates of daily survival rates with reduced bias. *Auk* **117**:92–109.

Schieck J (1997) Biased detection of bird vocalizations affects comparisons of bird abundance among forested habitats. *Condor* **99**:179–90.

Solymos P, Lele S, Bayne E (2012) Conditional likelihood approach for analysing single visit abundance survey data in the presence of zero inflation and detection errors. *Environmetric* (in press).

Solymos P, Moreno M (2010) Analyzing Single Visit Site Occupancy Data with Detection Error. *R Package Version 1.0-0*. http://cran.r-project.org.

Stauffer HB, Ralph CJ, Miller SL (2004) Ranking habitat for marbled murrelets: a new conservation approach for species with uncertain detection. *Ecol Appl* **14**:1374–83.

Tyre AJ, Tenhumberg B, Field SA, *et al.* (2003) Improving precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecol Appl* **13**:1790–801.