LOCATION-ONLY AND USE-AVAILABILITY DATA

A re-evaluation of a case–control model with contaminated controls for resource selection studies

Christopher T. Rota[1]*, Joshua J. Millspaugh[1], Dylan C. Kesler[1], Chad P. Lehman[2], Mark A. Rumble[3] and Catherine M. B. Jachowski[4]

[1]Department of Fisheries & Wildlife Sciences, University of Missouri, Columbia, MO, USA; [2]South Dakota Department of Game, Fish, and Parks, Custer State Park,Custer, SD, USA; [3]Rocky Mountain Research Station, U.S. Forest Service, Rapid City, SD, USA; and [4]Department of Fish & Wildlife Conservation, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Summary

1. A common sampling design in resource selection studies involves measuring resource attributes at sample units used by an animal and at sample units considered available for use. Few models can estimate the absolute probability of using a sample unit from such data, but such approaches are generally preferred over statistical methods that estimate a relative probability of use.

2. The case–control model that allows for contaminated controls, proposed by Lancaster & Imbens (1996) and Lele (2009), can estimate the absolute probability of using a sample unit from use-availability data. However, numerous misconceptions have likely prevented the widespread application of this model to resource selection studies. We address common misconceptions regarding the case–control model with contaminated controls and demonstrate its ability to estimate the absolute probability of use, prevalence and parameters associated with categorical covariates from use-availability data.

3. We fit the case–control model with contaminated controls to simulated data with varying prevalence (defined as the average probability of use across all sample units) and sample sizes ($n_1 = 500$ used and $n_a = 500$ available samples; $n_1 = 1000$ used and $n_a = 1000$ available samples). We then applied this model to estimate the probability Ozark hellbenders (*Cryptobranchus alleganiensis bishopi*) would use a location within a stream as a function of covariates.

4. The case–control model with contaminated controls provided unbiased estimates of all parameters at $N = 2000$ sample size simulation scenarios, particularly at low prevalence. However, this model produced increasingly variable maximum likelihood estimates of parameters as prevalence increased, particularly at $N = 1000$ sample size scenarios. We thus recommend at least 500–1000 used samples when fitting the case–control model with contaminated controls to use-availability data. Our application to hellbender data revealed selection for locations with coarse substrate that are close to potential sources of cover.

5. This study unites a disparate literature, addresses and clarifies many commonly held misconceptions and demonstrates that the case–control model with contaminated controls is a viable alternative for estimating the absolute probability of use from use-availability data.

Key-words: Bayesian analysis, data cloning, Markov chain Monte Carlo sampling, maximum partial likelihood estimator, optimization, presence-only, prevalence, pseudo-absence, radiotelemetry, use-availability

*Correspondence author. E-mail: ctr4g2@mail.missouri.edu

Introduction

The study of resource selection is essential for describing relationships between animals and their environment, understanding factors that determine the distribution of species and managing wildlife populations. Resource selection studies are often motivated by a need to understand what factors increase (or decrease) the probability an animal will use a sample unit. A use-availability design is a common sampling design in resource selection studies. We define 'use' as physical presence within a sample unit, which is often used synonymously with the term 'presence'. We define 'sample unit' as the basic unit from which data are collected. In a resource selection context, sample units can range from trees a woodpecker may forage on to resource patches of similar vegetation.

Under a use-availability sampling design, resource attributes (denoted x) are recorded from a random set of sample units that were used by an animal (denoted $z = 1$), and resource attributes are also recorded at a random set of sample units considered available to an animal. 'Available' sample units are synonymously called 'background' (Royle *et al.* 2012), 'contaminated controls' (Lancaster & Imbens 1996) or 'pseudo-absences' (Phillips, Anderson & Schapire 2006), though in practice it is unknown whether such sample units were used. Although these data are often referred to as 'use-availability' data (sensu Manly *et al.* 2002), some authors synonymously use the term 'presence-only' data. Estimating the absolute probability, a sample unit is used (i.e. a resource selection probability function; RSPF) from such data is difficult because the number of used sample units is not proportional to the occurrence of used sample units in the population of interest.

A common solution to this problem is to treat available sample units as if they were true absences. For example, Manly *et al.* (2002, p. 100) advocate fitting a logistic regression model to use-availability data. The resulting parameter estimates can then be substituted into a log-linear function that is assumed proportional to the absolute probability of use:

$$\Pr(z = 1|x) \propto \exp(\beta_1 x_1 +, \ldots, +\beta_p x_p).$$

.

This function is commonly referred to as a resource selection function (RSF), because it is assumed proportional to the absolute probability of use. Machine learning algorithms such as Maxent (Phillips, Anderson & Schapire 2006; Phillips & Dudík 2008) and Random Forests (Cutler *et al.* 2007) are also commonly used to construct RSFs from use-availability data. Machine learning methods focus primarily on maximizing predictive capability (Elith *et al.* 2006) rather than parametric estimation and can estimate highly complex relations between resource attributes and the relative probability a sample unit is used. We note that while some of the techniques outlined above, such as Maxent, are frequently referred to as species distribution models, they address problems identical to those encountered in resource selection studies, namely what environmental variables are associated with the spatial distributions of species. For more detailed reviews of RSFs (and species distribution models), see Guisan & Zimmermann (2000), Manly *et al.* (2002), Guisan & Thuiller (2005), and Pearce & Boyce (2006). An important problem with treating available sample units as true absences is an inability to estimate the absolute probability a sample unit is used. The resulting RSF is assumed proportional to the absolute probability of use, though such proportionality is not guaranteed (Keating & Cherry 2004; Royle *et al.* 2012). Additionally, relative probabilities may be meaningless if baseline probabilities are close to 0 or 1. For example, even if a sample unit is 5 times more likely to be used when a particular attribute is present, if the baseline probability of use is 0·0001, an animal is still highly unlikely to use that sample unit.

Given the shortcomings described above, practitioners tasked with wildlife management and ensuring biodiversity should prefer to build RSPFs that produce unbiased estimates of the absolute probability a sample unit is used. Recall that under a use-availability study design, resource attributes, x, are recorded at a random set of used locations, $z = 1$. The central statistical problem is then estimating $\Pr(x|z = 1)$. Applying Bayes rule, we get:

$$\Pr(x|z = 1) = \frac{\Pr(z = 1|x)\Pr(x)}{\Pr(z = 1)}. \qquad \text{eqn 1}$$

Notice that the right-hand side of equation 1 contains the term $\Pr(z = 1|x)$. This can be modelled via the logit link as:

$$\text{logit}(\Pr(z = 1|x)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

and is the RSPF that is typically of interest to practitioners. Notice also that the denominator of equation 1 denotes the average probability any available sample unit is used, commonly referred to as 'prevalence'. This equation, and the associated likelihood function, has been obtained by several authors (Lele & Keim 2006; Dorazio 2012; Royle *et al.* 2012). Maximizing the likelihood function with respect to the parameters involves approximating $\Pr(z = 1)$ with large samples of available sample units (e.g. Lele & Keim (2006) suggest recording resource attributes at $\geq$ 10 000 available sample units). Although the maximum likelihood estimator associated with equation 1 provides unbiased estimates of RSPF parameters, problems persist. Recording resource attributes from enough available sample units to adequately approximate prevalence may be difficult, particularly if a large spatial area is considered available and resource attributes are measured in person on the ground. Additionally, Lele (2009) described numerical maximization difficulties with the maximum likelihood estimator proposed by Lele & Keim (2006).

Instead, one can obtain maximum likelihood estimates (MLEs) of RSPF parameters using a partial likelihood estimator derived from equation 1. Lancaster & Imbens (1996) proposed this model in the context of case–control sampling (hereafter called the case–control model with contaminated controls), Lele (2009) proposed the same model in the context of resource selection studies and this model is also the 'observed' likelihood described by Ward *et al.* (2009). The primary difference between the case–control model with contaminated controls proposed by Lancaster & Imbens (1996) and Lele (2009) and the full likelihood derived from equation 1 is that prevalence is treated like a parameter in the case–control model with contaminated controls. Although Lele (2009) demonstrated that MLEs of RSPF parameters obtained by maximizing this model with respect to the parameters are unbiased, widespread misconceptions exist, which has likely precluded widespread implementation. Keating & Cherry (2004) encountered difficulties fitting the case–control model with contaminated controls, including failure of optimization algorithms to converge to a unique solution when using categorical covariates or if starting values were far from actual values and lack of commercial software for fitting this model. Unfortunately, the difficulties encountered by Keating & Cherry (2004) have led others to dismiss this model as unstable and difficult to implement (e.g. Johnson *et al.* 2006; Pearce & Boyce 2006; Li, Guo & Elkan 2011). Another common misconception is that prevalence cannot be estimated from use-availability data (Elith *et al.* 2011).

Solutions to all of these problems have been proposed in the literature, but widespread use of the case–control model with contaminated controls suffers from poor linkages among relevant advancements, a divergent terminology and thus continued misconceptions. For example, Lele & Keim (2006) describe the circumstances under which parameters associated with categorical covariates can be estimated. However, they do not reference the problems encountered by Keating & Cherry (2004), and thus, their solution may have gone widely unnoticed. Similarly, Royle *et al.* (2012) dispel the notion that prevalence cannot be estimated from use-availability data. However, Keating & Cherry (2004) refer to prevalence as the 'unconditional probability of use', and Lele (2009) simply refers to prevalence as 'α' (noting the constraint $\alpha \in (0, 1)$). Thus, it may be unclear to many readers that the advancement made by Royle *et al.* (2012) even applies to the models considered by Keating & Cherry (2004) and Lele (2009). Finally, there are few linkages among relevant literature. For example, Lele (2009) neither cites Lancaster & Imbens (1996) with the original formulation of the case–control model with contaminated controls, nor suggests the model he proposes is the same one evaluated by Keating & Cherry (2004). Thus, many practitioners may fail to notice that Lele (2009) provides solutions to many of the problems encountered by Keating & Cherry (2004).

Here, we address commonly held misconceptions regarding Lancaster & Imbens (1996) and Lele's (2009) case–control model with contaminated controls. Using simulations, we demonstrate that parameters associated with categorical covariates and prevalence can be estimated from use-availability data. We also show that modern computational advances can be used to obtain stable estimates of RSPF parameters. We go beyond demonstrating the basic feasibility of the case–control model with contaminated controls and evaluate model behaviour over a variety of realistic field conditions, which can help guide future studies. We also provide R and WinBUGS code (Appendix S1, Supporting information) to make the model accessible to potential users. By demonstrating the basic feasibility of this model, using simulations to help guide study design and providing model code, we hope to encourage widespread application of a promising model in studies of resource selection.

Materials and methods

THE CASE–CONTROL MODEL WITH CONTAMINATED CONTROLS

We begin by deriving Lancaster & Imbens (1996) and Lele's (2009) case–control model with contaminated controls from a basic case–control model. We believe developing this model in a case–control context will draw explicit links between use-availability sampling and more familiar logistic regression models. Case–control sampling involves collecting a random sample of n_1 used sample units from the population of N_1 used sample units and a second sample of n_0 unused sample units from the population of N_0 unused sample units. Note that case–control sampling assumes use and nonuse is known without error. Relevant covariates are recorded at all used and unused sample units. We denote $\eta_i = 1$ if sample unit i is included in either the used or unused samples. Additionally, we denote $z_i = 1$ if sample unit i possesses a trait of interest, $z_i = 0$ otherwise. Hereafter, the trait of interest is whether a sample unit is used by the study species. We can then write the probability of any used unit being included in the sample as $P_1 = P(\eta_i = 1|z_i = 1) = \frac{n_1}{N_1}$. Similarly, we can write the probability of any unused unit being included in the sample as $P_0 = P(\eta_i = 1|z_i = 0) = \frac{n_0}{N_0}$ (Hosmer & Lemeshow 2000; Keating & Cherry 2004). Knowing the relative frequencies of used and unused sample units in the population of interest, we can write the probability of using a sample unit, conditional on covariates and the probability that a unit is included in the sample, as:

$$P(z_i = 1|x_i, \eta_i = 1) = \frac{\exp(\beta_0 + \ln(\frac{P_1}{P_0}) + \beta_1 x_{i1} + \cdots + \beta_j x_{ij})}{1 + \exp(\beta_0 + \ln(\frac{P_1}{P_0}) + \beta_1 x_{i1} + \cdots + \beta_j x_{ij})} \qquad \text{eqn 2}$$

where β_0 is the intercept parameter and $\beta_1, \ldots \beta_j$ are the j parameters associated with the $x_{i1}, \ldots, x_{ij}$ unique covariates. The ratio $\frac{P_1}{P_0}$ is the case–control adjustment necessary to account for used and unused sample units not being sampled proportionally.

In studies of resource selection, we often know with certainty that particular sample units are used by a species of interest.

However, two problems are common to use-availability data. First, we do not know which sample units remain unused, since failure to detect use does not mean a sample unit remained unused. Second, we rarely know the proportion of all sample units that were used by a species of interest. In use-availability sampling, we still collect a random sample of n_1 used sample units. In addition, we collect a second sample of n_a sample units considered available to the species of interest, without regard to use. Relevant covariates are recorded at all used and available sample units. We denote prevalence as π, noting that $\pi = \frac{N_1}{N}$, where $N = N_1 + N_0$. We then rewrite the case–control adjustment introduced in equation 2 as:

$$\frac{P_1}{P_0} = \frac{\frac{n_1}{N_1}}{\frac{n_0}{N_0}}. \qquad \text{eqn 3}$$

While we do not know which of the available sample units are used, we expect πn_a sample units are used and $(1-\pi)n_a$ sample units are unused. Therefore, to accommodate our uncertainty in which available sample units are used, we redefine equation 3 as

$$\frac{P_1}{P_0} = \frac{\frac{n_1+\pi(n_a)}{N_1}}{\frac{(1-\pi)n_a}{N_0}} = \frac{n_1}{\pi n_a} + 1 \qquad \text{eqn 4}$$

which allows us to express the case–control adjustment in terms of the known values n_1 and n_a, and the unknown term π, which is to be estimated. We can now redefine the probability a sample unit is used, accounting for a use-availability sampling protocol, as:

$$\psi_i = P(z_i = 1 | x_i, \eta_i = 1) = \frac{\exp(\beta_0 + \ln(\frac{n_1}{\pi n_a} + 1) + \beta_1 x_{i1} + \cdots + \beta_j x_{ij})}{1 + \exp(\beta_0 + \ln(\frac{n_{11}}{\pi n_a}) + \beta_1 x_{i1} + \cdots + \beta_j x_{ij})}. \qquad \text{eqn 5}$$

Above, we have established a model for the probability a sample unit is used. However, in use-availability sampling, the z_is are only partially observed. We know the species used all n_1 used samples, but we do not observe z_i at the n_a available sample units. Therefore, we also need to develop a model for our observations, conditional on the latent state. Let $y_i = 1$ for all used samples and let $y_i = 0$ for all available samples. Conditional on a sample unit being used ($z_i = 1$) and selected for sampling, we can write the probability of observing use as:

$$\theta_i = P(y_i = 1 | z_i = 1, \eta_i = 1) = \frac{n_1}{n_1 + \pi n_a}. \qquad \text{eqn 6}$$

Observing use or availability can be considered a Bernoulli trial, and we can write the likelihood as:

$$L(\beta_0, \beta_1, \ldots, \beta_j, \pi | y_i, \boldsymbol{x}_i, n_1, n_a) = \prod_{i=1}^{n} (\psi_i \times \theta_i)^{y_i} (1 - [\psi_i \times \theta_i])^{1-y_i}. \qquad \text{eqn 7}$$

Finally, this model is amenable to Bayesian analysis (Lele 2009), since:

$$f(\beta_0, \beta_1, \ldots, \beta_j, \pi | y_i, \boldsymbol{x}_i, n_1, n_a) \propto L(\beta_0, \beta_1, \ldots, \beta_j, \pi | y_i, \boldsymbol{x}_i, n_1, n_a) \times f(\beta_0) f(\beta_1) \times \ldots \times f(\beta_j) f(\pi) \qquad \text{eqn 8}$$

where $f(\beta_0, \beta_1, \ldots, \beta_j, \pi | y_i, x_i, n_1, n_a)$ is the joint posterior distribution of the parameters, conditional on the data, and $f(\beta_0)$, $f(\beta_1)$, …, $f(\beta_j)$, $f(\pi)$ are the prior distributions of model parameters. R and WinBUGS code for fitting this model is provided in Appendix S1.

Simulation Study

We conducted a simulation study to evaluate several properties of the case–control model with contaminated controls. First, we evaluated whether parameters associated with categorical covariates and prevalence can be estimated from use-availability data. Second, we show that modern computational advances (e.g. data cloning, Lele, Dennis & Lutscher 2007) can be used to obtain stable estimates of model parameters. Finally, we evaluate the behaviour of parameter estimates in relation to the number of sample units (hereafter sample size) and prevalence. We evaluated models at low (π = 0·05), moderate (π = 0·45) and high (π = 0·75) prevalence and with sample sizes of $N = 1000$ and $N = 2000$ sample units. Sampling was evenly split between used and available sample units, so a sample size of $N = 1000$ represents $n_1 = 500$ used and $n_a = 500$ available sample units. We thus evaluated six unique combinations of prevalence and sample size and randomly simulated 100 data sets per scenario.

We modelled the probability of using a sample unit as a function of one continuous and one categorical variable. The continuous and categorical variables represent hypothetical resource attributes at each sample unit. We fixed the intercept parameter, $\beta_0^{true} = 0$, the parameter associated with the continuous covariate, $\beta_1^{true} = 3$ and the parameter associated with the categorical covariate $\beta_2^{true} = -3$. The entire simulated landscape was composed of $2{\cdot}8 \times 10^6$ sample units. For each sample unit, we drew random values of the continuous covariate, x_{i1}, from a normal distribution and we drew random values of the categorical covariate, x_{i2}, from a Bernoulli distribution. We varied prevalence by altering the mean of the continuous covariate. For example, the mean value of the continuous covariate was relatively small for low prevalence scenarios, resulting in a reduced average probability of use over all sample units. We calculated the 'true' probability of using a sample unit as $\text{logit}(\psi_i^{true}) = \beta_0^{true} + \beta_1^{true} \times x_{i1} + \beta_2^{true} \times x_{i2}$. A sample unit was then considered used if the 'true' probability of use at each sample unit was greater than a number randomly drawn from a *uniform*(0, 1) distribution. We considered all $2{\cdot}8 \times 10^6$ sample units available.

We fit the case–control model with contaminated controls to simulated data using Lele, Dennis & Lutscher's (2007) data-cloning algorithm. We refer the reader to Lele, Dennis & Lutscher (2007) for details, but note two important properties of this algorithm. First, data cloning provides numerically stable MLEs of parameters and associated variances and co-variances, even when starting values are far from MLEs. Second, data cloning provides Bayesian estimates when implemented with only one clone.

Hellbender Resource Selection

As an application to field data, we fit the case–control model with contaminated controls to Ozark hellbender (*Cryptobranchus alleganiensis bishopi*) resource selection data collected on the North Fork of the White River (NFWR) in southern Missouri. Hellbenders are relatively large, long-lived, fully aquatic salamanders adapted to streams with abundant rocky cover (Taber, Wilkinson & Topping 1975; Bodinof *et al.* 2012). The Ozark subspecies is endemic to a narrow region in extreme southern Missouri and northern Arkansas where it has declined precipitously since the 1980s, resulting in listing as an endangered species under the Endangered Species Act (Federal Register 2011). Fourteen captive-reared hellbenders were fitted with radiotransmitters and released into the NFWR. Hellbenders were relocated approximately every 24–36 h between 19 May and 14 November 2008, and again between 26 March and 18 August 2009. All telemetry locations were visually confirmed and were considered used sample units in our analysis. In this context, sample units are spatial coordinates within a stream. Available sample units were randomly selected from a 5-m-radius circle (79 m^2) centred on each used sample unit, which was representative of a typical home range size for the species. Distance to cover (continuous) and substrate type (fine, coarse or bedrock; categorical) were measured at each used and available sample unit. We thus modelled the probability a hellbender used a sample unit as a function of distance to cover and substrate type. The analysis consisted of 1749 used and available samples (N = 3498). See Bodinof *et al.* (2012) for further details on sampling methodology and Bodinof *et al.* (2013) for archived data.

We fit hellbender use-availability data to a Bayesian implementation of the case–control model with contaminated controls, which is equivalent to using the data-cloning algorithm with 1 clone (Lele, Dennis & Lutscher 2007). We fit this model in WinBUGS (Gilks, Thomas & Spiegelhalter 1994) via the R2WinBUGS interface (Sturtz, Ligges & Gelman 2005; see Appendix S1 for model code). We assumed independent *normal*($\mu = 0$, $\sigma^2 = 100$) prior distributions for the intercept parameter and regression coefficients and assumed an independent *uniform*(0, 1) prior distribution for π. We evaluated sensitivity to prior distributions by additionally specifying independent *normal*($\mu = 0$, $\sigma^2 = 1000$) prior distributions for the intercept parameter and regression coefficients. We simulated marginal posterior distributions from three chains, each of which ran for 101 000 iterations. We discarded the first 1000 iterations as burn-in and kept every 100th iteration thereafter. We thus kept 3000 samples from the marginal posterior distribution of model parameters. The Brooks–Gelman–Rubin convergence diagnostic (Brooks & Gelman 1998) suggested adequate convergence ($\hat{R} \approx 1$ for all parameters).

Results

SIMULATION STUDY

Our simulations demonstrated low overall bias in MLEs and highlighted scenarios where biases may occur. We found increasing variation in point estimates of the intercept parameter and regression coefficients (the β_1 and β_2 parameters) as prevalence increased (Fig. 1), suggesting increased 'contamination' of available sample units (i.e. available sample units are actually used) makes parameter estimation increasingly difficult. Increasing sample size from $N = 1000$ to $N = 2000$ substantially reduced variation and bias in point estimates of these parameters. Point estimates for the parameter associated with the categorical covariate (the β_2 parameter, Fig. 1) showed identical patterns of variation and bias as the parameter associated with the continuous covariate (the β_1 parameter, Fig. 1), demonstrating that parameters associated with categorical covariates can be estimated from use-availability data. Additionally, π (prevalence, Fig. 2) was estimated without bias over all sample size and prevalence scenarios.

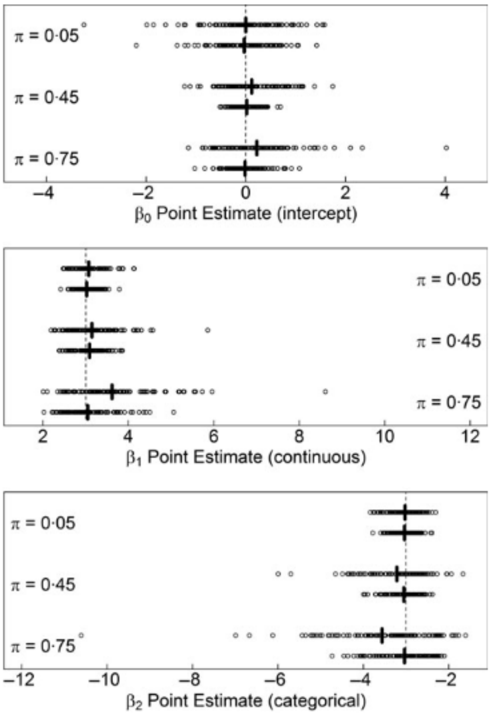

Fig. 1. Maximum likelihood estimates (MLEs) of model parameters (β_0, β_1, β_2) obtained from fitting Lancaster & Imbens's (1996) and Lele's (2009) case–control model with contaminated controls to simulated data. Each point represents MLEs from one simulated data set. Within each prevalence scenario (π = 0·05, π = 0·45 and π = 0·75), top lines indicate a sample size of 1000 ($n_1 = n_a = 500$) and bottom lines indicate a sample size of 2000 ($n_1 = n_a = 1000$). Vertical solid lines represent mean MLEs from each prevalence/sample size scenario, and the vertical dashed line represents the data-generating value for each parameter.

At high prevalence scenarios, the case–control model with contaminated controls tended to exaggerate estimates of β_1 and β_2 parameters. This exaggeration occurred because of a relatively flat 'ridge' in the likelihood surface at high prevalence scenarios (Fig. 3). This flat ridge led to occasional overestimation of the intercept parameter, β_0, and subsequent exaggeration of associated regression coefficients, which are correlated with the intercept parameter. Increasing the sample size created a steeper likelihood surface gradient, which reduced biases at high prevalence.

HELLBENDER RESOURCE SELECTION

Our Bayesian implementation of the case–control model with contaminated controls revealed a negative relation between hellbender use of sample units and distance to cover (mean $\hat{\beta}_1$ = −0·66, 95% CI = [−0·83, −0·50]), a positive relation between hellbender use of sample units and coarse substrate (mean $\hat{\beta}_2$ = 0·96, 95% CI = [0·76,

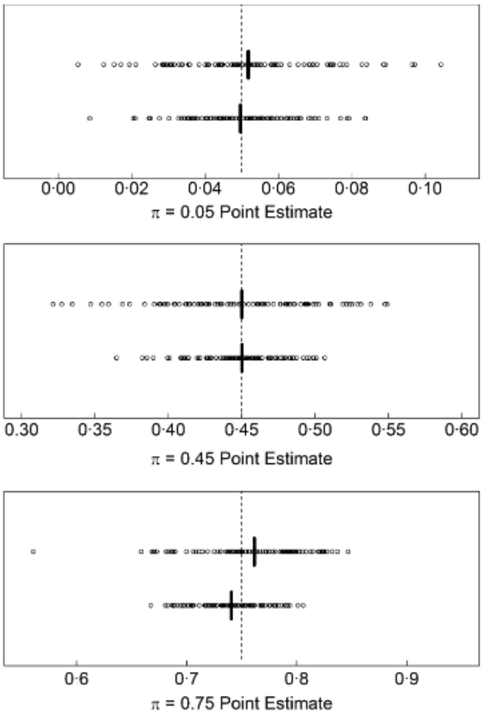

Fig. 2. Maximum likelihood estimates (MLEs) of prevalence (π) obtained from fitting Lancaster & Imbens's (1996) and Lele's (2009) case–control model with contaminated controls to simulated data. Each point represents MLEs from one simulated data set. Within each prevalence scenario ($\pi = 0{\cdot}05$, $\pi = 0{\cdot}45$ and $\pi = 0{\cdot}75$), top lines indicate a sample size of 1000 ($n_1 = n_a = 500$) and bottom lines indicate a sample size of 2000 ($n_1 = n_a = 1000$). Vertical solid lines represent mean MLEs from each prevalence/sample size scenario, and the vertical dashed line represents the data-generating value.

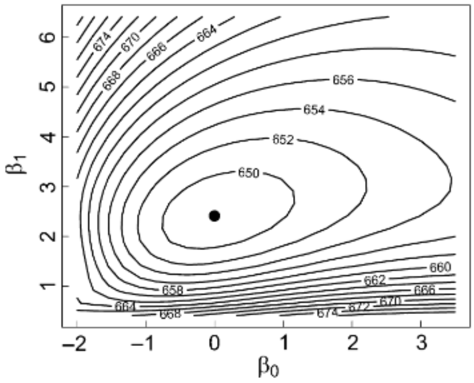

Fig. 3. Log-likelihood surface of the case–control model with contaminated controls calculated from a high prevalence ($\pi = 0{\cdot}75$), $n = 1000$ ($n_1 = n_a = 500$) simulation scenario. The single point represents the minimum $-1 \times$ log-likelihood. This log-likelihood surface demonstrates the increasingly shallow gradient at high values of β_0 and β_1.

1·18], Fig. 4) and no relation between hellbender use of sample units and bedrock substrate (mean $\hat{\beta}_3 = -0{\cdot}13$, 95% CI = [−0·43, 0·17]). Our model also estimated low hellbender prevalence (mean $\hat{\pi} = 0{\cdot}03$, 95% CI = [0·00, 0·10]), suggesting low probability of using any sample unit within the NFWR. This is consistent with current understanding of Ozark hellbenders, which remain rare throughout the NFWR. We found marginal posterior distributions of Bayesian models to be insensitive to choice of prior distributions selected for this analysis.

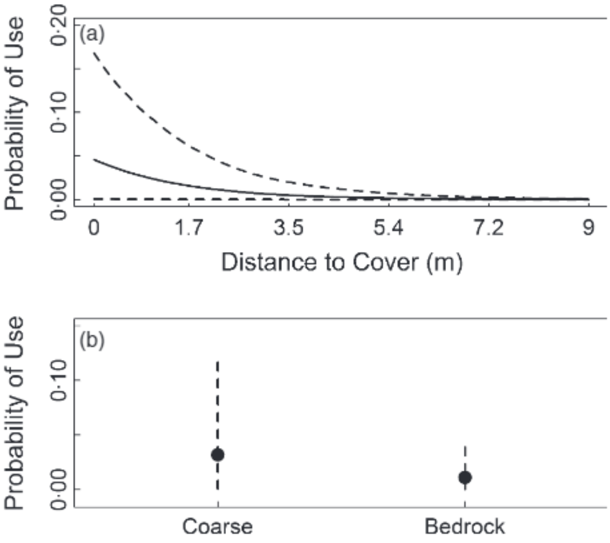

Fig. 4. Estimated probability a hellbender would use a location within a stream as a function of (a) distance to cover and (b) river substrate. The estimated distance to cover response assumes an underlying coarse substrate. The estimated river substrate response assumes a constant distance to cover of 0·60 m (the average distance to cover measured at all available sample units). Response curves were calculated from a Bayesian implementation of Lancaster & Imbens's (1996) and Lele's (2009) case–control model with contaminated controls. The solid line or point represents the mean estimated response for each level of a covariate, and the dashed lines represent 95% credible intervals.

Discussion

Our results demonstrate that the case–control model with contaminated controls originally proposed by Lancaster & Imbens (1996) and subsequently proposed by Lele (2009) is a stable and unbiased method for estimating the parameters of RSPFs from use-availability data. We overcame all of the previously reported shortcomings of this model, including sensitivity of optimization algorithms to starting values and an inability to estimate prevalence and parameters associated with categorical covariates. Keating & Cherry (2004) reported failure of optimization algorithms to converge to a unique value if starting values were far from MLEs. However, this result was a function of the optimization algorithm rather than a flaw in the model itself. If the likelihood surface contains local maxima, gradient-based optimization algorithms may converge on local maxima rather than global maxima if starting values are far from data-generating values. Modern computational advances, such as Lele, Dennis & Lutscher's (2007) data-cloning algorithm, help overcome these optimization issues. Data cloning relies on Markov chain Monte Carlo (MCMC) techniques often used for Bayesian estimation. As a result, data cloning will converge on MLEs, even if starting values are far from MLEs and local maxima exist (Gelman *et al.* 2004; Lele, Dennis & Lutscher 2007). Our results also address the commonly held misconceptions that neither categorical covariates (Keating & Cherry 2004) nor prevalence (Elith *et al.* 2011) can be estimated from use-availability data. Our simulations indicate that the case–control model with contaminated controls produces unbiased estimates of both categorical covariate parameters (the β_2 parameter) and prevalence (π). Although Lele & Keim (2006) described the conditions under which categorical covariate parameters can be estimated, and Royle *et al.* (2012) dispel the notion that prevalence cannot be estimated from use-availability data, we explicitly link these solutions to the problems encountered by Keating & Cherry (2004).

We believe the case–control model with contaminated controls offers several advantages when modelling resource selection of animals. Absolute probabilities are more intuitive to interpret than relative probabilities. Indeed, probabilistic interpretations are so intuitive that many software programs that construct RSFs from use-availability data (e.g. Maxent; Phillips, Anderson & Schapire 2006) produce output scaled between 0 and 1 (which is often erroneously interpreted as absolute probabilities). The case–control model with contaminated controls offers managers the ability to estimate the absolute probability a sample unit is used, facilitating straightforward comparisons between species and studies. Furthermore, models commonly used to estimate the parameters of RSFs, such as the exponential model (Manly *et al.* 2002, p. 100) or Maxent (Phillips, Anderson & Schapire 2006), produce resource selection 'indices', which may not be proportional to the absolute probability of use (Keating & Cherry 2004; Royle *et al.* 2012). In contrast, we demonstrated the case–control model with contaminated controls produces unbiased estimates of RSPF parameters. Finally, this model facilitates estimation of RSPF parameters with modest sample size requirements relative to alternative methods (e.g. Lele & Keim 2006; Royle *et al.* 2012), particularly if resource variables at available sample units are to be measured in the field. We thus believe the case–control model with contaminated controls will provide a practical method for estimating the parameters of RSPFs from field data.

Our simulations revealed potential sources of bias in the case–control model with contaminated controls. We expected some bias at high prevalence, since this leads to many available sample units that were actually used (i.e. 'contaminated controls'). In practice, we do not expect contamination rates at the level explored in simulated data ($\pi = 0{\cdot}75$) to be a problem, since common species (those with high prevalence) are more efficiently sampled using different protocols. For example, estimating the probability that a common species uses a sample unit may be more efficient by simply surveying a random selection of sample units and recording detection/nondetection. Indeed, a use-availability design is likely most efficient when the species of interest is relatively rare or difficult to detect, such that few observations would be made from a selection of sample units made without regard to use.

Our application of the case–control model with contaminated controls to hellbender use-availability data highlights the utility of this model when applied to a real data set. Recovery of Ozark hellbenders, like many rare habitat specialists, depends on conservation of specific resources that may naturally occur at low densities. In such circumstances, conservation planning can benefit from tools designed to identify habitat characteristics of high conservation priority, as well as species prevalence. For example, our application of this model was useful for identifying resource characteristics likely to be important to hellbenders as well as their rarity in a biologically relevant spatial extent (i.e. a river). Our estimates of the relation between probability of use and coarse substrate and distance to cover are consistent with Bodinof *et al.* (2012). However, our implementation had the advantage of estimating the absolute probability a hellbender would use a particular section of stream as a function of substrate and distance to cover. Estimating absolute probabilities of use is particularly useful for species that occur at low or high prevalence, since relative probabilities may be uninformative in this context. Indeed, we found that Ozark hellbenders were approximately 2·6 times as likely to use sections of stream that contain coarse substrate (because the odds ratio of using coarse substrate = $e^{\hat{\beta}_2} = e^{0.96} = 2.61$). However, the low prevalence estimated by our model indicates that they are still unlikely to use any portion of the NFWR. These findings emphasize the importance of identifying patches of densely arranged coarse substrate in NFWR as a conservation strategy for Ozark hellbenders.

In addition to estimating probabilities of use within a study area, parameters estimated using the case–control model with contaminated controls can also be used to predict the absolute probability of use at new sample units. This represents a major advantage of the case–control model with contaminated controls relative to modelling approaches that estimate the parameters of RSFs, since predictions of absolute probability of use are straightforward to interpret and compare across species. Accordingly, all of the tools commonly used to evaluate predictive performance (e.g. AUC, Fielding & Bell (1997), *k*-fold cross-validation, Boyce *et al.* 2002) can be used to validate RSPFs. Evaluating the predictive performance of a model with independent data is often the most useful way to evaluate that model's generality.

Our implementation of the case–control model with contaminated controls assumes independence of used and available samples. Assuming independence of used samples may be problematic if observations of space use are highly correlated. However, certain sampling protocols may help alleviate spatial autocorrelation in used samples. For example, ensuring that successive locations are adequately spaced in time may help alleviate concerns with spatial autocorrelation (Swihart & Slade 1985). Hellbender locations were separated by at least 24 h, which was assumed to be an adequate period for successive locations to be spatially independent. If spatial correlation is believed to be present in used samples, models that allow spatially correlated errors can be used. An autologistic model (Augustin, Mugglestone & Buckland 1996) may prove particularly useful in this context, since an autologistic model and the case–control model with contaminated controls rely on the same underlying RSPF. Another way to address spatial correlation in used samples is to model resource selection at the level of the individual animal and scale individual estimates to the population level (Marzluff *et al.* 2004; Thomas, Johnson & Griffith 2006). Spatial correlation represents a form of pseudoreplication (Hurlbert 1984), leading to overly precise, but unbiased, parameter estimates (Kutner *et al.* 2005). Thus, when population-level estimates are based on individual-level mean responses, spatial autocorrelation becomes irrelevant because individual-level means remain unbiased.

A critically important step in modelling use-availability data is defining what resources (or sample units) are available. In principle, all used resources represent a subset of available resources (Buskirk & Millspaugh 2006). Depending on the scale of the study, availability may be defined based on movement paths or home ranges of individual animals, up to the distributional limits of a species (Buskirk & Millspaugh 2006; Thomas & Taylor 2006). Additionally, availability is often defined by study site, political boundaries or by the limits of GIS coverage (e.g. when defining the ‘background’ in Maxent), though such arbitrary definitions can strongly affect inference regarding general patterns of resource selection (Johnson 1980). Definitions of what is available to an animal will necessarily differ to reflect study goals, though we recommend definitions that are biologically meaningful to a species rather than definitions based on convenience (e.g. conveniently available GIS layers).

Sample size should be considered when estimating the absolute probability of use from use-availability data. Even at sample sizes considered large for some field studies ($n_1 = n_a = 500$), the case–control model with contaminated controls exhibited nontrivial bias at high prevalence. A one-size-fits-all sample size recommendation is potentially problematic, since biases may operate as a function of underlying parameters such as prevalence or strength of resource selection. Nonetheless, we recommend samples no smaller than 500 or 1000 used sample units. We encourage potential users to conduct prospective simulations to guide appropriate sampling design when using this model, including exploration of various nonlinear response functions (e.g. quadratic, threshold) and link functions (e.g. probit link).

Given the relatively large-sample requirements, the case–control model with contaminated controls will probably be most useful when applied to data collected from animals fitted with radiotelemetry or satellite GPS technology. However, we note that this model is not restricted to such data. This model may also be suitable for large-scale survey efforts that generate reliable presence points, but do not generate reliable absences. For example, the case–control model with contaminated controls may prove useful for modelling breeding bird survey data, which generates reliable detections of breeding birds, but has been plagued by uncertain absences.

Our results tie together pieces of a disparate literature and demonstrate the unbiased nature of the case–control model with contaminated controls. We address the misconceptions that have prevented widespread use of this model and discuss how they can be overcome. Further, we identify conditions when the case–control model with contaminated controls may not be appropriate, helping guide the appropriate application of this model. Although presented in a resource selection context, this model can be extended to any context where a researcher wishes to compare a group with a known feature to the population as a whole. By demonstrating the unbiased nature of the case–control model with contaminated controls, we hope to spur further research into a model that promises to be a powerful tool in studies of resource selection.

Acknowledgements

This work was supported in part by the South Dakota Department of Game, Fish and Parks, State Non-Game Fund (Project W–75–R–132, No. 7599), State Wildlife Grant T39, study number 2439, Collection Agreement 08-CO-11221632, and the U.S. Forest Service, Rocky Mountain Research Station Joint Venture Agreement (07-JV-11221609-211). We thank J. Belant, R. Gitzen, F. Huettmann, B. Inouye, J. Stafford and D. Turner, and 3 anonymous reviewers for constructive comments on early versions of this manuscript.

Data accessibility

The data used in the Ozark hellbender resource selection analysis are archived in the Dryad repository (datadryad.org): doi:10.5061/dryad.2189s.

References

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.

Bodinof, C.M., Briggler, J.T., Junge, R.E., Beringer, J., Wanner, M.D., Schuette, C.D., Ettling, J. & Millspaugh, J.J. (2012) Habitat attributes associated with short-term settlement of Ozark hellbender (*Cryptobranchus alleganiensis bishopi*) salamanders following translocation to the wild. *Freshwater Biology*, **57**, 178–192.

Bodinof, C.M., Briggler, J.T., Junge, R.E., Beringer, J., Wanner, M.D., Schuette, C.D., Ettling, J. & Millspaugh, J.J. (2013) Data from: Habitat attributes associated with short-term settlement of Ozark hellbender (*Cryptobranchus alleganiensis bishopi*) salamanders following translocation to the wild. Dryad Digital Repository doi:10.5061/dryad.2189s. < http://datadryad.org >

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

Brooks, S.P. & Gelman, A. (1998) General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–455.

Buskirk, S.W. & Millspaugh, J.J. (2006) Metrics for studies of resource selection. *Journal of Wildlife Management*, **70**, 358–366.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J. & Lawler, J.J. (2007) Random forests for classification in ecology. *Ecology*, **88**, 2783–2792.

Dorazio, R.M. (2012) Predicting the geographic distribution of a species from presence-only data subject to detection errors. *Biometrics*, **68**, 1303–1312.

Elith, J., Graham, C., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillipa, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006) Novel methods improve prediction of species‘distributions from occurrence data. *Ecography*, **2**, 129–151.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E. & Yates, C.J. (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, **17**, 43–57.

Federal Register. (2011). *Endangered status for the Ozark hellbender salamander*. 76 (194): 61956–61978. United States Fish and Wildlife Service.

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.

Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2004) Posterior simulation. *Bayesian Data Analysis*, 2nd edn, pp. 283–310. Chapman & Hall/CRC, Boca Raton.

Gilks, W.R., Thomas, A. & Spiegelhalter, D.J. (1994) A language and program for complex Bayesian modelling. *Journal of the Royal Statistical Society Series D (The Statistician)*, **43**, 169–177.

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

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

Hosmer, D.W. & Lemeshow, S. (2000) *Applied Logistic Regression*. John Wiley & Sons, New York, NY, USA.

Hurlbert, S.H. (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, **54**, 187–211.

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

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*, **70**, 347–357.

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

Kutner, M.H., Nachtsheim, C.J., Neter, J. & Li, W. (2005) *Applied Linear Statistical Models*. McGraw-Hill, Boston, MA, USA.

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

Lele, S.R. (2009) A new method for estimation of resource selection probability function. *Journal of Wildlife Management*, **73**, 122–127.

Lele, S.R., Dennis, B. & Lutscher, F. (2007) Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters*, **10**, 551–563.

Lele, S.R. & Keim, J.L. (2006) Weighted distributions and estimation of resource selection probability functions. *Ecology*, **87**, 3021–3028.

Li, W., Guo, Q. & Elkan, C. (2011) Can we model the probability of presence of species without absence data? *Ecography*, **34**, 1096–1105.

Manly, B.F.J., McDonald, L.L., Thomas, D.L., McDonald, T.L. & Erickson, W.P. (2002) *Resource Selection by Animals: Statistical Design and Analysis for Field Studies*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

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.

Pearce, J.L. & Boyce, M.S. (2006) Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, **43**, 405–412.

Phillips, S., Anderson, R. & Schapire, R. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips, S.J. & Dudík, M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**, 161–175.

Royle, J.A., Chandler, R.B., Yackulic, C. & Nichols, J.D. (2012) Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution*, **3**, 545–554.

Sturtz, S., Ligges, U. & Gelman, A. (2005) R2WinBUGS: a package for running WinBUGS. *Journal of Statistical Software*, **12**, 1–16.

Swihart, R.K. & Slade, N.A. (1985) Influence of sampling interval on estimates of home range size. *Journal of Wildlife Management*, **49**, 1019–1025.

Taber, C.A., Wilkinson, R.F. Jr & Topping, M.S. (1975) Age and growth of hellbenders in the Niangua River, Missouri. *Copeia*, **4**, 633–639.

Thomas, D.L., Johnson, D. & Griffith, B. (2006) A Bayesian random effects discrete-choice model for resource selection: population-level selection inference. *Journal of Wildlife Management*, **70**, 404–412.

Thomas, D.L. & Taylor, E.J. (2006) Study designs and tests for comparing resource use and availability II. *Journal of Wildlife Management*, **70**, 324–336.

Ward, G., Hastie, T., Barry, S., Elith, J. & Leathwick, J.R. (2009) Presence-only data and the EM algorithm. *Biometrics*, **65**, 554–563.

Received 15 August 2012; accepted 3 April 2013
Handling Editor: Falk Huettmann

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. R and WinBUGS code for fitting the case-control model with contaminated controls.