New Idea

False negatives—A false problem in studies of habitat selection?

André Desrochers, Eliot J.B. McIntire, Steve G. Cumming, Josh Nowak, and Sapna Sharma

A. Desrochers (andre.desrochers@sbf.ulaval.ca), Centre d’étude de la forêt, Faculté de foresterie et de géomatique, Université Laval, Québec, CANADA G1V 0A6

E.J.B. McIntire (eliot.mcintire@sbf.ulaval.ca), Centre d’étude de la forêt, Faculté de foresterie et de géomatique, Université Laval, Québec, CANADA G1V 0A6

S.G. Cumming (stevec@sbf.ulaval.ca), Centre d’étude de la forêt, Faculté de foresterie et de géomatique, Université Laval, Québec, CANADA G1V 0A6

J. Nowak (Joshnk@yahoo.com), Centre d’étude de la forêt, Faculté de foresterie et de géomatique, Université Laval, Québec, CANADA G1V 0A6

S. Sharma (sapna.sharma23@gmail.com), Center for Limnology, University of Madison - Wisconsin, Madison, Wisconsin, USA, 53706

Abstract

Habitat use and resource selection studies based on presence-only data often use simulated locations as a benchmark for comparison with locations where a species has been observed. Such designs, termed “use-versus-availability”, are commonly analyzed with logistic regression, resulting in resource selection functions (RSFs). Statistical ecologists have recently expressed concerns about the appropriateness of case-control logistic regression for the calculation of RSFs, based on the claim that given enough time available locations may be used by the target species, thus creating “false negatives” and violating the assumption that the two outcomes, use and availability, are mutually exclusive. Accordingly, statistical ecologists propose alternative methods to address the latter concern. We argue that simulated absence data can be interpreted as true absences, hence amenable to case-control logistic regression, when absence data are considered as paired instantaneously in time with presence-only data.

Keywords: habitat selection, resource selection function, sampling design, radio-telemetry, logistic regression, false zeros

Introduction

Habitat selection is a process that results in disproportionate use by individual organisms of available habitats (Johnson 1980). For several decades, ecologists have compared observed species locations with random samples of locations within some reference area to infer habitat selection (Manly et al 2002). The random locations, in contrast to the observed, are deemed representative of what is available to the organism or population under study (for a discussion of “availability”, see Fortin et al 2005). Resource Selection Functions (RSF) were introduced in the early 1990s as a statistical means to evaluate the relative probability of occurrence in specific types of habitat by a set of animals in a given study region (Manly et al 2002).
Most commonly, a RSF is the result of a logistic regression examining two outcomes, “used” versus “unused” or “used” versus “available” locations against a set of habitat covariates measured at each location (e.g. landcover type, climate, topography, distance to some hypothesized critical feature). The sign and magnitude of regression coefficients are meant to inform us about whether organisms under study occur disproportionately in specific types of habitat. Designs treating locations as “used” versus “unused” are equivalent to case-control designs as defined by Hosmer and Lemeshow (2000). True case-control designs are considered rare in the habitat selection literature (Keating and Cherry 2004). Thus, RSFs are commonly fit using logistic regression under “use-available” designs to evaluate habitat selection when the only information available (apart from habitat covariates) is a set of locations where the individuals sampled were known to occur (see Keating and Cherry 2004 for statistical background). Such information is typical of individual-based studies based on VHF or GPS telemetry data (White and Garrott 1990).

Wildlife ecology relies increasingly on RSFs (Strickland and McDonald 2006). However, this enthusiastic use of RSFs has been challenged by some statistically minded ecologists. For example, in a frequently-cited paper, Keating and Cherry (2004), have questioned the interpretation of estimated and predicted quantities obtained through case-control logistic regression models of data obtained under a use-available design. Those concerns have led to the development of alternative ways to analyse designs comparing occurrence-only data to random locations throughout the area of interest (e.g., Johnson et al 2006, Lele 2009, Lele and Keim 2006) or to the next available locations, with a defined spatial constraint (Fortin et al 2005).

Here, we examine a key concern, the misinterpretation of available locations as “unused”, and argue that the logistic regression approach to RSFs and the interpretation of estimated parameters in terms of odds-ratios are in fact correct when the paired or stratified nature of so-called use-available designs is understood temporally. We illustrate this point with a simulation of animal movement in a heterogeneous landscape. We show that conditional or case-control logistic regression on the simulated data can recover the true selection coefficients without bias.

What do we mean by “unused”? Standard logistic regression can be used to estimate the unbiased parameters of an RSF in situations where all “used” (case) and “unused” (control) points are correctly attributed (Manly et al 2002: 83):

“One of the simplest ways of estimating a resource selection probability function (RSPF) involves taking a census of the used and unused units in a population of resource units, and fitting a logistic regression function for the probability of use as a function of variables that are measured on the units.”

Usually, these situations have been thought to arise primarily in census studies where the researchers sample the units of interest (e.g., plot) and identify whether the individual(s) of interest is (are) present or absent. In studies where the individual is the unit being followed (e.g., via GPS or VHF collar), the random sample of available point locations must be assumed to be unused, or more specifically “rejected for use”, to correctly use logistic regression. As traditionally formulated, this assumption can fail significantly for use-available designs, as discussed in detail by Keating and Cherry (2004). Problems are thought to arise because, contrary to true used-unused designs, some of the available points may in fact be used by sampled individuals at times when they are not observed. If a non-trivial proportion of “unused” are so contaminated (sensu Lancaster and Imbens 1996), or forced to be contaminated in simulations, logistic regression will naturally produce biased parameter estimates and probabilities of use (Johnson et al. 2006). Indeed, as pointed out by Lele and Keim (2006: 3026):

“(…) given enough time, eventually most of the study area gets visited and one could, albeit incorrectly, infer that probability of use is 1 for every type of habitat.”

How long must an available location be exposed in order to be used? Conversely, at what particular time is a location unused? To our knowledge, those questions have not been answered in theoretical or empirical RSF studies. Yet, if contamination of controls is cause for concern, then exposure of available sites must be sufficiently long such that sampled individuals have a non-negligible chance of visiting them. Who determines how long such exposure should be? Fortunately, logistic regression does not require exposure time for available locations to be above any specific threshold. Inverting the previously cited remark of Lele and Keim (2006), if we allow sampling time to be sufficiently short, then the probability of use, by monitored individuals, of all locations where presence was not recorded by telemetry approaches zero, and use-availability essentially becomes case-control. Simply stated, it is impossible for an individual to be at two places simultaneously. Samples of available locations will be uncontaminated provided that: 1) use data comes from a set of marked individuals, 2) available and used points are distinct,
...and 3) available and used points are paired in time, i.e., they are sampled simultaneously. Under those conditions, any sampled individual can be known to have been absent, at the time of observation, from all points other than where it was actually observed, and therefore those points were unused. Importantly, as the time interval decreases to zero, the bias from “false negatives” disappears and we are left with the true probability of use for that individual in that habitat. An analogous argument can be made if map resolution or extent are allowed to increase to the point where the probability of any pixel being used approaches zero (Aarts et al 2008).

Seen from a temporal perspective, habitat selection studies under the traditional framework of RSFs are amenable to case-control rather than use-availability sampling designs (sensu Keating and Cherry 2004). This is because they do not share the problem of studies with contaminated controls such as the one explored by Lancaster and Imbens (1996), for reasons worth discussing in some detail. In Lancaster and Imbens (1996), a set of workers with a certain common attribute were compared to a sample drawn from a larger population containing an unknown proportion of people having the attribute in question. A key aspect of Lancaster and Imbens (1996) study is that the attribute in question was invariant, akin to, for example, eye colour. In such cases, no matter what the temporal relationship between pairs of cases and controls, there remains a probability that some of the “controls” were in fact “cases”. Thus, in Lancaster and Imbens (1996) and similar situations, contamination cannot be avoided by controlling for time. However, in a habitat selection study the trait in question is an individual’s presence at a known location at a particular point in time and its absence from all other points, including a randomly selected control location. Thus, we argue that controlling for the time dimension removes the contamination problem.

“Shared time” and the meaning of pairing

Assigning specific times to random points is just a form of paired (one-to-one) or more generally, stratified (one-to-many) designs. In classic stratified designs, data from different treatments are grouped by levels of some meaningful shared factor such as county, forest stand or year. In the case of resource selection studies, each stratum will typically be a set of locations composed of one used and one or more available locations. To avoid pseudo-replication, the effect of stratum should be included as a factor in a mixed model or conditional model. The sample size for an individual is the number of used locations, regardless of how many random points were created to measure availability. What has not, to our knowledge, been previously considered is: can these locations within a given stratum share any non-spatial attribute? Traditionally, these shared attributes have been defined to constrain the inferential spatial scale of the question, such as home range, study site, or watershed. We can easily add shared time to these shared spatial attributes. If time, whether specified or not, is taken to be the precise moment at which the “used” location was obtained, then it follows that available points are in fact unused at a specific known time, at least in the case where marked individuals are monitored. If we accept the “shared-time” assumption, then the problem of contamination or false negatives in use-available designs vanishes, relative probabilities of occurrence are instantaneous, and unbiased RSFs at least can be estimated by case-control logistic regression. Instantaneous occurrence probabilities are the raw material for a variety of inferences, such as on the probability that animal “A” will occur in location “X” during time interval “T”. The only other information that is required to obtain true occurrence probabilities for the marked individuals is a set of locations where those individuals can be during the reference time-period. That defining this set of possible locations may prove difficult is not relevant to our interpretation of available locations.

Simulation

To demonstrate that our case-control approach provides unbiased estimates, we used a correlated random walk simulation model written in SELES (Fall and Fall 2001). Our model had two components: a randomly generated 5,000 by 5,000 pixel landscape of 7 discrete habitat types (“habitat polygons”), and 20 individuals moving within it. A correlated random walk is defined by the distributions of turn angles and movement distances. Here, turn angles were normally distributed, with a mean angle of 0 degrees relative to the movement direction at the previous time-step and a variance of 20 degrees. Turning angle was thus independent of habitat. Movement distance varied among habitat types, as follows. The seven habitat classes were coded as 1,2,…,7 and these values defined movement distance in pixels. Thus, a movement originating in a pixel of habitat type $i$ has length $i$ pixels and a direction randomly chosen, conditional on the previous movement. This formulation created a situation where, given an equal amount of time, an animal starting in habitat type 1 moved less far than those animals in habitat type 7, creating habitat selection for those habitat types with lower indices. Thus, habitat selection in our study is generated by differing move distances between habitats, emulating a potential causal mechanism for habitat selection by animals. This is distinctly different than generating our habitat selection from the same function as the analysis of resource selection, thus, we do not...
have true coefficients for the RSF. Instead, we calculated true frequencies of use of each habitat from all data points by counting the number of time steps the individuals were in each habitat divided by the total area (total number of pixels) of each habitat. We calculated true log odds from these true frequencies of use, and calculated known habitat preferences as the differences in log odds relative to a reference habitat type.

To emulate a common regular sampling strategy used with GPS collars, we first generated “used” points by taking a location for each individual every 20 units of time. To obtain a sample of unused points, we selected random points from the entire landscape excluding the paired used point—our conditional pairing in time. We performed 100 replicates of a simulation of 20 individuals with 13,500 samples from each. We then performed a conditional logistic regression and calculated estimated habitat preferences from the differences in estimated parameters. We found no bias in habitat preferences when contrasting known habitat preferences with estimated habitat preferences (Figure 1).

Inference beyond marked individuals

Wildlife telemetry samples the locations visited by a set of marked individuals. These samples may be systematic (e.g., hourly), random, or otherwise distributed in time. The data are appropriate for inference at the level of the set of marked individuals, given an appropriate specification of fixed and random effects. For the case of discrete habitat classes, the observed frequencies can be used to estimate the probabilities that a randomly selected individual at a randomly selected time will be in a particular habitat class. It also allows us to obtain true probabilities of use by one (or a set of) marked individuals for a given period of time. Given the “shared-time” assumption, a case-control logistic regression with a use-available design can produce unbiased estimates of the strength of habitat selection. As RSFs, the fitted models can also be used to predict the relative probabilities of use.

However, wildlife telemetry studies are usually designed to generate inferences at the population level, i.e., beyond the set of marked individuals. For this reason, researchers will normally assume that the monitored individuals are a random sample from the target population. If there is reason to believe that marked individuals used to measure resource selection constitute a random sample, true instantaneous probabilities of use of a given location by the species can be obtained if the proportion of the population that is marked is known or estimated. Probability of species’ use would be related to the time integral of this probability summed over all individuals in the population.

RSF studies are increasingly used to generate maps depicting absolute or relative probabilities of use by wildlife, and thus to help management decisions such as delineating areas for conservation or sustainable use. Absolute probabilities of selection by wildlife are arguably more useful to managers than relative probabilities (Lele 2009). Unfortunately, wildlife managers are often given relative selection probabilities, because the commonly held perspective in RSF studies is that random locations may be used by focal animals during a generally undefined time frame. Even though this may be true, we showed that this assumption is not necessary to generate simple RSFs. The “shared time” assumption presented here defines the time frame for unused locations and rids us of false negatives. It also enables us to obtain true probabilities simply from logistic analysis of use-available designs, and to apply those probabilities beyond the set of marked individuals under study.

![Figure 1](image_url)

**Figure 1.** Absence of bias in the estimation of habitat selection using logistic regression with simulated movement data. Bias represents the difference between estimated log odds and true log odds for a specific habitat (here, with movement rate = 2), relative to a reference habitat (rate = 4). Results for all habitats (movement rates) were similar to the one presented here.

**Acknowledgements**

This research was supported by grants from the National Scientific and Engineering Research Council of Canada. We thank Bryan Manly, Daniel Fortin, Thierry Duchesne, Mark S. Boyce, Pedro Peres-Neto, and Subhash Lele for comments on an early draft of the manuscript. S. Lele kindly provided thought-provoking...
papers for our reading group on statistical ecology, which led to this manuscript.

**Referees**

Pedro Peres-Neto – peres-neto.pedro@uqam.ca
Université du Québec à Montréal

Mark S. Boyce – boyce@ualberta.ca
University of Alberta

**References**


**Response to referee**

In his response to this paper, Boyce (2010) provides a clear and succinct summary of the rationale behind RSFs. We concur with his closing observation that it "should not be necessary to force the data into a statistical framework inappropriate for the data". Accordingly, even though the use of logistic regression has been shown to give unbiased estimates within a use-available framework, Boyce did not explicitly address the problem of inappropriate degrees of freedom for statistical tests for parameters. Using logistic regression, and associated tests of significance, gives equal degrees of freedom for each zero (available) and one (presence). Thus, logistic regression may be seen as inappropriate to the presence-only data generally collected by telemetry studies. This leaves us with the choice of re-thinking the analytical framework associated with this design (the predominant view) or revisiting the idea of use/availability (the motivation for our paper). We feel that the response by Boyce accurately depicts the current issues with the use of RSF as a basis for understanding habitat use, occupancy or selection, especially in a use/available framework. However, we think he has not fully addressed the fact that the underlying sampling scheme with presence-only data need not be the combination of a set of random locations from which a subset of presence locations is drawn. Another way to present our interpretation of the joint distribution of presence and availability data is to see it as a time series of spatial partitions, each of which is composed of mutually-exclusive case (presence)/control (absence) data.

Boyce appears to associate case-control sampling schemes with situations in which a limited-size buffer area around used locations is imposed so as to restrict the sampling of random locations. We agree that the spatial scale from which available or control locations are drawn is important to the interpretation of RSFs. However, we believe that spatial scale is not central to our argument. In essence, the case-control sampling scheme refers to the situations in which used and unused data points in space represent a partition of two mutually-exclusive sets, irrespective of spatial or temporal scale.

Part of the difficulty we have with the interpretation of RSFs and more generally with the arguments adduced in Boyce’s response is that the discourse in this field is riddled with terms such as "truly available for election", "use", and "occupancy." We find it difficult to
assign to these terms any precise, operational biological meanings. No amount of statistical sophistication will help us progress toward a better theory of habitat selection, if the underlying terms and concepts are vague.