Invited Paper

Resource Selection Functions Based on Use–Availability Data: Theoretical Motivation and Evaluation Methods

CHRIS J. JOHNSON,[1] *Ecosystem Science and Management Program, University of Northern British Columbia, 3333 University Way, Prince George, British Columbia, V2N 4Z9, Canada*

SCOTT E. NIELSEN, *Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9, Canada*

EVELYN H. MERRILL, *Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9, Canada*

TRENT L. McDONALD, *Western EcoSystems Technology, Inc., Cheyenne, WY 82001, USA*

MARK S. BOYCE, *Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9, Canada*

Abstract

Applications of logistic regression in a used–unused design in wildlife habitat studies often suffer from asymmetry of errors: used resource units (landscape locations) are known with certainty, whereas unused resource units might be observed to be used with greater sampling intensity. More appropriate might be to use logistic regression to estimate a resource selection function (RSF) tied to a use–availability design based on independent samples drawn from used and available resource units. We review the theoretical motivation for RSFs and show that sample "contamination" and the exponential form commonly assumed for the RSF are not concerns, contrary to recent statements by Keating and Cherry (2004; Use and interpretation of logistic regression in habitat-selection studies. Journal of Wildlife Management 68:774–789). To do this, we re-derive the use–availability likelihood and show that it can be maximized by logistic regression software. We then consider 2 case studies that illustrate our findings. For our first case study, we fit both RSFs and resource selection probability functions (RSPF) to point count data for 4 bird species with varying levels of occurrence among sample blocks. Drawing on our new derivation of the likelihood, we sample available resource units with replacement and assume overlapping distributions of used and available resource units. Irrespective of overlap, we observed approximate proportionality between predictions of a RSF and RSPF. For our second case study, we evaluate the classic use-availability design suggested by Manly et al. (2002), where availability is sampled without replacement, and we systematically introduce contamination to a sample of available units applied to RSFs for woodland caribou (Rangifer tarandus caribou). Although contamination appeared to reduce the magnitude of one RSF beta coefficient, change in magnitude exceeded sampling variation only when >20% of the available units were confirmed caribou use locations (i.e., contaminated). These empirically based simulations suggest that previously recommended sampling designs are robust to contamination. We conclude with a new validation method for evaluating predictive performance of a RSF and for assessing if the model deviates from being proportional to the probability of use of a resource unit. (JOURNAL OF WILDLIFE MANAGEMENT 70(2):347–357; 2006)

Key words

bias, contaminated control, habitat modeling, logistic discriminate, logistic regression, resource selection function, RSF, sampling design, use–availability, validation.

In their provocative review of logistic regression in habitat-selection studies, Keating and Cherry (2004) stated that using logistic regression with a use–availability design for estimating resource selection functions (RSF) "does not guarantee maximum-likelihood estimates, valid probabilities, or valid likelihoods." Indeed, Manly et al. (2002) clearly indicated that use–availability designs result in an RSF, which is proportional to the probability of use, not a resource selection probability function (RSPF). To obtain "valid" selection probabilities, it would be necessary to change designs or know sampling fractions so that an RSPF could be estimated. More important, Keating and Cherry's (2004) criticism of using logistic regression with a use–availability design centers on the fact that the statistical likelihood defined in Manly et al. (2002:100) can, in some cases, increase above 1.0 when the underlying RSPF is assumed to have an exponential form. They then correctly argue that the likelihood is not valid in these cases and that the resulting RSF might not provide values that are proportional to the true underlying RSPF. They suggest that these problematic cases arise when a large proportion of the resource units in the population are used, resulting in a contaminated sample with some resource units appearing both in the sample of used and sample of available units.

Despite some correctly qualified statements, Keating and Cherry's (2004) paper is easily misinterpreted to imply that estimating an RSF based on a use–availability design is generally flawed or inappropriate. Instead, for data collected in wildlife habitat studies, we argue that use–availability is often the most correct design. Applications of used–unused data are burdened by an asymmetry of errors meaning that used points are known with certainty but unused points are not and may become used if monitored more intensively or for a longer period. When constraints on sampling used and unused units exaggerate the asymmetry of errors and excessively restrict the domain of inference, it may be more appropriate to draw 2 independent samples: a sample of used resource units and a sample of random or available resource units, to estimate a RSF. Often, these RSFs assume an exponential form, a fact that was criticized by Keating and Cherry (2004). While careful interpretation is necessary, we show that the exponential form of the RSF is appropriate and that the method does not suffer from sample contamination, which we might find with case-control designs (Keating and Cherry 2004). We argue that the use–availability design for estimating an RSF remains a valuable method for studying wildlife–habitat relationships.

[1] *E-mail: johnsoch@unbc.ca*

In this paper, we briefly review the motivation and limitations of use–availability designs for estimating an RSF. We present a new derivation of the method that is different from that of Manly et al. (2002), but nonetheless arrives at the same conclusion. We then describe 2 case studies to illustrate a lack of bias in use–availability RSFs when confronted with sample overlap and contamination. We conclude by proposing a new evaluation method designed to detect poor fit of a RSF. Our validation method evaluates the predictive ability of the model and also whether model predictions deviate from being proportional to the probability of use as required for an RSF.

Theoretical Motivation

Logistic regression is a powerful, robust method that has seen much use in wildlife habitat studies (Manly et al. 2002). The interpretation of results from logistic regression must be considered carefully and are highly dependent on sampling design (Manly et al. 2002, Keating and Cherry 2004). If we have a single random sample of resource units that are inspected for use or nonuse, logistic regression clearly can be applied directly. We call these sampling situations 1-sample used–unused or presence–absence designs. For such designs, statistical inference procedures have been developed (Hosmer and Lemshow 2000) and logistic regression can provide estimates of the probability of use for a resource unit (i.e., the RSPF).

However, many wildlife habitat studies cannot provide unbiased assessments of which resource units are unused, or to do so the inferences from these studies are restrictive. Sampled resource units observed to be used can be identified with some certainty, but unused units can be difficult or impossible to identify within most practical sampling frames. For example, one might obtain a sample of resource units used by animals—say with radiotelemetry—but it might be impractical or impossible to know all of the resource units that could have been used to assign an unused designation with any certainty. One might still estimate a logistic regression model, but its application and interpretation must be constrained by the temporal and spatial domain of the sampling. We might be willing to state that estimates from surveys conducted during a certain week during specified times of day at particular sampling locations produce a RSPF, but the sampling constraints might severely limit the scope of application.

A more practical and perhaps more honest approach is to draw a sample of used resource units and a sample of available resource units (which might be either used or unused) and to estimate an RSF that is proportional to the probability of use. From this simple design, the likelihood of a unit being used given that it appears in either the used or available sample can be defined (see Appendix and Manly et al. 2002:eq. 5.8) and maximum likelihood estimation, with all its desirable properties, can be performed.

Maximum likelihood estimation of the RSF is the most desired solution except that maximizing the general likelihood requires specialized nonlinear maximization software such as SAS's Proc NLIN or S-Plus's NLMINB. The mathematics behind these procedures, and the ability to run them, is simply unapproachable for most biologists. This is the main reason why Manly et al. (2002) advocate a "short-cut," which allows the likelihood to be maximized using ubiquitous and easy-to-run logistic regression routines. The "short-cut," however, requires one to assume that the underlying RSPF (and RSF) has the exponential form

$$w^*(\boldsymbol{x}) = C\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k) = C\exp(\beta'\boldsymbol{x}) \quad (1)$$

for a vector, $\boldsymbol{x}$, of k predictor covariates with coefficients, β_i, and where C is a scaling constant to make $w^*(\boldsymbol{x})$ a valid probability. With this assumption, Manly et al. (2002) showed that standard logistic regression software can be used to maximize the likelihood to obtain estimates of the β's assuming a small fraction of the available population is used.

The exponential form for $w^*(\boldsymbol{x})$ is at the center of Keating and Cherry's (2004) criticisms of the method. They correctly pointed out that if the argument of the exponential is positive (i.e., $\boldsymbol{\beta}'\boldsymbol{x} > 0$), $w^*(\boldsymbol{x})$ is not a valid probability because it is >1.0, and that nothing in the method guarantees a positive argument in all cases. They then claimed the estimated RSF is not proportional to the RSPF in all cases. However, they failed to point out that the likelihood of Manly et al. (2003) is always valid provided $0 \leq w^*(\boldsymbol{x}) \leq 1$ for all $\boldsymbol{x}$, that a proper scaling constant C always exists even if we do not know its value, and that perfectly valid maximum likelihood estimates of a nonexponential $w^*(\boldsymbol{x})$, or of a function proportional to it, can be obtained from this likelihood by some maximization method other than logistic regression.

Over and above the theoretical issues surrounding the exponential form of $w^*(\boldsymbol{x})$, the only other significant objection raised by Keating and Cherry (2004) is that "contamination" of the available sample biases RSF estimates. Contamination is a legitimate issue in a case-control setting because it means that the control sample has a mixture of cases and noncases. Keating and Cherry (2004) claimed the same issue arises in use–availability designs, again implying systemic problems with logistic regression applied to use–availability designs.

In fact, there are no systemic problems in applying logistic regression to the use–availability design caused by the exponential form of $w^*(\boldsymbol{x})$ or contamination, even though the derivation of Manly et al. (2003) contains an assumption that is not satisfied in all situations. We provide an alternative derivation of the likelihood (Appendix) that closely parallels that of logistic discriminate analysis (Seber 1984:308–319) that does not require $\boldsymbol{\beta}'\boldsymbol{x} < 0$, yet arrives at the conclusion that logistic regression can be employed to maximize the use–availability likelihood when the RSF is assumed to have an exponential form. The price paid for eliminating the constraint on $\boldsymbol{\beta}'\boldsymbol{x}$ is that the RSPF cannot be estimated unless a census of both the available and used populations is taken. From this alternative derivation (Appendix) and subsequent empirical simulations, we conclude that applications of logistic regression to use–availability data yield useful, robust, and valid RSFs under relatively mild assumptions.

Contamination and Overlap

In this section, we address practical issues of contamination and the overlap in the distribution of used and available resource units. We define contamination to be obtaining a mixture of used and unused resource units in the sample of available units. In contrast, overlap occurs when a used resource occurs in both the sample of available units and sample of used units. For example, when monitoring the movements of woodland caribou, we might sample resource availability across a landscape and remove all animal locations that

correspond spatially with a sampled available unit. Following this design, the sample of available locations might contain future, past, or unrecorded locations of a caribou producing a contaminated sample. If we were to retain the animal locations that corresponded with the available unit, then we would observe overlap between the 2 samples.

Contamination is a concern for case-control studies, but case-control studies are designed to differentiate between cases with a characteristic (i.e., have a disease) and controls without a characteristic (i.e., do not have a disease). This objective is fundamentally different from the objective of use–availability designs. Use–availability designs are designed to estimate a function (i.e., $w^*[\boldsymbol{x}]$ or $w[\boldsymbol{x}]$), which when multiplied by the frequency of $\boldsymbol{x}$ in the available population produces the frequency of $\boldsymbol{x}$ in the used population. This function is the weighting function that transforms the available distribution into the used distribution as defined in the literature on weighted distribution theory (see Patil and Rao 1978).

Extending the contamination argument from case-control studies to use–availability studies is inappropriate. Use–availability designs allow for the possibility of contamination. In fact, prohibiting contamination by requiring that the available sample contain only unused units biases the RSF. To see this, consider the extreme example where the entire population of available units is used. In this case, the true RSPF is 1.0 and the sample of available units would consist entirely of used units (100% contamination). Apart from sampling error, use of logistic regression as advocated here will estimate a RSF that also is constant because the distribution of $\boldsymbol{x}$ on units in the used sample will be the same as the distribution of $\boldsymbol{x}$ on units in the available sample. This is the correct function.

A separate but related issue is sample overlap. The design envisioned by Manly et al. (2002:99) does not admit the possibility of overlap because the available sample is selected first without replacement. If a unit happens to be included in both the used and available samples, Manly et al. (2002:101–102) state that the unit should be dropped from the used sample. Envisioning that the used sample was drawn first without replacement, McDonald (2003) derived a similar likelihood assuming overlap units are dropped from the available sample. If sampling actually is done with replacement, it is not hard to imagine that application of either of these methods for dealing with overlap will result in bias if the probability of overlap is large (see case 2 below). Nonetheless, when sampling is with replacement, application of logistic regression to data where overlap units are retained in both the used and available samples yields unbiased estimates of coefficients (see Appendix). When overlap occurs, the variance estimates reported by the logistic regression procedure for coefficients are not correct, even though coefficients estimates are. If no overlap occurs, variance estimates are correct. Coefficient variances in the case of overlap should be estimated by bootstrap methods that resample distinct units. While application of logistic regression to cases with significant overlap is justified theoretically, we verify that estimates are unbiased in the following 2 case studies.

Case Study 1: Proportionality of RSFs and RSPFs

In our first case study, we use biological data to demonstrate that an estimated RSF remains approximately proportional to an RSPF over a wide range of overlap. For this case, we draw on unpublished 1-sample presence–absence data for 4 species of forest birds to estimate both RSFs and RSPFs. We then examine the proportionality of the resulting RSF–RSPF predictions under various levels of overlap. Data were observations of 4 bird species obtained at 1,637 randomly selected 6-ha sites in the Bighorn Mountains of Wyoming, USA. We chose to analyze 4 species that 1) showed selection for landscape covariates, and 2) for which the percentage of used sites ranged from low (4%) to high (67%; Table 1). We used a cover map developed from satellite TM imagery to estimate forest extent to the nearest ha within the 6-ha sample site and to derive 3 forest configuration measures: mean forest patch size (MPS), mean fractal dimension of forest patches (FMPFD), and distance to the nearest adjacent forest patch (MNN) using Fragstats (McGargil and Marks 1995). Elevation was recorded to the nearest 100 m at the center of the plot from a digital elevation model and ranged from 2,050 to 3,089 m. Occurrence (use) was defined as observing the bird species during at least 1 of 2 replicate visits to the site during the breeding season. Nonoccurrence (unused) for the RSPF analyses was defined as the failure to observe the species on any visit to the site. Available sites

Table 1. Parameters of resource selection functions (RSFs) and resource selection probability function (RSPFs) for 4 forest bird species sampled near Wyoming, USA, derived using logistic regression and either a used and unused (RSPF) or a used and available (RSF) design. Bird use of a site was indicated by its presence (used) during either of 2 visits or absence (unused). Available sites included both the used and unused sites. Sample size (n) used to develop the RSF included the number of available sites (used and unused) plus the number of used sites, while percent used of the available indicates contamination rates.

Species	RSPF			RSF		No. used sites	No. unused sites	No. available sites	RSF n	% Used of available
Western flycatcher	β_0	−3.7001								
	β_1	0.1418	Forest[a]	0.1467	Forest	68	1,569	1,637	1,705	4.2
Gray jay	β_0	−8.0747								
	β_1	0.3110	Forest	0.2820	Forest	165	1,472	1,637	1,802	10.1
	β_2	0.0019	Elevation	0.0017	Elevation					
Mountain chickadee	β_0	−2.2577								
	β_1	0.1554	Forest	0.1060	Forest	530	1,107	1,637	2,167	32.4
	β_2	1.1108	FMPFD	0.8922	FMPFD					
Ruby-crowned kinglet	β_0	−0.2086								
	β_1	0.4729	Forest	0.1239	Forest	1,097	540	1,637	2,734	67.0
	β_2	−0.2658	FMPS	−0.0517	FMPS					

[a]Forest is the extent (ha) of forest cover recorded to the nearest 1 ha in a 6-ha plot; elevation is m to nearest 100 m at center of plot; FMPS is the mean forest patch size; FMPFD is the mean forest patch fractal dimension.

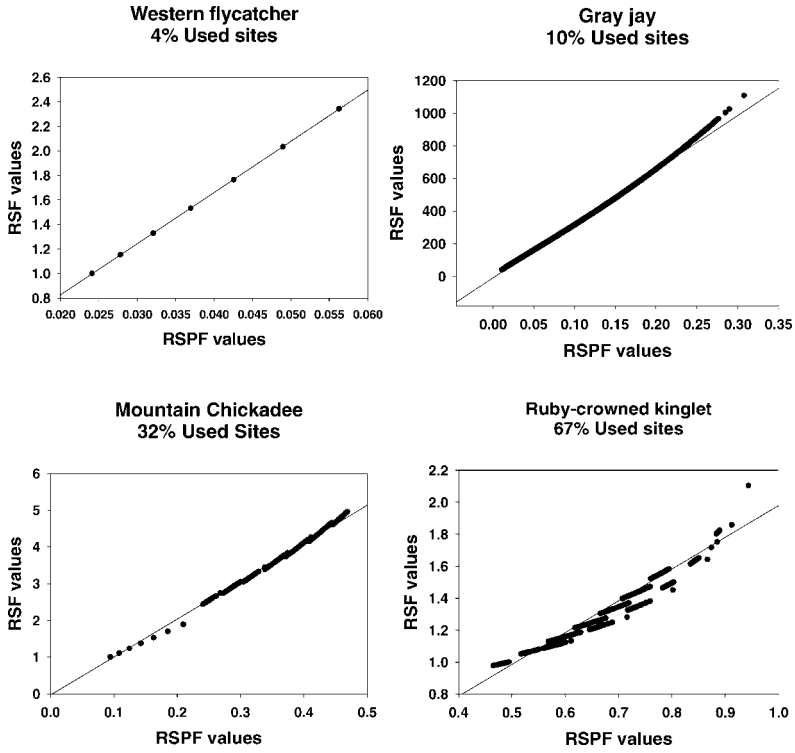

Figure 1. Relationship between predicted values for resource selection functions (RSF, y-axis) and resource selection probability functions (RSPF, x-axis) for bird species varying in occurrence sampled near Wyoming, USA. The least-squares straight line illustrates approximate proportionality between the RSF and RSPF values.

for all RSF analyses were defined to be the entire set of 1,637 sites. Treating the entire set of sites as available induced overlap between the used and available samples of between 4% and 67%; as a component of our evaluation, we ignored this overlap when fitting RSFs via logistic regression and all used resource units also were in the sample of available resource units.

We used logistic regression software to estimate both a RSF and RSPF model containing ≤ 2 covariates. The RSPF estimated assuming a used–unused design had the form

$$w^*(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots)}, \quad (2)$$

while the RSF estimated for a used–available design had the form

$$w(x) = \exp(\beta_1 x_1 + \beta_2 x_2 + \ldots). \quad (3)$$

We used AIC to choose the most parsimonious RSF from 10 a priori models that included variations of forest extent, elevation, and the 3 forest configuration indices (Anderson et al. 2000). When ΔAIC among models were <2, we used the most parsimonious model (Burnham and Anderson 2002). The same a priori models were used to select a RSF for each of the 4 species, and the covariates included in the chosen RSF then were used to develop the respective RSPF for each species. Following estimation, we plotted the predicted RSF for each site against the predicted RSPF for the same site (Fig. 1). We also plotted a least-squares linear regression line to evaluate proportionality. If the estimated RSF is proportional to the estimated RSPF, points in this plot will lie close to or along the straight line.

We found the extent (ha) of forest present within the 6-ha sample site to be the best predictor of the relative probability of selection for all 4 species (Table 1), and the only variable in the model for the western flycatcher (*Empidonax difficilis*). Mean forest patch fractal dimension (FMPFD) was included in the mountain chickadee (*Parus gámbeli*) model. Forest patch size (FMPS) was included in the ruby-crowned kinglet model (*Regulus calendula*). Elevation was included in the gray jay (*Perisoreus canadensis*) model (Table 1).

Despite overlap in used and available sites of 4%, 10%, and 32%, the estimated RSF was closely proportional to the RSPF for 3 of the 4 species analyzed except at the tails (Fig. 1), which is where the nonlinear effect of Eq. 2 compared to Eq. 3 is expected to be greatest. The remaining species (ruby-crowned kinglet) was very common, yielding an overlap of 67%. At high overlap, our range of observed RSPF values is wide compared to the previous examples, and again the deviation is expected to be greater because of the asymptotic nature of the RSPF. Nonetheless, even with this high overlap ($>50\%$), the RSF remained approximately proportional to the RSPF.

Case Study 2: Influence of Contamination on Beta Values

We recognize that past applications of RSF were formulated according to Manly et al.'s (2002) classic definition, where available resource units are drawn without replacement and the sample of used and available resource units are not permitted to overlap. Under such a design, extreme levels of contamination could bias RSF beta coefficients. To test this assertion, we develop a second case exploring the range of values in RSF β coefficients resulting from introduced contamination. As with the previous case study, we conducted our analyses using actual data typical of

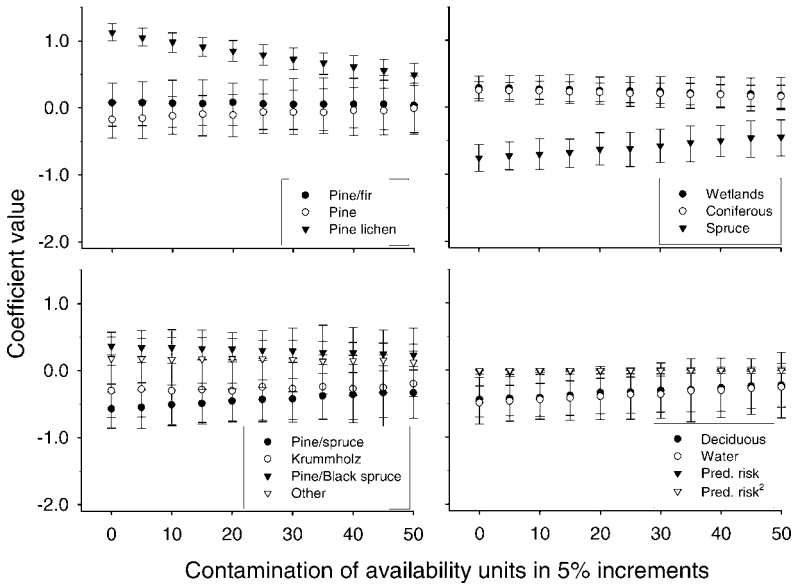

Figure 2. Effects of introduced contamination on magnitude of RSF coefficients calculated for woodland caribou of north-central British Columbia, Canada. Availability locations were contaminated in 5% increments from a hypothetical 0–50%.

use–availability studies of wildlife habitats, in this instance for data obtained from radiocollared animals.

We used location data previously reported for woodland caribou in central British Columbia, Canada (Johnson et al. 2002), to develop RSF models of the exponential form (eq. 3). Between December 1996 and March 1999, Johnson et al. (2002) fitted 8 female caribou with GPS radiocollars and recorded animal locations on a 3- or 4-hr schedule. To maintain simplicity of model interpretation, we used locations only from animals that spent the entire winter in forested habitats (Johnson et al. 2002). In total, we used 2,178 locations for estimating RSF models for woodland caribou.

For each animal location, we chose 1 random site to represent resource availability. Each random site was selected from a circular area (buffer) centered on the preceding animal location with a radius equal to the 95% movement distance for that GPS collar relocation interval (e.g., 4, 8, 12 hr, etc.; see Arthur et al. 1996). Resource use and availability were related to 12 categorical vegetation variables (Johnson et al. 2003) and 1 continuous variable for predation risk. We calculated predation risk as the weighted distance of a caribou location or random site from a patch of vegetation selected by radiocollared wolves (Johnson et al. 2002).

We assessed the impacts of contaminated availability data on RSF model parameters (i.e., β_is) by generating 10 new sets of use–availability data that ranged in contamination from 5% to a maximum of 50%, in 5% increments. Generation of these contaminated data sets required a number of manipulations of the original caribou use–availability data. First, we split our original sample of 4,356 caribou use and availability locations and units into 4 equal parts: 2 sets of caribou use locations and 2 sets of corresponding availability units. We designated 1 set of caribou locations and 1 set of available resource units (n = 2,178) for constructing the RSF and the remaining set of used locations (n = 1,089) for replacement. For each complete contamination data set, we randomly selected a percentage of caribou use locations (5–50%) and recoded them as available units, effectively contaminating the availability sample with locations that were visited by monitored caribou. As the final step in the simulation, we deleted the same number of availability sites as were contaminated and we randomly selected and added an equal number of used locations from the second set of 1,089 caribou locations. The randomly selected use locations served as replacements for the "contaminated" use locations and thus maintained an equal sample size of use locations and availability units across simulated levels of contamination. In reality, we might have observed this type of contamination if a GPS collar was unsuccessful in obtaining a location for a caribou occupying a resource unit, and that unit was then selected randomly as an available unit. Because this work is dependent on true not simulated data, our original sample of available resource units is burdened by some inherent contamination. Thus, our reported levels likely under-represent the true level of contamination in the caribou use data resulting in a relative not absolute comparison.

For each new data set with consistently higher contamination levels, we used a logistic model with the 12 categorical vegetation variables and 1 continuous variable for predation risk to estimate RSF coefficients. A discrete-choice model (Compton et al. 2002, Johnson et al. 2004) was more appropriate for our paired sampling design, but for consistency with the other case study, we used the more widely applied logistic model. When generating coefficients for the 10 RSF data sets containing introduced contamination, some sampling variation occurred due to the random selection of contaminated and replacement caribou use locations. We generated 500 data sets for each contamination level; calculated logistic regression models; and graphed the 5th, 50th (median), and 95th percentile coefficients. This procedure allowed a relative comparison of the potential range of coefficient values following 5–50% contamination of availability sites. We identified a significant effect when the 95th percentile coefficients of simulated data sets no longer overlapped the original data with an assumed contamination level of 0%.

Contamination of availability sites influenced the magnitude of the RSF β coefficients, although generally the deviation was small (Fig. 2). Change in the magnitude of coefficient values was greatest for pine lichen woodland, the covariate with the largest selection coefficient (Fig. 2). Given sampling variation, however, we observed significant differences for that covariate only following an introduction of 25% contamination. In all cases, change in coefficient value was oriented toward zero; thus, the maximum amount of change resulting from contamination was limited by the proximity of the coefficient to zero.

In an effort to place hypothetical and unobserved contamination rates into the context of actual use–availability data, we calculated the number of resource units (i.e., pixels) that would need to be visited by unobserved caribou over the course of our study to achieve a 5% contamination rate. Considering the 95th percentile movement radii for the caribou use locations (all caribou locations $n = 2{,}178$) the total area of available habitats equaled 7,200 km^2 or 11,520,000 25 × 25-m pixels. A 5% contamination rate would require the presence of at least 1 caribou in a minimum of 576,000 cells over the course of the study. Considering that our source population consisted of approximately 361 caribou (±136; Terry and Wood 1996) and that caribou travel in small groups during the winter, this extent of contamination is highly unlikely. As a final point of comparison, even if the approximate maximum number of caribou in the population (n = 497) visited a new and different pixel every hour for the 3-month winter period across the 3-year study duration, this population could contaminate a maximum of only 3,220,560 pixels (28%).

Model Evaluation

Resource selection function models are frequently used to predict maps of the relative probability of occurrence, especially since the integration of Geographic Information Systems (GIS) in wildlife ecology (e.g., Johnson et al. 2004, Treves et al. 2004). The predictive capacity or validation of maps produced by these models is often neglected, despite their widespread application in conservation and management. As Boyce et al. (2002) pointed out, however, there is a lack of statistical tests for assessment of model fit and accuracy for use–availability RSF models. The typical approaches for assessing logistic regression (e.g., ROC, Hosmer-Lemeshow goodness-of-fit, percent correctly classified, etc.) are inappropriate for the use–availability design. Boyce et al. (2002) suggested instead that RSFs should be evaluated on predictive performance using k-fold cross validation (Fielding and Bell 1997, Hastie et al. 2001). For each data fold, the withheld set can be assessed against the model predictions of the training data set using correlations between bin rank of the RSF values and the frequency of independent, withheld observations in the same bin rank standardized for area. Here, we modify this method to increase the precision of the evaluation technique and to assess the assumption that the RSF model is approximately proportional to probability of use.

Instead of relying on rank correlations between RSF bins and animal-use frequencies (Boyce et al. 2002), we propose the following approach:

1. Partition data into model-training and model-testing data (or k-fold groups).
2. Use logistic regression to estimate a RSF with the model-training data or alternatively for each training set in the k-folded data.
3. Predict RSF values in a GIS and reclassify pixels into ordinal classes or rank bins of a specified number.
4. Determine midpoint value of raw RSF scores for each ordinal RSF bin.
5. Determine the utilization $U(x_i)$ value for each bin i using the formula

$$U(x_i) = w(x_i)A(x_i)/\sum_j w(x_j)A(x_j) \quad (4)$$

where $w(x_i)$ is the midpoint RSF of bin i and $A(x_i)$ the area of bin i (Boyce and McDonald 1999).
6. Count the number of used observations in the withheld test data that fall in each RSF bin.
7. Estimate the expected number of validation observations within each bin (N_i) using,

$$N_i = N \times U(x_i) \quad (5)$$

where N is the total number of testing-data validation observations used and $U(x_i)$ the utilization function from eq. 4.
8. Compare expected (from step 7) to observed number (from step 6) of observations using linear regression and chi-square tests. First, assess the slope of the regression line for a significant difference from a slope of zero where use would equal availability and therefore indicate that the model is not different from that of a random or neutral model. Second, assess whether the slope is different from 1.0, which is the slope expected for a model that is proportional to the probability of use. Third, assess the constant for an intercept of zero, the intercept expected for a model that is approximately proportional to probability of use. And finally, use both the R^2 of the model and a χ^2 goodness-of-fit test to assess fit. A model that was proportional to probability of use would have a slope different from 0, but not different from 1, an intercept of 0, and a high R^2 value with a nonsignificant χ^2 goodness-of-fit value. Finally, χ^2 tests for each observed and expected proportion can be used to determine in which RSF bins the observed frequency differs from expected. If these conditions are not satisfied, the user might consider revisiting the process starting at step 3 (reclassify the RSF using a different model), rebinning the RSF values, or estimating a model with different environmental factors.

We demonstrate these metrics using the caribou RSF model with 0% contamination (see previous section) and an independent testing dataset of 267 caribou observations from 8 animals monitored by VHF radiotelemetry (Terry and Wood 1996). We used Spatial Analyst (Environmental Systems Research Institute 2004) to apply the coefficients from the caribou RSF model (Fig. 2) to the respective GIS covariate layers. We used quantile breakpoints to then reclassify the continuous RSF scores into 10 ordinal bins representing progressively more strongly selected

habitat classes. Total area (number of pixels) for each ordinal bin was queried from the resulting GIS map to estimate $A(x_i)$. Based on the breakpoint values of the reclassification we determined the midpoint RSF value $w(x_i)$ for each bin i, thus allowing us to calculate an expected utilization function $U(x_i)$. With $U(x_i)$ and known sample size of 267 caribou observations for independent validation (N), we calculated the number of expected observations within each ordinal bin (N_i). The number of observed observations, however, was simply the number of independent observations falling within each ordinal RSF bin on the predicted RSF map. We converted expected and observed numbers into proportions and assessed the relationship between expected and observed frequencies using linear regression. In addition, we assessed overall fit using a χ^2 goodness-of-fit test, as well as individual bin fit using χ^2 tests of observed to expected frequencies.

For the caribou example, the regression model suggested that the RSF model was reasonable overall, with a slope significantly different from 0, but not different from 1, and with an intercept close to 0 (Table 2, Fig. 3). However, model fit was lower than expected ($R^2 = 0.699$) suggesting that some RSF bins were different than expected from a model that was approximately proportional to the probability of use. Similarly, the chi-square goodness-of-fit test confirmed a relatively weak fit between observed and expected values ($\chi^2 = 15.81$, $p = 0.071$). Further assessments of individual bin fit using chi-square tests indicated that 5 of 10 bins were significantly different from expected (Fig. 3). Such differences might warrant the pooling of bins, or alternatively, an entirely different RSF model structure, because the model revealed inconsistencies and was not proportional to probability of use for an independent data set. Finally, assessments of individual animals revealed substantial variation suggesting that population-level predictions were not always useful for individual-level predictions (Table 3).

Discussion

Keating and Cherry (2004) raised 2 concerns over use of logistic regression for estimating resource selection functions. Their first was a theoretical concern stemming from application of logistic regression to maximize the general use–availability likelihood. Their second concern was that contamination of the samples can introduce bias into the parameter estimates and compromise the

Table 2. Resource Selection Function (RSF) bins where observed number of caribou locations from central British Columbia, Canada, differed (based on a χ^2 test) from the expected number of locations (indicated by x) based on $U(x_i)\times N$, with N being the total number of telemetry locations for an animal and $U(x_i)$ the expected proportion within bin of i.

	RSF bin (rank) number									
Caribou	1	2	3	4	5	6	7	8	9	10
041A	x	x			x	x	x	x		x
1D1A			x				x		x	
1D2B		x	x	x	x			x	x	x
771A								x		x
772B		x	x	x	x	x		x	x	x
832B	x	x			x	x	x	x	x	x
852B		x			x	x		x		
E41A	x		x			x		x	x	x

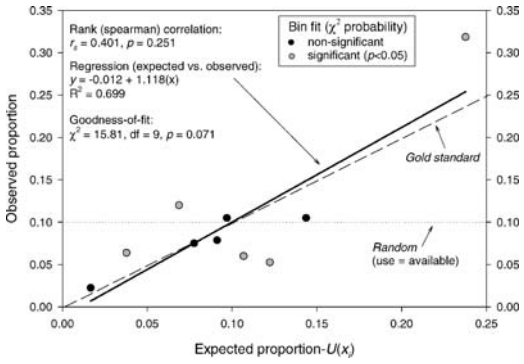

Figure 3. Expected versus observed proportion of telemetry observations (in 10 RSF bins) for an independent sample (n = 267) of caribou observations. A random (use = available) map would be depicted by observed values set at 0.1 (dotted line), while the gold standard (observed = expected) would occur along a line with a slope of 1 and intercept of 0 (dashed line). The fitted regression is shown as a dark line, while points are either black (bin observations are not significantly different than expected) or gray (significantly different from expected). Spearman rank correlation and overall goodness-of-fit are provided.

predictions from RSF and RSPF models. As a solution, they presented a novel likelihood developed by Lancaster and Imbens (1999) that explicitly accommodates contamination. Unfortunately, the likelihood of Lancaster and Imbens (1999) can be difficult to program and is unstable under certain conditions rendering it generally impractical (Keating and Cherry 2004).

We have made the point that Keating and Cherry's concerns are technically legitimate, but that these issues by no means render current or past RSFs estimated from use–availability designs invalid and useless. With regard to their first concern, we point out that the general use–availability likelihood usually is valid, although in some instances, when assuming an exponential RSPF, unbounded estimation via logistic regression might result in RSPF values >1.0. In these cases, the likelihood is no longer a true likelihood by classic definition. However, even when the likelihood is not a true likelihood in the classic sense, it can remain useful and yield coefficient estimates with negligible bias. There are many cases in statistics where nontrue or "quasi" likelihoods are useful and appropriate as the best or only analysis of collected data. McCullagh and Nelder (1989:325) note that most first-order, asymptotic theory connected with maximum likelihood is

Table 3. Characteristics (rank correlation and regression) of the accuracy of the global caribou Resource Selection Function (RSF) model for each individual caribou based on conditions local (Minimum Convex Polygon) to each animal. Caribou ranged across central British Columbia, Canada.

	Rank correlation		Expected vs. observed-regression		
Caribou	r_s	p	b_0	b_1	R^2
Pooled animals	0.401	0.251	–0.012	1.118	0.699
041A	0.820	0.005	–0.100[a]	2.001[b]	0.869
1D1A	0.869	0.001	0.000	0.996	0.921
1D2B	0.927	<0.001	–0.083[a]	1.827[b]	0.830
771A	0.716	0.020	–0.050[a]	1.499[b]	0.943
772B	0.665	0.036	–0.081[a]	1.809[b]	0.451
832B	0.842	0.002	–0.053[a]	1.534[b]	0.962
852B	0.651	0.042	0.015[a]	0.847	0.749
E41A	0.705	0.023	–0.071[a]	1.710[b]	0.840

[a] Significantly different from 0.0.
[b] Significantly different from 1.0.

founded on the well-known facts that 1) the expected value of the vector of log-likelihood partial derivatives with respect to all parameters is 0, and 2) that the negative of the expected value of the matrix of second-order partial derivatives of the log-likelihood is the covariance of parameters. If these properties are true for some function of the data and parameters, we may treat the function as if it were a true likelihood and maximize it to obtain maximum quasi-likelihood estimates with at least approximately the same characteristics as true maximum likelihood estimates. These properties probably hold, at least approximately, for the use–availability likelihood even in those cases where the argument to the exponential is >0. While more theoretical research is needed into the use of quasi-likelihoods in resource selection studies, it seems clear that negligible bias in parameter estimates can be expected even in extreme cases, and that this theoretical concern over the likelihood does not render the method generally useless and without merit. Furthermore, simply by interpreting the RSF as a logistic discriminant function between a distribution of used observations and a distribution of random sites, the $\boldsymbol{\beta}'\mathbf{x} \leq 0$ constraint is not necessary (Keating and Cherry 2004).

With regard to Keating and Cherry's (2004) second concern, it is clear that overlap or contamination of available resources by used resources does not always lead to inappropriate RSF models. We demonstrated that the RSF is approximately a linear function of the RSPF, even following large (>50%) amounts of sample overlap. Evaluating Manly et al.'s (2002) classic approach for sampling resource units, our second case study indicates that a contaminated sample of availability locations can alter model results potentially underestimating the true magnitude of selection or avoidance of a resource unit. In qualifying that general statement, our simulations reveals that contamination must be quite extreme before coefficients will deviate beyond sampling variation. Most coefficients were robust to contamination and only pine–lichen woodland, the most strongly selected resource, decreased to a level below sampling variation (5th percentile). Coincidentally, our observed contamination threshold is similar to the 20% value reported by Lancaster and Imbens (1999) as the level at which their modified likelihood was most appropriate for contaminated data. We note that our reported threshold is relative to the inherent contamination in the baseline data. We did not collar all caribou in the study population, nor did we monitor all animals from birth to death. Thus, our original and unperturbed sample of available resource units is certainly "contaminated" with unobserved locations.

In an effort to put a 5% contamination rate into the context of the study area and ecology of the woodland caribou that we monitored, we calculated the number of resource units that would need to be visited by unobserved animals during the study. Based on our analyses, a trivial contamination rate of 5% would require the presence of at least 1 caribou in an unrealistically large number of resource units. And of course, a 5% rate is well below the 20% threshold (at least for these data) where change in RSF coefficients might influence our conclusions. We admit that this comparison is crude and would be influenced by the definition of availability as well as the resolution (i.e., pixel size) of resource units.

Our results also revealed that percentage change from the true coefficient associated with contamination may be a function of the strength of selection or avoidance. We can envision this effect by considering habitat selection in the context of an uncontaminated sample where the sampled distribution of used resource units approximates that of the sample of available resource units, selection coefficients should approach zero. Likewise, where a sample of available resource units is completely contaminated by used locations, the distributions of resource units also should be near equal demonstrating no selection or avoidance of a resource.

Scale of observation, both temporal and spatial, should be an important consideration when evaluating the potential for contamination of animal use or plant occurrence data (Dungan et al. 2002). Although our results suggest that past applications of RSF are robust to contamination, practitioners should consider contamination on a case-by-case basis (Keating and Cherry 2004). The magnitude of β_i coefficient, spatial and temporal scale of observations, and the definition of resource availability might influence contamination. We suspect that study designs consisting of a small number of frequently used resource units relative to dense aggregations of animals or plants might result in levels of contamination sufficient to bias RSF models. However, based on our experience and the published literature, such cases are the small minority of RSF applications, not the majority.

We modified the *k*-fold cross validation method suggested by Boyce et al. (2002) to provide a technique that evaluates whether an estimated RSF is proportional to the probability of use. As with any statistical model, there can be no guarantee that the exponential model will yield a good RSF. Indeed, we show in our caribou example that an RSF might provide poor predictive capability. Additionally, it is not uncommon that the relationship between observed and predicted is nonlinear requiring an appropriate transformation of the RSF. Previous validation approaches (e.g., Boyce et al. 2002), although useful for evaluating the ranking of habitats, did not provide the necessary methods for testing the assumption of whether the RSF is proportional to probability of use. We suggest that validation, especially when making spatial predictions (e.g., maps), should be considered an essential element and step in the RSF modeling process.

We agree with Keating and Cherry (2004) that the form of the underlying model should be evaluated carefully; there is no reason to assume that the exponential model is the correct model in all situations, even though it is convenient and can work well. As demonstrated, we can plot the observed frequency of used resource units as a function of the predicted number of units to evaluate whether the selection function is of the correct shape (Boyce et al. 2002). If observed frequencies are linear relative to the predicted frequencies, then the RSF conforms to the important property of being "proportional to the probability of use" according to the Manly et al. (2002) definition. If not, a transformation, additional covariates, higher polynomial terms, or a different underlying model might be appropriate.

In practice, interpreting the exponential model as the logistic discriminant function, which contrasts a sample of used and a sample of available resource units is entirely consistent with the estimation of an RSF. Following this interpretation, resource units that appear in both the samples of used and available resource units creates no particular problems—the existence of both a one and a zero for these resource units means that they essentially

cancel each other and these resource units have little influence on the estimation of the logistic discriminant function contrasting the 2 distributions. Algorithms for estimating logistic regression permit the estimation of coefficients for eq. 1 even when resource units appear in both samples. When the RSF is evaluated using the validation method that we have described, we can ensure that the RSF is approximately proportional to the probability of use.

Management Implications

We urge researchers to consider carefully the sampling protocol and the choice of methods when developing and interpreting RSFs and RSPFs. For cases where one might encounter asymmetry of errors, we argue that a RSF constructed from use–availability data is the best choice. This sampling design is especially appropriate for studies that monitor mobile species discontinuously or, in general, where a census of all used units is difficult or impossible. Even where unused units can be reliably identified it might be impossible to confirm that those units will remain unused in the future. Thus, for some used–unused designs, researchers will need to constrain the scope of inference to the sampling period; such constraints will limit the general application of study findings.

Addressing the criticisms presented by Keating and Cherry (2004), we re-derived the use–availability likelihood and show that it can be maximized by logistic regression software. Application of this likelihood, when sampling used and available units with replacement, will produce RSFs that are robust to sample overlap and are proportional to the true probability of using a resource unit. As an example, we fit RSFs to use–availability data for a number of bird species with various levels of sample overlap. We demonstrate that the RSF can be an approximately linear function of the RSPF even at high rates of overlap. For past or current applications requiring model predictions that are proportional to the true probability of use, we present a technique that allows researchers to assess the proportionality of the predictions of RSFs relative to RSPFs.

Recognizing that past applications of RSF might have sampled available units without replacement (Manly et al. 2002), we evaluated the effect of sample contamination on β_i coefficients. Following the controlled and systematic contamination of available units for woodland caribou, some β_i coefficients converged toward zero. However, we observed a significant effect for only 1 covariate, and that was following contamination of approximately 25% of the available locations with use locations. These results suggest that most past applications of RSF are robust to sample contamination.

In summary, we demonstrate that the likelihood for calculating RSFs using logistic regression is valid, predictions from RSFs can be proportional to the true probability of use, and many past applications of RSFs are robust to sample contamination. Contrary to the conclusions of Keating and Cherry (2004), our results suggest that when carefully evaluated, RSFs estimated using logistic regression can be a powerful and useful tool for wildlife management and ecology.

Acknowledgments

We thank G. Seber for his help with the Lagrange multiplier technique in the appendix. This paper benefited from the comments and suggestions of B. Manly, D. Thomas, and 2 reviewers.

Literature Cited

Anderson, D. R., K. P. Burnham, and W. L. Thompson. 2000. Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wildlife Management 64:912–923.

Anderson, J. A., and V. Blair. 1982. Penalized maximum likelihood estimation in logistic regression and discrimination. Biometrika 69:123–136.

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

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

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

Burnham, K. P., and D. R. Anderson. 2002. Model selection and inference: a practical information–theoretic approach. Second edition. Springer-Verlag, New York, New York, USA.

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

Dungan J., J. N. Perry, M. R. T. Dale, P. Legendre, S. Citron-Pousty, M. -J. Fortin, A. Jakomulska, M. Miriti, and M. S. Rosenberg. 2002. A balanced view of scale in spatial statistical analysis. Ecography 25:626–640.

Environmental Systems Research Institute. 2004. Spatial Analyst Version 9.1. Redlands, California, USA.

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

Hastie, T., T. Tibshirani, and J. Friedman. 2001. The elements of statistical learning: data mining, inference, and prediction. Springer, New York, New York, USA.

Hosmer, D. W., and S. Lemeshow. 2000. Applied logistic regression. John Wiley and Sons, New York, New York, USA.

Johnson, C. J., N. D. Alexander, R. D. Wheate, and K. L. Parker. 2003. Characterising woodland caribou habitat in sub-boreal and boreal forests. Forest Ecology and Management 180:241–248.

Johnson, C. J., K. L. Parker, D. C. Heard, and M. P. Gillingham. 2002. A multiscale behavioral approach to understanding the movements of woodland caribou. Ecological Applications 12:1840–1860.

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

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

Lancaster, T., and G. Imbens. 1996. Case-control studies with contaminated controls. Journal of Econometrics 71:145–160.

Manly, B. F. J., L. L. McDonald, D. L. Thomas, T. L. McDonald, and W. P. Erickson. 2002. Resource selection by animals: statistical analysis and design for field studies. Second Edition. Kluwer, Boston, Massachusetts, USA.

McCullagh, P., and J. A. Nelder. 1989. Generalized linear models. Second edition. Chapman and Hall, London, United Kingdom.

McDonald, T. L. 2003. Estimation of resource selection functions when used and available samples overlap. Pages 35–39 *in* S. Huzurbazar, editor. Resource selection methods and applications. Omnipress, Madison, Wisconsin, USA.

McGarigal, K., and B. J. Marks. 1995. FRAGSTATS: spatial pattern analysis program for quantifying landscape structure. U.S. Forest Service General Technical Report PNW-351.

Patil, G. P., and C. R. Rao. 1978. Weighted distributions and size-biased sampling with applications to wildlife populations and human families. Biometrics 34:179–189.
Seber, G. A. F. 1984. Multivariate observations. John Wiley and Sons, New York, New York, USA.
Terry, E. L., and M. D. Wood. 1996. Seasonal movements and habitat selection by woodland caribou in the Wolverine Herd, North-Central B.C. Phase 2: 1994–1997. Peace/Williston Fish and Wildlife Compensation Program Report Number 204. Prince George, British Columbia, Canada.
Treves, A., L. Naughton-Treves, E. K. Harper, D. J. Mladenoff, R. A. Rose, T. A. Sickley, and A. P. Wydeven. 2004. Predicting human–carnivore conflict: a spatial model derived from 25 years of data on wolf predation on livestock. Conservation Biology 18:114–125.

Appendix

The following derivation of the use–availability likelihood and subsequent justification for use of logistic regression closely follows Seber's (1984:308–315) derivation of logistic discriminate function for the case of separate sampling. Seber (1984) in turn credits Anderson and Blair (1982) with a key part of the method.

We envision a finite population of N_a available resource units existing prior to animals making any selections. Each resource unit in this population has k measurable covariates associated with it that do not change through time. We denote these covariates as k-dimensional vectors $\boldsymbol{x}_i$ $(i = 1, \ldots, N_a)$. The frequencies of distinct $\boldsymbol{x}_i$ among all N_a available resource units defines a k-dimensional multivariate discrete probability distribution, which we denote $f_a(\boldsymbol{x})$. Let the set of distinct $\boldsymbol{x}_i$ in the available population be denoted $D_a(\boldsymbol{x})$, and we note that $f_a(\boldsymbol{x})$ is a $|D_a(\boldsymbol{x})|$–celled multinomial distribution. From this population, we take a random sample S_a of n_a units without replacement and observe $\boldsymbol{x}_i$ for all units in the sample. Following sampling, all selected units are replaced in the population and made available for use by the organism. We note that because use has not occurred yet, there is no concept of the available sample consisting of used and unused units. All units in the sample were simply available for use at one point in time.

We now envision animals selecting units in the available population during a fixed and finite time period, say T. After time period T, a second population of units exists that contains all units selected at least once during T. This is the population of used units and it contains $N_u \leq N_a$ units. The frequencies of distinct $\boldsymbol{x}_i$ among all N_u used resource units defines a k-dimensional multivariate discrete probability distribution, which we denote $f_u(\boldsymbol{x})$. From this population, we take a random sample S_u of n_u units without replacement and observe $\boldsymbol{x}_i$ for all units in the sample.

The RSF, denoted $w(\boldsymbol{x})$, is defined as,

$$w(\mathbf{x}) = \frac{f_u(\mathbf{x})}{f_a(\mathbf{x})},$$

or,

$$f_u(\mathbf{x}_i) = \frac{w(\mathbf{x}_i) f_a(\mathbf{x}_i)}{\sum_{\mathbf{x}_j \in D_a(\mathbf{x})} w(\mathbf{x}_j) f_a(\mathbf{x}_j)}$$

where the sum in the denominator assures that $f_u(\boldsymbol{x}_i)$ is a valid probability distribution. It is clear from the latter equation that $w(\boldsymbol{x})$ is the function that transforms the multivariate probability distribution of $\boldsymbol{x}$ among units in the available population into the multivariate probability distribution of $\boldsymbol{x}$ among units in the used population. A related function is the RSPF, denoted $w^*(\boldsymbol{x})$, defined as the actual unequal sampling probability that transforms $f_a(\boldsymbol{x}_i)$ into $f_u(\boldsymbol{x}_i)$. The magnitude of the RSPF depends upon characteristics of the actual selection that are likely to be unknown. For example, the RSPF depends upon whether the study design was set up to allow organisms a fixed or random number of choices. Despite the (at times) unknown nature of the RSPF, it must be proportional to the RSF as defined above, and estimation of the RSF will be sufficient in many applications. In the remainder, our objective is to estimate the RSF and we denote its dependence on $\boldsymbol{\beta}$ by writing $w(\boldsymbol{x}_i, \boldsymbol{\beta})$. At this point, we have not specified the form of w nor the relationship between $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$. For instance, w could be a complex non-linear function of $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ involving scatter plot smoothers or splines.

We now define random variables y_i to be 1 if the i-th unit in the composite sample of length $n_u + n_a$ appears in the sample of used units, and 0 otherwise. Note that if the same unit appears in both S_a and S_u, $y_i = 0$ for a particular $i \in S_a$, and $y_i = 1$ for a different $i \in S_u$. Here, we assume that either selection of S_a and S_u was with replacement, or that N_a and N_u are large enough that Pr(unit A and unit B both selected) is well approximated by Pr(unit A selected) Pr(unit B selected). The latter assumption on population sizes is parallel to assuming that the binomial distribution approximates the hypergeometric. Because S_a and S_u were drawn independently, the likelihood of observing the composite sample is,

$$\begin{aligned} L(\beta) &= \left[\prod_{i=1}^{n_a} f_a(\mathbf{x}_i)\right]\left[\prod_{i=1}^{n_u} f_u(\mathbf{x}_i)\right] \\ &= \prod_{i=1}^{n_a+n_u} f_u(\mathbf{x}_i)^{y_i} f_a(\mathbf{x}_i)^{1-y_i} \\ &= \prod_{i=1}^{n_a+n_u} C_\beta^{-y_i}\, w(\mathbf{x}_i, \beta)^{y_i} f_a(\mathbf{x}_i)^{y_i} f_a(\mathbf{x}_i)^{1-y_i} \\ &= C_\beta \prod_{i=1}^{n_a+n_u} w(\mathbf{x}_i, \beta)^{y_i} f_a(\mathbf{x}_i) \end{aligned}$$

where $C_\beta = \sum_{x_j \in D_a(\mathbf{x})} w(\mathbf{x}_j, \beta) f_a(\mathbf{x}_j)$. Note that we do not observe all distinct $\boldsymbol{x_i}$ and cannot compute C_β. We therefore condition on the composite sample and maximize $L(\boldsymbol{\beta})$ subject to the constraints

$$\sum_{\mathbf{x}_i \in D(\mathbf{x})} f_a(\mathbf{x}_i) = 1 \qquad (A1)$$

and

$$\sum_{\mathbf{x}_i \in D(\mathbf{x})} f_u(\mathbf{x}_i) = \sum_{\mathbf{x}_i \in D(\mathbf{x})} w(\mathbf{x}_i, \beta) f_a(\mathbf{x}_i) = 1 \qquad (A2)$$

where $D(\boldsymbol{x})$ is the set of distinct values of $\boldsymbol{x}$ in the composite sample. These constraints assure that f_u and f_a are probability functions on the composite sample, and we drop C_β from the likelihood because the second constraint assures $C_\beta = 1$ on $D(\boldsymbol{x})$.

To maximize $L(\boldsymbol{\beta})$ subject to constraints (A1) and (A2), we employ Lagrange multipliers. At this point, we must assume the

RSF has the exponential form $w(\boldsymbol{x},\boldsymbol{\beta}) = \exp(\beta_0 + \boldsymbol{x}'\boldsymbol{\beta})$. We also let $n_u(\boldsymbol{x}_i)$ be the number of units in the used sample with covariate vector equal to $\boldsymbol{x}_i$, $n_a(\boldsymbol{x}_i)$ be the number of units in the available sample with covariate vector equal to $\boldsymbol{x}_i$, $n(\boldsymbol{x}_i) = n_u(\boldsymbol{x}_i) + n_a(\boldsymbol{x}_i)$ be the number of units in the composite sample with covariate vector equal to $\boldsymbol{x}_i$. We then write $\log(L(\beta))$ as

$$\log(L(\beta)) = \sum_{\mathbf{x}_i \in D(\mathbf{x})} n_u(\mathbf{x}_i)[\beta_0 + \mathbf{x}'_i\beta] + \sum_{\mathbf{x}_i \in D(\mathbf{x})} n(\mathbf{x}_i)\log(f_a(\mathbf{x}_i))$$

and differentiate

$$\log(L(\beta)) - \lambda_1\left(\left[\sum_{\mathbf{x}_i \in D(\mathbf{x})} f_a(\mathbf{x}_i)\right] - 1\right) - \lambda_2\left(\left[\sum_{\mathbf{x}_i \in D(\mathbf{x})} \exp(\beta_0 + \mathbf{x}_i\beta) f_a(\mathbf{x}_i)\right] - 1\right) = 0 \qquad (A3)$$

with respect to $f_a(\boldsymbol{x}_i)$ for all i to obtain $d = |D(\boldsymbol{x})|$ equations of the form,

$$\frac{n(x_i)}{f_a(\mathbf{x}_i)} - \lambda_1 - \lambda_2\exp(\beta_0 + \mathbf{x}'_i\beta) = 0. \qquad (A4)$$

Multiplying through by $f_a(\boldsymbol{x}_i)$ and summing the d equations gives

$$\sum_{\mathbf{x}_i \in D(\mathbf{x})} n(\mathbf{x}_i) - \lambda_1 \sum_{\mathbf{x}_i \in D(\mathbf{x})} f_a(\mathbf{x}_i) - \lambda_2 \sum_{\mathbf{x}_i \in D(\mathbf{x})} \exp(\beta_0 + \mathbf{x}'_i\beta)\, \mathrm{f}_a(\mathbf{x}_i) = 0$$

which implies

$$n - \lambda_1 - \lambda_2 = 0$$

by the constraints. Differentiating (A3) with respect to β_0 gives,

$$\sum_{\mathbf{x}_i \in D(\mathbf{x})} \mathrm{n}_u(\mathbf{x}_i) - \lambda_2 \sum_{\mathbf{x}_i \in D(\mathbf{x})} \exp(\beta_0 + \mathbf{x}'_i\beta) f_a(\mathbf{x}_i) = 0.$$

Solving for λ_2 in this equation yields $\lambda_2 = n_u$ by constraint (A2), which in turn implies $\lambda_1 = n_a$ in the previous equation. Substituting λ_1 and λ_2 into the equations (A4) we obtain

$$\hat{f}_a(\mathbf{x}_i) = \frac{n(\mathbf{x}_i)}{n_u \exp(\beta_0 + \mathbf{x}'\beta) + n_a}.$$

Substituting $\hat{f}_a(\mathbf{x}_i)$ into $L(\boldsymbol{\beta})$ we obtain,

$$\begin{aligned} L^*(\beta) &= \prod_{i=1}^{n_a+n_u} n(\mathbf{x}_i) \frac{[\exp(\beta_0 + \mathbf{x}'_i\beta)]^{y_i}}{n_u\exp(\beta_0 + \mathbf{x}'_i\beta) + n_a} \\ &= \prod_{i=1}^{n_a+n_u} n(\mathbf{x}_i) \left[\frac{\exp(\beta_0 + \mathbf{x}'_i\beta)}{n_u\exp(\beta_0 + \mathbf{x}'_i\beta) + n_a}\right]^{y_i} \\ &\quad \times \left[\frac{1}{n_u\exp(\beta_0 + \mathbf{x}'_i\beta) + n_a}\right]^{1-y_i} \\ &== \prod_{i=1}^{n_a+n_u} \frac{n(\mathbf{x}_i)}{n_u^{y_i}} \left[\frac{\exp(\beta_0 + \ln(n_u/n_a) + \mathbf{x}'_i\beta)}{\exp(\beta_0 + \ln(n_u/n_a) + \mathbf{x}'_i\beta) + 1}\right]^{y_i} \\ &\quad \times \left[\frac{1}{\exp(\beta_0 + \ln(n_u/n_a) + \mathbf{x}'_i\beta) + 1}\right]^{1-y_i}. \end{aligned}$$

This is $L(\boldsymbol{\beta})$ assuming $w(\boldsymbol{x},\boldsymbol{\beta}) = \exp(\beta_0 + \boldsymbol{x}'\boldsymbol{\beta})$ subject to constraints (A1) and (A2), and is equivalent to a logistic regression likelihood in which the intercept is replaced by $\beta_0 + \ln(n_u/n_a)$. $L^*(\boldsymbol{\beta})$ (or $\log[L^*\{\boldsymbol{\beta}\}]$) can be maximized with respect to $\boldsymbol{\beta}$ by use of a logistic regression routine. In other words, maximum likelihood estimates of $\boldsymbol{\beta}$ can be obtained by estimating a logistic regression model that contains all x variables and an intercept, and recognizing that the estimated intercept is immaterial to the RSF. The estimated RSF is,

$$w(\mathbf{x}, \beta) = \exp(\mathbf{x}'\hat{\beta}) = \exp(x_1\hat{\beta}_1 + \cdots + x_k\hat{\beta}_k)$$

where $\hat{\beta}_1, \ldots, \hat{\beta}_k$ come from the logistic regression routine.

Despite the fact that we know n_u and n_a and could estimate β_0, the RSPF cannot be estimated in this case unless additional assumptions about selection are made. For example, if the total number of used units in the population can be assumed to be small, the methods of Manly et al. (2002) can be applied to estimate the RSPF in some cases.

Associate Editors: Strickland and McDonald.