*Ecology*, 90(12), 2009, pp. 3554–3565


# Accounting for animal movement in estimation of resource selection functions: sampling and data analysis

James D. Forester,[1,3] Hae Kyung Im,[1] and Paul J. Rathouz[1,2]

[1]*Center for Integrating Statistics and Environmental Science, University of Chicago, 5734 South Ellis Avenue, Chicago, Illinois 60637 USA*
[2]*Department of Health Studies, University of Chicago, 5841 S. Maryland Avenue, MC 2007, Chicago, Illinois 60637 USA*

*Abstract*. Patterns of resource selection by animal populations emerge as a result of the behavior of many individuals. Statistical models that describe these population-level patterns of habitat use can miss important interactions between individual animals and characteristics of their local environment; however, identifying these interactions is difficult. One approach to this problem is to incorporate models of individual movement into resource selection models. To do this, we propose a model for step selection functions (SSF) that is composed of a resource-independent movement kernel and a resource selection function (RSF). We show that standard case-control logistic regression may be used to fit the SSF; however, the sampling scheme used to generate control points (i.e., the definition of availability) must be accommodated. We used three sampling schemes to analyze simulated movement data and found that ignoring sampling and the resource-independent movement kernel yielded biased estimates of selection. The level of bias depended on the method used to generate control locations, the strength of selection, and the spatial scale of the resource map. Using empirical or parametric methods to sample control locations produced biased estimates under stronger selection; however, we show that the addition of a distance function to the analysis substantially reduced that bias. Assuming a uniform availability within a fixed buffer yielded strongly biased selection estimates that could be corrected by including the distance function but remained inefficient relative to the empirical and parametric sampling methods. As a case study, we used location data collected from elk in Yellowstone National Park, USA, to show that selection and bias may be temporally variable. Because under constant selection the amount of bias depends on the scale at which a resource is distributed in the landscape, we suggest that distance always be included as a covariate in SSF analyses. This approach to modeling resource selection is easily implemented using common statistical tools and promises to provide deeper insight into the movement ecology of animals.

*Key words: animal movement; case-control; conditional logistic regression; elk; relocation kernel; resource selection function; telemetry; Yellowstone.*

## Introduction

Resource use by animal populations is the product of patterns of movement and resource selection by many individuals. A variety of statistical methodologies have been used to gain insight into such resource use (Millspaugh and Marzluff 2001, Moorcroft et al. 2006, Strickland and McDonald 2006, Thomas and Taylor 2006). One approach popular in wildlife ecology compares characteristics of locations used by animals to locations deemed available to but unused by those same animals (Manly et al. 2002, Thomas and Taylor 2006). This is often accomplished by applying a form of logistic regression to fit a resource selection function (RSF) to "use-available" or "case-control" data composed of "used" or "case" locations where animals are observed and a set of "available" or "control" locations sampled uniformly from the domain of availability for those animals (Manly et al. 2002). The domain of availability is usually defined to include the entire area occupied by a population or individual under study (e.g., Manly et al. 2002, Boyce et al. 2003). The RSF is a function on that domain that is proportional to the density of locations used by that population or individual given the resources present at all locations in the domain. A key feature of RSF models therefore lies in the notion of "availability" (Manly et al. 2002, Boyce et al. 2003), which can be more challenging to define than it might first appear, especially when analysis focuses on individual animals.

The most common applications of RSF analyses examine habitat use at either the population level, where all observed individuals are assumed to have the same domain of availability, or at the home range level where each individual is assumed to have its own domain of

Manuscript received 8 May 2008; revised 9 December 2008; accepted 18 February 2009. Corresponding Editor: M. Fortin.
[3] Present address: Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, Massachusetts 02138 USA.
E-mail: jdforester@gmail.com

availability that is fixed over time. Population-level analyses provide insight into what broad-scale factors the animals select for, but can miss important interactions between individual animals and characteristics of their particular home ranges (Boyce et al. 2003). Home-range analyses account for the fact that individual animals in a population often exist in very different habitats; however, in typical home range analyses, an animal is assumed to be capable of moving between any two locations within its home range in the time interval between telemetry fixes. This assumption may be reasonable if the time intervals between telemetry fixes are very long; however, global positioning system (GPS) radio collars are now providing extremely fine-scale temporal telemetry data with individuals' locations recorded at intervals on the order of several hours or less. The assumption that an entire home range is available to a given individual in the next time interval, given the animal's current location, is therefore violated for many species; better approaches allow the domain of availability for each individual to be a subset of the home range that varies over time as a function of current location.

To account for availability that varies both among individuals and over time, Arthur et al. (1996) defined a circle around each animal location at each time and considered all resources to be uniformly available on the circumscribed disc during the next time interval. Cooper and Millspaugh (1999) suggested (in the context of a discrete-choice model) that all resources on the disc are not equally available and included distance as a covariate in their RSF analysis (see also Manly et al. 2002: chapter 8). Taking a different approach, Hjermann (2000) introduced a continuous availability function that allowed availability to decrease with distance from the animal; Rhodes et al. (2005) modified this procedure to account for home range behavior. By allowing availability to change along an animal's path, and including features such as an exponential movement kernel (e.g., Rhodes et al. 2005), these methods begin to approximate mechanistic features of the movement process (Moorcroft and Barnett 2008). Whereas mechanistic models of habitat use and movement are a major goal in ecology (Moorcroft et al. 2006, Moorcroft and Barnett 2008), a challenge with the methods described above is that the user must define a radius of availability in order to either describe a choice set (Cooper and Millspaugh 1999) or conduct a full likelihood analysis based on the entire domain of availability (Rhodes et al. 2005).

In an analysis of wolf influences on individual elk movements, Fortin et al. (2005) introduced the notion of a step selection function (SSF), which extended the idea of an RSF to account for distance and turning angle of animal movement. They used conditional logistic regression (CLR) with a matched case-control (use-available) design in which they sampled control locations for each animal step at each time point from the empirical distribution of step lengths (the distance between two locations) and turning angles (the angular deviation from the bearing of the previous step) relative to the starting location for that interval. Because the relocation distribution for a given step depends strongly on both the local distribution of resources and strength of selection, and because the empirical step distribution from which control samples are drawn is marginal over landscapes, it is unclear if this method produces unbiased estimates of resource selection coefficients. If the bias is small, however, it has great advantage over full likelihood methods because the sampling method is easy to program and resource selection coefficients are estimated very quickly using easily accessible statistical software.

In this article, we provide a formal statistical framework for studying these methods. In *Model and methods of estimation*, we formulate a SSF model for animal movement as a function of two components: (1) a resource-independent movement kernel that describes, for a fixed time interval, the probability of an animal moving from one location to any other location in the absence of selection (i.e., movement on a flat, homogeneous landscape), and (2) a resource selection function that describes the relative probability of occurrence at any given location without taking into account distance. We then embed this model in a sampling plan for matched use-available designs for individual animal telemetry data. The resulting model provides some theoretical justification for the SSF method (Fortin et al. 2005) and demonstrates the relationship to the Rhodes et al. (2005) approach. It also directly suggests (1) a new class of parametric case-control sampling plans, and (2) a simple approach to accounting for the sampling plan in standard CLR that involves including functions of movement distance in the CLR model. We show that in analyses that ignore distance, a parametric assumption about the resource-independent movement kernel is required in order to choose a sampling plan that will yield unbiased estimates; however, we provide a new analytic approach that requires weaker assumptions. In *Simulation study*, we show empirically via simulations that our approach is very robust to such assumptions. We simulate movement data using a variety of landscapes, selection strengths and resource-independent movement kernels. We then compare the ability of three plans for sampling control points and three methods of data analysis to produce unbiased estimates of the RSF coefficients. Finally, we apply these methods to real elk movement data and discuss the implications of coefficient bias for the application and interpretation of SSF models.

## Model and Methods of Estimation

### *Model formulation*

Following the framework of Rhodes et al. (2005), we define the density of location $b \in D_a$ to which an

individual animal moves from location $a$ in a given time interval, conditional on the individual starting the interval at location $a$ and on covariates $Z(b)$ found in the landscape at location $b$, to be

$$f(b \mid a, a_0, Z) = \frac{\phi(a_0, a, b; \theta)\omega\{Z(b); \beta\}}{\int_{c \in D_a} \phi(a_0, a, c; \theta)\omega\{Z(c); \beta\}\, dc}. \quad (1)$$

The numerator of Eq. 1 is the step selection function, $D_a$ is the domain of availability to the individual during the given time interval, and $a_0$ is the location at the beginning of the previous interval (required to determine the turning angle of a step). The first term in the step selection function, $\phi(a_0, a, b; \theta)$, is a two-dimensional resource-independent movement kernel that describes how the animal would move in a constant $Z(b)$ field (i.e., a homogeneous landscape), where $\theta$ is a vector of parameters governing the density $\phi(\cdot)$. The factor $\omega\{Z(b); \beta\}$ is the resource selection function for landscape $Z(b)$ and is therefore proportional to the location density if all locations in $D_a$ are equally available, i.e., if $\phi(\cdot)$ is uniform on $D_a$. We generally specify $\omega(\cdot)$ as a log-linear function of $\beta$:

$$\omega\{Z(b); \beta\} = \exp\{Z(b)\beta\}. \quad (2)$$

The resource-independent movement kernel $\phi(\cdot)$ may take parametric or nonparametric forms depending on the application and available data. A good parametric starting point is to specify $\phi(\cdot)$ as

$$\phi(a_0, a, b; \theta) = \frac{\nu\lambda(\lambda r_{ab})^{\nu-1}\exp[-(\lambda r_{ab})^{\nu}]}{2\pi r_{ab}} \qquad \theta = (\nu, \lambda). \quad (3)$$

In Eq. 3, $\phi(\cdot)$ describes the probability of moving distance $r_{ab}$ under a Weibull distribution with rate-of-travel parameter $\lambda^{-1}$ and a uniform turning angle distribution independent of distance. The Weibull distribution, which may have a mode greater than zero, reduces to an exponential distribution when shape parameter $\nu = 1$. We will use the assumption of uniform turning angles (Eq. 3) for the development of this paper and will thus omit formal conditioning on $a_0$; however, extensions that allow for nonuniform turning angles follow a similar theoretical development (Appendix A).

Full likelihood methods for estimation of $\beta$ and $\theta$ require evaluation of the integral in the denominator of Eq. 1 over all points in domain $D_a$ (Fig. 1). This integral may be intractable if $D_a$ is large (Rhodes et al. 2005), and in any case, requires numerical integration methods because there is generally no functional form for landscape $Z(b)$. Numerical integration in this context poses several problems: The methods are computationally expensive and may be difficult to implement if many large spatial databases must be accessed for all locations in the domain. If the landscape is very heterogeneous, numerical integration could be unreliable, as most methods are optimized for functions that are in some sense smooth. Additionally, $D_a$ may be infinite and therefore in practice requires an arbitrary cutoff to compute the integral. Finally, these methods often require custom programs that are difficult for nonspecialists to develop. An alternative to full likelihood analysis is to proceed by randomly sampling a small set of control points.

### *Case-control sampling*

Suppose instead of employing the likelihood in Eq. (1) for each observed (case) location $b$, we randomly sample $K$ control locations from the landscape around location $a$ using a resource-independent sampling function $\phi^*(a, b)$ on some $\tilde{D}_a$, where both $\phi^*(\cdot)$ and $\tilde{D}_a$ are defined by the user. We discuss three methods for generating such control points: uniform, empirical, and parametric sampling.

For uniform sampling (e.g., Boyce et al. 2003, Craiu et al. 2008), $\tilde{D}_a$ is defined as a disc with radius $d$ centered at $a$ and control locations are generated uniformly over $\tilde{D}_a$. Radius $d$ is often set to a summary statistic calculated from the empirical distribution of animal movement (e.g., a distance including 80% of observed step lengths). An alternative to uniform sampling that does not require the arbitrary selection of $d$ is empirical sampling (Fortin et al. 2005). In this approach, pairs of step lengths and turning angles are sampled (with replacement) from the empirical distribution of animal movement. Note that this distribution does not represent the true $\phi(\cdot)$ because it is marginal over $Z(b)$; the implications of this are explored below. A third method we have not yet seen in the literature is a parametric sample of the landscape generated from a known distribution $\phi^*(a, b; d)$ with support $\tilde{D}_a$ and governed by parameter $d$. For this method the researcher has absolute freedom to choose the functional form of $\phi^*(\cdot)$ as the specific form does not strictly matter; however, as we shall see, in practice it is helpful to use a distribution from the exponential family. As with uniform sampling, $\tilde{D}_a$ and $d$ may be chosen by the user or based on the data. E.g., $\tilde{D}_a$ may be the entire plane, and $\phi^*(a, b; d)$ be the exponential distribution given by Eq. 3, with $\nu = 1$ and $\lambda^{-1}$ equal to twice the sample mean of observed step lengths. As we will see, in all three approaches, it is preferable to chose $\tilde{D}_a$ to contain the unknown $D_a$ rather than the other way around.

We now develop the conditional likelihood that accounts for the control-point sampling design. For a given starting location $a$ and time interval, this design yields an unordered set $s = \{l_0, l_1, \ldots, l_K\}$, of $K$ locations sampled from the landscape. Per force, this set contains $K$ control locations and the case location $b$. The $K$-dimensional density of set $s$ given landscape $Z$, starting location $a$, and case location $b$ is given by

$$g(s \mid b, a, Z) = \frac{1}{\phi^*(a, b; d)}(K-1)! \prod_{l \in s} \phi^*(a, l; d). \quad (4)$$

Therefore, the conditional probability of observing case location $b$ given the observed set $s$, starting location $a$, and landscape $Z$ is

$$\begin{aligned}\Pr(\text{case} = b \mid s, a, Z) &= \frac{f(b \mid a, Z)g(s \mid b, a, Z)}{\sum_{l \in S} f(l \mid a, Z)g(s \mid l, a, Z)} \\ &= \frac{f(b \mid a, Z)/\phi^*(a, b; d)}{\sum_{l \in S} f(l \mid a, Z)/\phi^*(a, l; d)} \\ &= \frac{\phi(a, b; \theta)\exp\{Z(b)'\beta\}/\phi^*(a, b; d)}{\sum_{l \in S} \phi(a, l; \theta)\exp\{Z(l)'\beta\}/\phi^*(a, l; d)}.\end{aligned}$$

In the foregoing, the second equality is due to the fact that the product in Eq. 4 is the same in the numerator and all terms in the denominator. The third equality is due to the fact that the denominator of Eq. 1 is the same in the numerator and all terms in the denominator. Simplifying and rearranging this equation yields the conditional probability:

$$\Pr(\text{case} = b \mid s, a, Z) = \frac{\exp\left[Z(b)'\beta + \ln\{\phi(a, b; \theta)/\phi^*(a, b; d)\}\right]}{\sum_{l \in s} \exp\left[Z(l)'\beta + \ln\{\phi(a, l; \theta)/\phi^*(a, l; d)\}\right]} \quad (5)$$

which accounts for the case-control sampling design.

*Implications of case-control sampling and new models*

The development in *Case-control sampling* and the conditional probability (Eq. 5) have important implications for analysis of animal location data arising via case-control sampling, and suggest new estimators for resource selection parameter $\beta$ and resource independent movement parameter $\theta$. First, Eq. 5 shows that if the method for sampling control locations is equivalent to sampling from the resource-independent movement kernel, then the sampling ratio $\phi(a, b; \theta)/\phi^*(a, b; d)$ will be one and will cancel. In this case, Eq. 5 becomes

$$\Pr(\text{case} = b \mid s, a, Z) = \frac{\exp\{Z(b)'\beta\}}{\sum_{l \in s} \exp\{Z(l)'\beta\}} \quad (6)$$

which is the standard form of the conditional logistic regression (CLR) likelihood for one matched set in 1:$K$ matched case-control sampling (Hosmer and Lemeshow 2000: Chapter 7). The implication is that, for $T_i$ time intervals on the $i$th individual in a sample of $i = 1, \ldots, n$ individuals, $\beta$ can be consistently estimated by maximizing the CLR likelihood:

$$L(\beta) = \prod_{i=1}^{n} \prod_{t=1}^{T_i} \frac{\exp\{Z(b_{it})'\beta\}}{\sum_{l \in s_{it}} \exp\{Z(l)'\beta\}} \quad (7)$$

where $b_{it}$ and $s_{it}$ are the observed location and set of sampled locations for individual $i$ at time $t$. The corollary is that, when $\phi(a, b; \theta)/\phi^*(a, b; d)$ depends on distance, standard CLR may produce biased parameter estimates of $\beta$ because it does not take account of the sampling plan and its impact on the likelihood through the sampling ratio $\phi(a, b; \theta)/\phi^*(a, b; d)$.

Second, since $\phi^*(\cdot)$ is known to the user, it is possible to consistently estimate both $\beta$ and $\theta$ through inclusion of the sampling ratio $\phi(a, b; \theta)/\phi^*(a, b; d)$ in an extended CLR likelihood, i.e.,

$$L(\beta) = \prod_{i=1}^{n} \prod_{t=1}^{T_i} \{\exp[Z(b_{it})'\beta + \ln\{\phi(a, b_{it}; \theta)/\phi^*(a, b_{it}; d)\}]\} \div \left\{\sum_{l \in s_{it}} \exp[Z(l)'\beta + \ln\{\phi(a, l; \theta)/\phi^*(a, l; d)\}]\right\}. \quad (8)$$

A potential difficulty of this approach, however, is that standard CLR software may no longer be used unless the sampling ratio is of a special form. For example, if both $\phi(\cdot)$ and $\phi^*(\cdot)$ are exponential distributions with mean $1/\theta$ and $1/d$, respectively, then $\ln\{\phi(a, l; \theta)/\phi^*(a, l; d)\} = (d - \theta)r_{al} + \ln(\theta/d)$. In this case, the likelihood (Eq. 8) is of standard CLR form wherein distance $r_{al}$ between $a$ and sampled points $l \in s$ is included as a covariate with coefficient $\theta^* = (d - \theta)$, and constant term $\ln(\theta/d)$ drops out.

A third implication is that, if $\tilde{D}_a$ is chosen to be a strict subset of the true $D_a$, any analysis appears likely to fail because case locations $b \in D_a$ will arise that are not in $\tilde{D}_a$ and the sampling ratio $\phi(a, b; \theta)/\phi^*(a, b; d)$ will be infinity for those observations. The empirical and parametric (with infinite support $\tilde{D}_a$) sampling schemes discussed in *Case-control sampling* avoid this problem.

The form of conditional likelihood (Eq. 8) additionally suggests that it may be possible to easily account for the discrepancy between $\phi(\cdot)$ and $\phi^*(\cdot)$, at least approximately, and thereby to obtain improved estimators of $\beta$ relative to a traditional CLR that ignores sampling design. We propose to do so by modeling $\ln\{\phi(a, r_{al}; \theta)/\phi^*(a, r_{al}; d)\}$ as a function of $r_{al}$ that is linear in the parameters, thereby permitting use of standard CLR software. The idea is that, although the functional form of unknown $\phi(\cdot)$ may be complicated, the form of $\ln\{\phi(\cdot)/\phi^*(\cdot)\}$ may be well-approximated by a parametric model that makes no assumptions about the functional form of $\phi(\cdot)$. The first such model,

$$\ln\{\phi(a, r_{al}; \theta)/\phi^*(a, r_{al}; d)\} = \theta r_{al} \quad (9)$$

is an exact solution if $\phi(\cdot)$ and $\phi^*(\cdot)$ are both of exponential form. We have found that most CLR analyses of matched use–available data do not include distance as a predictor, although at least two discrete choice analyses have done so (Cooper and Millspaugh 1999, Manly et al. 2002: Chapter 8). The conditional likelihood proposed in

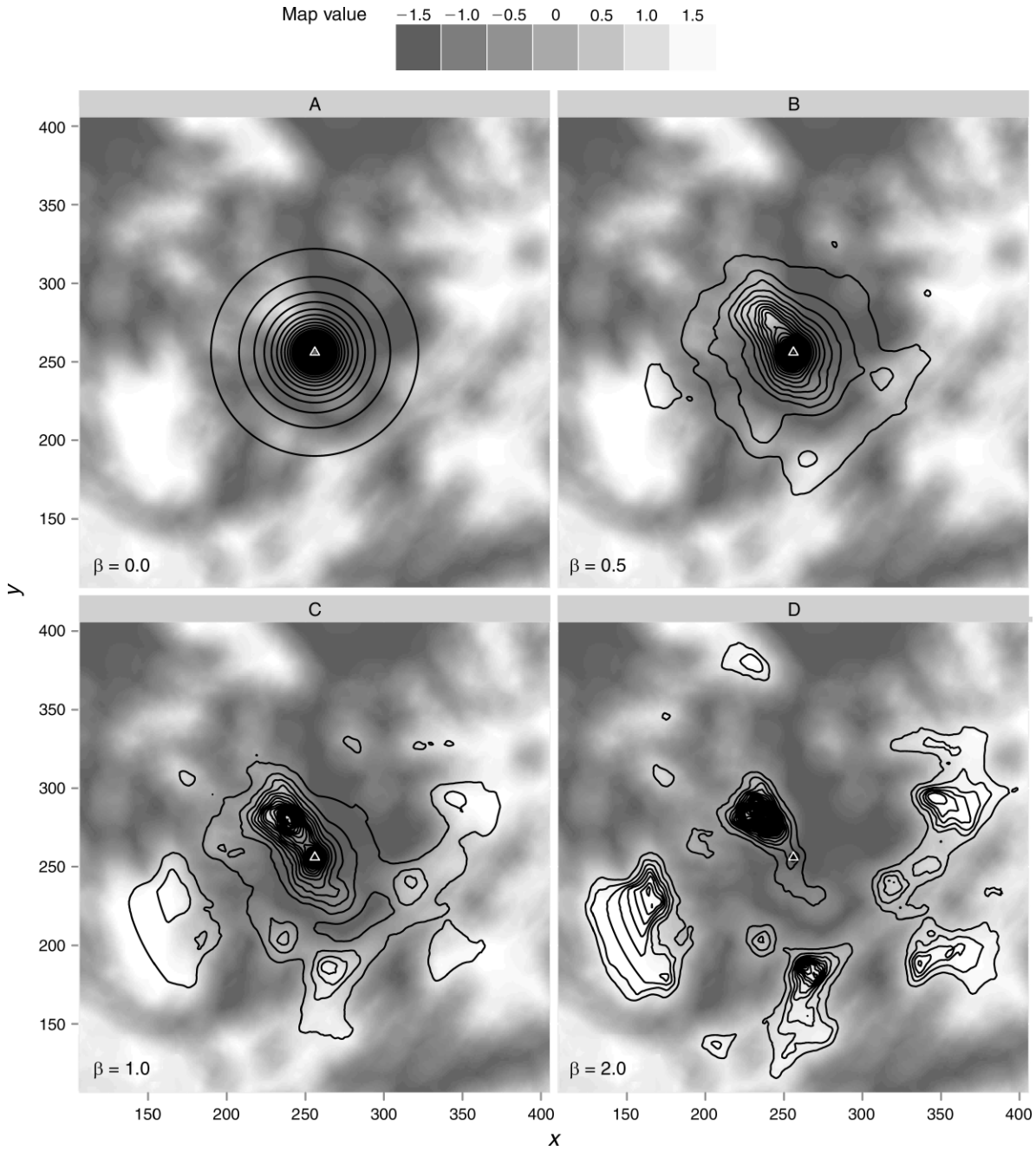


FIG. 1. Each panel shows a resource map and corresponding redistribution kernel for an animal starting at the white triangle in the middle of the map. The kernel depicted in each panel represents the product of the resource-independent movement kernel $\phi(\cdot)$, a Weibull distribution with rate parameter $\lambda = 0.02$ and shape parameter $w = 1.2$, and the resource selection function $\exp(\beta z)$, where $\beta$ is the selection coefficient for the resource $z$ shown on the map. (A) $\beta = 0$, (B) $\beta = 0.5$, (C) $\beta = 1$, (D) $\beta = 2$. The maximum density value was $3.0 \times 10^{-3}$; however, to show detail, kernel values were truncated at $7.6 \times 10^{-4}$. Each contour represents $(7.6 \times 10^{-4})/50$.

Eq. 8 along with the approximate model (Eq. 9) provides some theoretical justification for this approach.

Richer models for $\ln\{\phi(a, r_{al}; \theta)/\phi^*(a, r_{al}; d)\}$ arise by adding squared or other nonlinear functions of $r_{al}$. We propose as an easily implemented solution the use of linear splines of $r_{al}$ (Harrell 2001:18). This involves specifying a model for the sampling ratio as

$$\ln\{\phi(a, r_{al}; \theta)/\phi^*(a, r_{al}; d)\} = \theta_0 r_{al} + \theta_1 (r_{al} - \tau_1)_+ + \cdots + \theta_k (r_{al} - \tau_k)_+ \quad (10)$$

where $u_+ = u$ if $u > 0$ and 0 otherwise, and the $k$ $\tau_j$'s are

user-specified knots. The number and locations of knots may be chosen, for example, as the quantiles from the empirical step-length distribution. The knots can also be chosen based on substantive knowledge of the animals' movements. We explore the use of these two models with a variety of sampling plans $\phi^*(\cdot)$ in the simulation study in the next section.

## Simulation Study

Not accounting for the inequality of $\phi(\cdot)$ and $\phi^*(\cdot)$ may lead to bias in the $\hat{\beta}$ of conditional logistic regression models. To study the consequences of ignoring this aspect of case-control sampling, and to examine the performance of our proposed models (Eqs. 9 and 10) which account for this inequality, we simulated movements on maps with resources distributed at a variety of spatial scales. We then performed case-control sampling of landscapes using several specifications for $\phi^*(\cdot)$. Finally, we analyzed the simulated data using conditional logistic regression with and without accounting for the log ratio of $\phi(\cdot)$ and $\phi^*(\cdot)$. The aim was to measure and compare the severity of bias in estimation of the selection effect $\beta$, and the associated confidence interval coverage probabilities, for multiple estimators under a variety of selection strengths and spatial scales of resource distribution.

### *Data generation*

*Map generation.*—We generated four landscapes $Z$ (each composed of $1024 \times 1024$ square cells) representing a range of resource distribution scales. The resource values were generated from a mean-zero, variance-one Gaussian random field (GRF) based on an exponential covariance function (Stein 1999:section 2.7) with range parameter set to 0.1, 1, 5, or 10 times the mean movement distance ($\mu$) of the resource-independent movement kernel used for the animal movement simulation. A larger range indicates greater correlation between two points at a given distance and hence a coarser landscape (Fig. 2) . The value of $\mu$ was fixed at 21 map units (i.e., 21 times the length of one side of a map cell). Here, each cell can be thought of as representing a 28-m pixel because we have observed that elk in Yellowstone National Park show a similar rate of movement relative to the grain-size of available habitat maps (Forester et al. 2007).

*Movement models.*—Once the four maps were generated, we simulated animal movements based on one of two resource-independent movement kernels and one of four levels of habitat selection. The coefficients for the resource selection functions (Eq. 2) were set to $\beta = 0$, 0.5, 1, or 2. All three resource-independent movement kernels used uniform turning angles (i.e., no directional persistence). Step lengths were distributed exponentially [$\phi_1(\cdot)$] or as a mixture of two Weibull distributions [$\phi_2(\cdot)$]:

$$\phi_1(a,b) = \frac{\lambda_1 \exp(-\lambda_1 r_{ab})}{2\pi r_{ab}}$$

$$\phi_2(a,b) = 0.71\frac{\nu_1\lambda_2(\lambda_2 r_{ab})^{\nu_1-1}\exp[-(\lambda_2 r_{ab})^{\nu_1}]}{2\pi r_{ab}} + 0.29\frac{\nu_2\lambda_3(\lambda_3 r_{ab})^{\nu_2-1}\exp[-(\lambda_3 r_{ab})^{\nu_2}]}{2\pi r_{ab}}.$$

Parameters for the $\phi_j(\cdot)$'s were chosen such that the average distance $r_{ab}$ moved under each kernel is $\mu \approx 21$ map cells. For $\phi_1(\cdot)$, $\lambda_1 = 1/\mu$ while the parameters for $\phi_2(\cdot)$ approximate the marginal distribution of observed elk step lengths collected over 5-h intervals in Yellowstone (Forester et al. 2007), yielding $\lambda_2 = 1/14$, $\lambda_3 = 1/42$, $\nu_1 = 1.22$, $\nu_2 = 1.01$; see *Empirical example*. This parameterization for $\phi_2(\cdot)$ is a fairly strong deviation from the exponential example in that the mode is greater than zero and it has a fatter tail.

To simulate a movement to location $b$, given starting location $a$, we generated a set of $P$ (=2000) proposal locations based on a two-dimensional exponential distribution centered on $a$. The $x$ and $y$ coordinates of each proposal location $p$ were calculated as $p_x = a_x + r_p \cos(u_p)$ and $p_y = a_y + r_p \sin(u_p)$, where $a_x$ and $a_y$ are the coordinates of the current location $a$, $r_p$ is the step length drawn from the proposal density $\phi_p(r_p) = \lambda_p \exp(-\lambda_p d_p)$ with $\lambda_p = 1/45$, and $u_p$ is a bearing drawn from a uniform distribution over the interval $[0, 2\pi)$. Location $b$ was then chosen from the set of $P$ proposal locations based on the location density (Eq. 1) reweighted by the inverse of the proposal density and normalized over all $P$ locations. Specifically, this is calculated as $[\phi(a, p)\omega(Z(p))/\phi_p(r_p)]/\Sigma_{p' \in P}\, \phi(a, p')\omega(Z(p'))/\phi_p(r_{p'})$.

For each of the 32 combinations of simulation parameters, we randomly chose the starting locations for 100 independent and otherwise identical individuals in the middle one-ninth of the map (thus allowing the simulations to proceed with little edge effect). For each individual, we simulated 30 steps based on the above model with reflective map boundaries. This process was repeated 1000 times with different starting locations, so the simulation results are based on 1000 replicates of 3000 steps each.

### *Analysis*

For each replicate, we performed 1:$K$ case-control sampling under three methods of generating control locations (where $K = 20$). The first method was uniform sampling where, for each replicate, control locations were generated uniformly over a disc with radius $d$, $d = 1.2 \times \hat{m}$, and $\hat{m}$ was the maximum of 3000 observed step lengths for that replicate (i.e., $d$ captured 100% of the locations). The second method used empirical sampling, where pairs of step lengths and turning angles were jointly sampled with replacement from the empirical distribution of steps for each replicate. This approach is similar to that proposed by Fortin et al. (2005), except that those authors independently sam-

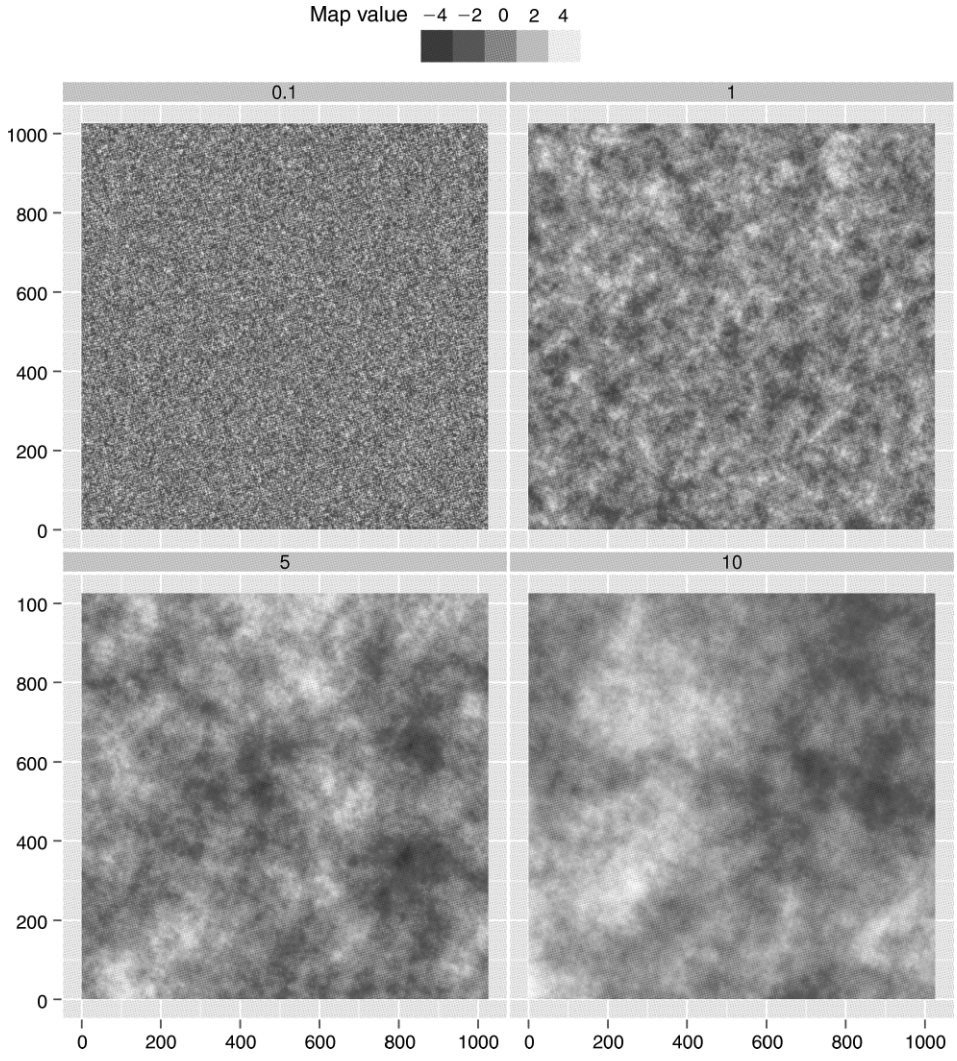


FIG. 2. The four resource maps used for the simulations. The scale of each map (0.1, 1, 5, and 10) is relative to the mean of the animals' resource-independent movement kernel μ (μ = 21 map units). Units are arbitrary.

pled step lengths and turning angles from their marginal empirical distributions (after confirming the lack of linear-angular correlation). The final method used a parametric control-point sample of the landscape generated from an exponential distribution with rate parameter $(2\hat{\mu})^{-1}$, where $\hat{\mu}$ is the sample mean of observed step lengths for each replicate. This choice of sampling mean allowed for the entire area considered "available" by the animal to be sampled without a substantial reduction of efficiency. Turning angles were generated uniformly on [1, $2\pi$), independently of step length.

After generating three sets of control locations for each observed location, we fit CLR models to each matched set using one of three model formulations. The first, which we call the "null" model, includes only the landscape covariate. The second model, referred to as "distance," includes distance ($r_{al}$) to account for the sampling ratio in Eq. 8 as described in Eq. 9. The final model, referred to as "splines," is the same as the second model except that the sampling ratio is modeled as a linear spline function of distance (Eq. 10) with knots at the first, second and third quartiles of observed step lengths for that replicate.

### *Results*

Before drawing broad conclusions comparing the performance of the various methods, we point out several interesting features of these results. First, all methods of sampling and analysis identify the absence of selection when $\beta = 0$, although 95% coverage probabilities are only acceptable under parametric sampling analyzed with the distance or spline models or under empirical sampling with any model. The optimal performance of the empirical sampling when $\beta = 0$ is to be expected because in this case $\phi^*(\cdot) = \phi(\cdot)$ (Fig. 3). Second, uniform sampling produced biased results for all scales of resource distribution and all levels of

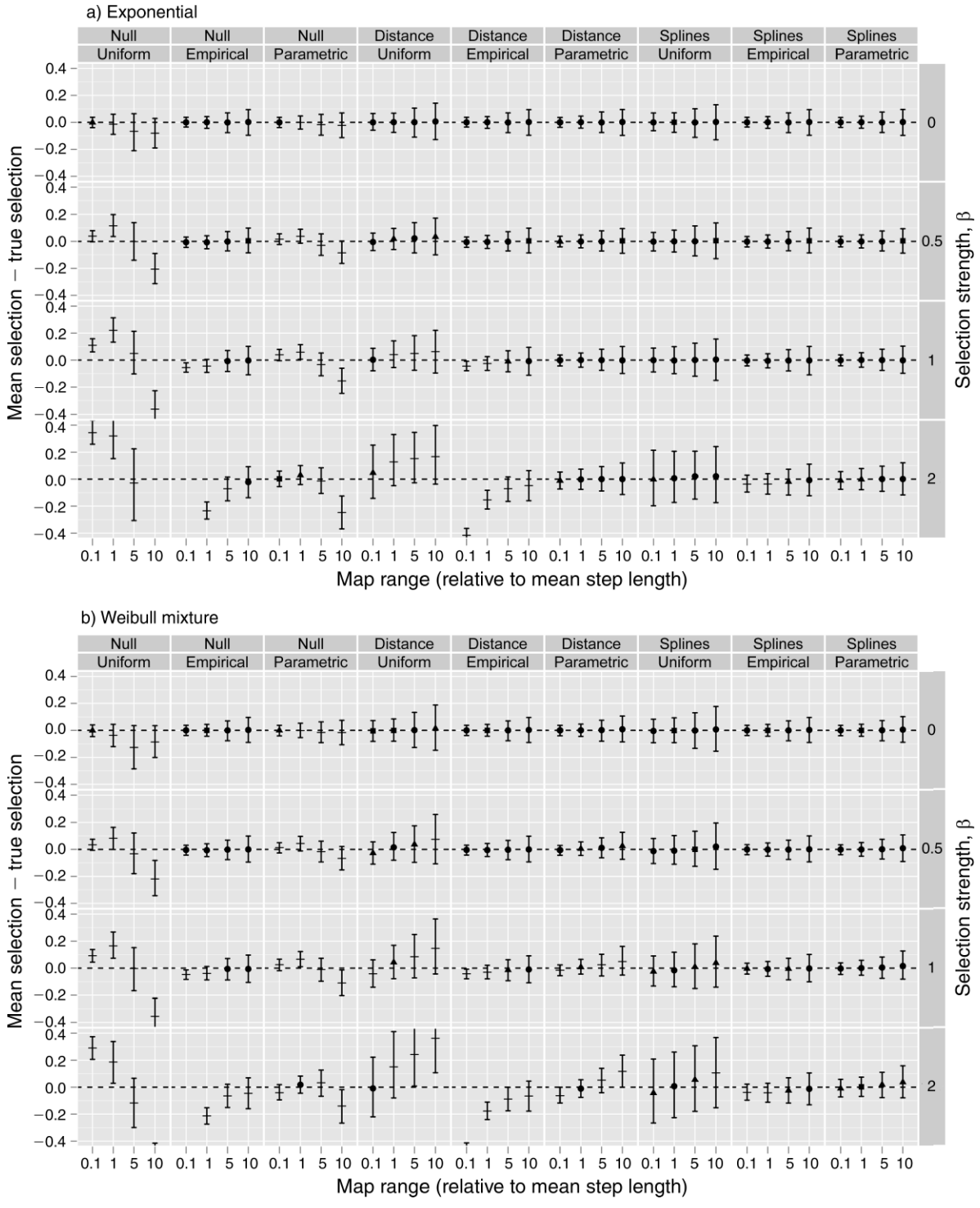


FIG. 3. A set of 1000 movement paths were simulated for a factorial set of parameters that included a gradient in map scale (0.1–10, shown on the *x*-axis) and selection strength ($\beta = 0$, 0.5, 1, 2, displayed on the right edge of each row of plots). The resource-independent movement kernel was either (A) a single exponential distribution or (B) a mixture of two Weibull distributions. Each of these paths was analyzed using conditional logistic regression based on three separate sampling schemes (uniform, empirical, and parametric). Models included habitat only (Null), habitat + distance (Distance), or habitat + 3-knot linear spline of distance (Spline). To show bias, we subtracted the true selection value from each estimated parameter and plotted the mean estimate (error bars represent the 0.025 and 0.975 quantiles of the 1000 $\hat{\beta}$'s). The symbols used for the mean estimates represent the percentage of simulations that produce theoretical confidence intervals that cover the true parameter value (plus sign, $\leq$90%; triangle, $\leq$94%; circle, $\leq$96%; square, >96%).

selection greater than zero. Although adding distance as a covariate ameliorated this problem somewhat and splines largely reduced the bias, the uniform parameter estimates were relatively inefficient (i.e., large error bars) compared to the other methods. Third, for $\beta \leq 1$, with $\phi(\cdot)$ set to the exponential distribution, the distance and splines models under parametric sampling performed very well; this is due to the fact that in these settings, the model for $\ln\{\phi(\cdot)/\phi^*(\cdot)\}$ is exact. Fourth, for $\beta = 2.0$, empirical sampling with null and distance analysis models did not perform well in terms of either bias or coverage probability, especially when the map range was low (i.e., greater variability in landscape per unit distance). A different result was seen under parametric sampling, where the null model performed poorly on the coarsest maps and the distance model was acceptable when the underlying resource-independent movement kernel was exponential. Note that it makes sense that empirical sampling would exhibit its worst performance for the largest value of $\beta$ because this setting creates the biggest difference between $\phi(\cdot)$ and $\phi^*(\cdot)$ (Appendix B).

Examining the results more broadly, empirical sampling produced relatively unbiased estimates when $\beta \leq 1$, and was the best performer under the null analysis model. When $\beta = 0$ (no resource selection), the empirical distribution corresponds exactly to an unbiased sample of the true distribution (the resource-independent movement kernel), so that the resulting estimates are unbiased. As selection levels increased, however, parameter estimates based on empirical sampling were routinely biased low—especially on the fine-grained landscapes (Appendix B). This is because, under strong selection, the empirical distribution is marginal over the variability in the landscape, whereas the true resource-independent movement kernel applies to a hypothetical landscape wherein $Z$ is constant over space. Adding distance to the model did not appreciably reduce the bias when selection was strong, but the addition of linear splines did help in this regard.

Parametric sampling typically produced parameter estimates that were more strongly biased than those fitted using empirical sampling under the null model. When distance was added to the model, however, this bias was largely removed. The spline model also removed most of the parameter bias and was more effective than the distance model when $\phi(\cdot)$ deviated from the exponential distribution (Fig. 3b). In summary, analysis using the null model was only acceptable in special cases, whereas the splines model under either empirical or parametric sampling performed well in all but a few extreme cases.

## Empirical Example

### *Data*

Data for this empirical example were collected in Yellowstone National Park (YNP) from 2001 to 2004. The park is located in northwestern Wyoming and extends into Idaho and Montana (USA). Elevation ranges from 1500 to >3000 m, but much of the park consists of subalpine plateaus covered primarily with lodgepole pine (*Pinus contorta* Loudon var. *latifolia* Engelm.) and, less commonly, other evergreen species and quaking aspen (*Populus tremuloides* Michx.). Sagebrush (*Artemisia* spp.) grasslands are also abundant in some areas of the landscape (Despain 1990). Telonics GPS radio collars (Telonics, Mesa, Arizona, USA), programmed to collect locations at 5-h intervals, were placed on 16 cow elk captured during the study. The collars were collected after one year or following the death of an animal. Details of the animal capture and the study area are described elsewhere (Cook et al. 2004, Forester et al. 2007).

For this example we considered only summer movements (15 June to 15 September). We also included only locations for which the previous and subsequent locations were also of high quality (i.e., three-dimensional location and position dilution of precision (PDOP) $< 8.0$), resulting in 3669 observations (136–270 observations per animal, median = 232). The mean step length was 635 m (SD = 821), while the median, 95th percentile, and maximum were 381 m, 1982 m, and 9759 m, respectively. Three landscape covariates were considered: proportion of regenerating (i.e., post fire) forest in a 350-m buffer around the index location (range = 0–10, mean = 3.9, SD = 3.0), proportion of open cover types (i.e., grassland or shrubland) within the same buffer (range = 0–10, mean = 2.6, SD = 2.7), and estimated herbaceous biomass at the index location (range = 0–333 g/m$^2$, mean = 39.4, SD = 35.7).

### *Analysis*

We used the same analysis methods described in *Simulation study*; however, here we examine two sets of uniform sampling models, one where the sampling radius includes 100% of the observed locations (uniform-100; $d = 1.2 \times 9759$ m) and the other where only 95% of the observed locations are included (uniform-95; $d = 1982$ m). For the parametric sampling models we set $\phi^*(\cdot)$ to an exponential distribution with $\lambda^{-1} = 1270$ m (i.e., twice the observed mean step length). All control points that fell outside of the study area or within lakes were discarded and re-drawn (this could lead to edge effects if many animals are close to the study-area border and areas outside the border are equally available to the animals). To capture diurnal variation in selection for the three landscape covariates, we included interaction terms between the landscape covariates and four harmonics of time of day ($t$ recorded in decimal hours at the endpoint of each step) calculated as $s_1 = \sin(2\pi t/24)$, $s_2 = \sin(4\pi t/24)$, $c_1 = \cos(2\pi t/24)$, and $c_2 = \cos(4\pi t/24)$. We also included interactions between distance and the four time harmonics. This model assumes that beyond the linear distance term ($\theta_0$) in Eq. 10, the higher-order spline terms are not subject to diurnal variability, and thus are not interacted with the harmonic terms. For details on how we

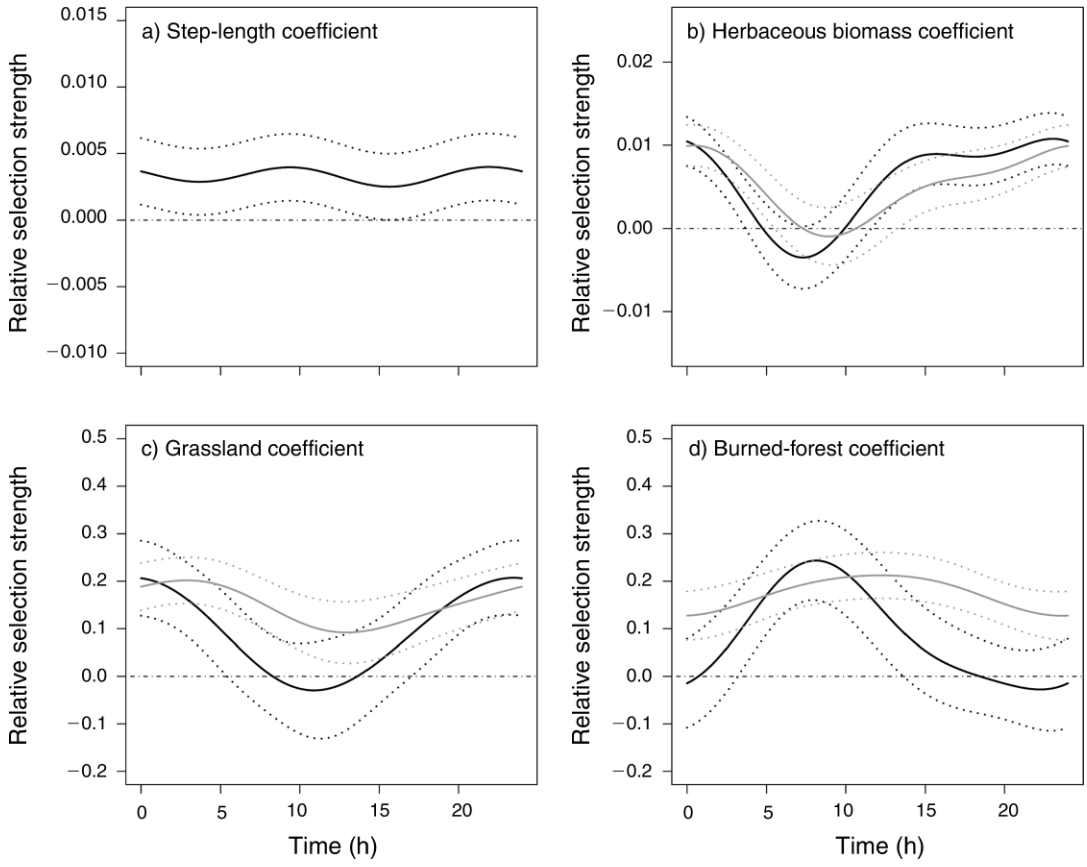


FIG. 4. Selection strengths for step length and each of the three landscape covariates for all hours of the day. Dotted lines are pointwise 95% confidence intervals around the mean estimates (solid lines). Black lines indicate the estimates based on the parametric sampling model (with splines), while gray lines show the estimates for the uniform-100 sampling model (without a distance function) where $d = 120\%$ of the maximum observed step length. The thin horizontal dash-dot lines in each graph mark the point of zero selection. Units for the covariates are: (a) m/5 h (step length), (b) g/m$^2$ (herbaceous biomass), (c) 10× the proportion of grassland in a 350 m radius (grassland), and (d) 10× the proportion of burned forest in a 350 m radius (burned forest).

accounted for autocorrelation in the regression procedure and for the computation of robust standard errors of the selection coefficient estimates, see Appendix C.

### *Results*

Diurnal changes in strength of selection are plotted in Fig. 4 and Appendix D (tables of robust parameter estimates for all models are also included in Appendix D). The models generally showed that selection for herbaceous biomass was nonsignificant between 05:00 hours and 12:00 hours but high during the night (Fig. 4b). Likewise, selection for open areas was low during the day and high at night, with an opposite pattern observed for burned forest (Fig. 4c, d). Including step length in the parametric sampling model was important at all times of the day (Fig. 4a). Further, the interactions between step length and harmonics of time of day accounted for diurnal changes in the average step length (Wald statistic = 50.77, df = 4, $P < 0.0001$).

When splines were included in models based on the three sampling regimes, the overall patterns of selection were similar (see Fig. D1). The parametric and empirical sampling models produced nearly identical estimates of selection. These patterns of selection were similar to those from the two uniform sampling models; however, estimates from the uniform-100 model had much larger standard errors.

There was very little difference among the selection estimates produced from the null, distance and spline versions of the empirical and parametric sampling models. In both cases, the only major difference was that the null model showed different hours of significant selection for the three resource variables (see Fig. D3 and Fig. D4). Including distance or splines in the uniform sampling models produced much more substantial changes in the parameter estimates (see Fig. D2 and Fig. D5).

The largest difference in selection estimates can be seen when comparing the parametric sampling model (with splines) to the null uniform-100 sampling model (see Fig. 4). The uniform sampling model shows different peaks of minimum and maximum selection for all three landscape covariates. Likewise, the overall strengths of selection for grassland and burned forest are

higher in the uniform sampling model (and positive for all hours of the day).

## Discussion

We have demonstrated the importance of accounting for characteristics of animal movement when developing models of resource selection based on telemetry data. The model developed by Rhodes et al. (2005) provides a general framework and we show that this model can be estimated using conditional logistic regression. To do this we wrote a SSF that extends the standard population RSF model to incorporate individual animal movements (distance and, if needed, direction) in the form of a resource-independent movement kernel. Additionally, we unified case-control sampling methods for estimation of this SSF under a single likelihood function that accounts for the sampling plan used to generate the control locations. We also proposed approximate models to account for animal movements and sampling. These methods are easy to implement and improve on existing methods.

Our results demonstrate that the coefficients produced in SSF analyses which ignore sampling may be biased when selection for a given covariate is strong, and this bias depends on the spatial distribution of that resource relative to the animal's average step length. In simulations, the uniform sampling method (based on a sampling radius that was equal to 1.2 times the maximum step length) yielded particularly biased estimates of selection. One possible reason this method performed poorly is that this radius may not have been appropriate to describe the movement process, highlighting the fact that an arbitrary choice of radius may have undesirable effects on the analysis. Empirical (e.g., Fortin et al. 2005) or, as we have proposed, parametric sampling methods avoid this arbitrary choice of the domain of availability, are just as easy to implement, and have overall superior performance in SSF case-control analyses compared to uniform sampling. Under these two sampling methods, bias is likely small when selection for covariates is relatively weak. Furthermore, when the bias is large, it can mostly be removed by including a spline function of distance moved by the animal as part of the regression model. Adding a linear distance term will also help remove bias; however, this works best when the functional form of the sampling distribution is similar to that of the resource-independent movement kernel (see Fig. 3).

The effects of this bias are clearly seen in the empirical example. While all models followed the same qualitative pattern when distance was included as a linear spline function, the uniform sampling models that did not account for distance showed very different levels of selection through the day. One reason that this may be occurring is that cervids are known to increase their movement during the crepuscular period (Green and Bear 1990, Ager et al. 2003). This change of movement rate emerges from changes in the allocation of time to different behaviors (e.g., bedding, foraging, and moving among habitat patches). Although selection levels are expected to change with those behaviors, we must also account for the fact that the changing movement rate alters the landscape available to an animal at any given point in time (i.e., $\phi(\cdot)$ is a function of time). Our solution to this problem was to interact distance moved with four harmonics of time of day. These interactions coupled with a three knot linear spline function allowed us to temporally correct for sampling bias without making any explicit parametric assumptions about the animals' resource-independent movement kernel.

We have identified a source of bias by focusing on sampling and analysis under essentially ideal conditions; however, there are other forms of bias associated with telemetry data that we did not deal with here. Two issues that are known to affect resource selection estimation are contamination and telemetry bias. Contamination occurs when available locations are also used locations (Keating and Cherry 2004) but the proportion of contaminated points must be high before the RSF coefficients are substantially affected (Johnson et al. 2006). The probability of explicit contamination is low when estimating a SSF because there is only one case (the actual location where the animal chose to go) per step. Still, other factors such as memory and the spatial distribution of conspecifics could affect selection and should be accounted for if possible. Telemetry bias (i.e., where the measurement error associated with the location of a data point or the probability of obtaining a data point are systematically affected by landscape covariates) and the timing of relocation intervals may interact with animal behavioral patterns to affect selection estimates (Nams 1989, Frair et al. 2004, Bradshaw et al. 2007). Telemetry bias should be tested for and corrected if large (Frair et al. 2004) and the choice of relocation interval should be based on the natural history of the animal under study and the temporal scale of the associated research questions.

Overall, we find that conditional logistic regression used with an appropriate control point sampling method and easily programmed distance terms is an efficient and easily accessible method for estimating animal resource selection. Further investigation is required to determine the implications of strongly nonuniform, correlated turning angles and selection for multiple covariates distributed at different spatial scales. However, the flexible formulation we describe allows for mechanistic approximations of the movement process and thus greater insight into the ecology and natural history of animals.

## Acknowledgments

This research was funded in part by the Center for Integrating Statistical and Environmental Science, a Center funded through a U.S. EPA STAR cooperative agreement with the University of Chicago (grant no. R-82940201-0). Funding

for the Yellowstone elk research was provided by the Integrated Research Challenges in Environmental Biology Program of the National Science Foundation (grant no. DEB-0078130). J. D. Forester was supported in part by NSF grant OCE 04-52678 to J. T. Wootton. This manuscript was improved based on input from M. Boyce, E. Merrill, D. Fortin, H. Beyer, P. Moorcroft, and three anonymous reviewers.

## Literature Cited

Ager, A. A., B. K. Johnson, J. W. Kern, and J. G. Kie. 2003. Daily and seasonal movements and habitat use by female rocky mountain elk and mule deer. Journal of Mammalogy 84:1076–1088.

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., J. S. Mao, E. H. Merrill, D. Fortin, M. G. Turner, J. Fryxell, and P. Turchin. 2003. Scale and heterogeneity in habitat selection by elk in Yellowstone National Park. Ecoscience 10:421–431.

Bradshaw, C. J. A., D. W. Sims, and G. C. Hays. 2007. Measurement error causes scale-dependent threshold erosion of biological signals in animal movement data. Ecological Applications 17:628–638.

Cook, R. C., J. G. Cook, and L. D. Mech. 2004. Nutritional condition of northern Yellowstone elk. Journal of Mammalogy 85:714–722.

Cooper, A. B., and J. J. Millspaugh. 1999. The application of discrete choice models to wildlife resource selection studies. Ecology 80:566–575.

Craiu, R. V., T. Duchesne, and D. Fortin. 2008. Inference methods for the conditional logistic regression model with longitudinal data. Biometrical Journal 50:97–109.

Despain, D. G. 1990. Yellowstone vegetation: consequences of environment and history in a natural setting. Roberts Rinehart, Boulder, Colorado, USA.

Forester, J. D., A. R. Ives, M. G. Turner, D. P. Anderson, D. Fortin, H. L. Beyer, D. W. Smith, and M. S. Boyce. 2007. State–space models link elk movement patterns to landscape characteristics in Yellowstone National Park. Ecological Monographs 77:285–299.

Fortin, D., H. L. Beyer, M. S. Boyce, D. W. Smith, T. Duchesne, and J. S. Mao. 2005. Wolves influence elk movements: behavior shapes a trophic cascade in Yellowstone National Park. Ecology 86:1320–1330.

Frair, J. L., S. E. Nielsen, E. H. Merrill, S. R. Lele, M. S. Boyce, R. H. M. Munro, G. B. Stenhouse, and H. L. Beyer. 2004. Removing GPS collar bias in habitat selection studies. Journal of Applied Ecology 41:201–212.

Green, R. A., and G. D. Bear. 1990. Seasonal cycles and daily activity patterns of Rocky-Mountain elk. Journal of Wildlife Management 54:272–279.

Harrell, F. E. 2001. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer series in statistics. Springer, New York, New York, USA.

Hjermann, D. Ø. 2000. Analyzing habitat selection in animals without well-defined home ranges. Ecology 81:1462–1468.

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

Johnson, C. J., S. E. Nielsen, E. H. Merrill, T. L. Mcdonald, and M. S. Boyce. 2006. Resource selection functions based on use-availability data: theoretical motivation and evaluation methods. Journal of Wildlife Management 70:347–357.

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

Manly, B. F., L. L. McDonald, D. L. Thomas, T. L. McDonald, and W. P. Erikson. 2002. Resource selection by animals: statistical design and analysis for field studies. Second edition. Chapman and Hall, New York, New York, USA.

Millspaugh, J. J., and J. M. Marzluff, editors. 2001. Radio tracking and animal populations. Academic Press, San Diego, California, USA.

Moorcroft, P. R., and A. Barnett. 2008. Mechanistic home range models and resource selection analysis: a reconciliation and unification. Ecology 89:1112–1119.

Moorcroft, P. R., M. A. Lewis, and R. L. Crabtree. 2006. Mechanistic home range models capture spatial patterns and dynamics of coyote territories in Yellowstone. Proceedings of the Royal Society B 273:1651–1659.

Nams, V. O. 1989. Effects of radiotelemetry error on sample-size and bias when testing for habitat selection. Canadian Journal of Zoology 67:1631–1636.

Rhodes, J. R., C. A. McAlpine, D. Lunney, and H. P. Possingham. 2005. A spatially explicit habitat selection model incorporating home range behavior. Ecology 86: 1199–1205.

Stein, M. L. 1999. Interpolation of spatial data: some theory for kriging. Springer, New York, New York, USA.

Strickland, M. D., and L. L. McDonald. 2006. Introduction to the special section on resource selection. Journal of Wildlife Management 70:321–323.

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

## APPENDIX A

An extension for nonuniform turning angles (*Ecological Archives* E090-249-A1).

## APPENDIX B

A comparison between empirical step-length distributions and the resource-independent movement kernel (*Ecological Archives* E090-249-A2).

## APPENDIX C

Details on using conditional logistic regressions to estimate Step Selection Functions (*Ecological Archives* E090-249-A3).

## APPENDIX D

Regression results (*Ecological Archives* E090-249-A4).