A 'How-to' Guide for Interpreting Parameters in Resource- and Step-Selection Analyses

John Fieberg, Johannes Signer, Brian Smith, Tal Avgar

Affilitiations

- JF: Department of Fisheries, Wildlife, and Conservation Biology, University of Minnesota, St. Paul, MN, 55112
- JS: Wildlife Science, Faculty of Forestry and Forest Ecology, University of Goettingen, Göttingen Germany
- BS, TA: Department of Wildland Resources and Ecology Center, Utah State University, Old Main Hill, Logan, UT 84322, USA

ABSTRACT

Resource-selection and step-selection analyses allow researchers to link animals to their environment and are commonly used to address questions related to wildlife management and conservation efforts. Step-selection analyses that incorporate movement characteristics, referred to as *integrated step-selection analyses*, are particularly appealing because they allow modeling of both movement and habitat-selection processes. Despite their popularity, many users struggle with interpreting parameters in resource-selection and step-selection functions. Integrated step-selection analyses also require several additional steps to translate model parameters into a full-fledged movement model, and the mathematics supporting this approach can be challenging for biologists to understand. Using simple examples, we demonstrate how weighted distribution theory and the inhomogeneous Poisson point-process model can facilitate parameter interpretation in resource-selection and step-selection

analyses. Further, we provide a "how to" guide illustrating the steps required to implement integrated step-selection analyses using the `amt` package. By providing clear examples with open-source code, we hope to make resource-selection and integrated step-selection analyses more understandable and accessible to end users.

Keywords: habitat selection, inhomogeneous Poisson point-process, integrated step-selection analysis, resource-selection function, spatial hazards, step-selection function, telemetry, relative selection strength

Introduction

New technologies (e.g., improved Global Positioning System [GPS] collars) and advances in remote sensing have made it possible to collect animal location data on unprecedented spatial and temporal scales [1, 2], which in turn has fueled the development of new methods for modeling animal movement and for linking individuals to their environments [3, 4]. Two of the most popular approaches for analyzing telemetry data, resource-selection and step-selection analyses, compare environmental covariates at locations visited by an animal ("used locations") to environmental covariates at a set of locations assumed available to the animal ("available locations") using logistic and conditional logistic regression, respectively [5–7]. These methods are widely available in most statistical software packages, and thus, they provide a robust and easy-to-implement framework for analyzing habitat-selection patterns; note, here and elsewhere, we have used the term *habitat-selection* rather than *resource-selection* to highlight our broader interest in modeling the effects of a diverse set of environmental variables (e.g., those capturing risks and environmental conditions in addition to resources), but we will often use these terms interchangeably. Despite their popularity, our collective experience has been that most users struggle to interpret parameters in these models. Further, it seems that papers attempting to address this issue have had limited success, and in some aspects may have increased confusion [8–11].

Here, we highlight how point-process models and weighted distribution theory provide simple and effective frameworks for interpreting regression parameters in resource-selection and step-selection analyses. In the sections that follow, we begin by reviewing recent research connecting resource-

selection functions to point-process models and weighted distribution theory. Using these connections, we demonstrate correct interpretation of parameters using simple examples of models fit to GPS locations of fisher (*Pekania pennanti*) from upstate New York [12]. We then provide a short review of step-selection analyses, including their history and methods for parameter estimation. Step-selection analyses are particularly appealing because: 1) they provide an objective method for defining habitat availability in terms of movement constraints; 2) they relax the assumption that locations are statistically independent; and 3) by including movement characteristics (e.g., functions of step length and turn angle) as predictors, they provide a means to model both movement and habitat selection processes [13]. Recognizing that many biologists may find the mathematics supporting integrated step-selection analyses intimidating, we aim to provide a "how to" guide demonstrating the steps required to implement the approach using the`amt` package [14]. We end with a short discussion highlighting challenges related to statistical dependencies and model transferability.

Resource-Selection Analyses

Logistic Regression

Much of the confusion surrounding the interpretation of parameters in resource-selection analyses can be attributed to the use of logistic regression in a non-standard way. Logistic regression is most easily understood as a model for binary random variables that can take on one of two values (0 or 1) with probability that depends on one or more explanatory variables [15].

Consider, for example, a prospective study designed to infer how various environmental characteristics influence whether a habitat patch will be used by one or more animals. In this case, we may randomly select n habitat patches and monitor them to determine if they are used ($y_i = 1$) or not ($y_i = 0$) for $i = 1, 2, \ldots, n$. Logistic regression allows us to model the probability that each patch will be used, $P(y_i = 1) = p_i$, as a logit-linear function of patch-level predictors $(X_{i1}, \ldots, X_{ip})$ and regression parameters $(\beta_1, \ldots, \beta_p)$:

$$y_i \sim \text{Bernoulli}(p_i)$$
$$\text{logit}(p_i) = \log\left(\frac{p_i}{(1-p_i)}\right) = \beta_0 + \beta_1 X_{i1} + \ldots \beta_p X_{ip}$$

After having fit a model, we can exponentiate the regression coefficients, $\exp(\beta_k)$, to quantify how the odds of use, $p/(1-p)$, change as we increase the kth predictor by 1 unit while holding all other predictors constant. We can also use the inverse-logit transformation (eq. (1)) to estimate the probability that a patch will be used, given its set of spatial predictors:

$$p_i = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \ldots \beta_p x_{ip})}{1 + \exp(\beta_0 + X_{i1}\beta_1 + \ldots X_{ip}\beta_p)} \tag{1}$$

The logit transformation ensures that p will be constrained between 0 and 1 for all values of the predictor variables.

Contrast this approach with how logistic regression is used to study habitat selection. In a typical habitat-selection study, logistic regression models are fit to separate samples of used and available sample units, groups that are not mutually exclusive (i.e., available habitat patches may also be used). We will refer to the combined locations as *use-availability* data. In this case, y_i is no longer a Bernoulli random variable since p_i depends on the ratio of used to available points (which is under control of the analyst). Further, most analyses of telemetry data involve point-level sampling in continuous space rather than discrete sample units. In this case, the probability associated with any point is necessarily 0 to ensure the distribution integrates to 1 over all of available space. Thus, it is perhaps not surprising that there has been considerable confusion and controversy surrounding the use of logistic regression with use-availability data [8, 9, 11].

Various arguments have been constructed to justify the use of logistic regression when analyzing use-availability data [9, 16, 17], but a significant breakthrough came when Warton et al. [18] made a connection between logistic-regression and the inhomogeneous Poisson point-process (IPP) model. An IPP is a model for random locations or events in space, where the expected spatial density of the locations depends on spatial predictors (see next section, **Inhomogeneous Poisson Point-process**

Model). Warton et al. [18] showed that as the number of available points is increased towards infinity, the slope parameters in logistic regression models will converge to the slope parameters in an IPP model. Interestingly, several other popular approaches for analyzing species distribution data, including MaxEnt [19, 20], weighted distribution theory with an exponential link function [21], and resource utilization functions [22], have been shown to be equivalent to fitting an inhomogeneous Poisson point-process model [18, 23–26].

Instead of focusing on p_i, as is typical in prospective studies, logistic regression applied to use-availability data should simply be viewed as a convenient tool for estimating coefficients in a *resource-selection function*, $w(x;\beta) = \exp(X_{i1}\beta_1 + \ldots X_{ip}\beta_p)$ [5, 27]. As we will see in the next section, this expression is equivalent to the intensity function of an IPP model, but with the intercept (the log of the baseline intensity) removed; the baseline intensity gives the expected density of points when all covariates are 0. Because resource-selection functions do not include this baseline intensity, they are said to measure "relative probabilities of use" or, alternatively, said to be "proportional to the probability of use" [16]. Although the term *probability of use* sounds appealing, it is important to remember the challenges with defining probability at the point-level. Further, although probability of use is easily defined for discrete sample units (e.g. grid cells), these probabilities should increase with the size of the spatial unit and also with the study duration [21, 28]. Thus, with telemetry studies, it seems more natural to model spatial (or spatio-temporal) "hazards" or rates of use in continuous space (and time), from which "probability of use" can be determined by integrating these hazards over whatever spatial (and temporal) unit is deemed appropriate. Point-process models allow us to do just that!

Inhomogeneous Poisson Point-Process (IPP) Model

The IPP model provides a simple framework for modeling the density of points in space as a log-linear function of spatial predictors through a spatially-varying intensity function, $\lambda(s)$:

$$\log(\lambda(s)) = \beta_0 + X_1(s)\beta_1 + \ldots X_p(s)\beta_p \tag{2}$$

where s is a location in geographic space, and $X_1(s), \ldots, X_p(s)$ are spatial predictors associated with

location s. The intercept, β_0, determines the log-density of points (within a small homogeneous area around s) when all $X_i(s)$ are 0, and the slopes, $\beta_1, \ldots, \beta_p$, describe the effect of spatial covariates on the log density of locations in space. The IPP model can be understood by listing its key features and assumptions, namely:

1. The number of locations in an area G, n_G, is given by a Poisson random variable with mean $\mu_G = \int_G \lambda(s)ds$.

2. Locations are independent (any clustering can be explained by spatial covariates).

If all available spatial predictors are measured only at a coarse scale (e.g., at a set of gridded or rasterized cells), then fitting the IPP model is equivalent to fitting a Poisson regression model [23]. Specifically, one may treat the counts, y_i, in n discrete spatial units ($i = 1, \ldots, n$), as a set of independent Poisson random variables with means $= \lambda(s_i)|G_i|$ where $\lambda(s_i)$ is given by eq. (2) and $|G_i|$ is the area of grid cell i. Note that $\log(E[y_i]) = \log(\lambda(s_i)|G_i|) = \log(\lambda(s_i)) + \log(|G_i|)$; thus, the log-link used in Poisson regression implies the area, $|G_i|$, should be included as an offset (a predictor variable with with regression coefficient fixed at a value of 1).

When spatial predictors are available at the point-level, as will be the case whenever constructing "distance to" predictors (e.g., distance to nearest road, water source, etc), it will be advantageous to model the locations in continuous space. In telemetry studies, the absolute density of points will be determined by the frequency and duration of data collection. Thus, β_0 will not be of biological interest, and it will be appropriate to focus efforts on estimating and interpreting the slope coefficients, $\beta_1, \ldots, \beta_p$, which determine relationships between the spatial covariates and the relative density of locations throughout the study area [24]. As is the case with linear and generalized linear models (e.g., Poisson regression), we can estimate parameters using maximum likelihood. This technique requires writing down an expression, called the *likelihood*, that captures the data generating mechanism in terms of one or more parameters. With telemetry data, it makes sense to work with the conditional likelihood of the IPP model [23], i.e., the likelihood of the observed locations in space, conditional on there being n observed locations in total. The conditional likelihood is given by:

$$L(\beta_1, \ldots, \beta_p | s_1, \ldots, s_n) = \prod_{i=1}^{n} \frac{\lambda(s_i)}{\int_{s \in G} \lambda(s) ds} \tag{3}$$

where the product is over the n observed locations, $\lambda(s_i)$ is the intensity function evaluated at observation i, and the integral in the denominator evaluates the intensity function over the spatial domain of interest [23, 29]. If we plug $\lambda(s_i) = \exp(\beta_0 + X_1(s_i)\beta_1 + \ldots X_p(s_i)\beta_p)$ into eq. (3), β_0 will cancel from the numerator and denominator, leaving us with:

$$L(\beta; s_1, \ldots, s_n) = \prod_{i=1}^{n} \frac{\exp(X_1(s_i)\beta_1 + \ldots X_p(s_i)\beta_p)}{\int_{s \in G} \exp(X_1(s_i)\beta_1 + \ldots X_p(s_i)\beta_p) ds} = \prod_{i=1}^{n} \frac{w(x(s_i); \beta)}{\int_{s \in G} w(x(s); \beta) ds} \tag{4}$$

where $w(x(s); \beta) = \exp(\beta_1 x_1(s) + \ldots \beta_p x_p(s))$ is our resource-selection function.

The binomial likelihood associated with logistic regression differs from eq. (4), but Warton et al. [18] showed that logistic regression estimators of slope coefficients converge to the those of the IPP model as the number of available points increases toward infinity. Thus, the connection to the IPP model addresses a common question that arises when estimating resource-selection functions, namely, "how many available points do I need?" The exact answer depends on how difficult it is to estimate the integral in the denominator of eq. (4); the recommendation we offer is to increase the number of available points until the estimated slope coefficients no longer change much. Fithian and Hastie [24] later showed that the convergence results of Warton et al. [18] hold only if the model is correctly specified, but assigning "infinite weights" to available points ensures the results hold more generally. Therefore, when fitting logistic regression or other binary response models (e.g., boosted regression trees) to use-availability data, we also suggest assigning a large weight (say 1000 or more) to each available location and a weight of 1 to all observed locations (larger values can be used to verify that results are robust to this choice). For a coded example in R [30], see section **Interpreting Parameters in Resource-Selection Functions** and Additional File 1.

Weighted Distributions

Weighted distribution theory provides another way to interpret parameters in resource-selection functions [21, 31]. Let:

- $u(x)$ = the frequency distribution of habitat covariates, x, at locations used by our study animals.
- $a(x)$ = the frequency distribution of habitat covariates, x, at locations assumed to be available to our study animals.

We can think of the *resource-selection* function, $w(x;\beta)$, as providing a set of weights that takes us from the distribution of available habitat to the distribution of used habitat:

$$u(x) = \frac{w(x,\beta)a(x)}{\int_{z\in E} w(z,\beta)a(z)dz} \tag{5}$$

The denominator of eq. (5) ensures that the left hand side integrates to 1 and thus, $u(x)$ is a proper probability distribution; the variable z here is just a dummy variable used to allow integration over the frequency distribution of our environmental covariates. Because these distributions are written in terms of the habitat covariates, x, instead of geographical locations, we say that model is parameterized in *environmental space* (E) [32, 33].

To show that weighted distribution theory is consistent with the IPP formulation discussed above, we can rewrite equation (5) in *geographic space* (G):

$$u(s) = \frac{w(x(s),\beta)a(s)}{\int_{g\in G} w(x(g),\beta)a(g)dg}, \tag{6}$$

where the denominator integrates over a geographic area, G, that is assumed to be available to the animal and g is a dummy variable for integration. Here $u(s)$ is equivalent to the utilization distribution encountered in the literature on probabilistic estimators of animal home ranges [34–36] and tells us how likely we are to find an individual at location s in geographic space. The utilization distribution, $u(s)$, depends on the environmental covariates associated with location s, through $w(x(s);\beta)$, and the distribution of available locations in geographic space, $a(s)$. Typically, $a(s)$ is assumed to be a uniform distribution within the geographical domain of availability, G (e.g., the individual's home range, the population's range, or the species range depending on the hierarchical level of habitat selection of interest; [37]), and all areas within G are assumed to be equally available

to the organism. Hence, $a(s)$ is typically a constant, $1/|G|$, that cancels from the numerator and denominator. Then, if we let $w(x(s);\beta) = \exp(x\beta)$, we end up with the conditional likelihood of the Inhomogeneous Poisson process model (eq. (4)) [23].

Interpreting Parameters in Resource-Selection Functions

To demonstrate how the IPP and weighted distribution theory frameworks help with interpreting parameters in fitted resource-selection functions, we now consider a simple example using 3,004 locations of a fisher named Lupe tracked as part of a larger telemetry study [12]. These data are publicly available and have been featured in a workshop highlighting Movebank's *Env-DATA* system for annotating locations with environmental covariates [38, 39]. The location data were combined with available points sampled randomly from within a minimum convex polygon (MCP) formed using Lupe's locations. The used and available locations were then transformed to a projected coordinate reference system (NAD83 / Conus Albers) and annotated with environmental variables measuring population density [40], elevation [41], and landcover class [42]. The original landcover data were grouped to form a variable named `landuseC` with the following categories: `forest`, `grass` and `wet` (Fig. 1). We created centered (mean = 0) and scaled (SD = 1) variables labeled `elevation` and `popden` from the original elevation and population density variables. We also created an indicator variable, `case_`, taking on a value of 1 for all used points and 0 for all available points (later, we discuss how to choose the number of available points).

For ease of interpretation, we will begin by assuming the effects of elevation, population density, and landcover class are additive and linear on the log scale (eq. (2)). Later, we will discuss how we can relax these assumptions using interactions to allow the effect of covariates to depend on the value of other habitat covariates and polynomials or splines to to relax the assumption of linearity. We assign a weight of 5000 to the available locations and a weight of 1 to all observed locations [24]. We can then fit a weighted logistic regression model using the `glm` function in R:

```
Lupe.dat$w <- ifelse(Lupe.dat$case_==1, 1, 5000)
```

```
RSF.Lupe <- glm(case_ ~ elevation + popden + landuseC,
                data = Lupe.dat,
```

```
                weight = w,
                family = binomial)
```

Before interpreting the coefficients, it is important to make sure we have included sufficient available points to allow parameter estimates to converge to stable values. To evaluate parameter stability, we fit logistic regression models to data sets with increasing numbers of available points (from 1 available point per used point to 100 available points per used point; see Additional File 1 for the code). As expected, estimates varied less from sample to sample as we increased the number of available points (Fig. 2). The intercept decreased as we increased the number of available points (as it is roughly proportional to the log difference between the numbers of used and available points), but the slope parameter estimates, on average, did not change much once we included at least 10 available points per used point. Thus, we conclude that, in this particular case, having 10 available points per used point is sufficient for interpreting the slope coefficients. Using more available points reduces Monte Carlo error, however, so we will proceed with 100 available points per used point.

Let's consider the interpretation of the continuous covariates reflecting elevation and population density (Table 1, *Model 1*). Qualitatively, we might infer from the positive coefficient for `elevation` and negative coefficient for `popden` that, all other things being equal, Lupe is likely to select locations at higher elevations and in areas of lower population density. But, how do we interpret these coefficients quantitatively? Consider the following two locations, both in the same landcover class and with the same associated population density, but differing by 1 unit in `elevation` (since we have scaled this variable, a difference of 1 implies that the two observations differ by 1 SD in the original units of elevation):

- location s_1: `elevation` = 3, `popden`=1.5, `landuseC` = `wet`
- location s_2: `elevation` = 2, `popden`=1.5, `landuseC` = `wet`

Using eq. (6), we can calculate the relative risk of Lupe using location 1 relative to location 2:

$$\frac{u(s_1)}{u(s_2)} = \frac{\exp(3\beta_{elevation} + 1.5\beta_{pop_den} + 0\beta_{grass} + 1\beta_{wet})a(s_1)}{\exp(2\beta_{elevation} + 1.5\beta_{pop_den} + 0\beta_{grass} + 1\beta_{wet})a(s_2)} \tag{7}$$

where we have dropped the integral from eq. (6) because it appears in both the numerator and

denominator (and thus, cancels out). Now, if both locations are *equally available*, then $a(s_1) = a(s_2)$, and we have:

$$\frac{u(s_1)}{u(s_2)} = \frac{\exp(3\beta_{elevation})\exp(1.5\beta_{pop_den} + \beta_{wet})}{\exp(2\beta_{elevation})\exp(1.5\beta_{pop_den} + \beta_{wet})} = \exp(\beta_{elevation}) = \exp(0.303) = 1.35 \quad (8)$$

In epidemiology, $\exp(\beta)$ is referred to as a *risk* or *hazard ratio*. In the context of habitat-selection analyses, Avgar et al. [10] refer to it as quantifying *relative selection strength* (RSS).

Note that we would arrive at the exact same expression if we chose *any* two locations that differed by 1 unit of `elevation` and had the same values for `popden` and `landuseC`. Thus, $\exp(\beta_{elevation})$ quantifies the risk (or hazard) ratio of 2 locations that differ by 1 SD unit of `elevation` but are otherwise equivalent (i.e., they are equally available and have the same values of all other habitat covariates). If Lupe were to be presented with two such hypothetical locations, the model suggests she would be 1.35 times more likely to choose the one with the higher elevation. A similar interpretation can be ascribed to `popden`. Given two observations that differ by 1 SD unit of `popden` but are otherwise equal, Lupe would be $\exp(-0.183) = 0.833$ times as likely to choose the location with higher population density (or, equivalently, $\exp(0.183) = 1.20$ times more likely to choose the location with the lower population density).

What about the coefficients for the landcover categories? Looking again at the regression output (Table 1, *Model 1*), we see that `grass` has a negative coefficient and `wet` has a positive coefficient. It is tempting to infer that Lupe spends most of her time in wet areas and rarely spends time in grassy habitats. As Figure 1 makes it clear, however, these inferences are not exactly correct. First, it is important to understand how categorical predictors are encoded in regression models. There are a number of different ways to parameterize the effect of categorical variables and unfamiliar readers may want to work through an introductory regression text (e.g., Chapter 6 of [43]). The default coding in R is to treat one of the levels (whichever comes first alphabetically) as a reference level and then to create a set of dummy variables that contrast the remaining levels of the categorical variable with this reference level. In our case, `forest` is the reference level. The coefficients associated with `grass` and `wet` represent contrasts between these land cover classes and the `forest` class.

Let's again consider 2 locations, this time assuming they have the same elevation and population

densities, but with one location in `wet` and the other location in `forest`:

- location s_1: `elevation` = 2, `popden`=1.5, `landuseC` = `wet`
- location s_2: `elevation` = 2, `popden`=1.5, `landuseC` = `forest`

The relative risk of an animal using location 1 relative to location 2 is given by (eq. (6)):

$$\frac{u(s_1)}{u(s_2)} = \frac{\exp(2\beta_{elevation} + 1.5\beta_{pop_den} + 0\beta_{grass} + 1\beta_{wet})a(s_1)}{\exp(2\beta_{elevation} + 1.5\beta_{pop_den} + 0\beta_{grass} + 0\beta_{wet})a(s_2)} = \exp(\beta_{wet})\frac{a(s_1)}{a(s_2)} \tag{9}$$

Thus, *assuming the two locations are equally available*, we might infer that Lupe would be $\exp(0.250) = 1.28$ times more likely to choose the `wet` location than the location in `forest`. Of course, we know from Figure 1 that `forest` and `wet` are not equally available on the landscape. The higher availability of `forest` habitat implies that Lupe is more likely to be in `forest` than `wet`. We could attempt to correct for differences in availability within the MCP surrounding Lupe's locations by multiplying our result by the ratio of habitat availability for `wet` relative to `forest` habitats (2.3% versus 95.7%; Fig. 1), giving us $\exp(0.250)(0.023)/(0.957) = 0.03$. This calculation suggests we are $(1/0.03 = 33)$ times more likely to find Lupe in `forest` than `wet` habitat. With this calcualtion we had to assume, perhaps naively, that the availability distributions for `popden` and `elevation` were the same in both `wet` and `forest` cover classes. In reality, if Lupe decides to move from `forest` to `wet`, it is likely that she will experience a change in `elevation` and `popden` too (i.e., these factors will not be held constant). To quantify the relative risk of finding Lupe in `forest` versus `wet` habitat, while also accounting for the effects other environmental characteristics that are associated with these habitat types, we can use integrated hazards – i.e., we can integrate the spatial utilization distribution, $u(s)$, over all `forest` and `wet` habitats:

$$\frac{u(s, s \in forest)}{u(s, s \in wet)} = \frac{\int_G u(s) I(s \in forest) ds}{\int_G u(s) I(s \in wet) ds} \tag{10}$$

where $I(s \in forest)$ and $I(s \in wet)$ are indicator functions equal to 1 when location s is in `forest` or `wet`, respectively (and 0 otherwise). We can estimate this ratio using:

$$\frac{\hat{u}(s, s \in forest)}{\hat{u}(s, s \in wet)} = \frac{\sum_{i=1}^{n_a} \hat{w}(x(s_i); \hat{\beta}) I(s_i \in forest)}{\sum_{i=1}^{n_a} \hat{w}(x(s_i); \hat{\beta})) I(s_i \in wet)}. \tag{11}$$

where the sum is over the distribution of available locations.

This ratio is also equal to 33, which agrees with the observed data; Lupe was found in `forest` habitat 33 times more often than in `wet` habitat (see Additional File 1 for code demonstrating how to calculate these quantities in R). Thus, we conclude Lupe is 33 times more likely to be found in `forest` than `wet` habitat, assuming she restricts her movements to the MCP surrounding her observed locations and all of this habitat is equally available to her.

Before moving on, it is important to note that naively-adjusted (multiplying by availability of `wet` and `forest` habitats) and integrated-hazards (i.e., adjusted) risk ratios will not always agree. In fact, we find that they differ when comparing the risk of finding Lupe in `wet` versus `grass` habitat, with the integrated-hazards risk ratio better agreeing with the observed data (see Additional File 1). Somewhat related, Avgar et al. [10] suggested calculating *average effects* for continuous predictors, X, by comparing the change in relative risk from increasing X by 1 unit (to $X = x + 1$) to the average value of $w(x(s); \beta)$ for all locations s with $x(s) = x$. These average effects will also be influenced by cross-correlations among predictor variables included in the model.

Instead of integrating $u(s)$ over discrete cover types, we could integrate over a specific geographic area. In addition, we could choose to change the area of interest (and thus, area of integration) from G to $\tilde{G}$. This approach makes it possible to use the same integrated hazards approach (i.e., eq. (11)) to project how Lupe would spend her time in a novel environment (referred to as an "out-of-sample" prediction). Out-of-sample predictions often suffer from poor accuracy, especially when compared to "in sample" predictions, i.e., predictions for the same area and time frame from which the original data were collected [44, 45]. We return to this important point in the discussion section.

Let's next consider what happens if we change the reference level of the land cover variable from `forest` to `wet` (Table 1, *Model 2*).

```
Lupe.dat <- within(Lupe.dat,
                   landuseC1 <- relevel(landuseC, ref = "other"))
RSF.Lupe2 <- glm(case_ ~ elevation + popden + landuseC1,
                 data = Lupe.dat,
                 weight = w,
```

```
                  family = binomial)
```

The coefficients for `elevation` and `popden` do not change. Note, however, that the coefficient for `forest` is negative despite Lupe using `forest` more than its availability (i.e., $u(s, s \in forest) > a(s, s \in forest)$) and Lupe spending more than 95% of her time in the forest! What is going on? Remember, the coefficients for categorical predictors reflect use:availability ratios for each level of the predictor relative to the use:availability ratio for the reference class. The coefficient for `forest` is negative because the use:availability ratio for `forest` is less than the use:availability ratio for the reference class, `wet` (see Fig. 1). Depending on the reference level, it is possible to have a positive (negative) coefficient even when that landcover class is used more (less) than its availability. Furthermore, it is possible for landcover class to be used frequently but have a negative coefficient. We have seen many ecologists, including some that are very quantitatively skilled and familiar with habitat-selection models, make mistakes when interpreting coefficients associated with categorical predictors! This example also highlights the importance of plotting your data (e.g., Fig. 1) and considering habitat availability when interpreting regression coefficients. Plotting distributions of covariates for both used and available locations is one of the best ways to understand fitted habitat-selection models, and is a good strategy to use for both continuous and categorical predictors [46, 47].

Interactions Between Environmental Predictors

Consider the distribution of `elevation` at used and available locations across the different habitat classes (Fig. 3). We see that there is a wider range of `elevation` in `forest` and `wet` habitat compared to `grass` habitat, and there is a clear association between `elevation` and `landuseC`, with higher median `elevation` at used locations in `forest` and `grass` habitat relative to `wet` habitat. Perhaps more importantly, we also see that values of `elevation` are higher, on average for used locations (compared to available locations) in `forest` and `grass`, whereas the opposite is true in `wet` habitat. Although we should be skeptical of interactions that we discover while exploring our data (i.e., interactions that were not specified *a priori*), an analyst may be tempted to include an interaction between `elevation` and `landuseC`. In *Model 3* (Table 1), we revert to having `forest` as the reference level and include the interaction between `elevation` and `landuseC`.

```r
Lupe.dat <- within(Lupe.dat , landuseC <- releval(landuseC, ref = "forest"))
RSF.Lupe3 <- glm(case_ ~ elevation + popden + landuseC + elevation:landuseC,
                 data = Lupe.dat,
                 weight = w,
                 family = binomial)
```

Using this syntax, R creates two new variables `elevation:landuseCgrass` equal to `elevation` when `landuseC` is `grass` and is 0 otherwise and `elevation:landuseCwet` equal to `elevation` when `landuseC` is `wet` and is 0 otherwise. The coefficients associated with these predictors quantify the change in slope (i.e., change in the effect of `elevation`) when the locations fall in `grass` or `wet` relative to the slope when the locations fall in `forest`. Starting from equation (6) and using the estimates for *Model 3* in Table 1, we can easily derive that the relative risk of choosing between two equally available locations that differ by 1 SD unit of `elevation` is equal to $\exp(0.313) = 1.37$ when the two locations are in `forest`, $\exp(0.313 + 0.112) = 1.53$ when the locations are in `grass`, and $\exp(0.313 - 0.499) = 0.83$ when the locations are in `wet` habitat. Thus, Lupe selects for higher elevations when in `forest` or `grass`, but avoids higher elevations when in `wet`. Alternatively, we can consider how `elevation` changes Lupe's view of the different landcover categories, noting that $\beta_{grass} = -1.471 + 0.112\texttt{elevation}$ and $\beta_{wet} = 0.183 - 0.499\texttt{elevation}$. Thus, we see that Lupe's strength of avoidance of `grass` (relative to `forest`) and selection for `wet` (relative to `forest`) both decline with `elevation`, and Lupe's inherent ranking of these 3 habitat types will change as `elevation` increases.

Non-Linear Effects and Other Considerations

When building models, it is important to consider the functional relationships between different environmental characteristics and habitat use. For example, we may classify available predictors based on whether they represent resources (higher values are generally preferable), risks (lower values are generally preferable) or conditions (values that are not too high or too low are preferable) [48]. It is often useful to allow for non-linear effects of conditions by including quadratic terms or using a set of spline basis functions. In either case, we end up requiring multiple coefficients to capture how relative risk changes with the environmental predictor. Consider, for example, that

we could include a quadratic term to model the effect of `elevation`. Estimating the relative risk for two locations, s_1 and s_2, that differ in their values of `elevation` but are otherwise equivalent would be straightforward using equation (6) - we would just need to calculate hazard ratios using coefficients for `elevation` and `elevation`^{2}:

$$\frac{u(s_1)}{u(s_2)} = \frac{\exp(elevation(s_1)\beta_{elevation} + elevation(s_1)^2\beta_{elevation^2})}{\exp(elevation(s_2)\beta_{elevation} + elevation(s_2)^2\beta_{elevation^2})} \tag{12}$$

Lastly, we note that Avgar et al. [10] provide simple formulas for calculating risk or hazard ratios under a number of different scenarios (e.g., models with quadratic polynomials, log-transformed covariates, and models with interactions). The `log_rss` function in the `amt` package [14] relies on R's generic `predict` function to aid the user in calculating the log of these hazard ratios for any combination of model structure and two alternative locations; its use is illustrated in Additional File 2. Understanding how these formulas are derived, however, helps build intuition and frees the user to construct estimators and estimation targets that capture relevant quantities of specific interest.

Statistical Independence

An important assumption of the IPP model, and hence, resource-selection functions, is that any clustering of spatial locations can be explained solely by spatial covariates. Strictly speaking, this assumption will almost never be met, particularly with modern-day telemetry studies that allow several locations to be collected on the same day. Telemetry observations close in time tend to also be close in space - i.e., telemetry observations exhibit serial dependence [49]. This serial dependence is likely to manifest itself in residual spatial autocorrelation that could be modeled using a spatial random effect or a spatial predictor constructed to account for the effects of movement constraints on habitat availability [50]. Models with spatial random effects are, however, more complicated and difficult to fit.

Alternatively, if telemetry observations are collected at regular time intervals, then the locations may be argued to provide a representative sample of habitat use from a specific observation window [51]. In these cases, it may be helpful to view our estimates of the parameters in our resource-selection function, $\hat{\beta}$, as useful summaries of habitat use for tagged individuals during these fixed time

periods. Nevertheless, the assumption of independence of our locations is clearly problematic and will lead to estimates of uncertainty that are on average too small. If we are primarily interested in population-level inferences, then we may choose to ignore within-individual autocorrelation when estimating individual-specific coefficients but use a robust form of SE that treats individuals as independent when describing uncertainty in population-level parameters, e.g., using a bootstrap [52] or generalized estimating equations approach [53–55].

Step-Selection Functions

Step-selection functions were developed to deal with serial dependence as well as temporally varying availability distributions resulting from movement constraints [6, 7]. Rather than treat locations as independent and identically distributed (with availability that does not depend on time), step-selection analyses model transitions or "steps" connecting sequential locations (Δt units apart) in geographical space:

$$u(s, t + \Delta t) = \frac{w(x(s); \beta(\Delta t))\phi(s, s'; \gamma(\Delta t))u(s', t)}{\int_{\tilde{s} \in G} w(x(\tilde{s}, s'); \beta(\Delta t))\phi(\tilde{s}, s'; \gamma(\Delta t))u(s', t)d\tilde{s}} \tag{13}$$

where $u(s', t)$ and $u(s, t + \Delta t)$ give the probability of finding the individual at location s and s' at times t and $t + \Delta t$, respectively, $w(x(s); \beta(\Delta t))$ is referred to as a step-selection function, and $\phi(s, s'; \gamma(\Delta t))$ is a selection-independent movement kernel that describes how the animal would move in homogeneous habitat or in the absence of habitat selection (i.e., when $w(x(s); \beta(\Delta t)) =$ a constant for all s). Note that we represent the parameter vectors (β and γ) as functions of the step duration (Δt). This notation reflects the fact that step-selection parameters are scale dependent (i.e., different Δt's will result in different estimates of β and γ; see [13] for more details). Thus, we generally require observations to be equally spaced in time, and care must be taken when comparing inference from models fitted at different temporal resolution (but see [56]).

As with resource-selection analyses, it is typical to model $w(x(s); \beta(\Delta t))$ as a log-linear function of spatial covariates and regression parameters, $w(x(s); \beta(\Delta t)) = \exp(X_1(s)\beta_1 + \ldots X_p(s)\beta_p)$. A key difference between resource-selection and step-selection analyses, however, is that the latter allow the available distribution to be time-dependent and equal to $a(s, t + \Delta T) = \phi(s, s', \gamma(\Delta t))u(s', t)$.

Consequently, step-selection analyses allow explicit consideration of temporally dynamic resources: $x(s,t)$. Further, step-selection parameters could be flexibly expressed as functions of temporal and/or spatial covariates, resulting in the more general formulation: $w(x(s,t);\beta(\Delta t, t, x(s,t)))\phi(s,s';\gamma(\Delta t, t, x(s,t)))$.

Models for $\phi(s,s';\gamma(\Delta t, t, x(s,t)))$

Step-selection approaches build on an early idea by Arthur et al. [57] to model time-dependent availability via a circular buffer with radius R centered on the previous location. Rhodes et al. [58] showed that this model is equivalent to assuming:

$$\phi(s,s';\gamma(\Delta t)) = \begin{cases} \frac{1}{\pi R^2}, & \text{if } ||s-s'|| \leq R \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

where $||s-s'||$ is the Euclidean distance between locations s and s', referred to as the *step length*. Rhodes et al. [58] also demonstrated that circular buffers imply that individuals are more likely to move large distances than short distances since there is more area, and thus probability, associated with outer rings of the circle. Instead, they suggested using an exponential distribution to accommodate right-skewed step-length distributions and a tendency for animals to make shorter rather than longer movements:

$$\phi(s,s';\gamma(\Delta t)) = \frac{\lambda \exp(-\lambda||s-s'||)}{2\pi||s-s'||} \tag{15}$$

Rather than specify a model directly in terms of $\phi(s,s';\gamma(\Delta t))$, it is more common to see movement kernels specified in terms of the distribution of step lengths, $d = ||s-s'||$, and turn angles (changes in direction from the previous bearing), θ. In the sections that follow, we will let $g(d;\gamma_d(\Delta t))$ and $f(\theta;\gamma_\theta(\Delta t))$ represent step-length and turn-angle distributions, respectively. Step-selection analyses frequently use either an exponential or gamma distribution for $g(d;\gamma_d(\Delta t))$. Turn angles may be assumed to be uniformly distributed as in [57] and [58]. Alternatively, circular distributions, such as the von Mises distribution or wrapped Cauchy or Weibull distributions, allow for a mode at 0

and can thus accommodate correlated movements (i.e., sequential steps are assumed, on average, to follow in the same direction as the previous step).

Although step-length and turn-angle distributions are typically assumed to be independent, animals commonly exhibit a mix of of temporally persistent movement behaviors, ranging between high-displacement movements (e.g., when traveling between habitat patches, migrating, or dispersing) and low-displacement movements (e.g., during foraging or resting bouts). If positional data are collected more frequently than the occurrence of behavioral switches, we might expect a negative cross-correlation between step lengths and turn angles (moving far is likely to coincide with moving straight) and a positive auto-correlation between the current and previous step lengths and turn angles. Moreover, as implied by the more flexible formulation, $w(x(s,t);\beta(\Delta t,t,x(s,t)))\phi(s,s';\gamma(\Delta t,t,x(s,t)))$, both step-length and turn-angle distribution may shift as a function of spatial and/or temporal covariates such as habitat permeability (e.g., elevation ruggedness, snow depth, or vegetation density), time of day, season, and predation risk [13, 59]. Thus, although ϕ is a "selection-independent" movement kernel, it may still depend on environmental or temporal covariates, and hence, may vary through space and time, resulting in both auto and cross correlations in step attributes.

Cross-correlation between step lengths and turn angles is difficult to model with common statistical distributions, but could be accommodated using copulae [60]. Alternatively, one could resample (i.e., bootstrap) step length and turn angle pairs, (d_t, θ_t), to preserve any correlation that is present in the data [6]. Although we generally find the bootstrap appealing [52], it has limitations in this context. In particular, the observed distribution of step lengths and turn angles will reflect both inherent movement characteristics of the species (captured by ϕ) as well as habitat selection (captured by w). Using the observed steps as a non-parametric model for ϕ without adjustment for the effect of w can result in biased estimates of β [61]. We will return to this point in the next section. As mentioned previously (see **Statistical Independence**), and regardless of the source of correlation, it may be preferable to calculate robust SEs by treating individuals as the relevant sampling unit when performing population-level inference [62]. Lastly, cross- and auto-correlations in step lengths and turn angles, as well as their dependencies on various temporal or environmental characteristics, could be modeled parametrically using an integrated step-selection function [13]. To do so, we need to include appropriate statistical interactions (e.g., between concurrent and previous step

lengths/turn angles and between these step-attributes and environmental or temporal covariates). We discuss this process further below, and provide examples in the Additional File 2. See also [63], [64], and [65].

Estimation of Movement and Habitat-Selection Parameters

Although it is possible to simultaneously estimate movement (γ) and habitat selection (β) parameters using maximum likelihood [58] or Bayesian methods [31], this is rarely done in practice as it would require custom-written code. Instead, it is common to use the following approach:

1. Estimate or approximate $\phi(s, s'; \gamma(\Delta t))$ using observed step lengths and turn angles, giving $\hat{\phi}(s, s'; \hat{\gamma}(\Delta t))$.
2. Generate time-dependent available locations by simulating potential movements from the previously observed location, $u(t, s')$. Similar to applications of RSFs, it is up to the user to decide how many available locations to sample for each used location, and, due to similar considerations (properly approximating the availability domain: $a(s, t + \Delta t) = \phi(s, s'; \gamma(\Delta t))u(t, s')$), the more points the merrier.
3. Estimate β using conditional logistic regression, with strata formed by combining time-dependent used and available locations.

If we knew $\phi(s, s', \gamma(\Delta t))$ and could simulate directly from it (skipping step 1), then this approach would provide unbiased estimates of β [61]. However, as mentioned in the previous section, estimating $\phi(s, s'; \gamma(\Delta t))$ from observed steps without adjusting for $w(x(s); \beta(\Delta t))$ can lead to biased estimates of γ and β.

Forester et al. [61] considered the case where $g(d, \gamma_d)$ is given by an exponential distribution with unknown parameter, λ. They showed that estimating λ directly from the observed distribution of step lengths (without adjusting for the effect of $w(x(s); \beta)$), and then proceeding with steps 2 and 3 results in a biased estimators of β, but that the bias (if $g(d, \gamma_d)$ is given by an exponential distribution) is eliminated if $\log(d_t)$ is included as a predictor in the model. Avgar et al. [13] further showed that the coefficient associated with $\log(d_t)$ could be used to modify $\hat{\lambda}$, leading to an unbiased estimator of λ and thus, $g(d, \gamma_d)$. In addition, they showed how similar adjustments could be used to obtain unbiased estimators of step-length (γ_d) and habitat-selection (β) parameters when the

distribution of step lengths is given by a gamma, half-normal, or log-normal distribution. Similarly, Duchesne et al. [66] showed that including $\cos(\theta)$ as a predictor can lead to unbiased estimators of turn angle parameters (γ_θ) when the distribution of turn angles follows a von Mises distribution. These adjustments are available in the `amt` package for the exponential, gamma, and von Mises distributions [14]. Avgar et al. [13] coined the term *integrated* step-selection analysis to emphasize that these results provide new opportunities to model both movement and habitat selection via tried and true statistical software for fitting conditional logistic regression models.

In Additional File 2, we provide a "How to" guide for implementing an integrated step-selection analysis using the `amt` package in R [14, 30]. Conducting an integrated step-selection analysis requires, in addition to the three steps outlined in this section, that we add a fourth step that re-estimates the movement parameters in $\phi(s, s'; \gamma(\Delta t))$ using regression coefficients associated with movement characteristics (e.g., $\log(d_t)$, $cos(\theta)$). This last step adjusts the parameters in $\phi(s, s'; \gamma(\Delta t))$ to account for the effect of habitat selection when estimating the movement kernel [13]. Importantly, interactions may be included between movement characteristics (e.g., $\log(d_t)$, $cos(\theta)$) and environmental covariates, x, to allow the movement kernel to depend on the environment. When interactions are included, step 4 results in a movement kernel, $\phi(s, s'; \gamma(\Delta t, t, x(s, t)))$, that depends on the habitat the animal is in at the start of the movement step (Fig. 4).

Interpretation of Parameters in an Integrated Step-Selection Analysis

The habitat-selection parameters can be interpreted in the same way as parameters in resource-selection functions (i.e., as spatial hazards, assuming locations are equally available and differing in terms of a single habitat covariate). Hence, the $ln(RSS)$ expressions in [10], and the `log_rss` function in `amt`, are suitable for calculating and interpreting the effects of the various habitat covariates. However, it is important to recognize that the used and available distributions in step-selection analyses are dynamic and non-uniform in space. In particular, they depend an individual's current location and movement tendencies (as well as the observed time scale determined by Δt; [67]; [68]). Thus, questions that require integrating hazards over space (e.g., eq. (10)) are more difficult to address, but may be computed using simulation modeling [68], by solving the master equation (eq. (13)) for its steady state [69, 70], or in some cases, by translating the fitted model

into a partial differential equation model with analytical steady-state distribution [71]. We also note that alternative modeling frameworks exist with parameters that directly describe long-term relative risk [72–74], but these methods are more computationally challenging to implement, and therefore, less likely to be widely used in applied settings. The `amt` package has a basic capacity to simulate the utilization distribution based on a parameterized integrated step-selection function [68], and we expect this approach to become more flexible in the near future.

Using an integrated step-selection approach, it is also possible to draw ecological inference using the selection-free movement kernel. For example, the fitted step-length and turn-angle distributions can tell us how much more likely an animal is to take large versus small steps or to turn left or right relative to moving straight. We can also calculate moments of these distributions under different environmental conditions, which can be informative when our models include interactions between movement characteristics and environmental predictors. For example, we could calculate the expected selection-free displacement rates (and/or directionality) as function of local snow depth (that is, if snow depth was included in our model as an interaction with step length). To calculate these expected values we must first adjust the 'tentative' parameters used to sample available steps (e.g., if we use a gamma distribution, the tentative shape and scale parameters) using the coefficient estimates obtained for step length (and/or its transformation) and cos(turn angle). The details of how to carry on these adjustments are provided in Additional File 3 and in [13]. Once the selection-free movement parameters are obtained, one can use them to calculate various aspects of the (theoretical) distributions of step lengths and turn angles, such as the mean, the median, or the 95% confidence bounds (see Additional File 2 for examples).

Discussion

We have highlighted how connecting resource-selection functions to IPP models and weighted distribution theory helps with interpreting parameters in resource-selection functions using simple examples. We have also reviewed step-selection analyses and demonstrated how to estimate movement and habitat-selection parameters when conducting an integrated step-selection analysis using the `amt` package. So far, we have focused on interpreting results when analyzing data from a single individual. We end with a brief discussion addressing statistical dependencies, particularly

when analyzing data from multiple individuals, along with issues related to model transferability and parameter sensitivity to changes in habitat availability and species population density.

Statistical Dependencies

Earlier, we highlighted the importance of statistical independence as it applies to individual locations when estimating resource-selection functions. We also noted that step-selection analyses typically assume step lengths and turn angles are independent of each other and also over time, though it is possible to account for these correlations using appropriate interactions (e.g., between step length at time t and time $t-1$, step length and turn angle both at time t). It would be nice to have multivariate distributions available that are capable of describing correlated step lengths and turn angles and any inherent autocorrelation. It is plausible, however, that models that allow movement parameters to vary by habitat type, using interactions between step length, turn angle, and habitat covariates, will be able to account for much of the autocorrelation and cross-correlation (between step lengths and turn angles) present in the data. Similarly, autocorrelation and cross-correlations may be accommodated by models that include a (possibly latent) behavioral state, with movement and habitat-selection parameters that are state-dependent [75, 76].

In addition to cross-correlation between step lengths and turn angles and serial dependencies, individuals living in different environments may exhibit different habitat-selection patterns and thus, repeated observations on the same set of individuals will induce further statistical dependencies. A simple strategy for dealing with repeated measures when individuals can be assumed to be independent is to fit models to individual animals and then treat the resulting coefficients as data when inferring population-level patterns [55, 77]. For example, sample means of the regression coefficients can be used to characterize average habitat-selection parameters. Estimating among-animal variability is a trickier due to sampling error; naively ignoring sampling error will lead to a positive bias in estimates of among-animal variability, but more formal two-step methods can address this issue [65, 78, 79]. Alternatively, generalized linear mixed models with random coefficients can be used to quantify among-animal variability in resource-selection and step-selection analyses [80].

Although it is possible to conduct integrated step-selection analyses with hierarchical models containing random effects, we have much to learn about how these approaches perform in practice. For

example, Muff et al. [80] found that parameters describing among-animal variability in habitat-selection parameters were biased low when movement characteristics were included in the model. Mixed-effect models with random coefficients are also "parameter hungry", requiring $p(p+1)/2$ variance and covariance parameters to be estimated, where p is the number of random coefficients. Models that allow all coefficients to be animal-specific and to covary are thus likely to be computationally challenging to fit and problematic for small data sets containing only a few individuals. For this reason, Muff et al. [80] assumed coefficients did not covary in their applied examples. In the context of our fisher analysis, this equates to assuming that knowing an individual's coefficient for `popden` tells us nothing about that animal's parameters for `elevation` or `landuseC` variables. For categorical variables, it is natural to expect parameters to have a negative covariance (since, for example, spending more time in `forest` must come at the expense of spending less time in other landuse categories). Research evaluating the performance of mixed-effect step-selection analyses under various data-generating scenarios would be helpful for evaluating robustness to assumption violations (e.g., those regarding the distribution of random parameters).

Sensitivity of Selection Coefficients to Species Population Density and Habitat Availability

Before concluding, we feel it is important to briefly discuss the oft observed pattern of density and availability dependence in habitat-selection inference [48, 81, 82]. Density-dependent inference may be observed when the same analysis is applied to individuals or populations of the same species, under similar environmental conditions, but at different population densities. Availability dependence (also referred to as a "functional response") may be observed when the same analysis is applied to individuals or populations of the same species, which experience different landscape-scale resource or habitat availabilities. For example, van Beest et al. [83], found that individual elk display availability-dependent resource-selection patterns (switching from selection to avoidance of certain habitats as function of the availability of these habitats within their home range), but that the strength of this functional response depended on elk population density. Such context dependencies are in fact so common that we do not know of a single instance where researchers were looking for them and failed to find them. Recently, Avgar et al. [84] showed that such context dependencies in habitat-selection patterns are expected to emerge even under the simplest theoretical

model of an Ideal Free Distribution [85]. Thus, habitat-selection models often have poor predictive capacity when transferred across different study areas, or even within the same area over time [44]. Yet, these differences may also be exploited; modeling frameworks that leverage data from multiple environments and across a range of population densities can potentially increase predictive capabilities [86]. As with any other attempt to model complex ecological data, critical evaluation of model performance for both within and out-of-sample data is essential [47].

Funding

JF received partial salary support from the Minnesota Agricultural Experimental Station. TA received partial salary support from the Utah Agricultural Experimental Station.

Authors' Contributions

JF developed the idea for the review, led the writing of the manuscript, and drafted the initial version of Additional File 1; B.S. and J.S. drafted the initial version of Additional File 2; B.S. and T.A. drafted the initial version of Additional File 3. All authors contributed critically to the manuscript text and Additional Files, and gave final approval for publication.

Supplementary Information

Additional File 1: `AppA_RSF_examples.html`, a tutorial demonstrating how to fit and interpret parameters in resource-selection functions.

Additional File 2: `AppB_SSF_examples.html`, a tutorial demonstrating how to fit and interpret parameters and output when conducting an integrated step-selection analysis.

Additional File 3: `AppC_iSSA_movement.html`, a description of methods used to adjust 'tentative' parameters in step-length and turn-angle distributions for the effects of habitat selection.

References

1. Kays R, Crofoot MC, Jetz W, Wikelski M. Terrestrial animal tracking as an eye on life and planet. Science. 2015;348:aaa2478.

2. Robinson OJ, Ruiz-Gutierrez V, Reynolds MD, Golet GH, Strimas-Mackey M, Fink D. Integrating citizen science data with expert surveys increases accuracy and spatial extent of species distribution models. bioRxiv. 2019;806547.

3. Guisan A, Thuiller W, Zimmermann NE. Habitat suitability and distribution models: With applications in r. Cambridge University Press; 2017.

4. Hooten MB, Johnson DS, McClintock BT, Morales JM. Animal movement: Statistical models for telemetry data. CRC press; 2017.

5. Boyce MS, McDonald LL. Relating populations to habitats using resource selection functions. Trends in Ecology & Evolution. 1999;14:268–72.

6. Fortin D, Beyer HL, Boyce MS, Smith DW, Duchesne T, Mao JS. Wolves influence elk movements: Behavior shapes a trophic cascade in Yellowstone National Park. Ecology. 2005;86:1320–30.

7. Thurfjell H, Ciuti S, Boyce MS. Applications of step-selection functions in ecology and conservation. Movement Ecology. 2014;2.

8. Keating KA, Cherry S. Use and interpretation of logistic regression in habitat-selection studies. Journal of Wildlife Management. 2004;68:774–89.

9. Johnson CJ, Nielsen SE, Merrill EH, McDonald TL, Boyce MS. Resource selection functions based on use-availability data: Theoretical motivation and evaluation methods. Journal of Wildlife Management. 2006;70:347–57.

10. Avgar T, Lele SR, Keim JL, Boyce MS. Relative selection strength: Quantifying effect size in habitat-and step-selection inference. Ecology and evolution. 2017;7:5322–30.

11. Chamaille-Jammes S. A reformulation of the selection ratio shed light on resource selection functions and leads to a unified framework for habitat selection studies. bioRxiv. 2019. doi:10.1101/565838.

12. LaPoint S, Gallery P, Wikelski M, Kays R. Animal behavior, cost-based corridor models, and real corridors. Landscape Ecology. 2013;28:1615–30.

13. Avgar T, Potts JR, Lewis MA, Boyce MS. Integrated step selection analysis: Bridging the gap

between resource selection and animal movement. Methods in Ecology and Evolution. 2016;7:619–30.

14. Signer J, Fieberg J, Avgar T. Animal movement tools (amt): R package for managing tracking data and conducting habitat selection analyses. Ecology and Evolution. 2019; July 2018:880–90.

15. Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. John Wiley & Sons; 2013.

16. Manly B, McDonald L, Thomas D, McDonald TL, Erickson WP. Resource selection by animals: Statistical design and analysis for field studies. Springer Science & Business Media; 2002.

17. Aarts G, MacKenzie M, McConnell B, Fedak M, Matthiopoulos J. Estimating space-use and habitat preference from wildlife telemetry data. Ecography. 2008;31:140–60.

18. Warton DI, Shepherd LC. Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. The Annals of Applied Statistics. 2010;4:1383–402.

19. Phillips SJ, Dudík M. Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. Ecography. 2008;31:161–75.

20. Elith J, Phillips SJ, Hastie T, Dudik M, Chee YE, Yates CJ. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions. 2011;17:43–57.

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

22. Millspaugh JJ, Nielson RM, McDONALD L, Marzluff JM, Gitzen RA, Rittenhouse CD, et al. Analysis of resource selection using utilization distributions. The Journal of Wildlife Management. 2006;70:384–95.

23. Aarts G, Fieberg J, Matthiopoulos J. Comparative interpretation of count, presence–absence and point methods for species distribution models. Methods in Ecology and Evolution. 2012;3:177–87.

24. Fithian W, Hastie T. Finite-sample equivalence in statistical models for presence-only data. The Annals of Applied Statistics. 2013;7:1917.

25. Hooten MB, Hanks EM, Johnson DS, Alldredge MW. Reconciling resource utilization and resource selection functions. Journal of Animal Ecology. 2013;82:1146–54.

26. Renner IW, Elith J, Baddeley A, Fithian W, Hastie T, Phillips SJ, et al. Point process models for presence-only analysis. Methods in Ecology and Evolution. 2015;6:366–79.

27. Boyce MS, Vernier PR, Nielsen SE, Schmiegelow FKA, Mark S.Boyce, Pierre R.Vernier, et al. Evaluating resource selection functions. Ecological Modelling. 2002;157:281–300.

28. Lele SR, Merrill EH, Keim J, Boyce MS. Selection, use, choice and occupancy: Clarifying concepts in resource selection studies. Journal of Animal Ecology. 2013;82:1183–91.

29. Cressie N. Statistics for spatial data. Terra Nova. 1992;4:613–7.

30. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019. https://www.R-project.org/.

31. Johnson DS, Thomas DL, Ver Hoef JM, Christ A. A general framework for the analysis of animal resource selection from telemetry data. Biometrics. 2008;64:968–76.

32. Hirzel AH, Le Lay G. Habitat suitability modelling and niche theory. Journal of Applied Ecology. 2008;45:1372–81.

33. Elith J, Leathwick JR. Species distribution models: Ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics. 2009;40:677.

34. Van Winkle W. Comparison of several probabilistic home-range models. The Journal of wildlife management. 1975;118–23.

35. Worton BJ. Kernel methods for estimating the utilization distribution in home-range studies. Ecology. 1989;70:164–8.

36. Signer J, Fieberg J. A fresh look at an old concept: Home-range estimation in a tidy world. bioRxiv. 2020.

37. Johnson DH. The comparison of usage and availability measurements for evaluating resource preference. Ecology. 1980;61:65–71.

38. Dodge S, Bohrer G, Weinzierl R, Davidson SC, Kays R, Douglas D, et al. The environmental-data automated track annotation (env-data) system: Linking animal tracks with environmental data. Movement Ecology. 2013;1:3.

39. Fieberg J, Bohrer G, Davidson SC, Kays R. Short course on analyzing animal tracking data. Presented at the north carolina museum of natural sciences, raleigh, nc, usa. May 21–23, 2018. 2018.

40. University C for IESIN.C.C, CIAT CI de AT. Gridded population of the world, version 3 (gpwv3): Population density grid. 2005. https://doi.org/10.7927/H4XK8CG2.

41. U. S. / Japan ASTER Science Team. ASTER global digital elevation model data set. NASA eosdis land processes daac. 2009.

42. Defourny P, Schouten L, Bartalev S, Bontemps S, Cacetta P, De Wit A, et al. Accuracy assessment of a 300 m global land cover map: The globcover experience. 2009.

43. Kéry M. Introduction to winbugs for ecologists: Bayesian approach to regression, anova, mixed models and related analyses. Academic Press; 2010.

44. Torres LG, Sutton PJ, Thompson DR, Delord K, Weimerskirch H, Sagar PM, et al. Poor transferability of species distribution models for a pelagic predator, the grey petrel, indicates contrasting habitat preferences across ocean basins. PLoS One. 2015;10:e0120014.

45. Yates KL, Bouchet PJ, Caley MJ, Mengersen K, Randin CF, Parnell S, et al. Outstanding challenges in the transferability of ecological models. Trends in ecology & evolution. 2018;33:790–802.

46. Merow C, Smith MJ, Silander JA. A practical guide to maxent for modeling species' distributions: What it does, and why inputs and settings matter. Ecography. 2013;36:1058–69.

47. Fieberg J, Forester JD, Street GM, Johnson DH, ArchMiller AA, Matthiopoulos J. Used-habitat calibration plots: A new procedure for validating species distribution, resource selection, and step-selection models. Ecography. 2018;41:737–52.

48. Matthiopoulos J, Fieberg J, Aarts G, Beyer HL, Morales JM, Haydon DT. Establishing the link between habitat selection and animal population dynamics. Ecological Monographs. 2015;85:413–36.

49. Fleming CH, Calabrese JM, Mueller T, Olson KA, Leimgruber P, Fagan WF. From fine-scale foraging to home ranges: A semivariance approach to identifying movement modes across spatiotemporal scales. The American Naturalist. 2014;183:E154–67.

50. Johnson DS, Hooten MB, Kuhn CE. Estimating animal resource selection from telemetry data using point process models. Journal of Animal Ecology. 2013;82:1155–64.

51. Fieberg J. Kernel density estimators of home range: Smoothing and the autocorrelation red herring. Ecology. 2007;88:1059–66.

52. Fieberg JR, Vitense K, Johnson DH. Resampling-based methods for biologists. PeerJ. 2020;8:e9089.

53. Fieberg J, Rieger RH, Zicus MC, Schildcrout JS. Regression modelling of correlated data in ecology: Subject-specific and population averaged response patterns. Journal of applied ecology. 2009;46:1018–25.

54. Koper N, Manseau M. Generalized estimating equations and generalized linear mixed-effects models for modelling resource selection. Journal of Applied Ecology. 2009;46:590–9.

55. Fieberg J, Matthiopoulos J, Hebblewhite M, Boyce MS, Frair JL. Correlation and studies of habitat selection: Problem, red herring or opportunity? Philosophical Transactions of the Royal Society of London B: Biological Sciences. 2010;365:2233–44.

56. Munden R, Borger L, Wilson RP, Redcliffe J, Brown R, Garel M, et al. Why did the animal turn? Time-varying step selection analysis for inference between observed turning points in high frequency data. bioRxiv. 2020.

57. Arthur SM, Manly BF, McDonald LL, Garner GW. Assessing habitat selection when availability changes. Ecology. 1996;77:215–27.

58. Rhodes JR, Mcalpine CA, Lunney D, Possingham HP, Ecology S, May N. A spatially explicit habitat selection model incorporating. Ecology. 2015;86:1199–205.

59. Avgar T, Mosser A, Brown GS, Fryxell JM. Environmental and individual drivers of animal movement patterns across a wide geographical gradient. Journal of Animal Ecology. 2013;82:96–106.

60. Durante F, Sempi C. Copula theory: An introduction. In: Copula theory and its applications. Springer; 2010. pp. 3–31.

61. Forester JD, Im HK, Rathouz PJ. Accounting for animal movement in estimation of resource

selection functions: Sampling and data analysis. Ecology. 2009;90:3554–65.

62. Prima M-C, Duchesne T, Fortin D. Robust inference from conditional logistic regression applied to movement and habitat selection analysis. PloS one. 2017;12.

63. Prokopenko CM, Boyce MS, Avgar T. Characterizing wildlife behavioural responses to roads using integrated step selection analysis. Journal of Applied Ecology. 2017;54:470–9.

64. Scrafford MA, Avgar T, Heeres R, Boyce MS. Roads elicit negative movement and habitat-selection responses by wolverines (gulo gulo luscus). Behavioral Ecology. 2018;29:534–42.

65. Dickie M, McNay SR, Sutherland GD, Cody M, Avgar T. Corridors or risk? Movement along, and use of, linear features varies predictably among large mammal predator and prey species. Journal of Animal Ecology. 2020;89:623–34.

66. Duchesne T, Fortin D, Rivest L-P. Equivalence between step selection functions and biased correlated random walks for statistical inference on animal movement. PLoS One. 2015;10.

67. Barnett AH, Moorcroft PR. Analytic steady-state space use patterns and rapid computations in mechanistic home range analysis. Journal of mathematical biology. 2008;57:139–59.

68. Signer J, Fieberg J, Avgar T. Estimating utilization distributions from fitted step-selection functions. Ecosphere. 2017;8.

69. Potts JR, Mokross K, Stouffer PC, Lewis MA. Step selection techniques uncover the environmental predictors of space use patterns in flocks of amazonian birds. Ecology and Evolution. 2014;4:4578–88.

70. Potts JR, Bastille-Rousseau G, Murray DL, Schaefer JA, Lewis MA. Predicting local and non-local effects of resources on animal space use using a mechanistic step selection model. Methods in Ecology and Evolution. 2014;5:253–62. doi:10.1111/2041-210X.12150.

71. Potts JR, Schlägel UE. Parametrising diffusion-taxis equations from animal movement trajectories using step selection analysis. Methods in Ecology and Evolution. 2020.

72. Michelot T, Gloaguen P, Blackwell PG, Étienne M-P. The langevin diffusion as a continuous-time model of animal movement and habitat selection. Methods in Ecology and Evolution. 2019;10:1894–

907.

73. Michelot T, Blackwell PG, Matthiopoulos J. Linking resource selection and step selection models for habitat preferences in animals. Ecology. 2019;100.

74. Michelot T, Blackwell PG, Chamaillé-Jammes S, Matthiopoulos J. Inference in mcmc step selection models. Biometrics. 2020;76:438–47.

75. Nicosia A, Duchesne T, Rivest L-P, Fortin D, others. A multi-state conditional logistic regression model for the analysis of animal movement. The Annals of Applied Statistics. 2017;11:1537–60.

76. Suraci JP, Frank LG, Oriol-Cotterill A, Ekwanga S, Williams TM, Wilmers CC. Behavior-specific habitat selection by african lions may promote their persistence in a human-dominated landscape. Ecology. 2019;100:e02644.

77. Murtaugh PA. Simplicity and complexity in ecological data analysis. Ecology. 2007;88:56–62.

78. Craiu RV, Duchesne T, Fortin D, Baillargeon S. Conditional logistic regression with longitudinal follow-up and individual-level random coefficients: A stable and efficient two-step estimation method. Journal of Computational and Graphical Statistics. 2011;20:767–84.

79. Craiu R, Duchesne T, Fortin D, Baillargeon S. TwoStepCLogit: Conditional logistic regression: A two-step estimation method. R package version 1.2. 5. 2016.

80. Muff S, Signer J, Fieberg J. Accounting for individual-specific variation in habitat-selection studies: Efficient estimation of mixed-effects models using bayesian or frequentist computation. Journal of Animal Ecology. 2020;89:80–92.

81. Mysterud A, Ims RA. Functional responses in habitat use: Availability influences relative use in trade-off situations. Ecology. 1998;79:1435–41.

82. Matthiopoulos J, Hebblewhite M, Aarts G, Fieberg J. Generalized functional responses for species distributions. Ecology. 2011;92:583–9.

83. Beest FM van, McLoughlin PD, Mysterud A, Brook RK. Functional responses in habitat selection are density dependent in a large herbivore. Ecography. 2016;39:515–23.

84. Avgar T, Betini GS, Fryxell JM. Habitat selection patterns are density-dependent under the

ideal free distribution. Journal of Animal Ecology. 2020.

85. Fretwell SD. On territorial behavior and other factors influencing habitat distribution in birds. Acta biotheoretica. 1969;19:45–52.

86. Matthiopoulos J, Field C, MacLeod R. Predicting population change from models based on habitat availability and utilization. Proceedings of the Royal Society B. 2019;286:20182911.

Tables

Table 1: Regression coefficients (SE) in fitted resource-selection functions fit to data from Lupe the fisher. Models 1 and 3 use forest as the reference level, Model 2 uses wet as the reference level. Model 3 includes interactions between elevation and landcover classes.

	Model 1	Model 2	Model 3
(Intercept)	-13.168	-12.918	-13.171
	(0.019)	(0.107)	(0.020)
elevation	0.303	0.303	0.313
	(0.017)	(0.017)	(0.017)
popden	-0.183	-0.183	-0.186
	(0.021)	(0.021)	(0.021)
landuseCgrass	-1.477		-1.471
	(0.278)		(0.278)
landuseCwet	0.250		0.183
	(0.108)		(0.116)
landuseC1forest		-0.250	
		(0.108)	
landuseC1grass		-1.727	
		(0.297)	
elevation:landuseCgrass			0.112
			(0.380)
elevation:landuseCwet			-0.498
			(0.127)

Figures

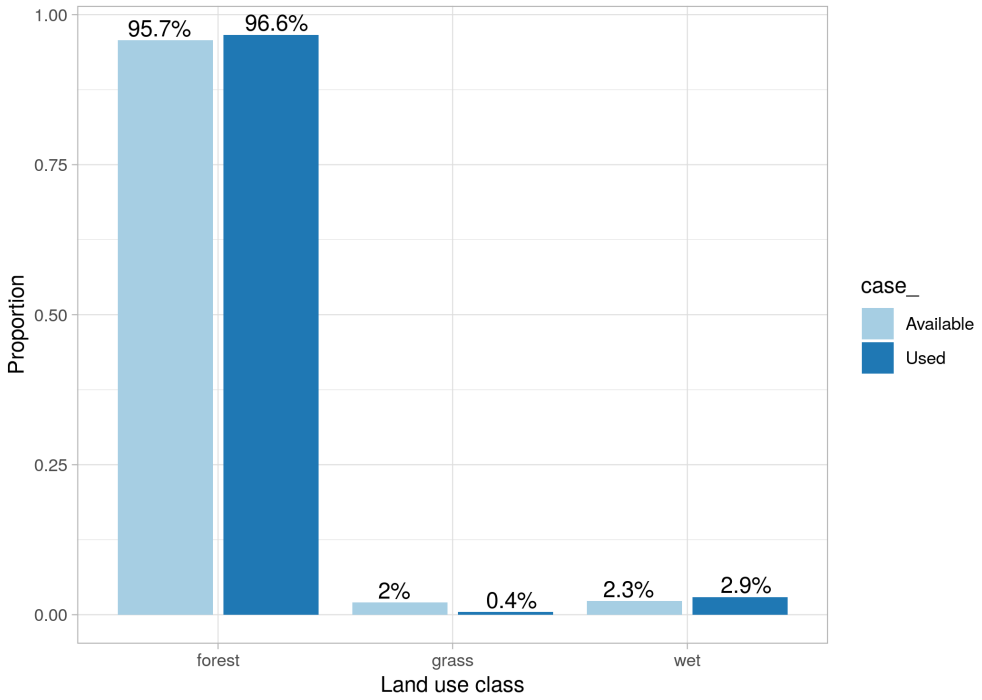

Figure 1: Distribution of used and available locations among different landscape cover classes for a fisher in upstate New York [12].

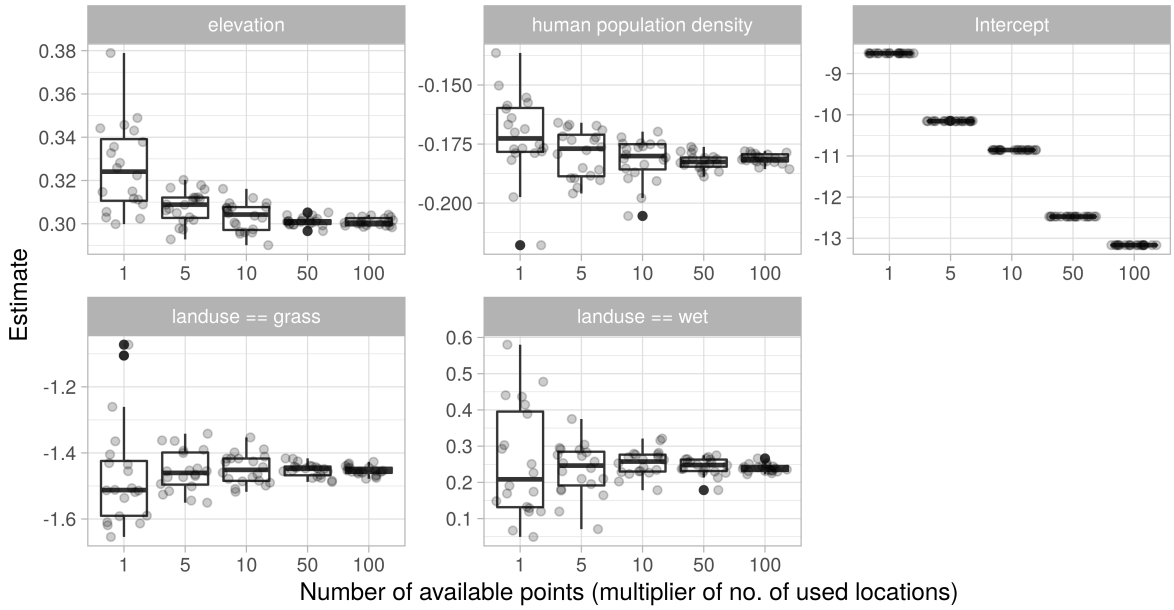

Figure 2: Estimated parameters in fitted resource-selection functions using increasing numbers of available points. Each dot represents an estimate from fitting a logistic regression model to 3004 GPS telemetry locations combined with a random sample of available points, with sample size given by the x-axis (where 1 means 3004 available points and 100 means 300,400 available points).

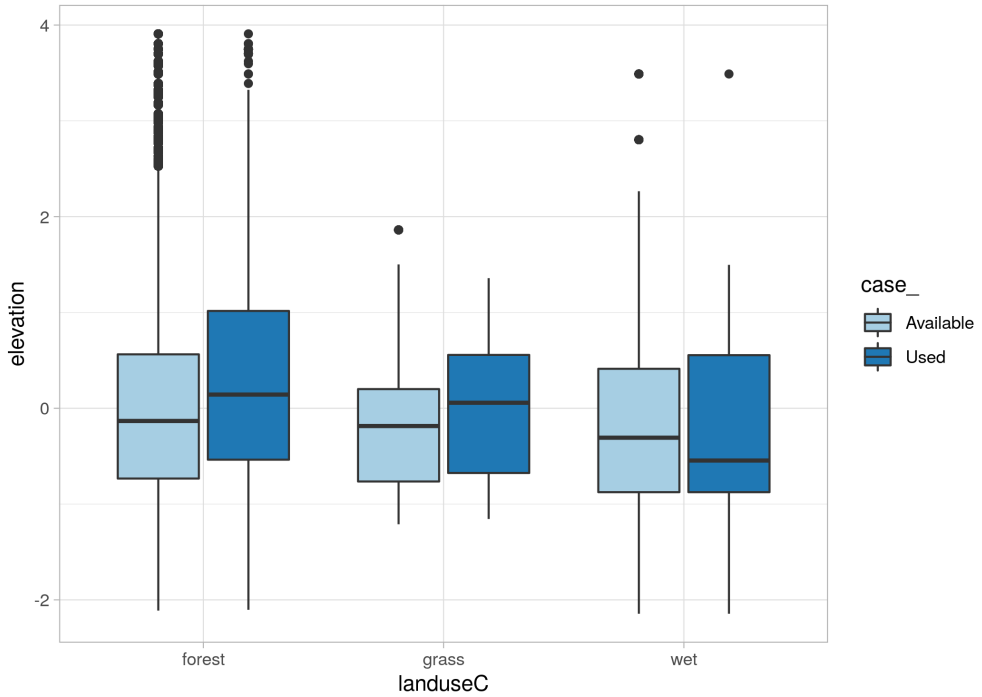

Figure 3: Distribution of elevation at used and available locations within each of 3 landcover types.

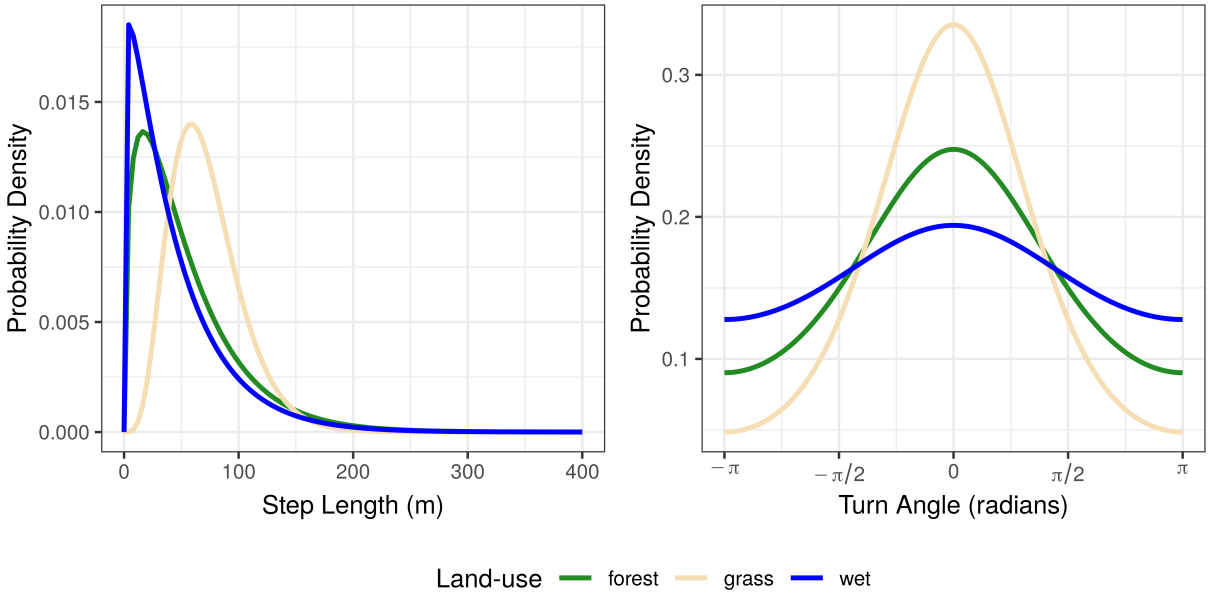

Figure 4: Step-length and turn-angle distributions from an integrated step-selection analysis applied to Lupe's location data (see Additional File 2). The conditional logistic regression model included interactions between movement characteristics (step length, log step length, and cosine of of the turn angle) and the landuse category Lupe was in at the start of the movement step. We see that Lupe tends to take larger, more directed steps when in `grass` and slower and more tortuous steps in `wet` habitat.