

Habitat Selection

Brian D. Gerber

November 5, 2024

1 Habitat Selection with independent spatial locations

Habitat selection modeling has a long history. It is also fairly confusing as the terminology and understanding of the statistical modeling has evolved.

A common habitat selection model will use animal-borne telemetry location data. We will consider these spatially locations independent, such that consecutive locations do not depend on each other. A major part of this dependence is due to movement constraints of the animal. If the locations are independent, there has been a long enough time between locations that the animal could have accessed any part of its range. The important word here is **could**, rather than did. Essentially, we are interested in these independent behavioral decisions of where the animal chooses to be located. As part of this type of model, we need to decide on all the locations on a landscape that are available to the animal to choose from. The boundary of this available area is often chosen to be the home range of the individual. It could also be a larger landscape though.

The simplest habitat selection model is one where we assume all locations are equally available to the individual animal within a defined area (e.g. home range). In the literature, this is the ‘Traditional Resource Selection Function’. We want to estimate whether animals are using certain landscape/spatial features in this area more than they are available to the animal (selection) or whether they are using these them less than they are available (avoidance). The statistical framework that combines use, selection, and availability is the weighted distribution formulation of a point process model (Hooten et al. 2017).

The components:

- $\boldsymbol{\mu}_i$ is a 2 x 1 vector of the animal’s true geographic coordinate for relocation i
- $\mathbf{x}(\boldsymbol{\mu}_i)$ are the environmental/spatial features hypothesized to influence animal selection at relocation $\boldsymbol{\mu}_i$.
- the function $g()$ is called the selection function and depends on $\boldsymbol{\beta}$
- $\boldsymbol{\beta}$ are the associated selection coefficients associated with $\mathbf{x}(\boldsymbol{\mu}_i)$

- the function $f()$ is called the availability function and depends on θ
- θ are the associated availability coefficients

To put this together, we can define the probability density function of the locations (μ_i) as,

$$[\mu_i|\beta, \theta] \equiv \frac{g(\mathbf{x}(\mu_i, \beta))f(\mu_i, \theta)}{\int g(\mathbf{x}(\mu, \beta))f(\mu, \theta)d\mu}.$$

However, we stated that we will make all locations and associated environmental/spatial features equally available to the individual within a defined area (called the support of the point process, \mathcal{M}). Therefore, the availability function is uniform and be removed as it would be equivalent to multiplying the selection function for each location i by one. We can then simplify our model to

$$[\mu_i|\beta] \equiv \frac{g(\mathbf{x}(\mu_i, \beta))}{\int g(\mathbf{x}(\mu, \beta))d\mu}.$$

But what is this g function? Generally, it can be any deterministic mathematical function that has positive support. Specifically though, it is commonly defined as exponential, $g() = \exp()$. Therefore, we can define our final model as,

$$[\mu_i|\beta] \equiv \frac{\exp(\mathbf{x}'(\mu_i)\beta)}{\int \exp(\mathbf{x}'(\mu)\beta)d\mu}.$$

Lastly, lets connect this model across N independent spatial points and define the complete likelihood as,

$$\prod_{i=1}^N \frac{\exp(\mathbf{x}'(\mu_i)\beta)}{\int \exp(\mathbf{x}'(\mu)\beta)d\mu}.$$

1.1 How do we fit this model?

It happens that we can fit this model with logistic regression.

The main idea in the logistic regression approach is to convert the individual spatial locations into a binary data set where the observed locations are represented by ones and the available locations are represented by zeros. That is, a background sample is typically taken from the availability distribution f in such a way that it represents a large but finite set of possible locations the individual could have occupied. The environmental covariates (\mathbf{x}_i) at the individual locations (μ_i) are associated with each of the ones for $i = 1, \dots, N$, and similarly, the covariates are recorded for the background sample of $M - N$ locations. The response variable is then specified as $y \equiv (1, \dots, 1, 0, \dots, 0)'$ and used in a standard

logistic regression with the complete set of covariates. That is, the fitting model becomes $y_i \sim \text{Bern}(p_i)$ where $\text{logit}(p_i) = \mathbf{x}'\boldsymbol{\beta}$ for $i = 1, \dots, m$ total binary observations.

The background sample (the zeros of length M) is what is providing the approximation of the integral in the above formulas (i.e., denominator). It is important we have enough locations of zeros to make sure this integral is well approximated (Northrup et al. 2013). The two ways we can ensure this are that we obtain many locations in the spatial areas of interest (\mathcal{M}) and assign them the response of zero. The second thing we can do is weight our observations in the logistic regression model. In R and in the function `glm`, we can do this with `weight` argument. Specifically, we want a vector that corresponds to y , where we assign a 1 for each value of 1 in y and a 1000 (or large number) to each value of zero in y . To test whether we have enough of a background sample, we can increase the number of zeros until the estimated coefficients ($\hat{\boldsymbol{\beta}}$) no longer change.

2 References

Hooten, M. B., Johnson, D. S., McClintock, B. T., & Morales, J. M. (2017). *Animal movement: statistical models for telemetry data*. CRC press.

Northrup, J. M., Hooten, M. B., Anderson Jr, C. R., & Wittemyer, G. (2013). Practical guidance on characterizing availability in resource selection functions under a use-availability design. *Ecology*, 94(7), 1456-1463.

3 Habitat Selection with temporally dependent spatial locations

Modern tracking technology allows us to collect high frequency relocation data on individuals. Therefore, for many species and relocation intervals (time between telemetry fixes/relocations) what is available to the animal is constrained by how the animal moves (e.g., speed and direction). Ignoring this constraint and assuming the data are independent will lead to pseudoreplication and thus unreliable inference. As such, we need to build this movement constraint into our habitat selection model, thereby connecting the spatial process model with a movement model. We can depict this model by defining the probability density function of the relocations and the availability distribution in terms of relocations from $i-1$ to i and the time between these (Δ_i). For example, we can generally specify it as,

$$[\boldsymbol{\mu}_i | \boldsymbol{\mu}_{i-1}, \boldsymbol{\beta}, \boldsymbol{\theta}] \equiv \frac{g(\mathbf{x}(\boldsymbol{\mu}_i, \boldsymbol{\beta}))f(\boldsymbol{\mu}_i | \boldsymbol{\mu}_{i-1}, \Delta_i, \boldsymbol{\theta})}{\int g(\mathbf{x}(\boldsymbol{\mu}_i, \boldsymbol{\beta}))f(\boldsymbol{\mu}_i | \boldsymbol{\mu}_{i-1}, \Delta_i, \boldsymbol{\theta})d\boldsymbol{\mu}}.$$

3.1 How do we fit this model?

Commonly, this model is not fit directly, but rather we setup the data to create occurrence (1) and available (0) samples for each step of a movement process. This is a movement-based habitat selection function or a step-selection function.