Appendix 1 – Explanation of the methods (plain language)

Modelling habitat selection using tracking data from central place foraging species:
A practical guide for ecologists

***

Habitat selection is fundamentally about how animals use their environment *relative* to what is available to them, rather than about where they occur in absolute terms. Just as a diner's restaurant choice depends on which venues are open and within commuting distance, an animal can only "prefer" a habitat if it had multiple habitats to choose from and selected some over, or more often than, others.

A simple way to express this is to write:

$$\text{Selection} = \frac{\text{Use}}{\text{Availability}}$$

Here, "use" refers to the locations where animals were recorded, whereas "availability" refers to the parts of the landscape they could have realistically accessed. Because availability is almost never known exactly, it must be approximated statistically.

Resource selection functions (RSFs)

Much of the early habitat selection literature used resource selection functions (RSFs) to carry out this approximation. RSFs are statistical models that compare the habitat characteristics (or environmental conditions) at used locations with those at a set of randomly chosen "available" locations (also known as "controls", "pseudo-absences", "quadrature", or "background points"). Typically, it is up to the analyst to decide how many available points should be used and where these should be placed. Because of this, the absolute probability of a habitat being chosen cannot be determined and RSFs only describe *relative* preference. To see why, consider a simple example where:

- *An animal truly prefers habitat A over habitat B.*
- *In the tracking data recorded from that animal, we observe 80 locations that fall within habitat A and 20 locations that fall within habitat B.*

Suppose that we generate 100 available points, which for simplicity fall in each habitat in equal proportions (i.e., 50 in habitat A and 50 in habitat B). The RSF ratios are therefore:

- *Habitat A: 80 / 50 = 1.6*
- *Habitat B: 20 / 50 = 0.4*

This suggests that habitat A is selected four times more strongly than habitat B (as 1.6 / 0.4 = 4). If we repeat the same analysis with 10,000 available points instead of 100 (so that 5,000 fall in each habitat), the ratios now become:

- *Habitat A: 80 / 5000 = 0.016*
- *Habitat B: 20 / 5000 = 0.004*

The numerical values (i.e., absolute probabilities of selection) have drastically changed, even if the behaviour of the animal is identical. However, habitat A is still selected four times more often than habitat B (0.016 / 0.004 = 4), so the ranking of habitats remains the same.

RSFs often assume that relative selection varies in response to the combined effects of multiple habitat covariates, each contributing positively or negatively depending on its estimated coefficient. In other words, habitats become more or less attractive depending on the combined influence of the environmental variables included in the model:

Relative selection = effect of habitat 1 + effect of habitat 2 + effect of habitat 3 + ... etc.

This is the same structure used across many statistical models.

In practice, RSFs are commonly estimated using a form of regression model called case-control logistic regression, where used points ("cases") are coded as one and available points ("controls") as zero. For statistical reasons, logistic regression works on a transformed scale called the log-odds scale, but the fitted coefficients still correspond directly to the relative selection strengths in the RSF. The intercept has no biological meaning because it depends entirely on the ratio of used to available points, which is subjective.

Inhomogeneous Poisson point process (IPP)

To avoid arbitrary choices about the number and placement of available points, more sophisticated statistical frameworks have been proposed that model availability directly in continuous space. The simplest and most widely used of these is the **inhomogeneous Poisson point process (IPP)**, which models data arising from a spatial point process. A *spatial point process* is a statistical model of the distribution of points scattered across space — for example, nest locations, sightings, or GPS fixes. Instead of modelling the presence or absence of a species at a set of fixed sites, a point process treats all observed points themselves as the data. The expected density of observations across different parts of the landscape is described by what is called an *intensity* surface, which is typically labelled as λ in the literature. Put simply, this intensity gives the expected number of points per unit area and broadly speaking, a higher intensity value at a location means that that location is more likely to contain observations.

In an IPP, this intensity varies across space depending on habitat conditions:

$\log(\lambda)$ at location s = baseline density + habitat effects

The baseline term represents the expected density of points (on the log scale) in the absence of habitat effects, or under average habitat conditions if covariate values have been mean-centred.

In short:
The IPP is useful when the goal is to estimate how the expected density of points varies across a landscape while avoiding the arbitrary definition of available habitat required by RSFs.

IPP approximations (gridded, down-weighted, and infinitely weighted Poisson)

Because the IPP describes a continuous landscape, we need a way to approximate that landscape using a finite number of calculations. A practical solution is to divide the study region into many small grid cells and treat the number of points in each of those cells as a Poisson count. A Poisson count describes how many events occur within an area when events are relatively rare and independent. In this context, it represents the number of observed locations falling inside each grid cell, given the expected number of points for that cell. This expected count comes from the intensity surface, and the Poisson distribution captures the natural variation around that expectation.

For each grid cell, the expected number of points is given as:

Expected count in a cell = surface area of the cell x intensity in that cell

Taking the logarithm gives:

log (expected count) = log (cell area) + log(intensity)

This is known as a **gridded Poisson approximation**. The smaller the grid cells, the more closely the results of this approximation will match those of the continuous-space IPP.

It is worth noting that the IPP can be approximated in other ways that use background points instead of grid cells. Those include:

- **Down-weighted Poisson regression,** whereby each background point is viewed as a tiny portion of the landscape. Because many such points are used, each one receives only a small weight (hence the term 'down-weighted'). Adding up all these weighted points approximates the full amount of space available across the study area. An intuitive way to think about this is that each point in down-weighted Poisson regression is like a small tile in a mosaic: every tile contributes a little, and together they reconstruct the whole picture.
- **Infinitely weighted logistic regression,** in which background points are instead treated as a whole sample drawn from the entire available area. As the sample becomes larger, the model places increasing emphasis on it so that it provides an increasingly accurate representation of the available landscape it is meant to represent.

While their underlying mathematics differ, both methods use the same ingredients and converge on the same IPP as the number of background points grows; both therefore yield the same habitat selection coefficients and ecological interpretation. The main difference is that infinitely weighted logistic regression only indicates which habitats are preferred relative to others, whereas down-weighted Poisson regression can additionally estimate the expected density of observations across the landscape. This is because the former is based on the same logistic model as an RSF, whereas the latter uses a Poisson formulation that explicitly accounts for the amount of available space. Typically, choosing between the two is however more a question of computational convenience or software availability than biological reasoning.

However, a key assumption of the IPP is that habitat alone explains why some areas are used more than others. After accounting for the effects of habitat in the model, any remaining clustering of locations is assumed to be random. In reality, ecological data often show *additional* clustering because animals revisit the same places over time, interact socially, aggregate as part of key behaviours (e.g., swarming, flocking) or simply respond to (unmeasured) features of the environmental which exhibit spatial patterns or are unevenly distributed across space (e.g., prey occur in patches, temperature varies along gradients, etc.).

Log-Gaussian Cox processes (LGCPs)

Log-Gaussian Cox process (LGCP) models address this by including an additional term called a *spatial random effect*, which takes the form of a smooth surface that increases or decreases λ locally, allowing the model to separate habitat-driven patterns from residual clustering. LGCP models are therefore direct extensions of the IPP:

LGCP intensity = habitat effects (as in the IPP) + unexplained spatial structure (via random effect)

In short: LGCPs are useful when habitat variables do not fully explain spatial patterns in the data.

Space-time point process models (STPPs)

The models above work well for data collected at fixed sites (e.g., camera-traps, acoustic sensors, opportunistic sightings), but tracking data pose additional challenges. In particular, successive GPS locations from the same individual tend to be strongly autocorrelated as they represent repeated snapshots of the animal moving continuously. Ignoring the way in which animals move can cause movement-driven clustering to be mistaken for habitat selection:

- Distant habitats may appear "avoided" simply because they were never reachable.
- Habitats near the animal's track may appear strongly "selected" because they were always accessible.

In other words, availability is dynamic and changes through time. Space-time point process models (STPPs) address this by defining availability as the set of locations an animal could realistically reach during each time interval.

The intensity of an STPP is the product of

1. A habitat-selection component, and
2. A movement-based availability component.

In an STPP:

Intensity at location s and time t = habitat effects at location s and time t x movement-based availability at location s and time t

The movement-based availability term is defined by a formal movement model and depends on:

- the previous location,
- the time since the last observation,
- and the animal's movement behaviour.

Different movement models (e.g., Brownian motion, ecological diffusion, Ornstein–Uhlenbeck) produce different shapes for this availability surface. STPPs allow fine-scale inference because they link each data point to the exact environmental conditions prevailing at the time of observation; however, they tend to be computationally expensive as the movement component must be evaluated at every time step.

In short:
STPPs are useful when availability changes through time because animals can only move limited distances between observations.

Marginalised space-time point processes (mSTPPs)

To reduce computational burden and improve generalisability, marginalised STPPs (mSTPPs) average availability over time. In an mSTPP:

Intensity at location s = habitat effects x time-averaged availability

The time-averaged availability is computed by summing the movement-based availability surfaces from all movement intervals. This approach:

- removes the need to evaluate a likelihood at every time step
- broadens the range of environmental conditions represented
- and improves transferability to new areas or time periods

In short:
mSTPPs are useful when the goal is broad-scale habitat selection inference while still accounting for movement constraints.

Conclusion

The progression of models described in this paper reflects a gradual relaxation of simplifying assumptions about how animals move and interact with their environment. Early models treated space as freely accessible, whereas more recent models recognise that animals experience their environment step by step along their paths. All of these methods are fundamentally trying to estimate an intensity surface, and they differ mainly in how they represent availability and account for spatial or temporal dependence.

In short:

- RSF = habitat selection relative to availability
- IPP = habitat selection expressed as an intensity surface
- LGCP = IPP + residual spatial structure
- STPP = IPP + movement-based availability
- mSTPP = IPP + time-averaged availability