
Article:
Michelot, T., Gloaguen, P., Blackwell, P.G. orcid.org/0000-0002-3141-4914 et al. (1 more author) (2019) The Langevin diffusion as a continuous-time model of animal movement and habitat selection. Methods in Ecology and Evolution, 10 (11). pp. 1894-1907. ISSN 2041-210X

https://doi.org/10.1111/2041-210x.13275

The Langevin diffusion as a continuous-time model of animal movement and habitat selection

Théo Michelot[1,2]*, Pierre Gloaguen[3], Marie-Pierre Étienne[4], Paul G. Blackwell[1]

[1]*University of Sheffield* , [2]*University of St Andrews*, [3]*AgroParisTech*, [4]*AgroCampus Ouest*

Abstract

1. The utilisation distribution of an animal describes the relative probability of space use. It is natural to think of it as the long-term consequence of the animal's short-term movement decisions: it is the accumulation of small displacements which, over time, gives rise to global patterns of space use. However, many estimation methods for the utilisation distribution either assume the independence of observed locations and ignore the underlying movement (e.g. kernel density estimation), or are based on simple Brownian motion movement rules (e.g. Brownian bridges).
2. We introduce a new continuous-time model of animal movement, based on the Langevin diffusion. This stochastic process has an explicit stationary distribution, conceptually analogous to the idea of the utilisation distribution, and thus provides an intuitive framework to integrate movement and space use. We model the stationary (utilisation) distribution with a resource selection function to link the movement to spatial covariates, and allow inference about habitat preferences of animals.
3. Standard approximation techniques can be used to derive the pseudo-likelihood of the Langevin diffusion movement model, and to estimate habitat preference and movement parameters from tracking data. We investigate the performance of the method on simulated data, and discuss its sensitivity to the time scale of the sampling. We present an example of its application to tracking data of Steller sea lions (*Eumetopias jubatus*).
4. Due to its continuous-time formulation, this method can be applied to irregular telemetry data. The movement model is specified using a habitat-dependent utilisation distribution, and it provides a rigorous framework to estimate long-term habitat selection from correlated movement data. The Langevin movement model can be written as a linear model, which allows for very fast inference. Standard tools such as residuals can be used for model checking.

Keywords: animal movement, continuous time, resource selection, step selection, Langevin diffusion, potential function, utilisation distribution

1 Introduction

A crucial concept in animal ecology is the utilisation distribution, "the probability density function that gives the probability of finding an animal at a particular location" (Anderson, 1982). In recent decades, improvements in tracking technologies have produced large amounts of animal location data, at a high

*Corresponding author: tm75@st-andrews.ac.uk

spatio-temporal resolution. Statistical methods have been developed to estimate the utilisation distribution from telemetry observations, and to link the movements of individual animals to habitat preferences and space use (Hooten et al., 2017).

In those approaches, a (generally two-dimensional) density function is estimated. It is of particular interest for ecological conservation to relate the utilisation distribution to environmental covariates, to understand how animals use space in response to their habitat (Millspaugh et al., 2006; Long et al., 2009; Nielson and Sawyer, 2013; Zhang et al., 2014). For this purpose, the utilisation function can be formulated in terms of spatial covariates of interest, typically using a resource selection function (Manly et al., 2002). A resource selection function links the distribution of observed locations of animals to the distribution of resources (or other spatial covariates), to infer habitat characteristics that are preferred (or "selected") by the animals. It is based on the idea that, knowing the habitat composition of a spatial unit, we can predict its long-term utilisation. However, resource selection functions rely on the assumption that telemetry observations are independent, which is unrealistic for high-frequency movement data.

Other popular approaches to estimate the utilisation distribution from tracking data include empirical histograms (Nielson and Sawyer, 2013), kernel density estimators (Anderson, 1982; Worton, 1989), and Brownian bridges (Horne et al., 2007; Kranstauber et al., 2012; Fleming et al., 2016). Similarly to resource selection functions, a limitation of such methods is that the estimation of the utilisation distribution is disconnected from the movement of the animal. Indeed, they often ignore the sequential structure of the data (Anderson, 1982; Worton, 1989; Nielson and Sawyer, 2013), or make unrealistic Brownian assumptions about the movement (Horne et al., 2007; Kranstauber et al., 2012), although see Fleming et al. (2015) for a kernel density estimator that corrects for the autocorrelation in animal telemetry data. Those models of space use do not estimate the utilisation distribution as a function of covariates, and two-stage approaches are required to link space use to habitat preferences (Millspaugh et al., 2006; Péron, 2019).

It is natural to think of the utilisation distribution as a consequence of the movement, which itself depends on the environment. Short-term movement decisions, based on habitat selection, give rise to long-term space use. This idea motivates the development of more mechanistic approaches that link the animal's movement to its environment, and, ultimately, to an explicit steady-state distribution, representing the utilisation distribution.

Following this idea, step selection functions model the likelihood of a step between two points in space as a combination of a movement kernel and a habitat selection function (Fortin et al., 2005; Forester et al., 2009; Thurfjell et al., 2014). The parameters of a step selection function describe preference at a local (step-by-step) scale, and strongly depend on the temporal scale of the data, and on the choice of the movement kernel. As such, their parameters cannot directly be linked to global space use. Recently, numerical methods have been proposed to approximate the utilisation distribution underlying a step selection function model. In particular, Potts et al. (2014) derived an equation for the evolution of the distribution of an animal's location in a step selection function model which, when simulated forward, converges to the utilisation distribution. Similarly, Avgar et al. (2016) and Signer et al. (2017) suggested that simulations from a fitted step selection function (as implemented by Signer et al., 2019) can be used to obtain its steady-state distribution. These methods are useful to derive long-term space use from short-term habitat selection, but the utilisation distribution cannot be expressed as a simple parametric function of the spatial covariates.

Hanks et al. (2015) proposed a continuous-time discrete-space model to link movement to environmental drivers. In their framework, the movement is considered as a continuous-time Markov process on a discrete grid of spatial cells. The spatial grid is usually chosen as the grid on which the spatial covariates are measured, and the observed locations are binned in the cells. Wilson et al. (2018) argued that the limiting distribution of that movement model can be interpreted as a utilisation distribution, and proposed a method

to estimate it on a discrete grid. A drawback of that approach is that it describes movement on a discrete spatial grid, and its formulation is therefore tied to a particular space discretization.

Recently, Michelot et al. (2018) proposed a step selection model, formulated in terms of an explicit utilisation distribution. Their approach describes individual movement as a Markov chain in continuous space, whose stationary distribution is the utilisation distribution. In particular, they suggest that Markov chain Monte Carlo (MCMC) algorithms, which are used to construct Markov chains with a given stationary distribution, can be viewed as movement models.

Others have described the position of an individual animal as a diffusion process which follows the gradient of a potential surface (Brillinger, 2010; Preisler et al., 2013; Gloaguen et al., 2018). The surface measures the potential interest for the individual, and it can be linked to habitat variables. These approaches offer a wide variety of flexible models to describe movement, but their link to the utilisation distribution is unclear in the existing literature. Indeed, potential-based models are often based on diffusion processes that are not stationary (Gloaguen et al., 2018), or lead to unrealistically simple utilisation distributions. For example, the stationary distribution is uniform over the study region for Brownian motion movement models (Skellam, 1951), and it is a normal distribution for the Ornstein-Uhlenbeck model (Blackwell, 1997). Including behavioural switching, with movement parameters that depend on behavioural state, gives more flexibility in the resulting utilisation distribution, but the details depend on the relationship between movement and behavioural parameters, and are not straightforward to interpret (Blackwell, 1997; Harris and Blackwell, 2013).

In this work, we describe a new mechanistic movement model, continuous in time and space. We model the animal's position as a diffusion process with a drift towards the gradient of its stationary (utilisation) distribution, bringing together the ideas of Brillinger (2010) and Michelot et al. (2018). As in Wilson et al. (2018), the limiting distribution of the process is the utilisation distribution. The movement model that we propose is based on the Langevin diffusion, which has also been used to construct an MCMC algorithm (Roberts and Rosenthal, 1998). As this model belongs to the class of potential-based models, inference can be performed from movement data using different estimation methods for stochastic differential equations (SDEs), such as pseudo-likelihood methods which are simple to implement (Gloaguen et al., 2018). We show here how this parametric model can also be linked to step selection approaches when the utilisation distribution is parameterized as a simple function of environmental covariates. Point estimators and confidence intervals of habitat selection parameters can easily be derived in a classical approximated inference framework.

In Section 2, the proposed movement model is formulated in its general form, and we explain how it can be used to model habitat selection. Section 3 describes a pseudo-likelihood method based on the Euler discretization scheme, to estimate the habitat selection parameters from telemetry data. In Section 4, we assess the performance of the inference methods in simulations, and we discuss conditions under which the model parameters can be recovered. In Section 5, we present the analysis of three trajectories of Steller sea lions (*Eumetopias jubatus*), with four environmental covariates as potential drivers of their movement.

2 Langevin movement model

2.1 General formulation

We denote by $\boldsymbol{X}_t \in \mathbb{R}^d$ the location of an individual animal in d-dimensional space at time $t \geq 0$, and $\pi : \mathbb{R}^d \to \mathbb{R}$ its utilisation distribution (Worton, 1989). In a steady-state regime, the utilisation distribution

is the probability density function π which satisfies

$$\mathbb{P}(\boldsymbol{X}_t \in A) = \int_A \pi(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}, \tag{1}$$

for any area $A \subset \mathbb{R}^d$. The two-dimensional case ($d = 2$) is by far the most common in movement ecology, although the framework works for any value of d.

We propose to describe the continuous-time location process of the animal $(\boldsymbol{X}_t)_{t \geq 0}$ with a Langevin diffusion for the density π, defined as the solution to the stochastic differential equation

$$\mathrm{d}\boldsymbol{X}_t = \frac{1}{2} \nabla \log \pi(\boldsymbol{X}_t) \mathrm{d}t + \mathrm{d}\boldsymbol{W}_t, , \quad \boldsymbol{X}_0 = \boldsymbol{x}_0. \tag{2}$$

where $\boldsymbol{W}_t$ stands for a d-dimensional standard Brownian motion, ∇ is the gradient operator, and with initial condition $\boldsymbol{X}_0 = \boldsymbol{x}_0$. Under some easily-satisfied technical conditions (that can be found in Dalalyan, 2017), Equation 2 has a unique solution. Crucially, the solution is a continuous-time Markov process with stationary distribution π, as defined in Equation 1 (Roberts and Tweedie, 1996). The Langevin diffusion is thus a natural choice to link a continuous-time model of animal movement with a steady-state distribution. Indeed, the process describes the animal's movements as the combination of a drift towards higher values of its utilisation distribution π (informed by the gradient of $\log \pi$), and a random component given by the Brownian motion.

In its simplest formulation, however, the Langevin diffusion cannot readily be used to model animal movement. Indeed, the speed of the process described above is only determined by the shape of the underlying utilisation distribution π, whereas it should be possible for two individuals to move at different speeds on the same long-term distribution of space use. A similar issue arises in an MCMC context, where Roberts and Rosenthal (1998) were interested in a more flexible class of Langevin-based algorithms to improve performance. To allow for this flexibility, following Roberts and Rosenthal (1998), we introduce an additional parameter γ^2 and we define the Langevin movement model (with speed) as the solution to

$$\mathrm{d}\boldsymbol{X}_t = \frac{\gamma^2}{2} \nabla \log \pi(\boldsymbol{X}_t) \mathrm{d}t + \gamma \mathrm{d}\boldsymbol{W}_t, \quad \boldsymbol{X}_0 = \boldsymbol{x}_0. \tag{3}$$

Note that if the solution to Equation 2 is denoted by $\boldsymbol{X}^*_t$ and the solution to Equation 3 by $\boldsymbol{X}_t$ then they are related by $\boldsymbol{X}_t = \boldsymbol{X}^*_{\gamma^2 t}$. In the following, γ^2 will be referred to as the speed parameter. We generally specify the model in terms of γ^2 (rather than γ) because it has a direct interpretation as the variance parameter of the random Brownian motion component. Xifara et al. (2014) described an even more general process, replacing the speed parameter in Equation 3 by a matrix. They showed that, in that case too, the stationary distribution of the process is π. It should be noted that, although we call γ^2 the speed parameter, the speed of the process described in Equation 3 also depends on the amplitude of the local gradient of the target distribution π. The process will tend to move more slowly in areas where π is flat than where it is steep.

Figure 1 shows two tracks simulated from the Langevin movement model on an artificial utilisation distribution, for two different values of γ^2. Although the two tracks explore space at very different speeds, they have the same equilibrium distribution.

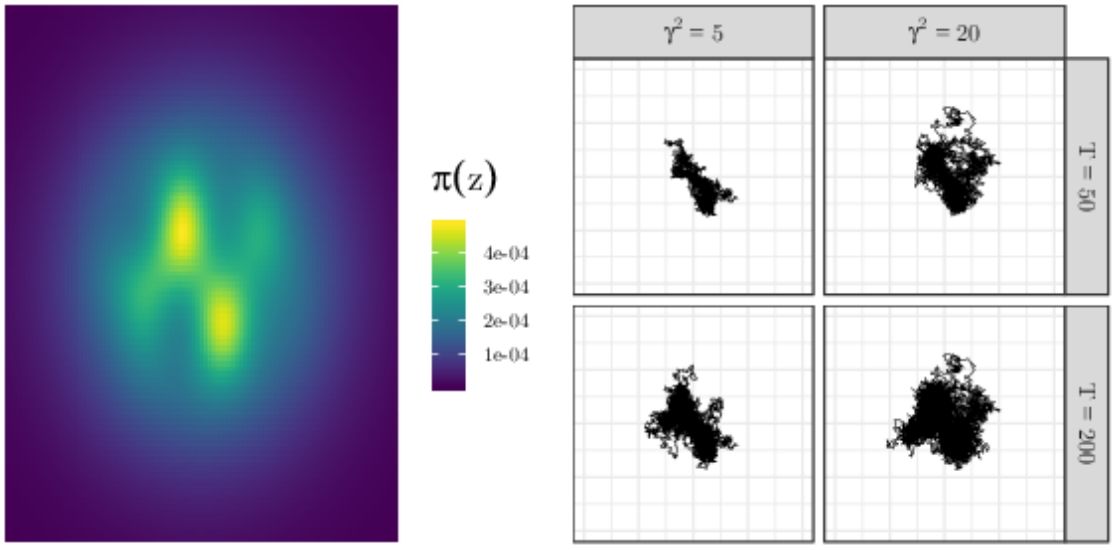

Figure 1: *Left: Artificial utilisation distribution π. Right: Trajectories simulated from the Langevin movement model on π, with two different values of the speed parameter γ^2 (5 and 20), after $T = 50$ and $T = 200$ time units. Although the process with $\gamma^2 = 5$ is much slower to explore space, the properties of the Langevin equation guarantee that both processes have the same stationary distribution π.*

2.2 Including covariates

We link the utilisation distribution of the individual to spatial covariates with the standard parametric form of resource selection functions (RSF),

$$\pi(\boldsymbol{x}|\boldsymbol{\beta}) = \frac{\exp\left(\sum_{j=1}^{J} \beta_j c_j(\boldsymbol{x})\right)}{\int_\Omega \exp\left(\sum_{j=1}^{J} \beta_j c_j(\boldsymbol{z})\right) d\boldsymbol{z}}, \tag{4}$$

where $c_j(\boldsymbol{x})$ is the value of the j-th covariate at location $\boldsymbol{x}$, $\Omega \subset \mathbb{R}^d$ is the study region, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)'$ is a vector of unknown parameters. The value of β_j measures the strength of the selection (attraction or avoidance) for the j-th covariate. The denominator in the right-hand side of Equation 4 is a normalizing constant, and is necessary to ensure that $\pi(\boldsymbol{x}|\boldsymbol{\beta})$ is a probability density function with respect to $\boldsymbol{x}$.

We consider the Langevin diffusion process with the target distribution π given in Equation 4. This defines a continuous-time movement model, such that the stationary (utilisation) distribution of the process is a normalized RSF. In this approach, the movement of an animal is modelled in response to the environmental features in its surroundings. At any instant, the animal tends to move towards better habitat, i.e. in the direction of the gradient of the RSF. This is formulated in continuous time, unlike other models of local habitat selection such as step selection functions (Forester et al., 2009) or the MCMC movement model of Michelot et al. (2018). Those models describe habitat selection at the scale of the time step of observations, whereas the Langevin movement model captures continuous-time habitat selection, independently of the time step of observations. In this respect, the approach we propose is similar to methods based on potential functions (Brillinger, 2010). In potential-based models, the movement process is also formulated in continuous time, and it is affected by the shape of the potential function in its surroundings. Many potential-based movement models are not stationary, so they do not capture long-term utilisation. However, Preisler et al. (2013) described assumptions under which a potential-based model is stationary, and for which the station-

ary distribution of the movement process can be derived from the potential function. The approach that we present can be seen as an special case of the model of Preisler et al. (2013), and we model the stationary distribution as a function of spatial covariates.

Note that Equation 3 requires $\log \pi$ to be a smooth function, i.e. with continuous first-order partial derivatives. If π is modelled by a resource selection function (Equation 4), then

$$\nabla \log \pi(\boldsymbol{x}|\boldsymbol{\beta}) = \sum_{j=1}^{J} \beta_j \nabla c_j(\boldsymbol{x}). \tag{5}$$

Therefore, it is supposed here that all covariates c_j are differentiable, and that their gradients are continuous at each point $\boldsymbol{x}$ and can be computed, either analytically, or by numerical approximation. In most real data sets, the covariate functions c_j are measured at discrete points in space. There is generally no analytical form for the gradient, and it is necessary to interpolate the covariate fields so that its gradient can be approximated. In Sections 4.2 and 5, bilinear interpolation is considered to obtain continuous covariate functions. In the special case of bilinear interpolation, the gradient can be derived analytically, which greatly speeds up the computations (Appendix D).

As a consequence of the interpolation, the Langevin movement model is not well suited to discrete or categorical covariates. While such a covariate field can be interpolated into a continuous function—using dummy indicator variables corresponding to the levels of a categorical covariate—its gradient will be zero except between points with different levels of the covariate where the value will only depend on the chosen interpolation method. Thus over much of the space, the movement model will simply be Brownian motion, and the utilisation distribution will be uniform. This issue is further explored in Section 6.

3 Inference

The continuous-time location process $(\boldsymbol{X}_t)_{t\geq 0}$ of the individual is observed discretely at times $t_0 < t_1 < \cdots < t_n$, and these observations are denoted by $(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Here, we assume that the locations are observed without error, but we discuss methods to account for measurement error in Section 5 and 6. We consider J spatial covariates $c_1, \ldots, c_J$, measured on a grid over the study region. $\boldsymbol{\theta}$ denotes the vector of all parameters of the Langevin movement model defined in Section 2, i.e. $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_J, \gamma^2)$. This section describes an inference method to estimate $\boldsymbol{\theta}$, from telemetry and habitat data. We focus on one individual animal, but the method can also be applied to obtain joint inferences from several individuals, as presented in the case study in Section 5.

3.1 Euler approximation of the likelihood

The likelihood of the observed locations, given $\boldsymbol{\theta}$, can be expressed using the *transition density* of the process $(\boldsymbol{X}_t)_{t\geq 0}$. The transition density is the probability density function of $\boldsymbol{X}_{t+\Delta}$ given $\boldsymbol{X}_t = \boldsymbol{x}_t$, and we denote it by $q_\Delta(\boldsymbol{x}|\boldsymbol{x}_t, \boldsymbol{\theta})$. Following the Markov property satisfied by the Langevin diffusion process, and assuming that the first position $\boldsymbol{x}_0$ is deterministic, the likelihood function is

$$L(\boldsymbol{\theta}; \boldsymbol{x}_{0:n}) = \prod_{i=0}^{n-1} q_{\Delta_i}(\boldsymbol{x}_{i+1}|\boldsymbol{x}_i, \boldsymbol{\theta}), \tag{6}$$

where $\boldsymbol{x}_{0:n}$ is shorthand for the set of observations, and $\Delta_i = t_{i+1} - t_i$.

As discussed in Gloaguen et al. (2018), in many practical cases, there is no closed-form expression for the density q_Δ, and the likelihood $L(\boldsymbol{\theta}; \boldsymbol{x}_{0:n})$ cannot be evaluated. To circumvent this problem, pseudo-likelihood approaches can be used as approximations. In these approaches, the diffusion process is approximated by a simpler, tractable process. The intractable transition density in Equation 6 is then replaced by that of the simpler process (usually, a Gaussian density), with moments given by a discretization scheme. The "pseudo-likelihood" then refers to the likelihood of the approximate diffusion process (Iacus, 2009; Gloaguen et al., 2018).

The most common pseudo-likelihood approach for discretely observed diffusion is the Euler discretization scheme (Iacus, 2009). In the Euler discretization, the transition density of the Langevin diffusion is approximated by the following Gaussian density between t_i and t_{i+1}, for $i = 0, \dots, n-1$,

Conditionally on $\{\boldsymbol{X}_i = \boldsymbol{x}_i\}$,

$$\boldsymbol{X}_{i+1} = \boldsymbol{x}_i + \frac{\gamma^2 \Delta_i}{2} \nabla \log \pi(\boldsymbol{x}_i|\boldsymbol{\beta}) + \varepsilon_{i+1}, \quad \varepsilon_{i+1} \overset{ind}{\sim} N\left(\boldsymbol{0},\ \gamma^2 \Delta_i \boldsymbol{I}_d\right), \tag{7}$$

where $\boldsymbol{I}_d$ is the $d \times d$ identity matrix. Under this approximation, the transition density of the process can then be written

$$q_{\Delta_i}(\boldsymbol{x}_{i+1}|\boldsymbol{x}_i, \boldsymbol{\theta}) = \phi\left(\boldsymbol{x}_{i+1} \middle| \boldsymbol{x}_i + \frac{\gamma^2 \Delta_i}{2} \nabla \log \pi(\boldsymbol{x}_i|\boldsymbol{\beta});\ \gamma^2 \Delta_i \boldsymbol{I}_d\right),$$

where $\phi(\cdot|\boldsymbol{\mu};\ \boldsymbol{\Sigma})$ is the p.d.f. of the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. This expression can be plugged into Equation 6 to obtain the approximate likelihood of a track $\boldsymbol{x}_{0:n}$.

The Euler discretization can also be used to simulate (approximately) from the Langevin movement model, as illustrated in the simulations of Section 4. The quality of the scheme decreases for longer time steps of simulation (Kessler et al., 2012, Chapter 1). The pseudo likelihood approach can also be used to derive an approximate AIC, to perform model selection, as we demonstrate in the analysis of Section 5. This is similar to the approximate AIC described by Uchida and Yoshida (2005), although they use a different discretization scheme.

3.2 Maximum likelihood estimation

The pseudo-likelihood function could be optimised numerically to obtain estimates of all model parameters. However, if π is modelled with the resource selection function of Equation 4, the discretised movement model can be written in terms of a linear model, and the pseudo maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ can simply be obtained using standard estimators for linear models. From now on, we focus on the two-dimensional case ($d = 2$), for definiteness, but derivation for other values of d is straightforward.

Plugging Equation 5 into Equation 7, we can write the model in the following matrix form. Let $\boldsymbol{Y}_i = (\boldsymbol{X}_{i+1} - \boldsymbol{X}_i)/\sqrt{\Delta_i}$ be the two-dimensional normalized random increment of the process between t_i and t_{i+1}, and denote

$$\boldsymbol{Y} = \begin{pmatrix} Y_{0,1} \\ \vdots \\ Y_{n-1,1} \\ Y_{0,2} \\ \vdots \\ Y_{n-1,2} \end{pmatrix}, \qquad \boldsymbol{D} = \frac{1}{2} \begin{pmatrix} \frac{\partial c_1(\boldsymbol{x}_0)}{\partial z_1} & \frac{\partial c_2(\boldsymbol{x}_0)}{\partial z_1} & \dots & \frac{\partial c_J(\boldsymbol{x}_0)}{\partial z_1} \\ \vdots & & & \vdots \\ \frac{\partial c_1(\boldsymbol{x}_{n-1})}{\partial z_1} & \frac{\partial c_2(\boldsymbol{x}_{n-1})}{\partial z_1} & \dots & \frac{\partial c_J(\boldsymbol{x}_{n-1})}{\partial z_1} \\ \frac{\partial c_1(\boldsymbol{x}_0)}{\partial z_2} & \frac{\partial c_2(\boldsymbol{x}_0)}{\partial z_2} & \dots & \frac{\partial c_J(\boldsymbol{x}_0)}{\partial z_2} \\ \vdots & & & \vdots \\ \frac{\partial c_1(\boldsymbol{x}_{n-1})}{\partial z_2} & \frac{\partial c_2(\boldsymbol{x}_{n-1})}{\partial z_2} & \dots & \frac{\partial c_J(\boldsymbol{x}_{n-1})}{\partial z_2}, \end{pmatrix},$$

where $\boldsymbol{Y}_i = (Y_{i,1}, Y_{i,2})$, and where $\partial/\partial z_i$ denotes the partial derivative with respect to the i-th spatial coordinate.

Moreover, let $\boldsymbol{T}_\Delta$ be the $(2n) \times (2n)$ diagonal matrix with i-th and $(n+i)$-th diagonal terms equal to $\sqrt{\Delta_{i-1}}$, for $i = 1, \ldots, n$, and write $\boldsymbol{Z} = \boldsymbol{T}_\Delta \boldsymbol{D}$. The matrix $\boldsymbol{Z}$ is known, since $\boldsymbol{T}_\Delta$ depends only on Δ_{i-1} $(i = 1, \ldots, n)$, and $\boldsymbol{D}$ depends only on the covariates c_j, $(j = 1, \ldots, J)$. Then the Euler approximation of the Langevin movement model can be rewritten as

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\nu} + \boldsymbol{E}, \tag{8}$$

where $\boldsymbol{E}$ is a $2n$-vector of independent $N(0, \gamma^2)$ variables, and where $\boldsymbol{\nu} = \gamma^2 \boldsymbol{\beta}$. The estimators for $\boldsymbol{\nu}$ and γ^2 are derived from standard linear model theory, and their expressions are given in Appendix A. In Appendix A, we also use linear model theory to derive confidence intervals for all the parameters of the model. Under the Euler approximation, the computation time for $\hat{\beta}$ and $\hat{\gamma}$ is equivalent to that for fitting a linear regression, thus very fast for standard data sets sizes. Another appeal of the formulation given in Equation 8 is that standard linear model residuals can be calculated for the Langevin movement model, and used to assess goodness-of-fit.

For the Langevin movement model based on a RSF, as defined in Section 2, the Euler approximation therefore provides explicit estimates and confidence intervals. Note that the Euler estimator is *biased* due to the approximation made in Equation 7 (see Kessler et al., 2012, Chapter 1). Therefore, both the estimate and the confidence interval must be interpreted with caution, as they depend on the quality of the Euler scheme. The potential use of other discretization schemes is discussed in Section 6.

3.3 The Metropolis-adjusted Langevin algorithm

The accuracy of the approximation, for the Euler scheme presented in the previous section, depends on the time interval of discretization. Here, we propose a method to measure the discretization error, based on the ideas of the so-called Metropolis-adjusted Langevin algorithm.

Simulations based on the Euler discretization of the Langevin diffusion process are not exact. Roberts and Tweedie (1996) called this simulation algorithm the "unadjusted Langevin algorithm", and they showed that it may not converge to the correct stationary distribution π. In the context of MCMC sampling, they described a "corrected" version of the discretized Langevin diffusion, to sample exactly from the target distribution. They considered the transition density of the discretized Langevin process as the proposal distribution for a Metropolis-Hastings algorithm. This "Metropolis-adjusted Langevin algorithm" (MALA) is a special case of Metropolis-Hastings, such that the limiting distribution of samples is the correct target distribution. We propose to use the MALA indirectly, to assess the accuracy of the Euler approximation in the context of inference presented in Section 3.2.

We suggest using the acceptance rate of the MALA to measure the discrepancy between the true Langevin diffusion and the discretized process. As the time step of discretization decreases, the discretized process becomes a better approximation, and the acceptance rate of the algorithm increases. In Appendix C, we present simulations from the MALA at different time steps of discretization, and show that the acceptance rate tends to 1 when the time step is small. This criterion becomes very valuable to assess a model fitted to real data. In the case of real data, the time step of discretization is given by the time step of observation, and it cannot be adjusted to improve the approximation. Then, the problem is to determine whether the time step of observation leads to a good approximation of the process, in the context of the analysis. This may depend on the speed of the process (i.e. the speed of movement of the animal), and on the spatial autocorrelation structure of the target distribution (i.e. of the covariates when modelled with a RSF). In

Section 5, we use the acceptance rate of simulations from the MALA to assess a Langevin movement model fitted to tracking data from three Steller sea lions.

4 Simulation study

In this section, we assess the performance of the inference method described in Section 3 in two simulation scenarios. In both cases, we simulate movement tracks from the Langevin process, using a very fine Euler discretization given in Equation 7. We simulate covariates and define an artificial utilisation distribution, expressed as a resource selection function, as shown in Equation 4. The objective is to recover the habitat selection parameters $\{\beta_1, \ldots, \beta_J\}$ and the speed parameter γ^2. The method presented in the previous sections is provided in the R package Rhabit, available on Github: `github.com/papayoun/Rhabit`. This simulation study, and the analysis of the next section, can be implemented using the package.

4.1 Scenario 1

We first consider a fully controlled simulation scenario, where the covariate fields are given by smooth analytical functions. In this idealized case, the gradient of the covariate functions, and thus of the utilisation distribution, can be calculated exactly at any point of the region of interest. The utilisation distribution π is defined as a RSF (Equation 4) of three covariates c_1, c_2 and c_3, given by

$$
\begin{aligned}
c_j(\boldsymbol{z}) &= \alpha_j \exp(-(\boldsymbol{z} - \boldsymbol{a}^j)^\intercal \boldsymbol{\Sigma}^j (\boldsymbol{z} - \boldsymbol{a}^j)) \times \sin\left(\omega_1^j (z_1 - a_1^j)\right) \times \sin\left(\omega_2^j (z_1 - a_2^j)\right),\ j = 1, 2, \\
c_3(\boldsymbol{z}) &= \| \boldsymbol{z} \|^2,
\end{aligned}
$$

where $\alpha_j, \boldsymbol{a}^j = (a_1^j, a_2^j)$, $\boldsymbol{\omega}^j = (\omega_1^j, \omega_2^j)$, and $\boldsymbol{\Sigma}^j = \text{diag}\{\sigma_1^j, \sigma_2^j\}$ are known simulation parameters whose values are given in Appendix B . For the simulations, we choose the resource selection parameters $\beta_1 = -1$, $\beta_2 = 0.5$, and $\beta_3 = -0.05$, and the speed parameter $\gamma^2 = 1$.

The first two covariates are smooth functions, for which the gradient can easily be derived. The third covariate is the squared distance to the centre of the map, and is used to include a weak force of attraction towards the centre (here, the point $(0, 0)$, somewhat related to the home range of the individual). These three covariates functions are shown in Figure 2. The resulting utilization distribution is the one shown on Figure 1.

Inference was performed independently on 600 data sets. Each data set was a trajectory of 300 points, simulated from the Langevin movement model. The tracks were first generated at a fine time resolution ($\Delta = 0.01$), to minimise the effect of the Euler approximation, and they were then thinned to time intervals of 0.5 time units.

We estimated all model parameters using the Euler method, presented in Section 3.2. We considered two different settings: (i) the true analytic gradient is used in the estimation, and (ii) the covariates are discretized on a 8×8 regular grid, and the gradient is obtained through the interpolation of the covariates. This second setting corresponds to the more realistic case where covariates are only observed on a discrete grid, and the gradient needs to be approximated. The gradient approximations were performed for the covariates c_1 and c_2 using standard bilinear interpolation (see Appendix D for details). The gradient of the Euclidean distance c_3 is computed exactly in both cases, as it could be in a real analysis.

Boxplots of the parameter estimates in the 600 replications are shown in Figure 3. All parameters were correctly estimated in this benchmark scenario, even when the covariates were discretized to a coarse grid.

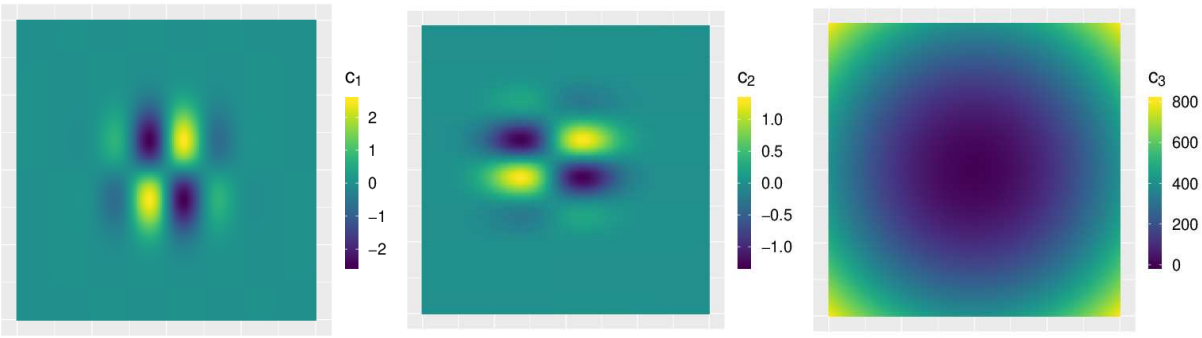

Figure 2: *Artificial covariates fields for the simulation scenario of Section* 4.1. *The resulting utilization distribution is the one shown on Figure* 1.

One can see a slight underestimation of the speed parameter. This is due to the chosen sampling time step, as discussed in the next section.

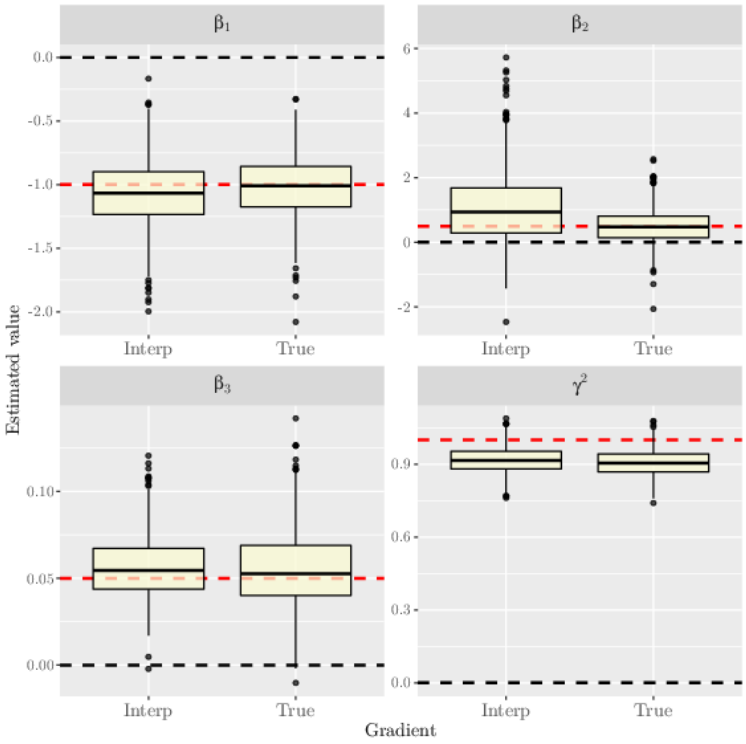

Figure 3: *Estimates for model parameters on* 600 *experiments replications of scenario* 1. *The dotted lines show the real values used in the simulations.*

4.2 Scenario 2

We considered a second simulation scenario, with randomly-generated covariate fields on a discrete grid, more similar to real environmental data. The main objective of this scenario is to investigate the effect of the sampling frequency on the estimation.

We simulated two covariates c_1 and c_2 as random fields over the study region $[-100, 100] \times [-100, 100]$, with a resolution of 1. We used the function `RMmatern` from the R package RandomFields to generate the

random covariates (Schlather et al., 2015). We also included the squared distance to the centre of the map as a covariate, c_3, to ensure that the simulations did not go near the boundaries of the map, where the gradient of the covariates is undefined. Then, we defined the target (utilisation) distribution as the (normalized) RSF, with coefficients $(\beta_1, \beta_2, \beta_3) = (4, 2, -0.1)'$. Plots of the simulated covariates, and of the utilisation distribution used in the simulations, are shown in Figure 4.

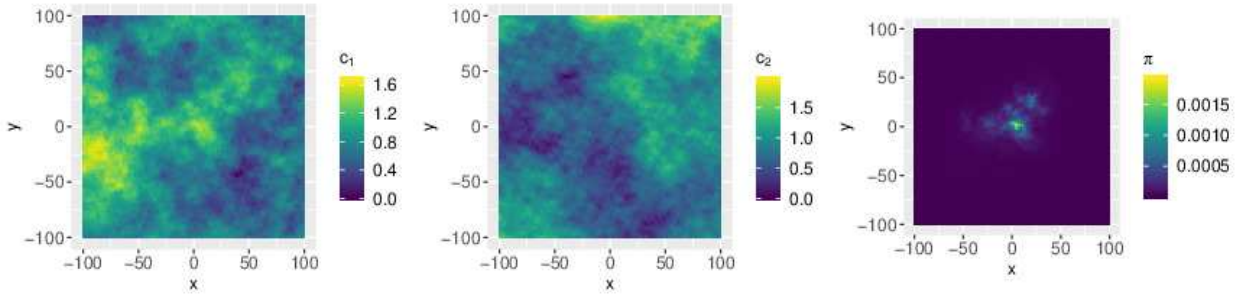

Figure 4: *Simulated covariate fields c_1 and c_2, and utilisation distribution obtained with $\boldsymbol{\beta} = (4, 2, -0.1)'$, used in the second simulation scenario. Note that the utilisation distribution also includes the effect of the squared distance to the centre of the map, not shown here.*

We simulated 100 trajectories from the Langevin movement model with target distribution π, and with speed parameter $\gamma^2 = 5$, at a temporal resolution of $\Delta = 0.01$h. (The time unit is arbitrary here, but we include it for readability.) At this time step of simulation, the Metropolis-adjusted Langevin algorithm has an acceptance rate around 99.5%, which indicates that the discretized process is a good approximation of the true process (Appendix C). We then subsampled each trajectory, for different time resolutions $\Delta \in \{0.01, 0.02, 0.05, 0.1, 0.25, 0.5, 1\}$, to emulate data sets obtained at different observation rates.

From each thinned data set, we kept the first 5000 locations of each of the 100 trajectories. We fitted the Langevin movement model independently to each thinned track, using the estimators given in Section 3.2. We evaluated the gradients of the covariates at each simulated location using bilinear interpolation. We obtained 100 point estimates of each model parameter, for each time step of observation (one for each track). The results are displayed in Figure 5.

There was a lot of variability in the accuracy and precision of habitat selection parameter estimates. The uncertainty on the estimates of the habitat selection parameters decreased as the time interval increased. This is not surprising: all trajectories had the same number of locations, such that those with longer time intervals explored a larger proportion of the study region. For example, a track of 5000 locations covers a time period of 5000 hours if the time interval is $\Delta = 1$h, but it only covers 50 hours if $\Delta = 0.01$h. Tracks with longer time intervals therefore covered a larger range of covariate values. Like in standard linear model analyses, the uncertainty on the coefficients is larger when the observed range of explanatory variables in Equation 8 is narrow.

To offset this effect, we considered a second analysis, in which all tracks covered the same period of time. We thinned each of the 100 tracks as before, but we then kept the locations over the time period from $t = 0$ to $t = 500$h, regardless of the time interval of observation. At a resolution of $\Delta = 0.01$h, each track comprised 50000 locations; for $\Delta = 1$h, each track comprised 500 locations. We fitted the Langevin movement model to each track separately, for each time resolution. The estimates are shown in Figure 6.

When the tracks were all truncated to the same interval of time, the variability of the estimates of the habitat selection parameters was the same for all time intervals. This suggests that the uncertainty on the estimates of the habitat selection parameters depends on the extent of spatial exploration, rather than on

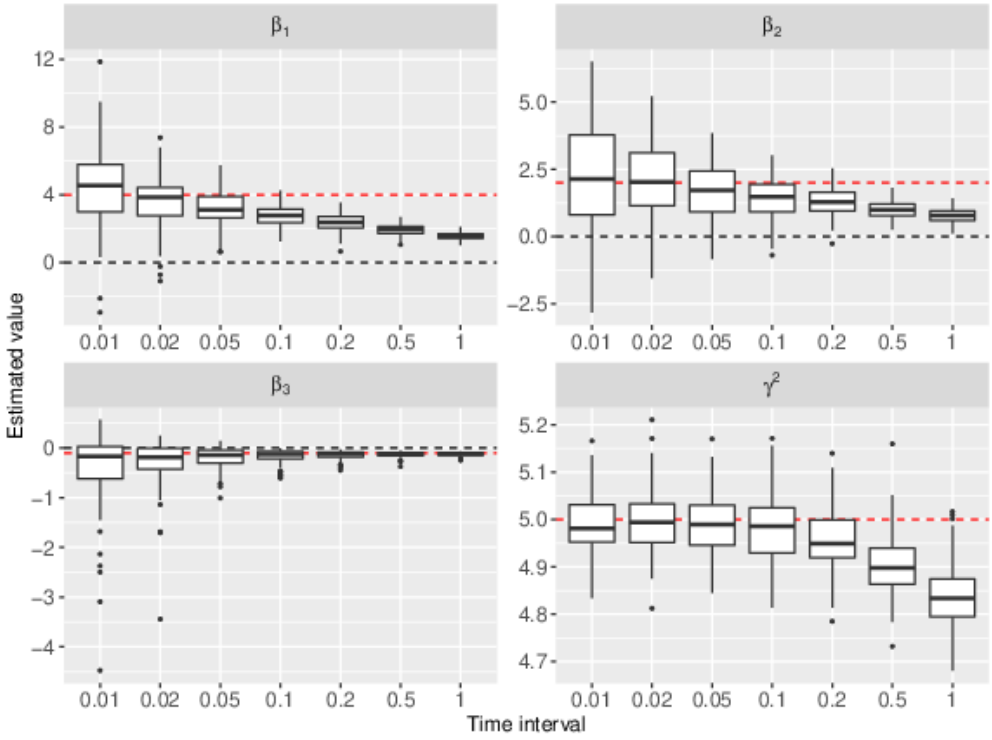

Figure 5: *Boxplots of* 100 *estimates of the habitat selection parameters* ($\beta_1, \beta_2, \beta_3$)*, and of the speed parameter* (γ^2)*, for different time intervals of observation, when the number of observations is the same for all thinned tracks. The red dotted lines show the true values of the parameters. The x axis is on the log scale.*

the number of observations. However, in this case, the variance in the estimates of the speed parameter γ^2 increased as the number of observations decreased (i.e., in this case, as the time interval increased).

In both Figures 5 and 6, the estimates of β_1 and β_2 decreased (on average) as the time interval increased, leading to an underestimation of the parameters for longer time intervals. This is a common problem for the estimation of discretely observed diffusion processes, because the consistency of the estimators requires Δ to tend towards 0 (for more details, see Kessler et al., 2012). For long time intervals, the habitat selection parameters are underestimated in absolute value, i.e. the strength of the (positive or negative) effect is underestimated. A possible interpretation of this bias, in the context of the estimation of space use and habitat selection, is the following. As the time interval increases, the estimated utilisation distribution becomes flatter, to reflect our growing uncertainty about the effect of the covariates on the short-term movement. In the extreme, for very long time intervals, we would have no information about the selection process, and the estimated utilisation distribution would be flat, corresponding to a uniform distribution of space use over the study region. In this respect, our approach differs from other methods of estimation of the utilisation distribution, such as resource selection functions or kernel density estimators. With those methods, locations collected at a coarse resolution are still informative about long-term habitat selection and space use, and they could be used to recover the utilisation distribution. However, in the Langevin movement model, space use is not estimated directly. Instead, the short-term habitat selection is estimated, as captured by the effect of the local gradient of the covariates on the movement of the animal. Therefore, since the utilisation distribution is a by-product, obtained as the stationary distribution of the short-term movement process, the Langevin model may fail to capture both the short-term habitat selection and the long-term space use if the time intervals between observations are too long. In the case of very coarse data,

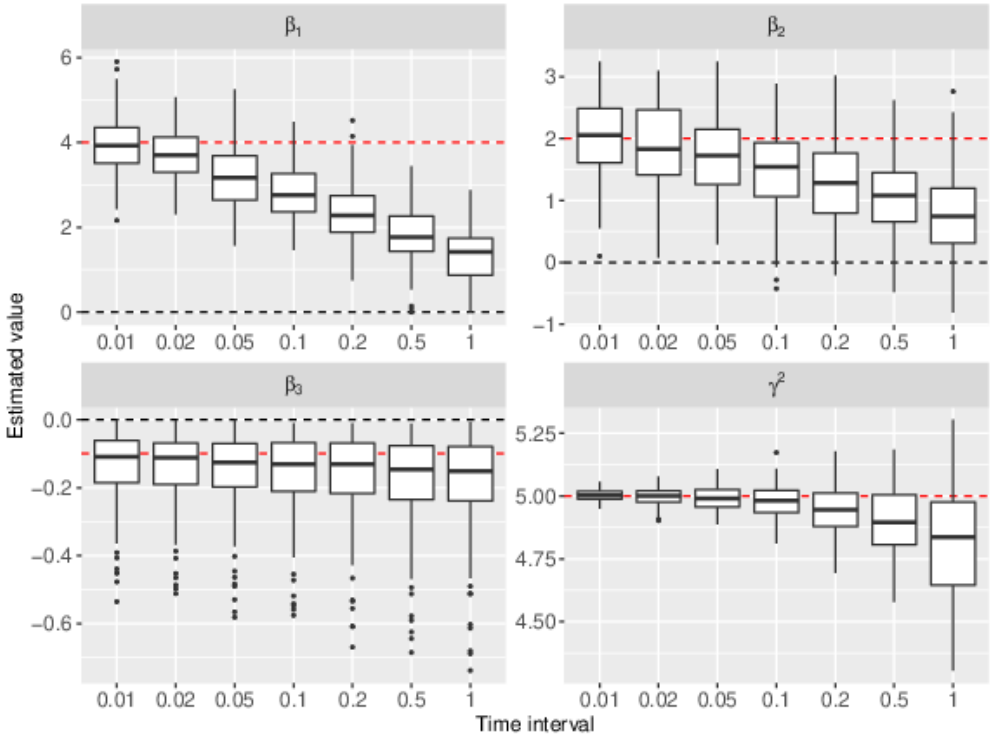

Figure 6: *Boxplots of* 100 *estimates of the habitat selection parameters* ($\beta_1, \beta_2, \beta_3$)*, and of the speed parameter* (γ^2)*, for different time intervals of observation, when the duration is the same for all thinned tracks. The red dotted lines show the true values of the parameters. The x axis is on the log scale.*

the correlation between successive observed locations would be small, and the RSF could be estimated using the standard generalized linear model approach based on use-availability data (Johnson et al., 2006).

Note that, although the strength of selection was underestimated in the simulations with long time intervals, the sign of the effect – i.e. selection or avoidance – was always estimated correctly. The estimates of the speed parameter γ^2 were very close to the true value in all simulation experiments with $0.01 \leq \Delta \leq 0.1$. It seems to be underestimated for longer time intervals of observation, because the total distance travelled by the process is underestimated when the discretization is coarse.

To investigate the performance of the method for the analysis of data sets collected at irregular time intervals, we ran a similar experiment where the observations were thinned at random. The results were very similar to the simulations with regular intervals, and are presented in Appendix E. These findings confirm that, due to its continuous-time formulation, the Langevin movement model can directly be used on tracking data collected irregularly.

5 Illustration

In this section, we fit the Langevin movement model to a data set described by Wilson et al. (2018), collected on Steller sea lions in Alaska. The data set comprises three trajectories, obtained from three different individuals, for a total of 2672 Argos locations. The time intervals were highly irregular, with percentiles $P_{0.025} = 6$min, $P_{0.5} = 1.28$h, $P_{0.975} = 17.4$h. In addition to the locations, Wilson et al. (2018) provided four spatial covariates over the study region, at a resolution of 1km: bathymetry (c_1), slope (c_2), distance to sites of interest (c_3), and distance to continental shelf (c_4). The sites of interest were either haul-out or rookery

sites. Maps of the covariates are shown in Figure 7, and we refer the readers to Wilson et al. (2018) for more detail about the data set.

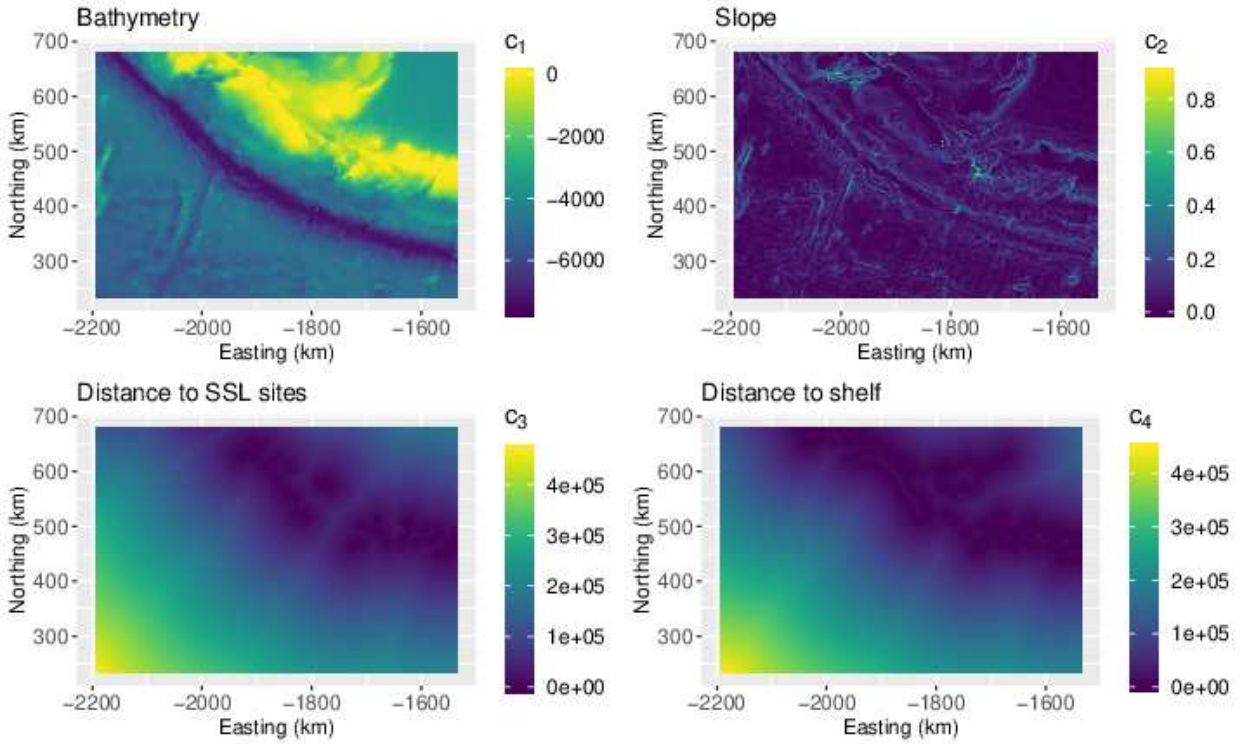

Figure 7: *Covariate maps for the sea lion analysis.*

Covariates As can be seen in Figure 7, there is strong collinearity between the distance to sites of interest and the distance to the shelf (i.e. between c_3 and c_4). We derived the correlation matrix R for the four covariates,

$$R = \begin{pmatrix} 1 & 0.05 & -0.60 & -0.61 \\ \cdot & 1 & -0.17 & -0.19 \\ \cdot & \cdot & 1 & 0.98 \\ \cdot & \cdot & \cdot & 1 \end{pmatrix}.$$

This confirms that the correlation between the covariates c_3 and c_4 is high (0.98). This is because sites of interest are rookeries and haul-out sites, which are on the island shelf. The effects of those two covariates therefore cannot be estimated separately, and we decided to exclude the distance to the shelf c_4 from the analysis.

Data pre-processing To correct for the measurement error in the locations, and to follow the data preparation of Wilson et al. (2018), we first fitted a continuous-time correlated random walk to the tracks, using the R package crawl (Johnson et al., 2008; Johnson and London, 2018). The continuous-time correlated random walk is a state-space model, that can be used on irregular and noisy telemetry data. The package crawl implements the Kalman filter for this model, to estimate the true location of an animal from observations made with measurement error. We used the code provided by Wilson et al. (2018) to fit the continuous-time model to each track, and obtained predicted locations for the times of the observations.

Results We then fitted the Langevin movement model to the filtered tracks, using the inference method of Section 3. To investigate inter-individual heterogeneity, we fitted a model to each track separately, and then a joint model to the three tracks. In the following, we call the three individuals "SSL1", "SSL2", and "SSL3". In each model, we estimated four parameters: the three habitat selection parameters (β_1, β_2, β_3), and the speed parameter γ^2. In our approach, most of the computation time is needed to evaluate the gradient of each covariate at all observed locations, which took less than one second on a 2GHz i5 CPU. Like in the simulation study of Section 4.2, the covariates were interpolated, so that their gradient could be evaluated at each filtered location. The point estimates and 95% confidence intervals of all model parameters, obtained from the equations given in Section 3.2, are presented in Table 1. For the joint model fitted to the three trajectories, the estimated utilisation distribution, and its logarithm (for comparison with Wilson et al., 2018), are plotted in Figure 8.

Parameter	SSL1	SSL2	SSL3	All individuals
β_1 ($\times 10^4$)	3.67 (-1.96, 9.31)	8.17 (2.21, 14.1)	0.41 (-0.86, 1.68)	1.34 (0.004, 2.72)
β_2 ($\times 10$)	-2.77 (-15.7, 10.1)	1.81 (-7.67, 11.3)	0.67 (-1.61, 2.94)	0.76 (-1.74, 3.25)
β_3 ($\times 10^5$)	-1.14 (-2.89, 0.60)	-4.49 (-8.73, -0.25)	-2.38 (-3.80, -0.96)	-2.06 (-3.07, -1.05)
γ^2	8.97 (8.47, 9.51)	7.49 (6.84, 8.23)	18.2 (17.1, 19.3)	12.4 (11.9, 12.8)

Table 1: *Estimates and 95% confidence intervals in the Steller sea lion analysis. We fitted a Langevin movement model to each individual separately ("SSL1", "SSL2", and "SSL3"), and then jointly to the three individuals ("All individuals").*

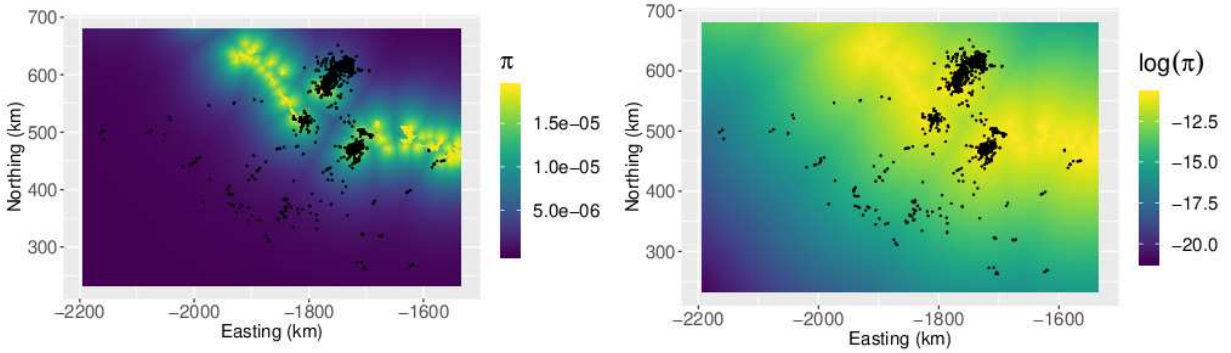

Figure 8: *Estimated utilisation distribution for the sea lion analysis (left), and its logarithm, for comparison with Wilson et al. (2018) (right). This figure shows the results of the model fitted jointly to the three individuals. The black dots are the filtered sea lion locations.*

There are clear differences in the estimated parameters for the four fitted models. To select between the individual models and the joint model, we compared the AIC of the joint model to the *sum* of the AICs of the three individual models. Here, the AIC of the joint model was 30281, and the sum of the individual AICs was 29902, which indicates that the individual models are strongly favoured.

In all models, the 95% confidence interval of the parameter β_2 for the slope covariate included zero, i.e. the covariate did not have a clear effect on the sea lions' movement. However, for the other two covariates, the estimated effects varied across models. The estimate of β_1, corresponding to the effect of the bathymetry covariate, was positive in the model fitted to SSL2 and in the joint model for the three tracks. This suggests that SSL2 tended to move towards areas of shallow water. However, there was no clear effect of bathymetry

for SSL1 and SSL3. The effect of the distance to sites of interest, β_3, was estimated to be negative for SSL2 and SSL3, and in the joint model. This indicates that the model captured the attraction of these two sea lions towards the sites of interest (rookeries and haul-out sites). In the joint model and in the model for SSL2, the estimated effects of the bathymetry and of the distance to sites of interest are consistent. Indeed, the sites of interest are haul-out sites, or rookeries, which are located in areas of shallow water. The speed parameter γ^2 was also estimated for the three individuals, and is given in Table 1. The speed parameters of SSL1 and SSL2 were quite similar, but the estimate for SSL3 was more than twice larger, suggesting faster movement. The speed parameter should be interpreted with care because, in general, the actual speed of movement also depends on the habitat selection parameters (as described in Section 2.1). Here, the estimated speed parameters indicate that, in the absence of covariate effects (e.g. in a large area of homogeneous habitat), SSL3 will tend to move about twice as fast than SSL1 and SSL2.

Model checking We can use linear model residuals to assess the goodness of fit. In Appendix F , we show a quantile-quantile plot of the residuals against the normal distribution, which indicates some clear lack of fit. Following the usual checking procedure for the linear model, we derived the predicted steps and inspected persistent structure that was not captured by the model. The bivariate predicted steps along the trajectory of the seal SSL2 are shown in Appendix F .

From the map of the predicted steps, we can see that the model fails to predict long steps, which occur when the animal is at sea (e.g. in transit between sites of interest). This may be due to the lack of flexibility of the model to capture phases of movement with different speeds. The speed parameter γ^2 is assumed to be constant in time so, even if it captures the average speed of movement, it may fail to account for either very slow or very fast movement. This motivates a more flexible formulation for the speed parameter, for instance a state-switching model where each state i is characterised by a different parameter γ_i^2. This extension would not be straightforward to implement in continuous time, however, and would lead to a more complex inference framework.

We also see (Figure S3 in Appendix F) that the predicted steps always point towards shallow water or towards the closest site of interest, because their direction is determined by the estimated habitat selection parameters. As a result, the model fails to predict displacements away from sites of interest, for example. This suggests that, to fully understand the drivers of the sea lion movements, additional covariates may need to be included in the analysis.

Euler approach validity As illustrated in the simulations of Section 4.2, we computed our MALA index by bootstrap to assess the validity of the Euler method. Overall, the acceptance rate of the algorithm ranged between 93.2% and 99.1%, with a mean of 97.4%, which seems to indicate that the application of the Langevin movement model is appropriate for this data set.

6 Discussion

This work introduces a new model of animal movement, based on the Langevin diffusion process, that integrates the movement with space use and habitat selection. Our model follows the idea of potential-based movement models proposed by Preisler et al. (2004), and it is explicitly connected to the utilisation distribution of the individual, from stationarity properties of the Langevin diffusion process. If spatial covariates are available, the utilisation distribution can be modelled with a resource selection function, embedded in the movement process, to infer habitat preferences. The Langevin movement model therefore describes animal movement in response to spatial covariates, i.e. step selection. Pseudo-likelihood methods

can be used to obtain estimates of the habitat selection parameters in a linear model framework, from which an estimated utilisation distribution can be computed. The Langevin movement model is formulated in continuous time, and it can deal with location data collected at irregular time intervals, without the need to interpolate them. Similarly, because it models movement in continuous space (unlike the method presented by Wilson et al. 2018), the interpretation of the results is not tied to a particular space discretization.

In this paper, we used the Euler discretization scheme to obtain pseudo maximum likelihood estimators. This scheme is the most widely-used method to carry out inference for discretely-observed diffusion processes, when the transition density is not analytically tractable (see Preisler et al., 2004; Brillinger, 2010; Russell et al., 2018, for applications in ecology). There exist other pseudo-likelihood approaches, and Gloaguen et al. (2018) argued that better inferences could be obtained with more refined schemes. In particular, they found that the Ozaki discretization provided more reliable results in their applications. However, the Ozaki scheme requires the evaluation of the partial derivatives of the drift, i.e. the (partial) second derivatives of $\log \pi$ in the Langevin movement model. To compare the Euler and the Ozaki scheme, we repeated the simulation study of Section 4.1, using the Ozaki scheme for the estimation (the results are not shown here), and found out that the theoretical advantages of the Ozaki scheme were counterbalanced by the need of a second-order interpolation, and the Euler scheme provided more reliable estimates. Therefore, in the context of the Langevin movement model, the Euler scheme is typically more robust as it requires fewer numerical approximations.

In the case study of Section 5, we used a two-stage approach to deal with the measurement error. We first fitted a state-space model, the continuous-time correlated random walk, to filter the Argos locations. Then, we fitted the Langevin movement model to the filtered tracks. There are several drawbacks to the two-stage approach. Indeed, it is difficult to propagate the uncertainty from the measurement error to the final parameter estimates (although multiple imputation could be used; see e.g. Scharf et al., 2017). Besides, the two stages are not consistent, because the first stage ignores the environmental effects that are estimated in the second stage. To avoid this issue, the two steps could be integrated into a state-space model that incorporates measurement error directly on top of the Langevin movement process. The state equation of the full model is given by the transition density of the Langevin movement model, or a discretization of it (like the one given in Equation 7). A natural choice for the observation equation would be $\tilde{\boldsymbol{X}}_i = \boldsymbol{X}_i + \boldsymbol{\eta}_i$, where $\tilde{\boldsymbol{X}}_i$ is the noisy observed location, $\boldsymbol{X}_i$ is the true location, and $\boldsymbol{\eta}_i \sim N(\boldsymbol{0}, \sigma^2_{\text{obs}}\boldsymbol{I}_2)$ models the measurement error. Under the Euler scheme, the approximate transition density is normal, and a Kalman filter can be used to compute the pseudo-likelihood of this hierarchical state-space model.

As in Michelot et al. (2018), the approach taken here is fundamentally a local one. One consequence of this, touched on in Section 2.2, is that within regions where covariates are constant, there is no selection, as the drift term in Equation 3 is zero. This is not necessarily unrealistic; in fact it follows from the assumption that the utilisation density at a point depends only on the values of covariates at that point, as in Equation 4. However, it does suggest a more general framework, in which the movement model is based on smoothed versions of covariates, with the spatial scale of smoothing acting as a proxy for the perception or decision-making scale of the animal, as distinct from the movement scale. However, the estimation of this unknown spatial scale in this more general model should be addressed.

The inspection of the residuals in the sea lion case study suggested that a potential improvement would be to allow the speed parameter γ^2 to vary in time. This would not break the stationarity property of the Langevin movement model, as long as γ^2 does not depend on the utilisation distribution π at the current location. However, the linear model formulation would not apply in that case. In the analysis of Section 5, we found that habitat selection and movement parameters varied between individuals. Another extension of the presented work would be to incorporate a random effect in the model, to account for individual

deviation from the overall population model. Using the Euler scheme, this extension could be written as a mixed linear model.

Acknowledgements

TM was supported by the Centre for Advanced Biological Modelling at the University of Sheffield, funded by the Leverhulme Trust, award number DS-2014-081. We thank the associate editor and two reviewers whose insightful comments greatly improved the paper.

Authors contribution

All authors conceived the ideas and designed methodology. Théo Michelot and Pierre Gloaguen led the code development and the simulation studies. All authors contributed critically to the drafts and gave final approval for publication.

Data accessibility

The Steller sea lion data set used in Section 5 is provided by Wilson et al. (2018). The code for the simulations and the case study will be made available on Zenodo.

References

Anderson, D. J. (1982). The home range: A new nonparametric estimation technique: Ecological archives e063-001. *Ecology*, 63(1):103–112.

Avgar, T., Potts, J. R., Lewis, M. A., and Boyce, M. S. (2016). Integrated step selection analysis: bridging the gap between resource selection and animal movement. *Methods in Ecology and Evolution*, 7(5):619–630.

Blackwell, P. G. (1997). Random diffusion models for animal movement. *Ecological Modelling*, 100(1-3):87–102.

Brillinger, D. (2010). *Handbook of Spatial Statistics*, chapter 26. Handbooks of Statistical Methods. Chapman and Hall/CRC Press.

Dalalyan, A. S. (2017). Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676.

Fleming, C. H., Fagan, W. F., Mueller, T., Olson, K. A., Leimgruber, P., and Calabrese, J. M. (2015). Rigorous home range estimation with movement data: a new autocorrelated kernel density estimator. *Ecology*, 96(5):1182–1188.

Fleming, C. H., Fagan, W. F., Mueller, T., Olson, K. A., Leimgruber, P., and Calabrese, J. M. (2016). Estimating where and how animals travel: An optimal framework for path reconstruction from autocorrelated tracking data. *Ecology*. DOI: 10.1890/15-1607.

Forester, J. D., Im, H. K., and Rathouz, P. J. (2009). Accounting for animal movement in estimation of resource selection functions: sampling and data analysis. *Ecology*, 90(12):3554–3565.

Fortin, D., Beyer, H. L., Boyce, M. S., Smith, D. W., Duchesne, T., and Mao, J. S. (2005). Wolves influence elk movements: behavior shapes a trophic cascade in Yellowstone National Park. *Ecology*, 86(5):1320–1330.

Gloaguen, P., Etienne, M.-P., and Le Corff, S. (2018). Stochastic differential equation based on a multimodal potential to model movement data in ecology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(3):599–619.

Hanks, E. M., Hooten, M. B., Alldredge, M. W., et al. (2015). Continuous-time discrete-space models for animal movement. *The Annals of Applied Statistics*, 9(1):145–165.

Harris, K. J. and Blackwell, P. G. (2013). Flexible continuous-time modelling for heterogeneous animal movement. *Ecological Modelling*, 255:29–37.

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

Horne, J. S., Garton, E. O., Krone, S. M., and Lewis, J. S. (2007). Analyzing animal movements using Bownian bridges. *Ecology*, 88(9):2354–2363.

Iacus, S. M. (2009). *Simulation and inference for stochastic differential equations: with R examples*. Springer Science & Business Media.

Johnson, C. J., Nielsen, S. E., Merrill, E. H., McDONALD, T. L., and Boyce, M. S. (2006). Resource selection functions based on use-availability data: theoretical motivation and evaluation methods. *The Journal of Wildlife Management*, 70(2):347–357.

Johnson, D. S. and London, J. M. (2018). crawl: an R package for fitting continuous-time correlated random walk models to animal movement data.

Johnson, D. S., London, J. M., Lea, M.-A., and Durban, J. W. (2008). Continuous-time correlated random walk model for animal telemetry data. *Ecology*, 89(5):1208–1215.

Kessler, M., Lindner, A., and Sorensen, M. (2012). *Statistical methods for stochastic differential equations*. Chapman and Hall/CRC.

Kranstauber, B., Kays, R., LaPoint, S. D., Wikelski, M., and Safi, K. (2012). A dynamic Brownian bridge movement model to estimate utilization distributions for heterogeneous animal movement. *Journal of Animal Ecology*, 81(4):738–746.

Long, R. A., Muir, J. D., Rachlow, J. L., and Kie, J. G. (2009). A comparison of two modeling approaches for evaluating wildlife-habitat relationships. *The Journal of Wildlife Management*, 73(2):294–302.

Manly, B., McDonald, L., Thomas, D., McDonald, T. L., and Erickson, W. P. (2002). *Resource selection by animals: statistical design and analysis for field studies, Second Edition*. Kluwer Academic Publishers, Dordrecht.

Michelot, T., Blackwell, P. G., and Matthiopoulos, J. (2018). Linking resource selection and step selection models for habitat preferences in animals. *Ecology*. DOI: 10.1002/ecy.2452.

Millspaugh, J. J., Nielson, R. M., McDonald, L., Marzluff, J. M., Gitzen, R. A., Rittenhouse, C. D., Hubbard, M. W., and Sheriff, S. L. (2006). Analysis of resource selection using utilization distributions. *The Journal of Wildlife Management*, 70(2):384–395.

Nielson, R. M. and Sawyer, H. (2013). Estimating resource selection with count data. *Ecology and evolution*, 3(7):2233–2240.

Péron, G. (2019). Modified home range kernel density estimators that take environmental interactions into account. *Movement ecology*, 7(1):16.

Potts, J. R., Bastille-Rousseau, G., Murray, D. L., Schaefer, J. A., and Lewis, M. A. (2014). Predicting local and non-local effects of resources on animal space use using a mechanistic step selection model. *Methods in ecology and evolution*, 5(3):253–262.

Preisler, H. K., Ager, A. A., Johnson, B. K., and Kie, J. G. (2004). Modeling animal movements using stochastic differential equations. *Environmetrics*, 15(7):643–657.

Preisler, H. K., Ager, A. A., and Wisdom, M. J. (2013). Analyzing animal movement patterns using potential functions. *Ecosphere*, 4(3):1–13.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363.

Russell, J. C., Hanks, E. M., Haran, M., and Hughes, D. (2018). A spatially varying stochastic differential equation model for animal movement. *The Annals of Applied Statistics*, 12(2):1312–1331.

Scharf, H., Hooten, M. B., and Johnson, D. S. (2017). Imputation approaches for animal movement modeling. *Journal of Agricultural, Biological and Environmental Statistics*, 22(3):335–352.

Schlather, M., Malinowski, A., Menck, P. J., Oesting, M., and Strokorb, K. (2015). Analysis, simulation and prediction of multivariate random fields with package RandomFields. *Journal of Statistical Software*, 63(8):1–25.

Signer, J., Fieberg, J., and Avgar, T. (2017). Estimating utilization distributions from fitted step-selection functions. *Ecosphere*, 8(4).

Signer, J., Fieberg, J., and Avgar, T. (2019). Animal movement tools (amt): R package for managing tracking data and conducting habitat selection analyses. *Ecology and evolution*, 9(2):880–890.

Skellam, J. G. (1951). Random dispersal in theoretical populations. *Biometrika*, 38(1/2):196–218.

Thurfjell, H., Ciuti, S., and Boyce, M. S. (2014). Applications of step-selection functions in ecology and conservation. *Movement ecology*, 2(1):4.

Uchida, M. and Yoshida, N. (2005). AIC for ergodic diffusion processes from discrete observations. *preprint MHF*, 12.

Wilson, K., Hanks, E., and Johnson, D. (2018). Estimating animal utilization densities using continuous-time Markov chain models. *Methods in Ecology and Evolution*, 9(5):1232–1240.

Worton, B. J. (1989). Kernel methods for estimating the utilization distribution in home-range studies. *Ecology*, 70(1):164–168.

Xifara, T., Sherlock, C., Livingstone, S., Byrne, S., and Girolami, M. (2014). Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statistics & Probability Letters*, 91:14–19.

Zhang, Z., Sheppard, J. K., Swaisgood, R. R., Wang, G., Nie, Y., Wei, W., Zhao, N., and Wei, F. (2014). Ecological scale and seasonal heterogeneity in the spatial behaviors of giant pandas. *Integrative Zoology*, 9(1):46–60.

A Estimators

In this appendix, we derive unbiased estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\gamma}^2$ for the Euler discretization of the Langevin movement model.

A.1 Unbiased estimators

The Euler discretization of the Langevin movement model can be written as a linear model,

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\nu} + \boldsymbol{E}$$

where $\boldsymbol{\nu} = \gamma^2\boldsymbol{\beta}$. From standard results of linear models, we obtain estimators for $\boldsymbol{\nu}$ and γ^2, as

$$\hat{\boldsymbol{\nu}} = \left(\boldsymbol{Z}^\intercal \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^\intercal \boldsymbol{Y},$$

and

$$\hat{\gamma}^2 = \frac{1}{2n-J}\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2,$$

where $\hat{\boldsymbol{Y}} = \boldsymbol{Z}\hat{\boldsymbol{\nu}}$ is the predicted value of $\boldsymbol{Y}$.

Following standard linear model properties, $\hat{\boldsymbol{\nu}} \sim \mathcal{N}\left(\gamma^2\boldsymbol{\beta}, \gamma^2\left(\boldsymbol{Z}^\intercal\boldsymbol{Z}\right)^{-1}\right)$. From Cochran's theorem, $\hat{\boldsymbol{\nu}}$ and $\hat{\gamma}^2$ are independent, and $(2n-J)\hat{\gamma}^2/\gamma^2 \sim \chi^2(2n-J)$, where $\hat{\gamma}^2$ is an unbiased estimator of γ^2.

Considering $\tilde{\boldsymbol{\beta}}$ defined as $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\nu}}/\hat{\gamma}^2$, we have

$$\begin{aligned}
\mathbb{E}[\tilde{\boldsymbol{\beta}}] &= \mathbb{E}[\hat{\boldsymbol{\nu}}]\mathbb{E}\left[\frac{1}{\hat{\gamma}^2}\right] \\
&= \gamma^2\boldsymbol{\beta}\mathbb{E}\left[\frac{1}{\hat{\gamma}^2}\right] \\
&= \gamma^2\boldsymbol{\beta} \times \frac{2n-J}{\gamma^2}\mathbb{E}\left[\frac{1}{\frac{2n-J}{\gamma^2}\hat{\gamma}^2}\right]
\end{aligned}$$

From the properties given above, we have

$$\frac{1}{\frac{2n-J}{\gamma^2}\hat{\gamma}^2} \sim \text{Inv-}\chi^2(2n-J),$$

and we obtain

$$\mathbb{E}[\tilde{\boldsymbol{\beta}}] = \frac{2n-J}{2n-J-2}\boldsymbol{\beta}.$$

Thus, to obtain an unbiased estimator of $\boldsymbol{\beta}$, we define

$$\hat{\boldsymbol{\beta}} = \frac{2n-J-2}{2n-J}\tilde{\boldsymbol{\beta}}.$$

A.2 Confidence intervals

As said above, $(2n-J)\hat{\gamma}^2/\gamma^2 \sim \chi^2(2n-J)$, so that a confidence interval for γ^2 is given by

$$CI_\alpha(\gamma^2) = \left[\hat{\gamma}^2\frac{2n-J}{q_{1-\alpha/2,2n-J}};\ \hat{\gamma}^2\frac{2n-J}{q_{\alpha/2,2n-J}}\right],$$

where $q_{\alpha/2,2n-J}$ stands for the quantile of order $\alpha/2$ from a χ^2 distribution with $2n-J$ degrees of freedom.

It is also possible to derive confidence intervals (or confidence ellipsoids) using the distribution of $\hat{\boldsymbol{\beta}}$. We first use the covariance matrix of $\tilde{\beta}$,

$$\begin{aligned}
\text{Cov}(\tilde{\beta}_j, \tilde{\beta}_k) &= \mathbb{E}\left[\frac{\hat{\nu}_j\hat{\nu}_k}{\hat{\gamma}^4}\right] - \mathbb{E}\left[\frac{\hat{\nu}_j}{\hat{\gamma}^2}\right]\mathbb{E}\left[\frac{\hat{\nu}_k}{\hat{\gamma}}^2\right] && \text{by definition} \\
&= \mathbb{E}[\hat{\nu}_j\hat{\nu}_k]\mathbb{E}\left[\frac{1}{\hat{\gamma}^4}\right] - \mathbb{E}\left[\frac{1}{\hat{\gamma}^2}\right]^2\mathbb{E}[\hat{\nu}_j]\mathbb{E}[\hat{\nu}_k] && \text{by independence of the estimators } \hat{\gamma} \text{ and } \hat{\boldsymbol{\nu}}.
\end{aligned}$$

We now note that

$$\begin{aligned}
\mathbb{E}[\hat{\nu}_j\hat{\nu}_k] &= \text{Cov}(\hat{\nu}_j, \hat{\nu}_k) + \mathbb{E}[\hat{\nu}_j]\mathbb{E}[\hat{\nu}_k] \\
&= \gamma^2\Lambda_{jk} + \beta_j\beta_k\gamma^4,
\end{aligned}$$

where $\Lambda_{jk} := \left[(\boldsymbol{Z}^\intercal\boldsymbol{Z})^{-1}\right]_{jk}$. Moreover,

$$\begin{aligned}
\mathbb{V}\left[\frac{1}{\hat{\gamma}^2}\right] &= \frac{2(2n-J)^2}{\gamma^4(2n-J-2)^2(2n-J-4)} && \text{as } \frac{\gamma^2}{(2n-J)\hat{\gamma}^2} \sim \text{Inv-}\chi^2(2n-J) \\
\mathbb{E}\left[\frac{1}{\hat{\gamma}^4}\right] &= \mathbb{V}\left[\frac{1}{\hat{\gamma}^2}\right] + \mathbb{E}\left[\frac{1}{\hat{\gamma}^2}\right]^2 = \frac{(2n-J)^2}{\gamma^4(2n-J-2)^2}\left(\frac{2}{2n-J-4}+1\right).
\end{aligned}$$

Finally, we can rewrite

$$\begin{aligned}
\text{Cov}(\tilde{\beta}_j, \tilde{\beta}_k) &= \mathbb{V}\left[\frac{1}{\hat{\gamma}^2}\right]\mathbb{E}[\hat{\nu}_j]\mathbb{E}[\hat{\nu}_k] + \mathbb{E}\left[\frac{1}{\hat{\gamma}^4}\right]\text{Cov}(\hat{\nu}_j, \hat{\nu}_k) \\
&= \frac{(2n-J)^2}{(2n-J-2)^2}\left\{\frac{2\beta_j\beta_k}{2n-J-4} + \frac{\Lambda_{jk}}{\gamma^2}\left(1+\frac{2}{2n-J-4}\right)\right\}.
\end{aligned}$$

B Simulation parameters

The simulation parameters used in Section 4.1 are given in Table S1.

	$j=1$	$j=2$
α_j	6	6
a_1^j	0	-2
a_2^j	0	$\frac{\pi}{2}$
ω_1^j	0.6	0.1
ω_2^j	0.2	0.5
σ_1^j	0.4	0.4
σ_2^j	0.4	0.4

Table S1: *Simulation parameters for Section* 4.1.

C MALA simulation study

Simulations based on a discretization of the Langevin diffusion process are not exact. In Section 3.3, we suggested that the acceptance rate of the Metropolis-adjusted Langevin algorithm (MALA) could be used to measure the discrepancy between the true and the approximated processes. In the MALA, the transition density of the discretized Langevin diffusion is used as the proposal distribution of a Metropolis sampler. At each iteration, a point is drawn from the proposal distribution, and it is accepted or rejected with some probability. In this appendix, we use simulations to evaluate the acceptance rate of the MALA, for different time steps of discretization.

We considered the utilisation distribution used in the simulations of Section 4.2, as the target distribution. We simulated from the MALA on the target distribution, at nine different (regular) time steps of discretization,

$$\Delta \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5\}.$$

For each, we simulated ten tracks of length $T = 1000$ from the MALA, with speed parameter $\gamma^2 = 5$. We counted the average proportion of rejected steps over the ten tracks. Figure S1 shows the acceptance rates for the different time discretizations.

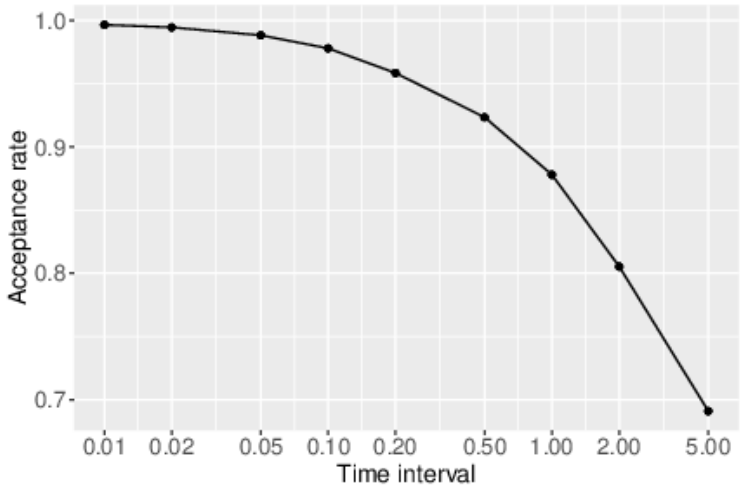

Figure S1: *Acceptance rates in simulations from the Metropolis-adjusted Langevin algorithm. The x axis is on the log scale.*

As expected, the acceptance rate decreased as the time interval of discretization increases. This is because the Euler discretization is only valid at a fine time resolution, and the approximation becomes worse on a coarse time scale. The acceptance rate tends to 1 when the step of discretization decreases, i.e. when the transition density becomes a better approximation of the Langevin diffusion process. In this simulation scenario, the average acceptance rate was around 99.7% for $\Delta = 0.01$, and around 69.1% for $\Delta = 5$.

D Bilinear interpolation gradient

In this appendix, we derive the analytical expression of the gradient for the bilinear interpolation of a surface. We used the bilinear interpolation to interpolate the piecewise-constant covariate functions, in the Langevin movement model.

Let f be the two-dimensional function of interest. We want to approximate f and its gradient at a point (x, y). The point is in a grid cell delimited by x_1 (lower) and x_2 (upper) along the x axis, and by y_1 (lower)

and y_2 (upper) along the y axis. The function f is known at the four corner points of the grid cell, and we write

$$\begin{aligned} f(x_1, y_1) &= f_{11} \\ f(x_1, y_2) &= f_{12} \\ f(x_2, y_1) &= f_{21} \\ f(x_2, y_2) &= f_{22}. \end{aligned}$$

The bilinear interpolation of f at the point (x, y) is

$$\hat{f}(x, y) = \frac{y_2 - y}{y_2 - y_1}\left(\frac{x_2 - x}{x_2 - x_1} f_{11} + \frac{x - x_1}{x_2 - x_1} f_{21}\right) + \frac{y - y_1}{y_2 - y_1}\left(\frac{x_2 - x}{x_2 - x_1} f_{12} + \frac{x - x_1}{x_2 - x_1} f_{22}\right).$$

For convenience, we rewrite it as

$$\hat{f}(x, y) = \frac{1}{(y_2 - y_1)(x_2 - x_1)}\Big\{(y_2 - y)\big[(x_2 - x)f_{11} + (x - x_1)f_{21}\big] + (y - y_1)\big[(x_2 - x)f_{12} + (x - x_1)f_{22}\big]\Big\}.$$

We can then derive the partial derivatives of the interpolated function,

$$\begin{aligned} \frac{\partial \hat{f}}{\partial x}(x, y) &= \frac{(y_2 - y)(f_{21} - f_{11}) + (y - y_1)(f_{22} - f_{12})}{(y_2 - y_1)(x_2 - x_1)} \\ \frac{\partial \hat{f}}{\partial y}(x, y) &= \frac{(x_2 - x)(f_{12} - f_{11}) + (x - x_1)(f_{22} - f_{21})}{(y_2 - y_1)(x_2 - x_1)}. \end{aligned}$$

E Simulation study with temporally irregular data

In this appendix, we investigate the performance of the method for locations collected at irregular time intervals. We followed the procedure described in Section 4.2 of the paper to simulate data. Instead of thinning the 100 simulated tracks regularly, we thinned them randomly to obtain irregular time grids. We then fitted the Langevin movement model to all simulated tracks, and compared the results with those obtained for regular data in Section 4.2. The results are presented in Table S2. The estimates of all model parameters are very similar in the regular and irregular simulation scenarios, for similar interval lengths. This confirms that the method of inference presented in this paper can be applied similarly to data collected at regular or irregular time intervals, due to the continuous-time formulation of the Langevin movement model.

F Checking residuals in the Stellar sea lion study

In this appendix, we investigate the goodness of fit of the Langevin movement model, for the Steller sea lion analysis of Section 5. We use the standardized linear model residuals, obtained as

$$\tilde{r}_{i,d} = (Y_{i,d} - \hat{Y}_{i,d})/\sqrt{\hat{\gamma}^2},$$

where $d \in \{1, 2\}$ is the dimension, and $\hat{Y}_{i,d}$ is the predicted value of $Y_{i,d}$. If all assumptions of the model are satisfied, the residuals follow a standard normal distribution in each dimension.

	regular		irregular	
	$\bar{\Delta} = 0.05$	$\bar{\Delta} = 0.2$	$\bar{\Delta} = 0.05$	$\bar{\Delta} = 0.2$
β_1	3.15 (1.85, 4.32)	2.31 (0.94, 3.77)	2.97 (1.84, 4.25)	2.18 (0.87, 3.47)
β_2	1.69 (0.38, 2.91)	1.29 (0.02, 2.52)	1.53 (0.42, 2.73)	1.23 (-0.27, 2.52)
β_3	-0.16 (-0.51, -0.01)	-0.18 (-0.58, -0.02)	-0.16 (-0.53, -0.02)	-0.18 (-0.61, -0.03)
γ^2	4.99 (4.90, 5.08)	4.95 (4.74, 5.17)	4.99 (4.89, 5.09)	4.96 (4.77, 5.15)

Table S2: *Comparison of estimates of the Langevin model parameters, for regular and irregular time intervals of simulation.* $\bar{\Delta}$ *is the mean time interval of the simulated data. For each scenario,* 100 *tracks were simulated, so* 100 *estimates were obtained for each parameter. This table gives the mean of the* 100 *estimates, as well as the* 2.5% *and* 97.5% *quantiles* (*in brackets*).

Figure S2 shows a quantile-quantile plots of the standardized residuals against the standard normal distribution, for the individual model fitted to SSL2. We present the results for SSL2 because it is the shortest track, and the plots are more easily readable, but similar observations can be made for all fitted models. The quantile-quantile plots do not align on the $y = x$ line, which indicates that the standardized residuals do not follow a standard normal distribution.

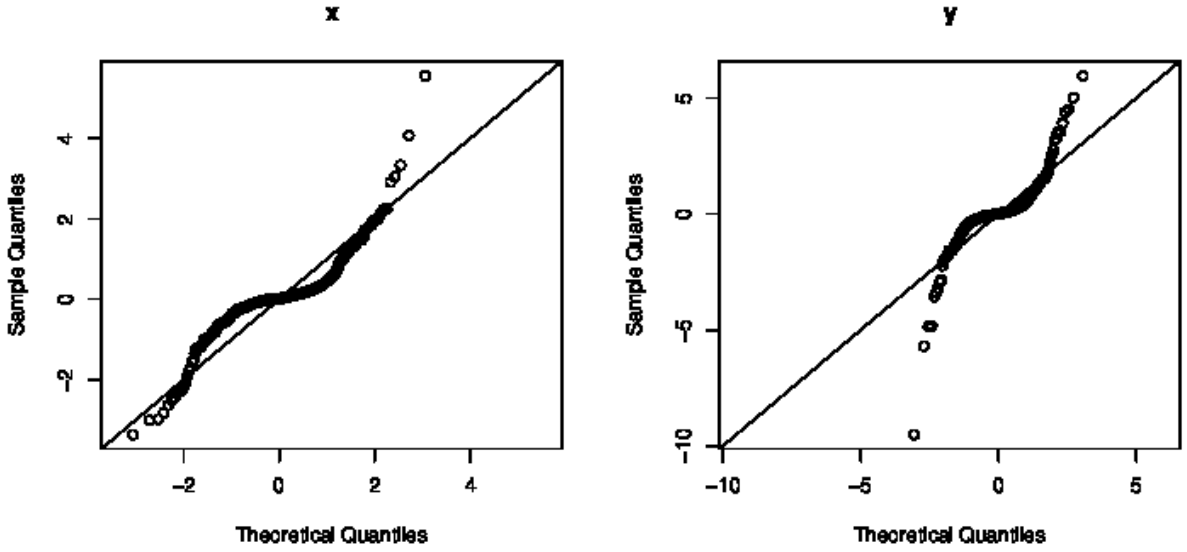

Figure S2: *Quantile-quantile plots of the standardized residuals against the standard normal distribution.*

To investigate the origin of this non Gaussian feature, we produced a map of the predicted steps and the observed steps, for comparison. This is shown in Figure S3. The amplitude of the residuals is often large in periods of transit between sites of interest, when the animal is moving faster. That is, the model fails to predict long steps. This could be improved with a more flexible formulation for the speed parameter γ^2, which we discuss in the manuscript (Section 5). We also see that the predicted direction of movement is determined by the gradient of the utilisation distribution. In this model, the predicted steps therefore point towards shallower areas, and towards sites of interest, due to the estimated habitat selection parameters (Table 1). As a result, the model fails to predict displacements away from the sites of interest, for example. This suggests that the covariates included in this model do not capture some of the structure found in the observed movement tracks.

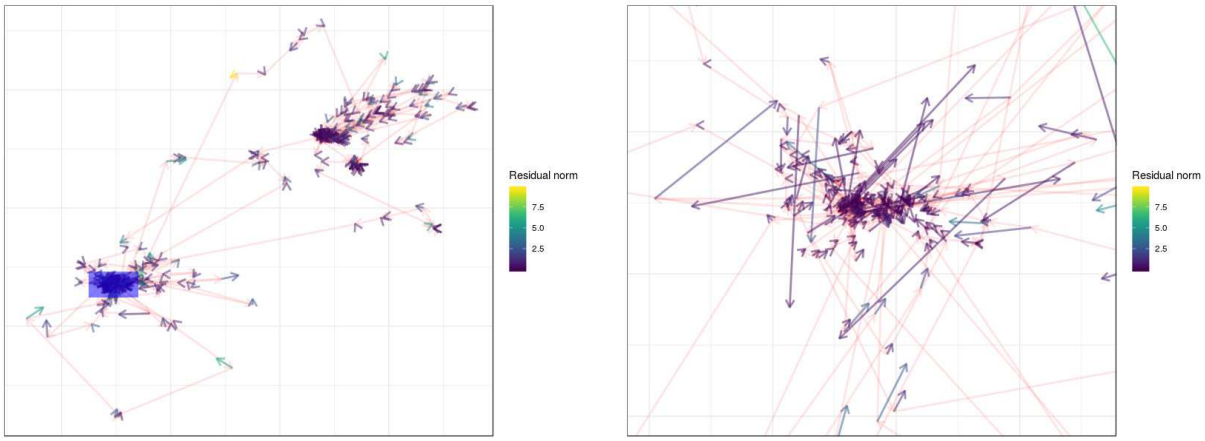

Figure S3: *Predicted and observed steps for SSL2. The red segments are the observed steps of the individual (filtered using the R package crawl), and the arrows are the predicted steps starting from each sampled point. Their colour shows the amplitude of the residuals (the error in the predicted step length). The right plot magnifies the region delimited by a blue rectangle in the left plot.*