Novel statistical methods for inferring human impacts on animal movement and migration from large-scale datasets

Majaliwa M. Masolele

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Mathematics and Statistics
College of Science and Engineering
University of Glasgow

July 2025

Abstract

Multiple stressors contribute to the decline of numerous animal species within and outside protected areas worldwide. While our understanding of anthropogenic habitat loss and degradation, climate change, and anthropogenic pressures as potential drivers of these declines is improving, we still lack a mechanistic understanding of how their fine-scale effects translate into animal movement decisions and how these decisions ultimately influence survival and, in turn, population dynamics at a broad scale. Unravelling these patterns requires associations of spatial covariate fields and fine-scale movement data, typically collected using Global Positioning System (GPS) tags deployed on animals. These tags provide a bivariate time series of coordinates at defined intervals, facilitating insights into how animals move, where and when they forage, and the nature of both intra- and interspecific interactions.

Despite the availability of such movement data alongside the corresponding environmental data, significant analytical challenges persist. Habitat selection models, particularly resource selection functions (RSFs) and step selection functions (SSFs), represent a fundamental tool to identify the characteristics of suitable habitats for animals at both broad and fine scales. The core concepts underlying these methods are based on the ratio between habitat availability and habitat use by the animal. However, while these models enhance our understanding of habitat suitability, they often yield divergent conclusions

even when applied to the same datasets, likely due to differences in their spatial and temporal scales of operation. A pressing question, therefore, is how parameters derived from fine-scale movement models can be reconciled to produce the patterns similar to those from broad-scale models and thereby improving our understanding of how animals' use of space relates to the distribution of resources, risks, and environmental conditions. Addressing this challenge requires a modelling framework that enables parameter scalability, quantifies uncertainty, and remains computationally efficient while capturing the influence of spatial covariate fields, such as human-made infrastructure.

The objective of this thesis is to advance our understanding of how animal space use relates to the distribution of resources, risks, and environmental conditions by integrating and developing state-of-the-art multiscale statistical methods within a Bayesian framework while maintaining computational efficiency. This will enhance our ability to assess how animals respond to changing landscapes and climate conditions, predict future spatial distributions based on current patterns, identify the key drivers that displace or restrict animals from otherwise suitable habitats, and pinpoint critical habitats that should be preserved from human alteration. Throughout this thesis, I will focus on models of animal movement, particularly habitat selection models, and contribute to expanding the array of statistical methods available for analysing movement data. An overarching goal of the thesis is to develop methods that can be applied to the study of the Serengeti wildebeest migration, a vital ecological process in one of the most biodiverse ecosystems on earth. I will begin by reviewing existing and widely used methods in the literature. Subsequently, I introduce a multiscale step selection model that facilitates the estimation of long-term animal space use without requiring simulations from the fitted model, and I will leverage variational inference within a Bayesian framework to estimate selection and avoidance parameters from movement observations and environmental data while

demonstrating the importance of formally quantifying uncertainty in these estimates.

The focus then shifts to examining the effects of anthropogenic structures, such as buildings, on the spatial distribution of migratory wildebeest using multiscale inference from the previous chapter. This analysis will provide insight into whether wildebeest select or avoid areas near buildings and how these selection patterns influence their space use at the population level within the ecosystem. These findings will be essential for a later chapter, where I simulate how wildebeest space use is expected to change in response to the introduction of new additional buildings in the ecosystem.

In Chapter 5, I use hierarchical sparse Gaussian processes to estimate the mean migration routes of the Serengeti wildebeest population. These modelled routes form the basis for improving spatial predictions of where wildebeest are likely to spend most of their time during critical life-history stages such as calving, weaning, rutting, or migration. This is achieved by integrating wildebeest space use patterns derived from local environmental features such as anthropogenic structures, as detailed in Chapter 4 with the population mean migration routes inferred here. The latter are used as a proxy for the influence of long-term spatial memory on movement decisions. This integrative modelling framework offers a more ecologically grounded understanding of wildebeest spatial distribution across specific days of the year and during key life-history events.

In Chapter 6, I will develop a novel simulation approach to model the placement of buildings in different scenarios and explore the impact of different allocation strategies on wildebeest space use. This will be achieved by simulating hypothetical building distributions using a nonlinear preferential attachment rule to place buildings at specific locations and incorporating an accept-reject mechanism to increase and decrease building clustering. Then I will estimate the new patterns of wildebeest space use using the methodology introduced in chapter 4 and quantify the shift from observed space use by

employing the Kullback-Leibler divergence.

This thesis demonstrates that multiscale animal movement models provide valuable insights into how animal space use is shaped by the distribution of resources and risks in changing landscapes. A key finding is that considerable uncertainty can persist even in large telemetry datasets, underscoring the importance of quantifying uncertainty in resource selection analyses. The study on the spatial distribution of migratory wildebeest reveals that while these animals tend to avoid areas near anthropogenic structures, this behavior does not lead to complete exclusion. Instead, it results in a reduced duration of time spent in the vicinity of such structures. Furthermore, the study incorporating local environmental responses with long-term spatial memory effects reveals that spatial predictions of wildebeest distribution during key life-history stages, such as calving, are improved by reducing uncertainty about where populations are most likely to spend time on specific days or during particular events. Finally, a simulation study indicates that the impact on wildebeest space use is more pronounced when new developments occur in previously undeveloped regions or in isolation from existing infrastructure, highlighting the importance of strategic spatial planning in conservation efforts.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my deepest gratitude to my PhD supervisors, Prof. Colin J. Torney and Prof. J. Grant C. Hopcraft, whom I first met eight and nine years ago, respectively. Since those early encounters, you have both been unwavering in your support, and I still struggle to find adequate words to convey the depth of my appreciation. Your mentorship has profoundly shaped the trajectory of my academic and professional life, and I would not have had the courage or confidence to pursue this path without your guidance. Your generosity, encouragement, and insightful feedback have continuously pushed me beyond what I believed possible and made my PhD journey an extraordinary experience. Thank you for your patience, your invaluable teaching, and your steadfast commitment to my academic development. I am especially grateful for the countless hours you devoted to reviewing ideas, refining analyses, and shaping this thesis, often rescuing me from the mire of confusion. I consider myself incredibly fortunate to be part of your academic lineage.

I am also sincerely thankful to the University of Glasgow for granting me the opportunity to pursue my PhD and for the financial support provided through the Engineering and Physical Sciences Research Council (EPSRC) scholarship.

I also thank Professor Juan Morales and Drs. Christina Faust, Thomas Morrison, and Peter Stewart for their valuable time, insight, and constructive discussions on some part

of this thesis.

I would like to thank my examiners, Professor Rafael de Andrade Moral and Dr. Jafet Belmont Osuna for their dedication and insightful discussion of my PhD thesis.

I extend heartfelt thanks to Drs. Jared Stabach, Lacey Hughey, and Katherine Mertes at the Conservation Ecology Center, Smithsonian's National Zoo and Conservation Biology Institute, USA. I am deeply grateful for your mentorship, enthusiasm and the opportunity to contribute to research on animal movement, remote sensing, and spatial analysis while working remotely from Tanzania after completing my Master's degree. This experience was both formative and unforgettable.

My sincere appreciation goes to my PhD assessors, Prof. Christina Cobbold and Dr. Robert Teed, for their critical engagement with my work and for consistently reinforcing the importance of precise mathematical notation during annual progress reviews.

I would also like to acknowledge the camaraderie and support of my fellow PhD students, both current and former, from the School of Mathematics and Statistics and the School of Biodiversity, One Health, and Veterinary Medicine. In particular, I am grateful to Cyrus Kavwele, Zabibu Kabalika, Ronald Vicent, Dennis Minja, Andrea Kipingu, Mecklina Michael, Houssein Kimaro, Elkana Maghembe, Pietro Colombo, the Levehume Scholars, office mates in room 217, and many others. The friendships and memories we shared throughout this journey are truly invaluable.

Finally, and most importantly, I wish to thank my family: my parents, brothers, and sisters, for their boundless love, patience, and unwavering support throughout this journey. I am very lucky to have you in my life !!!!

Declaration

I hereby affirm that the contents of this thesis is valid, original, and accurate, except where external sources have been cited and properly referenced. The thesis has been completed under the supervision of Professor Colin J. Torney and Professor J. Grant C. Hopcraft. This thesis has not been submitted, in whole or in part, for consideration toward any other degree or qualification at this or any other university.

Chapter 1

Introduction

Movement is a fundamental ecological and biological process that influences the life histories, physiology, morphology, and behaviour of animals. This process is typically driven by an animal's ability to navigate, experiential knowledge, physical fitness, decision-making related to movement, and the abiotic or biotic factors it encounters (Nathan et al., 2008). Animal movement encompasses a wide range of behaviours and activities, which subsequently elucidate complex relationships between animals and their environments. Animal movement behaviours can be categorised into two major groups: reproduction-related movement (dispersal) and resource-related movement (which includes movement behaviours such as migration, sedentarism, and nomadism). Dispersal refers to moving away of matured individuals from an origin (birthplace) to a new site for reproduction. Sedentarism involves an individual inhabiting a relatively small area with a stable home range or territory where the resources have little variability. Nomadism involves seasonal movement within a defined territory without permanent settlement and there is spatiotemporal unpredictability of resources, while migration involves seasonal movements between spatially distinct areas where there is

seasonal variation of resources in space and time (Mueller and Fagan, 2008). These categories differ in movement pathways, movement distance, timing, and the shapes and sizes of home ranges, which are crucial for planning and implementing various wildlife management strategies. Therefore, investigating ecological questions concerning the timing, mechanisms, reasons and locations of animal movements is essential to enhance our understanding of their mobility, fitness, survival, and broader ecological processes and functions (Fagan and Calabrese, 2014).

In the last decade, there has been significant progress in our capacity to collect animal movement data at a very high spatial and temporal resolution that improves our understanding of the dynamics of animal movement (Joo et al., 2020; Nathan et al., 2022). This progress is coupled with advances in mathematical and statistical methodologies, facilitating the extraction of critical movement characteristics and the understanding of the factors driving observed movement patterns (Hooten et al., 2017). At the heart of movement ecology, random walk models applied in both continuous and discrete time frameworks (Kareiva and Shigesada, 1983), have been playing a crucial role in the advancement and development of various methods and models that are being applied in studying animal movement. These models serve various purposes, including the identification of distinct behavioural states (e.g. stationary versus exploratory) within movement data (Morales et al., 2004), deriving estimates of animal home ranges by leveraging trajectory autocorrelation (Fleming et al., 2015), delineating spatially or temporally dynamic migration routes (Gurarie et al., 2017), and assessing the role of social interactions on movement decisions (Haydon et al., 2008; Torney et al., 2018).

Likewise, statistical modelling approaches such as step selection functions (Forester et al., 2009; Fortin et al., 2005; Thurfjell et al., 2014), state-switching step selection functions (Klappstein et al., 2023; Pohle et al., 2024), and hidden Markov models (Langrock

et al., 2012) have been used to understand the drivers of animal movement and behavioural state at the fine scale by incorporating environmental covariates. More recently, non-parametric methodologies have emerged, enabling the integration of continuous and time-varying movement parameters into models (Torney et al., 2021). An important next step in advancing these models is quantifying spatially varying parameters that can be translated across scales and identifying the corresponding environmental factors that influence the dynamics of animal movement. Achieving this requires the development of flexible, data-driven statistical models capable of capturing intricate spatial patterns and the nonlinear behavioural responses animals exhibit, while also efficiently handling the substantial volume of data required for accurate prediction and inference (Paun et al., 2022).

1.1 Modeling animal movement and its drivers

Addressing ecological questions such as when, how, why and where animals move is vital to advance our knowledge about their movement, habitat use, behavioural patterns, and survival, particularly in the context of global landscape and climate change (Fagan and Calabrese, 2014; Nathan et al., 2008). The variability in movement strategies among animals, even within the same species, can be attributed to differences in personalities, fitness levels, motivations, environmental conditions (Hooten et al., 2017), past experiences, and interactions with conspecifics and other species (Majaliwa et al., 2022). Consequently, there is significant interest in the mathematical and statistical fields in quantifying, classifying, and measuring this heterogeneity in movement behaviour. However, the rapid increase in fine-scale movement data, facilitated by advances in telemetry technology (Kays et al., 2015), has outpaced the development of appropriate models and methods, posing a new mathematical challenge in measuring and disentangling pat-

terns from these detailed animal movement data.

Over the past four decades, movement ecology has experienced several transformative step changes in the modelling of animal movement and its underlying drivers. These advances have been primarily driven by progress in data collection technologies, theoretical frameworks, and computational methods. Collectively, these paradigm shifts have fundamentally changed the way ecologists and wildlife biologists collect, process, and analyse movement data to address key ecological questions. Rather than detailing the full historical progression of modelling approaches and methods used to understand the patterns and drivers of animal movement, this thesis focuses specifically on habitat selection models. These models provide ecological insights into how animals select or avoid resources, using data derived from spatial surveys and telemetry.

The early era of habitat selection models in movement ecology relied primarily on descriptive analyses based on direct observations, mark-recapture techniques, and basic telemetry data. For instance, to infer animal habitat use, ecologists and wildlife biologists often compared the number of observations within each habitat category using non-parametric tests such as the Friedman test (Friedman, 1937) and Quade's weighted ranking test (Quade, 1979). Other approaches involved comparing observed versus expected occurrences in habitat categories using parametric tests such as the Chi-square goodness-of-fit test in combination with Bonferroni confidence intervals (Neu et al., 1974), and Manly's selection ratio (Manly et al., 2007). However, these methods were limited in both temporal and spatial resolution and could only accommodate categorical variables, thereby restricting the depth and scope of ecological inference.

The advent of radio-tracking technologies, such as GPS tags, marked a significant advancement in animal movement studies by enabling ecologists and wildlife biologists to access continuous, fine-scale trajectories of animal movement. This revolution in data

collection facilitated more robust analyses of spatial behaviour and resource use. In parallel, the development of resource selection functions (RSFs) provided a statistical framework to analyse habitat selection using a use–availability design, with model parameters typically estimated via logistic regression (Manly et al., 2007). RSFs allowed for the integration of categorical and continuous environmental variables. However, as the temporal resolution of tracking data increased, the assumptions underlying RSFs, particularly the lack of temporal autocorrelation among observations of the same individual, limited their applicability. These limitations led to the development of step selection functions (SSFs), which account for spatial and temporal autocorrelation by conditioning available steps on the animals movement characteristics (Forester et al., 2009; Fortin et al., 2005; Thurfjell et al., 2014). While SSFs offer improved modelling of fine-scale resource use, they treat movement and resource selection as separate processes, which can introduce bias in the estimation of selection and avoidance parameters. To overcome this issue, integrated step selection functions (iSSFs) were introduced (Avgar et al., 2016). This framework jointly models movement and resource selection, enabling simultaneous inference of both components. iSSFs provide insights into key ecological questions, such as when and where an animal moves faster or is more likely to remain in or leave a particular habitat. More recently, multiscale step selection functions (MSSFs) in discrete-time (Michelot et al., 2020, 2019a) and continuous-time (Michelot et al., 2019b) have been proposed, extending this approach to infer parameters that are translated across multiple spatial scales. These methodological advancements continue to transform the field of movement ecology, offering novel tools to understand how animals interact with dynamic landscapes and respond to anthropogenic change.

1.2 Scaling to populations

Understanding the distribution of animals across landscapes at various spatial and temporal scales requires scaling up from individual-level or fine-scale movement decisions to long-term, broad-scale patterns observable at the population level. A significant challenge in modern movement ecology is to bridge this gap, translating individual or fine-scale movement decisions into population-level inferences (Hawkes, 2009; Holdo and Roach, 2013; Torney et al., 2018). This difficulty largely stems from the fact that commonly used standard models, such as step selection functions (SSFs), are typically designed to capture fine-scale movement decisions or local habitat selection, making it challenging to extrapolate findings to broader spatial patterns such as annual migrations that are representative of the population level.

To address this issue, species-habitat association studies often employ hierarchical (or random effects) models that account for inter-individual variation when inferring population-level responses. Other common strategies include averaging model coefficients across individuals (Hooten et al., 2017) or employing resampling techniques such as bootstrapping (Fieberg et al., 2020). However, averaging coefficients can introduce bias, particularly when a subset of individuals exhibit strong selection responses that overshadow weaker or opposing trends among others (Holloway and Miller, 2014). This may lead to an inaccurate representation of population-level responses, either overestimating or underestimating the true effect.

Moreover, the predicted population-level space use derived from fine-scale models such as SSFs often fails to match the spatial distributions generated by broad-scale models such as resource selection functions (RSFs), even when their underlying habitat selection functions appear similar (Barnett and Moorcroft, 2008; Moorcroft and Barnett, 2008;

Signer et al., 2017). This discrepancy highlights the limitations of using standard parametric long-term space use equations to capture broad-scale patterns, i.e. population space use. Therefore, there is a critical need for multiscale modelling approaches in the parameterisation of SSFs, specifically methods that ensure fine-scale movement decisions translate directly into broader-scale space-use patterns, without requiring further analyses such as simulation from the fitted model. While SSF and RSF models are often formulated as Poisson point processes over continuous space ($\mathbb{R}^2$), this formulation alone does not guarantee that the parameters estimated from SSFs will yield the correct long-term utilisation distribution. In other words, coefficient estimates from SSFs do not automatically scale to match the distribution implied by the RSF parameters (Michelot et al., 2019a). Multiscale approaches that have model structures that ensure that local movement steps align with long-term utilisation distribution can resolve this mismatch. By ensuring consistency between the step selection process and the implied RSF, these frameworks produce stationary distributions that directly reflect the intended habitat preferences. Such frameworks would enable transferable parameter estimates across spatial scales, improve the robustness, predictive accuracy, and generalisability of movement models across different landscapes and population contexts.

Several statistical approaches, including multiscale inference frameworks, have been proposed to ensure that parameters derived from fine-scale models yield spatial distributions consistent with those from broad-scale models. One such approach conceptualises animal movement as analogous to a Markov chain Monte Carlo (MCMC) sampler operating in parameter space, an idea formalised as MCMC step selection (Michelot et al., 2020, 2019a). This framework enables direct application of parameters estimated from SSFs to broad-scale space-use predictions, owing to its scalability across the temporal and spatial scale. While this method facilitates the estimation of space-use patterns without requir-

ing computationally intensive numerical simulations, the original formulation relies on the numerical approximation of the Hessian matrix to estimate parameter uncertainty. Consequently, it does not yield full posterior distributions for model parameters. This limitation is nontrivial: neglecting to formally quantify uncertainty in movement models can result in overconfident habitat selection estimates and unreliable predictions of long-term space use. Addressing this limitation is therefore essential for advancing statistical methods in habitat selection modelling, as it highlights a critical methodological gap.

1.3 Thesis outline

This thesis is composed of seven chapters. In summary, the current chapter provides a general introduction, while Chapter 2 presents the background theory. Chapter 3 focuses on the development of efficient methods for movement data analysis, Chapters 4 and 5 on analysis of real telemetry data of animals, and Chapter 6 on simulations. Finally, Chapter 7 summarises the conclusions and outlines potential future research directions.

Chapter 2 offers a comprehensive overview of the background theory, including the methodological foundations necessary to understand the material discussed in subsequent chapters. Given that the primary focus is on habitat selection models in animal movement, I begin by introducing random walk models, which serve as the foundation for most models used in movement ecology. This is followed by the discussion of specific methods, such as resource selection functions (RSFs), and step selection functions (SSFs), all of which are well-suited for this type of analysis. In addition, I discuss the challenges associated with these approaches and introduce a relatively novel class of multiscale models, such as the multiscale step selection function (MSSF), which has opened new avenues for research in movement ecology. Lastly, I describe the Greater Serengeti-Mara ecosystem and its contemporary threats, and present modelling techniques used to un-

derstand animal responses to these threats.

In Chapter 3, I develop an approximate Bayesian multiscale step selection function that enables the prediction of broad-scale movement patterns using selection and avoidance parameters inferred from a model that is fit to fine-scale movement observations and environmental data. This method leverages GPU-enabled machine learning and is inspired by optimisation techniques that approximate the true posterior distribution of the parameters using a simple distribution.

In Chapter 4, I leverage wildebeest telemetry data and spatial data on buildings to investigate the effects on the spatial distribution of migratory wildebeest in the Greater Serengeti-Mara ecosystem. Specifically, I examine whether migratory wildebeest select or avoid buildings and how these structures influence their movement and space use within the ecosystem. I introduce a novel approach for quantifying nonlinear responses, identifying the area of influence (i.e. distance thresholds), and assessing the interactive effects of multiple anthropogenic structures.

In Chapter 5, I use hierarchical sparse Gaussian processes to estimate the mean migration routes of the Serengeti wildebeest population. These modelled routes form the basis for improving spatial predictions of where wildebeest are likely to spend most of their time during critical life-history stages such as calving, weaning, rutting, or migration. This is achieved by integrating wildebeest space use patterns derived from local environmental features such as anthropogenic structures, as detailed in Chapter 4 with the population mean migration routes inferred here. The latter are used as a proxy for the influence of long-term spatial memory on movement decisions. This integrative modelling framework offers a more ecologically grounded understanding of wildebeest spatial distribution across specific days of the year and during key life-history events.

The multiscale inference presented in Chapters 4 and 5 provides critical insights into how fine-scale effects propagate into broad-scale movement patterns. However, as the landscape continues to be modified by the addition of new anthropogenic structures, inferences drawn from these studies become increasingly limited, particularly in predicting future animal distributions based on observed responses. Consequently, a simulation-based approach becomes essential for assessing potential future distribution patterns. In Chapter 6, I therefore develop a simulation-based framework to quantify changes in migratory wildebeest space use resulting from the addition of new anthropogenic structures in the ecosystem. This framework integrates a combination of a nonlinear preferential attachment rule and an accept-reject mechanism to simulate the placement of new buildings under three future development scenarios. The simulation allows new buildings to be allocated both in areas with existing structures and in previously undeveloped regions while also accounting for variations in building clustering patterns, including both increased and decreased clustering. Following this, I estimate the simulated migratory wildebeest space use and compare it to the observed space use presented in Chapter 4 using the Kullback-Leibler divergence.

In the final chapter, Chapter 7, I summarise the key findings of my research and discuss their implications for future research directions.

Chapter 2

Background

In this chapter, I review key methods and theoretical findings from the literature, describe the Greater Serengeti-Mara ecosystem and its contemporary threats, and present modelling techniques used to understand animal responses to these threats, providing a foundation for the material presented in the following chapters.

2.1 Introduction

Understanding the factors that contribute to how and why animals move and select resources in a landscape is crucial for ecological research and wildlife management decision-making (Nathan et al., 2008). However, conservation ecologists and wildlife biologists frequently encounter difficulties in obtaining reliable population-level estimates of the movement and selection of resources by animals. These challenges stem from the inherent variability in individual behaviours within the same species or population, as habitat selection and movement patterns can differ substantially under varying environmental conditions. Furthermore, the implementation of movement models is often hindered by the complexity of computational procedures and the lack of standardised mathematical notation used to describe these models (McClintock et al., 2014). The growing number of methodological approaches for analysing animal movement data also introduces uncertainty regarding model selection for specific movement observations, thereby presenting an ongoing challenge in the mathematical modelling of animal movement.

Animal movement plays a fundamental role in shaping ecological processes, species interactions, and population dynamics. Accurately modelling this movement requires careful selection of an appropriate conceptual framework that represents how animals move through space and time. The choice of framework is primarily guided by the spatial and temporal resolution of the movement data collected, as well as the ecological questions being addressed. Animal movement data are typically obtained through technologies such as GPS collars and satellite tags, which provide high-resolution individual trajectories, or through camera traps and aerial surveys, which offer information on the spatial distribution of animals at specific locations. Trajectory-based data, which track

individual animals through space and time, are particularly well-suited for investigating fine-scale behaviours and decision-making processes, such as foraging strategies or migratory routes. In contrast, data collected at fixed locations, such as presence-absence records or estimates of animal density, are more appropriate for population-level analyses aimed at understanding spatial and temporal patterns of animal occurrence. Recognising the conceptual distinctions between these approaches is essential for selecting appropriate modelling tools and accurately interpreting movement data in ecological research. Broadly, animal movement modelling frameworks can be categorised into two primary approaches: Lagrangian, which focuses on tracking individuals over time, and Eulerian, which examines changes in population-level patterns at fixed spatial locations (Hooten et al., 2017; Smouse et al., 2010). The Lagrangian approach uses stochastic differential equations to represent the changes in an animal's position, $\mathbf{r}(t)$, within a two-dimensional space (x, y) at a given time t. Specifically, the changes in the animal's location in the two-dimensional space $(dx(t), dy(t))$ at time t are given as follows:

$$\begin{bmatrix} dx(t) \\ dy(t) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_x(\mathbf{r}(t), t) \\ \boldsymbol{\mu}_y(\mathbf{r}(t), t) \end{bmatrix} dt + \mathbf{D}(\mathbf{r}(t), t) \begin{bmatrix} d\boldsymbol{\Psi}_x(t) \\ d\boldsymbol{\Psi}_y(t) \end{bmatrix} \tag{2.1}$$

where $\boldsymbol{\mu}_{x,y}$ is the drift values, $\mathbf{D}$ is the diffusion matrix which represent animal motility in the context of animal movement, and $\boldsymbol{\Psi}$ represents the infinitesimal increments of a Wiener process in each dimension. Several special cases can be derived from the general stochastic movement model in eqn. 2.1, depending on the properties of the drift and diffusion terms. For example, if the drift components are zero and the diffusion terms along the x and y axes are independent, the model reduces to an uncorrelated random walk, in which the movement is purely stochastic with no preferred direction. In contrast, if the drift components are not independent, the model becomes a correlated

random walk, typically due to the dependence in drift or persistence of directionality. If the diffusion components remain independent and the drift components are non-zero, the result is a biased random walk, where an animal exhibits a tendency of moving in a particular direction over time (Smouse et al., 2010).

The Eulerian approach, which is place-based and typically continuous-time, is employed for population-level inferences. Unlike the Lagrangian approach, which involves tracking individual animal movement over time, the Eulerian approach instead focuses on specific locations in the environment and observes how animals move through or interact with those locations (habitats) over time. This includes modelling the spatial intensity of animals over space and time, modelling how animals or groups of animals move through a certain habitat, and the probability of animal occupancy in an N-dimensional geographical space (Hooten et al., 2017).

2.2 Random walk models

A random walk model is a stochastic process that is the building block of many mathematical models developed in studying and modelling animal movement. In animal movement, a discrete-time random walk is typically characterised by three components which are: (i) the distance an animal moves in each step (distance), (ii) the time interval between each step, and (iii) the direction in which an animal moves. In this model, animals are presumed to take a series of successive steps in random directions, where the direction or distance of movement is determined probabilistically. The movement trajectory of an individual animal is broken down into various movement characteristics such as step length (the distance between consecutive steps) and turn angles (the direction change between consecutive steps) (Kareiva and Shigesada, 1983). These movement characteristics can be translated into behavioural patterns (e.g., resting, foraging, mov-

ing) using summary statistics like mean step length and mean cosine of turn angles. Furthermore, this model enables the association of movement characteristics or states with landscape features, facilitating the understanding of behavioural changes or shifts in movement states across spatiotemporal scales (Benhamou, 2014; Morales et al., 2004). Integrating empirical movement data with ecological theory provides insight into how animals perceive and respond to their environment.

Random walk models have significantly advanced and transformed the field of movement ecology. However, as landscapes change, animals often exhibit movement behaviours that routinely repeat within defined intervals such as nesting or denning behaviour, daily foraging routes, and seasonal migration, thus violating some assumptions of random walk models. One such violated assumption is the short-term autocorrelation in individual movement paths, which random walk models typically accommodate. Due to the increasing environmental heterogeneity, animals may need to use long-term memory to recall areas with stable foraging resources or to avoid recently encountered risks, thus introducing long-term autocorrelations into their movement tracks (Fagan and Calabrese, 2014). Therefore, advancements in these models are necessary to incorporate long-term autocorrelation. This integration will not only align animal movement models more closely with biological processes, but also expand the scope of movement ecology beyond purely random movement patterns (Fagan and Calabrese, 2014).

In movement ecology, random walk models can be put into two categories; discrete-time movement and continuous-time movement models. Discrete-time movement models involve discretising movement observation into regular intervals and often lead into the results that are dependent on the choice of the discretisation step. Continuous-time movement models incorporate movement observation recorded at irregular intervals (irregular sampling intervals) in the modelling approach because they represent animal

movement as a continuous-time stochastic process, rather than being constrained to discrete time steps. This formulation allows for the estimation of an animal's location at any time point, even when no direct observation is available, by capturing the temporal autocorrelation and modelling the underlying movement dynamics between observed locations.

The two categories can be extended to describe various types of random walk model used to describe the different types of animal movement observed in ecology, such as simple random walk (uncorrelated random walk), correlated random walk, Lévy walk, and Brownian motion.

2.2.1 Simple random walk

This is the memoryless random walk model, in which the animal typically moves a constant distance and in random directions at each successive step. It is memoryless because the distance and direction of the next step in which the animal takes depend only on the current state and not on the previous step (the decision to move next is not influenced by where the animal has been in the past). In ecology, a simple random walk model is typically used when there is no clear pattern in the direction of movement of an animal. Mathematically, the location of the animal after the n steps in one-dimensional space is given by,

$$X_n = X_0 + \sum_{i=1}^{n} \epsilon_i \tag{2.2}$$

where X_n is the animal location after n steps, X_0 is the initial location, and ϵ_i is a random variable representing the i-th step. If the distribution of random variables ϵ_i is not symmetric around zero, the walk is biased.

2.2.2 Correlated random walk

This is a type of random walk model that incorporates memory, where the subsequent step and direction of the animal are influenced by the previous step, specifically the magnitude and direction of the previous step (Kareiva and Shigesada, 1983). This model is generally categorised into unbiased and biased correlated random walks. The walk is considered unbiased when there is no long-term directional preference, whereas it is biased when there exists a sustained persistence of the directional over time. The correlated random walk model is commonly applied to describe animals that exhibit a degree of directional persistence for a period before altering their course while navigating a landscape. As a result, there is often a correlation in the direction between consecutive steps taken by the animal. Such movement patterns are typically observed during activities such as foraging, transitioning between habitat patches in fragmented landscapes, and migration (Fagan and Calabrese, 2014). Mathematically, the location $\mathbf{x}_{t+1}$ of the animal at time step $t+1$ given it is at location $\mathbf{x}_t$ at time step t can be expressed as,

$$\begin{aligned} sl_t &\sim \text{Gamma}(k, \psi), \\ \phi_t &\sim \text{von Mises}(0, \kappa), \\ \theta_t &= \theta_{t-1} + \phi_t, t = 2, \cdots, n, \\ x_{t+1} &= x_t + sl_t \cos(\theta_t), \\ y_{t+1} &= y_t + sl_t \sin(\theta_t) \end{aligned} \tag{2.3}$$

where sl_t is the step length and ϕ_t is the turning angle which controls how much the direction changes from one step to the next drawn from a distribution centred around 0, and θ is the heading of the step.

2.3 Modelling animal habitat preference

Recent advances in tracking technology have resulted in an increasing availability of high-resolution spatiotemporal animal movement data (Kays et al., 2015). This volume of data provides new opportunities to address various ecological questions, such as those related to species' energy budgeting, behavioural patterns (e.g., encamped vs. migratory), and the influence of environmental covariates on species distributions (species-habitat associations) (Matthiopoulos et al., 2020). Various modelling methods are employed to study species-habitat preferences. One widely used method is the habitat selection function, which evaluates the ratio of used to available resource units of the animal. This includes resource selection functions (RSFs) (Manly et al., 2007), step selection functions (SSFs) (Forester et al., 2009; Fortin et al., 2005; Thurfjell et al., 2014), and recently developed integrated step selection functions (iSSFs) (Avgar et al., 2016). While RSFs are typically used to model broad-scale (global) habitat selection inferences, SSFs and iSSFs are applied at fine scales, particularly with high-resolution data that exhibit spatiotemporal autocorrelation.

2.3.1 Resource selection function (RSFs)

Resource Selection Functions (RSFs) are spatial statistical tools that are primarily used to calculate the probability of an animal occupying a specific location $(x \in \mathcal{D} \subseteq \mathbb{R}^2)$ by evaluating the ratio of the resource units used by the animal to alternative resource units that are theoretically available (Manly et al., 2007). This method is applied to answer ecological questions, such as which landscape features wildlife avoids or selects when moving through a landscape. RSFs employ a use-availability framework and are typically fitted using weighted logistic regression to identify the features of the landscape that the animals select or avoid. The used locations are the observed locations (e.g., ani-

mal telemetry data and spatial survey data), while the available locations are randomly generated within the range of the animal's home (home range) or study domain (Manly et al., 2007). The selection of the domain for sampling the available locations depends on the ecological questions to be addressed, which can influence the RSFs results (Northrup et al., 2013).

RSFs are evaluated across three spatial scales of habitat selection: first-order, second-order, and third-order (DeCesare et al., 2012; Johnson, 1980). First-order habitat selection involves observed telemetry locations of many individuals of a species, with availability sampled within the entire study area. Second-order habitat selection involves observed locations of a single individual, with availability sampled within that population home range. Third-order habitat selection involves observed individual telemetry locations, with availability sampled within the individual's home range. One of the strengths of RSFs is their flexibility in using both animal tracking data and spatial survey data when fitting the model. However, high-resolution animal telemetry data often suffer from positive spatial and temporal autocorrelation (Noonan et al., 2019), violating the independence and identically distributed (IID) assumption of RSFs. Traditional RSFs assume data arise from a Poisson point process, which requires non-dependent sampled data, thus not allowing for autocorrelation. To avoid this violation, telemetry data are often thinned or sub-sampled, leading to data loss, biased parameter estimates, statistical inefficiency (typically measured via autocorrelation time), and potentially missing important information on animal resource usage (Alston et al., 2023).

To remedy the issues of spatial and temporal autocorrelation in RSFs, Alston et al. (2023) proposed the Integrated Resource Selection Function (iRSF). This approach mitigates sampling bias in irregular tracking data by applying likelihood weighting, assigning weights to each observed location in an animal's movement track based on its

level of autocorrelation. This allows the inclusion of autocorrelated data through down-weighting rather than removing or thinning observations. While the iRSF avoids data thinning, it does not address ecological questions of temporal changes in resource selection, such as seasonal (i.e., dry and wet) or diurnal (i.e., day and night) variations without the data being segmented. However, such questions may be what ecologists are interested in when they focus on short-term scale conservation strategies such as implementing prescribed burning, water provisioning, and habitat manipulation. To detect temporal changes in resource selection, Dejeante et al. (2024) proposed dynamic RSFs which use a discrete-time state-space model to estimate time-varying coefficients from logistic regression (Fahrmeir, 1992). This approach allows for the identification of periods in which resource selection or avoidance coefficients are homogeneous or heterogeneous with respect to landscape features, such as woodland habitats. Despite its utility, this method produces downward-biased standard errors in the presence of auto-correlated movement data and cannot estimate long-term space use (static distribution) because the parameters are continuously changing. The next step in the evolution of RSFs is to combine the approaches of (Alston et al., 2023) and (Dejeante et al., 2024) to develop time-varying RSFs that account for both temporal autocorrelation and the influence of animal behaviour on resource selection. Conventional RSFs models rely on simplifying assumptions namely that animal locations are temporally independent, that all available habitats are equally accessible, and that the inferred parameters reflect broad-scale patterns of resource selection or avoidance. These assumptions constrain their applicability, particularly when movement data are collected at a high frequency or when geographic barriers constrain habitat accessibility. To address these limitations, alternative methods such as step selection functions (SSFs) have been developed. SSFs explicitly account for spatial and temporal autocorrelation by modelling habitat availability conditional on an animal's movement characteristics, offering a more ecologically realistic framework to

analyse resource selection at finer spatial and temporal scales.

2.3.2 Step selection function (SSFs)

SSFs are spatial statistical modelling tools that account for spatio-temporal autocorrelation when assessing the effects of environmental covariates on animal movement decisions at a fine scale. SSFs originated from the locally biased correlated random walk, where an animal's movement decisions are influenced by the local environmental conditions in the immediate vicinity. This means that the alternative steps theoretically available for the animal in each used step are generated and constrained based on the direction and distance travelled during that time step, as shown in Fig. 2.1. A common assumption in SSFs is that the animal telemetry locations are recorded in discrete time (regular intervals) (Thurfjell et al., 2014). This assumption often leads to the exclusion of observations that do not fit within the predefined time intervals, as data outside this tolerance are omitted to regularise the movement tracks.

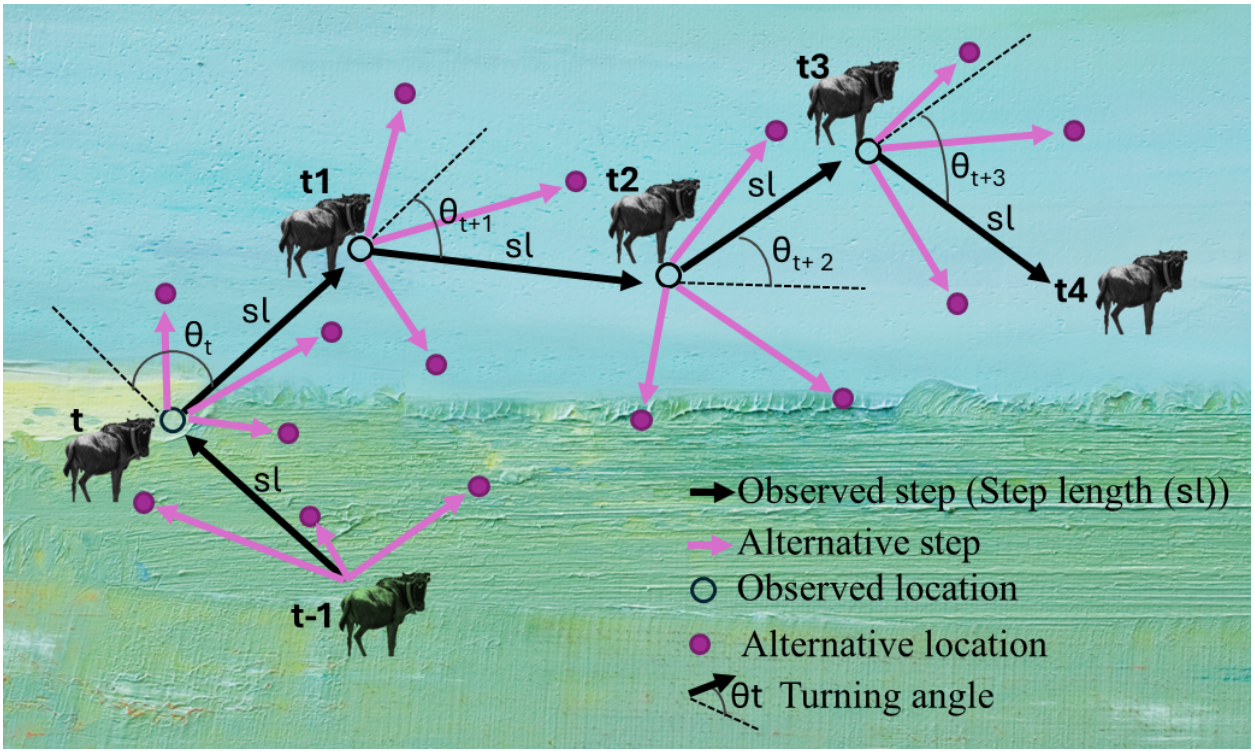

Figure 2.1: An illustration of an animal movement track with observed steps (sl)(black in color), and alternative steps (pink in color) as is commonly done in SSF analysis.

However, recent advances proposed by Hofmann et al. (2024) introduced a method for modelling incomplete animal tracking data (that is, data with irregularities) using SSFs. This approach allows for the inclusion of all animal telemetry locations in the model, thus improving the accuracy of selection and avoidance estimates. Standard SSF formulations also overlook the influence of behaviour on habitat selection, which could introduce bias into model estimates that can lead to underestimation or overestimation of uncertainty (Roever et al., 2014). Traditionally, a two-stage approach has been used to model behaviour-dependent habitat selection. First, a discrete-time hidden Markov model classifies the states of animal behaviour, and then the SSFs are fitted to each dataset corresponding to the classified behavioural states (Clontz et al., 2021; Picardi et al., 2022). This approach fails to account for the uncertainty in behavioural states and

does not allow for the dependence of state on resource selection, as it separates animal behaviour, movement, and habitat selection.

To address these limitations, Klappstein et al. (2023) and subsequently Pohle et al. (2024) have proposed the state-switching step selection function. This method unifies behavioural state dynamics and resource selection within a single-model formulation. The model is structured as a hidden Markov model, where the observation process is defined by an SSFs. Unlike the two-stage approach, the state-switching step selection jointly estimates behavioural state transitions, resource selection parameters, and state classification, allowing these components to inform one another within an integrated statistical framework.

In general, SSFs are typically fitted using standard conditional logistic regression in an exponential form with a matched case-control design, where observed steps are the cases, and alternative sampled steps are the controls, conditional on the set of available alternative steps at each movement decision point. Alternatively, maximum likelihood estimation and numerical integration can be used (Michelot et al., 2024). The likelihood function to estimate selection coefficients $\boldsymbol{\beta}$ using the maximum likelihood approach is given as,

$$L\left(\boldsymbol{\beta}|\mathbf{s}_1,\cdots,\mathbf{s}_t,\cdots,\mathbf{s}_T\right) = \prod_{t=1}^{T} \frac{\exp\left(\boldsymbol{\beta}\mathbf{X}\left(\mathbf{s}_t,t\right)\right)}{\sum_{i=1}^{n}\exp\left(\boldsymbol{\beta}\mathbf{X}(\mathbf{s}_{t,i},t)\right)} \tag{2.4}$$

where, $\boldsymbol{\beta}$ represents the vector of habitat selection parameters and $\mathbf{X}$ the matrix of covariate values at the end of step $\mathbf{s}_t$.

SSFs have become a mainstream and popular tool for ecologists in modelling animal resource selection and movement. Despite their utility in movement ecology, SSFs have several limitations. First, the parameters inferred from the SSFs vary with changes in the sampling interval of the movement observations (dependent on sampling) (Avgar

et al., 2016; Fieberg et al., 2021). In other words, if the same movement track were sampled at different intervals (i.e., 1 hour, 2 hours, 4 hours, 1 day, and so forth), the step selection functions (SSFs) parameterised from these different samples would yield varying estimates of selection. Second, SSFs provide narrow confidence intervals for selection estimates if the sampling interval (Δt) is shorter than the time required to ensure statistical independence of observed steps τ, that is, when $\Delta t < \tau$. Third, SSF-estimated parameters cannot be directly used to predict long-term space use by animals without adjustments such as simulation from the fitted model (Barnett and Moorcroft, 2008; Michelot et al., 2019a; Signer et al., 2017).

To address these issues, additional methodologies have been employed. One approach incorporates spatially structured random effects to account for unobserved spatial variation in SSFs (Arce Guillen et al., 2023). Likewise, conceptualising animal movement as a Markov chain Monte Carlo (MCMC) sampler in parameter space (MCMC step selection) has been proposed (Michelot et al., 2020, 2019a). This approach allows inferred parameters from SSFs to be used directly for broad-scale space use predictions because of their scalability in time and space. However, this framework does not account for the influence of behaviour on resource selection, assuming instead that an animal's selection or avoidance of resources remains constant regardless of the behavioural motivations behind movement. Additionally, the current multiscale modelling framework does not formally quantify the uncertainty in the movement and resource selection and avoidance estimates. Addressing this limitation is crucial for advancing the statistical methodologies used in habitat selection modelling, as it represents a significant methodological gap.

2.4 Bayesian statistics

Bayesian methods belong to a statistical modelling inference framework that interprets probability as a measure of belief or confidence, not just as a limiting frequency of outcomes, and uses Bayes rule to update the initial prior belief or hypotheses with new evidence. The initial prior may be based on prior belief, domain-specific assumptions, and expert knowledge. Inference is typically performed using sampling-based approaches such as Markov chain Monte Carlo (MCMC) methods or approximate inference techniques such as variational inference (VI).

2.4.1 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a class of powerful computational algorithms designed to generate samples from complex probability distributions, particularly when direct sampling is infeasible (Brooks et al., 2011; Gilks et al., 1995). These methods are especially powerful in settings where the distribution is only known up to a normalising constant, that is, where the functional form of the probability density is known but difficult to integrate analytically. The fundamental idea is to generate samples from a target distribution $\pi(\mathbf{z})$, defined over a state space $\mathcal{Z}$, by constructing a Markov chain whose stationary distribution coincides with $\pi(\mathbf{z})$. This approach effectively defines a random walk over $\mathcal{Z}$, so that as the chain evolves, its marginal distribution converges to the target distribution. In the context of Bayesian inference, the target distribution of interest is the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{D})$, where $\mathbf{D}$ denotes the observed data and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ represents the model parameters. Consequently, the state space becomes the parameter space $\boldsymbol{\Theta}$.

To ensure that the Markov chain converges to the desired stationary distribution, it is

necessary to satisfy the detailed balance condition. Let $\rho(s,t) = p(X_{n+1} = t \mid X_n = s)$ denote the transition probability of the Markov chain, where $s,t \in \mathcal{Z}$ and X_n is the stochastic process defined by the chain. Then, the detailed balance condition holds if there exists a probability distribution κ on $\mathcal{Z}$ such that:

$$\kappa(s)\rho(s,t) = \kappa(t)\rho(t,s), \quad \forall s,t \in \mathcal{Z}. \tag{2.5}$$

When the detailed balance condition is satisfied, κ becomes a stationary distribution of the Markov chain, which means the transition probability can be defined to guarantee that the chain will converge to the target distribution $\pi(\mathbf{z})$. The samples drawn from this chain are then used to approximate expectations, quantiles, and other statistical summaries of interest. The Monte Carlo component refers to using these samples to compute numerical estimates, while the Markov chain component ensures that each sample depends only on the previous one, satisfying the Markov property.

In Bayesian inference, MCMC methods provide a general framework to obtain information on the distributions and estimate posterior distributions of a set of unknown parameters using a stochastic sampling process (Gelman et al., 2013). One of the strengths of MCMC is its ability to sample directly from the posterior distribution for a wide class of models, where the shape of the posterior is determined by both the likelihood function and the prior distribution.

However, this advantage comes at a computational cost. Generating a sufficient number of samples to approximate the posterior accurately can be time-consuming, particularly in high-dimensional or hierarchical models. Nevertheless, MCMC remains a cornerstone of modern Bayesian analysis due to its ability to handle analytically intractable posteriors. The posterior distribution is typically defined via Bayes' rule, which combines the

prior distribution and the likelihood function as follows:

$$p(\boldsymbol{\theta}|\mathbf{D}) = \frac{p(\mathbf{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{D})} \tag{2.6}$$

where $p(\boldsymbol{\theta}|\mathbf{D})$ is the posterior distribution, $p(\mathbf{D}|\boldsymbol{\theta})$ is the likelihood function, $p(\mathbf{D})$ is the marginal likelihood, and $p(\boldsymbol{\theta})$ is the prior probability, reflecting our prior beliefs about the parameters $\boldsymbol{\theta}$ before seeing the data. The selection of an appropriate prior distribution may be based on information from previous published studies, past experience, expert opinion, or theoretical understanding. While priors help constrain models appropriately, they also introduce assumptions that should be examined critically. Regardless of type, all priors carry some information and influence the resulting posterior distribution, particularly when data are limited.

Priors are typically classified as non-informative, weakly informative, informative, or conjugate (Gelman et al., 2013):

Non-informative priors (also called vague, flat, or diffuse) are intended to exert minimal influence on the posterior distribution. A common example is a uniform distribution, such as $\theta \sim \text{Unif}(0,1)$, which assigns equal probability across a range. While used when prior knowledge is limited, these priors still encode assumptions and must be specified within an ecologically or scientifically plausible range.

Weakly informative priors moderately constrain parameters based on general domain knowledge. They help stabilise estimates, reduce overfitting, and prevent implausible inferences particularly useful in small or noisy datasets. These priors balance flexibility with realism and contribute information without overwhelming the data or heavily influencing the posterior.

Informative priors are based on existing data, pilot studies, or expert knowledge. When

well specified, they improve precision without introducing bias and are most useful when prior understanding of the ecological system studied is strong. Their assumptions must be transparent and justified. Sensitivity analyses should be used to assess how different informative priors influence the results. For example, informative priors can be used to rule out biologically implausible parameter values, such as a negative relationship between weight and length in humans. When multiple plausible priors exist, exploring a range from weakly to strongly informative can provide valuable insights into model robustness.

Finally, a **conjugate prior** belongs to the same distribution family as the posterior (e.g., a gamma prior for exponential likelihood yields a gamma posterior or a gamma prior for Poisson likelihood yields a gamma posterior). Conjugate priors simplify analytical and computation in complex or high-dimensional models but should only be used when they realistically represent prior beliefs.

Within this subsection, I now briefly introduce the Metropolis-Hastings (MH) algorithm and Hamiltonian Monte Carlo (HMC), the latter being one of the most widely used and efficient MCMC sampling algorithms in modern Bayesian computation, as it uses gradient information to efficiently explore complex posterior distributions, enabling large, informed moves through low probability regions and reducing sample autocorrelation.

Metropolis-Hastings algorithm

The Metropolis-Hastings (MH) algorithm is a widely used Markov Chain Monte Carlo (MCMC) method for obtaining a sequence of random samples from complex, high-dimensional probability distributions, especially when direct sampling is computationally prohibitive (Hastings, 1970). The MH algorithm generates a sequence of samples by proposing candidate points from a proposal (or jumping) distribution and then deciding

whether to accept or reject each candidate based on a ratio that ensures the stationary distribution of the chain matches or converges to the target distribution.

In particular, the algorithm allows simulation from a parameter's posterior distribution without requiring the calculation of the intractable normalising constant of the probability density. This is particularly useful in Bayesian inference, where the posterior distribution is proportional to the product of the likelihood and the prior, and the normalising constant (often referred to as the marginal likelihood) is typically difficult to compute directly, especially for complex models.

The MH algorithm works as follows (Hastings, 1970):

1. **Proposal Step:** Given the current state x_t, a new candidate x' is proposed by sampling from a *jumping (or proposal) distribution* $q(x'|x_t)$, which is typically chosen based on the target distribution's properties. This proposal distribution generates candidate samples in the parameter space.

2. **Acceptance Step:** The proposed candidate x' is accepted with probability given by the *Metropolis acceptance ratio*:

$$\alpha(x_t, x') = \min\left(1, \frac{p(x')q(x_t|x')}{p(x_t)q(x'|x_t)}\right)$$

   where $\alpha(x_t, x')$ is the is the probability of accepting x', $p(x_t)$ is the target distribution, and $q(x'|x_t)$ is the proposal distribution. The ratio compares the relative probability of the proposed sample to the current sample, ensuring that states with higher target probability are more likely to be accepted.

3. **Repeat:** If the candidate x' is accepted, the chain moves to x'. Otherwise, the chain remains at x_t. This process is repeated for many iterations, and the resulting

chain of samples converges to the target distribution as the number of iterations increases.

The efficiency of the MH algorithm heavily rely on the choice of the proposal distribution $q(x'|x_t)$. If $q(x'|x_t)$ is poorly chosen, the algorithm can suffer from high rejection rates, which slows the convergence. When the proposal distribution is symmetric, that is, $q(x'|x_t) = q(x_t|x')$, the acceptance ratio simplifies. In this case, the acceptance criterion reduces to the following simpler form:

$$\alpha(x_t, x') = \min\left(1, \frac{p(x')}{p(x_t)}\right)$$

This simplification occurs because the proposal distribution terms cancel out, and the acceptance probability only depends on the ratio of the target distribution's values at the proposed and current states. This is the *Metropolis algorithm*, a special case of MH when the proposal distribution is symmetric. The Metropolis algorithm is often more efficient in high-dimensional settings because it avoids the complications of an asymmetric proposal distribution, simplifying the acceptance criterion.

In general, the Metropolis-Hastings algorithm, and the Metropolis algorithm in the case of symmetric proposals, provide powerful tools for sampling from posterior distributions in Bayesian inference, especially when the normalising constant is not available. However, the choice of proposal distribution remains critical to the efficiency of the algorithm, and careful consideration must be given to its properties to ensure fast convergence and minimise computational cost.

Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) is considered one of the fastest Markov chain Monte Carlo (MCMC) sampling algorithms, leveraging gradients and momentum to generate more efficient Metropolis proposals when sampling from the posterior distribution (Neal, 2012). A key feature of HMC is its use of gradient information of the log-posterior to inform proposed moves in the parameter space, thereby reducing random walk behaviour and improving exploration. To generate proposals, HMC augments the parameter space with auxiliary momentum variables. Auxiliary momentum variables are temporary variables, typically drawn from a multivariate normal distribution, that pair with model parameters in HMC to form a physical system and help simulate Hamiltonian dynamics, enabling efficient exploration of the parameter space. Let $\mathbf{q} \in \mathbb{R}^d$ denote the vector of parameters of interest (position variables) and $\mathbf{p} \in \mathbb{R}^d$ the auxiliary momentum variables. HMC defines a Hamiltonian function as (Neal, 2012):

$$H(q, p) = U(q) + K(p) \tag{2.7}$$

where, $U(\mathbf{q})$ is the potential energy, defined as the negative log of the unnormalised posterior distribution:

$$U(q) = -\log[\pi(\mathbf{q})L(\mathbf{q} \mid \mathcal{D})] \tag{2.8}$$

where $\pi(\mathbf{q})$ is the prior density and $L(\mathbf{q} \mid \mathcal{D})$ is the likelihood function given data $\mathcal{D}$.

The other component $K(\mathbf{p})$ in eqn. 2.7 is the kinetic energy, often defined as a quadratic function:

$$K(\mathbf{p}) = \sum_{i}^{d} \frac{p_i^2}{m_i} \tag{2.9}$$

where m_i is the variance.

The joint distribution over $(\mathbf{q}, \mathbf{p})$ then becomes;

$$P(\mathbf{q}, \mathbf{p}) \propto \exp(-H(\mathbf{q}, \mathbf{p})) = \frac{1}{Z} \exp\left(\frac{-U(\mathbf{q})}{T}\right) \exp\left(\frac{-K(\mathbf{p})}{T}\right) \tag{2.10}$$

where Z is the normalising constant needed for this function to sum or integrate to one and T is the temperature that is typically set to 1 for mathematical convenience in MCMC and to define the canonical distribution, therefore allowing for the exact marginalisation of the marginal distribution $\mathbf{q}$ which is the desired target posterior distribution.

At each iteration of the algorithm, the state $(\mathbf{q}', \mathbf{p}')$ is proposed using Hamiltonian dynamics, implemented with the leapfrog method. Then the proposed state is accepted or rejected as the next state of the Markov chain based on an acceptance probability defined as:

$$min\left[1, \exp\left(-H(q', p') + H(q, p)\right)\right] = min\left[1, \exp\left(-U(q') + U(q) - K(q') + K(q)\right)\right] \tag{2.11}$$

This ensures that the resulting Markov chain satisfies detailed balance with respect to the joint distribution $P(\mathbf{q}, \mathbf{p})$, and therefore preserves it as the invariant distribution. Furthermore, the chain transitions from one state to another via proposals informed by Hamiltonian dynamics, and the accept-reject mechanism ensures that the transition kernel satisfies reversibility and ergodicity property, key conditions for MCMC convergence.

Using gradient information and momentum, HMC consistently makes intelligent proposals and explores the parameter space much more effectively, thereby increasing the acceptance rate, improving sampling efficiency in high-dimensional parameter spaces, and enhancing performance in complex models, such as multilevel models. Further-

more, HMC reduces the autocorrelation of samples. Implementing the HMC algorithm requires a thorough specification of several parameters. The most important parameters are (Neal, 2012): (i) the negative logarithmic probability of the data at the current parameter values, (ii) the gradient of the negative logarithmic probability at the current parameter values, (iii) the step size, (iv) a number of leapfrog steps, and (v) a vector of initial parameter values. In practice, however, the No-U-Turn sampler (NUTS), an adaptive extension of HMC is almost universally used in place of standard HMC with manually tuned leapfrog steps, as it automatically determines an optimal path length, thereby improving efficiency and reducing the need for parameter tuning.

2.4.2 Variational inference

Variational inference (VI) is an approximate Bayesian machine learning method employed to approximate the true posterior distribution using a simpler distribution, known as the variational distribution. Unlike standard Bayesian approaches, which typically sample directly from the posterior distribution using the MCMC algorithm, VI approximates the target distribution through optimisation techniques. The goal is to optimise the parameters and any hyperparameters of the proposed member of the family of the variational distribution so that it closely matches the true posterior, which is achieved by minimising the Kullback-Leibler (KL) divergence between the true posterior $p(\boldsymbol{\theta}|\mathbf{X})$ and the approximating distribution $q_{\lambda}(\boldsymbol{\theta})$, defined as,

$$KL[q_{\lambda}(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{X})] = \int q_{\lambda}(\boldsymbol{\theta}) \log\left(\frac{q_{\lambda}(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{X})}\right) d\boldsymbol{\theta}. \tag{2.12}$$

This optimisation-based approach often results in faster performance and better scalability for large datasets compared to MCMC, which can struggle when sampling from

both unimodal and multimodal posteriors. Furthermore, VI supports dividing data into batches (mini-batches) during optimisation, enabling it to handle large datasets more efficiently by avoiding memory limitations, whereas MCMC generally requires processing the entire dataset at each step of the MCMC sampler to evaluate the likelihood. In addition to the choice of variational family, VI requires also the choice of the optimisation method in order to minimise the objective function by iteratively updating the variational parameters.

VI offers significant advantages for high-dimensional or computationally demanding models, such as those commonly encountered in ecological systems. Despite its potential, VI has rarely been applied in the field of animal movement modelling, where MCMC and HMC methods have traditionally been the standard approaches. For example, VI has been used in a hierarchical continuous-time velocity model to identify key wildebeest migration pathways (Paun et al., 2022), demonstrating its capacity to capture complex, spatially varying movement patterns and effectively manage large-scale movement data. By incorporating VI into animal movement models, this thesis leverages existing computational tools in innovative ways to address persistent challenges in movement ecology, including uncertainty quantification, scalable inference, and the need for efficient algorithms capable of processing large tracking datasets within practical time constraints. This approach not only advances methodological development, but also provides practical solutions to pressing issues in the field.

2.4.3 Gaussian processes

Gaussian processes (GPs) are a class of non-parametric models rooted in Bayesian inference and formalised as continuous-time stochastic processes (Rasmussen, 2006). A defining characteristic of GPs is that any finite collection of random variables drawn

from the process follows a multivariate normal distribution. This property enables GPs to model complex relationships between inputs and outputs through a mean function and a covariance function (or kernel), the latter of which captures the spatial or temporal similarity between observations based on their locations. In a standard GP formulation, the relationship between observed outputs and their associated inputs can be expressed as

$$y_i = f(x_i) + \epsilon \tag{2.13}$$

where y_i are the outputs observed at the input points or locations x_i, $f(x)$ is an unobserved latent function, and $\epsilon \sim \mathcal{N}(0, \sigma_m^2)$ is an independent additive Gaussian white noise term often associated with measurement error. The goal of GPs is to infer a posterior distribution over possible functions $f(x)$ conditioned on the observed data. To achieve this, a GP prior is placed over the unobserved latent function $f(x)$

$$f(x) \sim \mathcal{GP}\left(m(x), K(x, x')\right) \tag{2.14}$$

where $m(x)$ is a mean function, often assumed to be zero (Murphy, 2012) and $K(x, x')$ is a covariance function (or kernel) that encodes prior assumptions about the smoothness, periodicity, or other properties of the process being modelled.

Taking the assumption that the unobserved function $f(x)$ is a realisation of a GP, Bayesian inference is applied using Bayes' theorem and Gaussian identities to calculate the posterior distribution over functions. This posterior quantifies both the expected outputs and the associated uncertainty, given the input (observed data) and prior beliefs.

GPs are particularly powerful and flexible for modelling ecological phenomena, especially when the underlying functional relationships are complex or poorly understood. In movement ecology, for instance, GPs have been used to reconstruct animal trajec-

tories (Rieber et al., 2024), identify periodic activity patterns, and detect deviations in the migratory behaviour of animals (Torney et al., 2021). More broadly, GPs are widely employed in ecological applications, including modelling spatio-temporal distributions of migratory populations (Piironen et al., 2022), temporal trend analysis, and environmental monitoring (Wang et al., 2021), as well as in general machine learning tasks such as optimisation (Snoek et al., 2012), regression (Williams and Rasmussen, 2006), and classification (Nickisch et al., 2008).

2.5 The Greater Serengeti-Mara Ecosystem and its contemporary threats

In the remainder of this chapter, an overview of the ecology of the Greater Serengeti-Mara region is presented. Understanding how animal space use relates to the distribution of resources, risks, and environmental conditions constitutes a primary goal of this thesis and serves to motivate the analyses developed and presented in the subsequent chapters.

The Greater Serengeti-Mara Ecosystem is a transboundary conservation area of global significance, straddling the border between Tanzania and Kenya ($33°30'$–$35°30'E$ and $1°15'$–$3°30'S$) in East Africa (Sinclair et al., 2008). Spanning approximately 37,516 km^2, some wilderness areas of this ecosystem offer a rare glimpse of what the world's landscape looked like a million years ago. At its core lies the Serengeti National Park, adjoined by several other protected areas, including the Ngorongoro Conservation Area, Maasai Mara National Reserve, and the Maswa, Grumeti, Ikorongo, Kijereshi, and Pololeti Game Reserves. In addition, it encompasses a network of community-managed wildlife conservancies in Kenya and wildlife management areas in Tanzania. These areas

are classified into various International Union for Conservation of Nature (IUCN) protected area categories based on their management objectives. The vegetation within the ecosystem is mainly characterized by wooded savanna, which cover approximately 60% of the area predominantly in the northern region while the remaining 40% consists of short grasslands, mainly in the southern part (McNaughton, 1985). The dominant grass species in these grasslands include *Digitaria macroblephera* and *Sporobolus ioclades*. The ecosystem hosts approximately seventy (70) species of mammals, with the wildebeest (*Connochaetes taurinus*) being the most abundant and functionally dominant species. In addition to its migratory behaviour, wildebeest play a critical role in driving key ecological processes such as nutrient cycling and storage (Sinclair et al., 2008).

Globally, approximately 33% of protected land, including designated conservation areas (Jones et al., 2018), is under increasing pressure from extensive anthropogenic activities, such as agriculture and livestock incursions. These pressures result in fragmentation and degradation of habitats, leading to loss of structural and functional heterogeneity in these ecosystems. The Greater Serengeti-Mara ecosystem is no exception and is currently experiencing human-induced disturbances at an unprecedented scale. This highlights the urgent need to understand how such disturbances impact the resilience of the ecosystem and how wildlife respond. Human-induced threats in the ecosystem can be broadly categorised into two types: First, non-local threats originating outside the ecosystem include rapid human population growth, livestock incursions, conversion of land for agriculture (Veldhuis et al., 2019), and large-scale irrigation projects for commercial farming (Kihwele et al., 2021). These external pressures have led to habitat loss, fragmentation, and degradation, particularly along the western boundary of the ecosystem, where hard edges have emerged due to increasing human activities (Kavwele et al., 2022; Veldhuis et al., 2019). In addition, large-scale irrigation upstream has significantly

reduced the flow of key rivers, such as the Mara River, which wildlife depend on during the dry season (Kihwele et al., 2021). This reduction in water availability can disrupt essential ecological processes, including nutrient transport across different regions of the ecosystem, as well as the spatial and temporal distribution patterns of wildlife.

Secondly, the ecosystem is increasingly impacted by local threats, particularly those associated with unsustainable mass tourism. The leverage of natural capital for tourism has led to a substantial increase in the development of hard infrastructure within what was once a largely pristine landscape. Examples of such infrastructure include road networks and tourist accommodations, such as campsites, lodges, and hotels built within the ecosystem itself (Larsen et al., 2020). These developments, along with other anthropogenic pressures, are gradually transforming the natural habitat into a human-modified landscape. This transformation poses significant risks to resident, nomadic, and migratory species, including wildebeest, by altering their spatial distribution, migration timing, and increasing mortality through incidents such as wildlife-vehicle collisions (Lyamuya et al., 2022). A critical and urgent challenge, therefore, is determining how to strike a balance between the infrastructure development necessary to support tourism in ecologically significant areas such as the Serengeti and the imperative to conserve biodiversity and maintain ecological integrity.

Given that most contemporary threats to the ecosystem are human-induced and are the primary drivers of ecological disturbances, it is crucial to understand how wildlife responds and adapts to these changes within and around the ecosystem. In addition, it is equally important to predict the potential impacts of these disturbances on animal populations and assess the extent to which they can compromise the resilience of the ecosystem. Achieving this requires the collection of fine-scale ecological data, along with the application of multiscale models and simulation studies. Such approaches are

essential for translating fine-scale animal responses and behaviours into ecosystem-wide predictions, providing a more comprehensive understanding of the broader ecological consequences.

2.5.1 Modelling wildlife responses to ecosystem threats

Various ecological methods are employed to collect field data, which can involve either invasive or non-invasive approaches such as tissue sampling or live trapping (invasive) or and camera trapping, collecting fecal samples, or acoustic monitoring (non-invasive). In the Greater Serengeti-Mara ecosystem, numerous studies have used Lagrangian and Eulerian individually or in combination to gain deeper insights into how wildlife interact with their environment. These approaches have supported the development and application of advanced animal movement modelling techniques such as species-habitat preference frameworks to understand how animals respond to both local and non-local threats in the ecosystem. For example, to investigate the influence of human presence on wildlife, Hopcraft et al. (2012) analysed long-term data obtained through systematic aerial censuses covering five species of Serengeti mammalian herbivores: African buffalo (*Syncerus caffer*), topi (*Damaliscus korrigum*), Coke's hartebeest (*Alcelaphus buselaphus*), Grant's gazelle (*Gazella granti*), and Thomson's gazelle (*Gazella thomsoni*). The study used resource selection functions to estimate the probability of each species' occurrence in relation to human proximity. This study notably illustrates how large-scale spatial survey data can be effectively analysed using habitat selection models, even at relatively coarse spatial resolutions.

At a broader spatial scale, Kavwele et al. (2022) employed camera traps to investigate whether the formation of hard edges resulting from expanding human activities influences the spatial distribution of migratory wildebeest and zebra (*Equus burchelli*) within

the ecosystem. Using a resource selection function that was conditioned on the presence of wildebeest or zebra along the transect at each camera trap location and at some point during the day, the study revealed that these migratory species exhibit a strong avoidance of hard edges, with displacement occurring up to 6-8 km into the core protected area. This finding highlights the utility of technological tools, such as camera traps, in ecological research. Camera traps allow for continuous, long-term data collection at a relatively low cost, making them especially valuable in contexts where fine-scale individual tracking technologies, such as GPS tags, are cost-prohibitive for sampling large numbers of individuals (Caravaggi et al., 2017; Rowcliffe, 2017). Thus, camera traps provide an efficient alternative for assessing population-level responses of wildlife to environmental changes (Beaudrot et al., 2020; Pettorelli et al., 2010), while retaining the advantages of applying the same modelling frameworks used with data derived from the Lagrangian approach.

At the individual level, animal responses to both local and non-local threats can be assessed using movement data collected through tracking devices such as GPS tags. These data allow for the reconstruction of movement trajectories over time, which can be decomposed into key metrics such as step length and turning angle. Analysing these components enhances our understanding of behavioural states, such as when animals are likely to move quickly and in a directed manner versus slowly and with less directional persistence. For example, Hopcraft et al. (2014) conducted a hierarchical analysis of movement trajectories in wildebeest and zebra to examine their responses to human presence across different spatial scales in the Serengeti ecosystem. Their findings revealed that both species display similar behavioural adjustments: increased displacement and changes in movement direction when near areas of high human density, compared to regions with lower human presence. This approach illustrates how behavioural

changes can be inferred at the individual level, revealing patterns that may be obscured in population-level analyses. Such insights improve our understanding of species adaptive strategies and offer a valuable tool to inform conservation planning in increasingly anthropogenic landscapes.

Furthermore, Veldhuis et al. (2019) used a combined methodological approach using remote sensing imagery and GPS collars affixed to animals to assess the impact of non-local threats on the spatial distribution of wildlife within the ecosystem. GPS-based movement data revealed that migratory large herbivores actively avoid peripheral areas of the ecosystem, resulting in a reduced use of these zones. This pattern was further supported by remote sensing imagery, which showed a high density and extensive network of livestock paths, indicating that illegal livestock incursions into protected areas may be displacing wildebeest toward the ecosystem's core due to increased competition for forage. This integrative approach illustrates the value of combining multiple methodologies to better understand the drivers of animal spatial distribution in landscapes shaped by complex and interacting non-local threats.

In summary, these studies highlight the value of integrating multiple modelling approaches to improve our understanding of how wildlife respond to environmental pressures from human activity, particularly in human-modified ecosystems. By leveraging tools such as GPS collars, camera traps, and remote sensing, wildlife biologists and ecologists can capture fine-scale and broad-scale patterns of animal movement, as well as resource selection and avoidance behaviours across the landscape. This integrative approach is crucial for guiding effective conservation strategies, especially in complex and dynamic ecosystems where local and non-local threats interact to influence species distributions at spatial and temporal scale. Importantly, such models can also provide a foundation for developing predictive models to assess the potential ecological impacts of

both planned and unplanned infrastructure and economic development, thereby informing conservation management decisions that balance development needs with long-term conservation goals.

This chapter has provided an overview of background theory that forms the foundation of the thesis as well as describing the challenges associated with the conservation of the Serengeti ecosystem. Each of the chapters that follow will make use of some or all of the theoretical foundations described above and all chapters are motivated by a central goal: to develop statistical methods that can be applied to the study of the Serengeti wildebeest migration. In particular, these methods seek to understand how migratory wildebeest space use within the ecosystem and how their behaviour and movement patterns are influenced by anthropogenic threats, particularly infrastructure development, which currently represents one of the most significant non-lethal risks to their migration.

Chapter 3

Efficient approximate Bayesian inference for quantifying uncertainty in multiscale animal movement models

Note:

The content of this chapter has been published in the journal Ecological Informatics, 84, p.102853, https://doi.org/10.1016/j.ecoinf.2024.102853.

Abstract

It is becoming increasingly important for wildlife managers and conservation ecologists to understand which resources are selected or avoided by an animal and how to best predict future spatial distributions of animal populations in the long term. However, inferring the patterns of space use by animals is a challenging multiscale inference problem, and formal uncertainty quantification of parameter estimates is an essential component of models that provide useful predictions across scales. In this study, we develop an approximate Bayesian inference framework for step selection models of animal movement which quantifies the uncertainty in estimates of resource selection and avoidance parameters within the Bayesian paradigm. The framework allows joint inference of movement and resource selection parameters of animals and is multiscale in that parameters inferred from fine scale movement steps scale to produce predictions of long-term patterns of space use. Our analysis focuses on simulated movement data in which we test the performance of our framework by altering movement parameters in the data-generating process. In our simulations, individuals respond to two environmental covariates and we employ all combinations of positive and negative selection coefficients corresponding to attraction to an environmental feature and avoidance of an environmental feature, respectively. In all scenarios, we recover the movement parameters used for the simulation of synthetic movement data using variational inference, an approximate Bayesian method, allowing us to formally quantify the uncertainty associated with each parameter for varying data set sizes. Our framework successfully recovered all combinations of movement parameters of the simulated data and accurately captured their posterior distributions given the available data suggesting that the framework is reliable and suitable for inferring how animals select resources and move on a landscape.

Notably, our analysis shows that even for reasonably large data sets (circa 10,000 observations) there can still be considerable uncertainty associated with resource selection parameters which can in turn lead to inaccurate predictions of long term space use if not properly incorporated into the modelling approach. To further illustrate the utility of our approach, we also present a case study of its application to an example data set consisting of GPS locations of a fisher (*Martes pennanti*). Our approach will be of interest to ecologists looking to address conservation questions such as when and where animals are likely to spend most of their time. Furthermore, the approach could be used to predict new suitable areas for conservation based on how GPS collared animals use or avoid resources while including uncertainty around the predictions, thereby helping to make informed management decisions.

3.1 Introduction

Conservationists and applied ecologists frequently need to determine the response of animals to different landscape features, predation risks, and human-driven disturbances in order to effectively manage and protect mobile species. Species-habitat association studies (Matthiopoulos et al., 2020) provide a framework for modelling observed patterns of resource selection and risk avoidance using data on environmental covariates and animal locations. In order to identify the behavioural drivers of species-habitat associations and to make accurate and generalisable predictions of animal space use, statistical methods should be able to quantify uncertainty and provide inferences that can be translated across scales (Torney et al., 2018), from the scale of observation, which is typically the individual, to the scale of interest, which for conservation applications is most often the population.

Various approaches have been employed in species-habitat association studies that focus on different spatio-temporal scales (Fieberg et al., 2021; Michelot et al., 2019a). Commonly used methods either summarize an animal's response at a broad scale, for example resource selection functions or RSFs (Manly et al., 2007), or at a fine scale by incorporating movement characteristics using step selection functions (SSFs) (Fortin et al., 2005; Thurfjell et al., 2014) and integrated step selection functions (iSSFs) (Avgar et al., 2016). While these methods all model the relationship between animal movement and environmental covariates, they differ in complexity and in their underlying assumptions. For example, RSFs assume that the telemetry locations of the same individual are independent of each other (Fieberg et al., 2010), and all areas within the home range or study site are equally accessible by the animal (Beyer et al., 2010; Matthiopoulos, 2003). When these assumptions do not hold due to the high temporal resolution of location data then

SSFs may be used instead to account for spatial and temporal auto-correlation of the animal's locations and the restricted availability of resources that varies among individual animals in space and at each time step (Forester et al., 2009; Fortin et al., 2005; Thurfjell et al., 2014). However, due to the separation between the movement process and the resource selection process SSFs are known to produce biased estimates of the selection coefficients. To counter this issue, including movement characteristics (step length, or natural logarithm of step length or turning angle) is required (Fieberg et al., 2021; Forester et al., 2009) leading to the simultaneous estimation of movement and resource selection parameters using iSSFs (Avgar et al., 2016).

Despite the usefulness that RSFs, SSFs, and iSSFs have in species-habitat association studies and the apparent similarity in their model structure, the parameters estimated by the different approaches when applied to the same movement data do not have the same ecological meaning, even though each of the methods seek to improve our understanding of animals' resource selection and avoidance patterns (Barnett and Moorcroft, 2008; Moorcroft and Barnett, 2008; Signer et al., 2017). While the RSF modelling approach has several disadvantages and should not be applied to auto-correlated movement data, a fitted RSF corresponds to a prediction about the long-term space use of an animal and is therefore highly relevant to many questions of interest for ecologists and conservationists. Conversely, while an iSSF models does provide unbiased estimates of resource selection and avoidance coefficients at the fine-scale, these coefficients cannot be directly employed to make predictions over longer time scales. Since long term population-level patterns of space use are the consequence of the selection patterns that occur, and are observed, at the fine scale, in principle a model that is fit to these fine scale data should be able to make accurate predictions across scales. To achieve this aim, different approaches have been proposed to tackle the multiscale inference problem in animal movement and

habitat selection.

Using a continuous-time animal movement model based on the Langevin diffusion equation, Michelot et al. (2019b) model the animal's positions using a diffusion process, ensuring that, over time, the model converges to a limiting distribution regardless of the animal's initial location. This convergence occurs due to the animal's inherent tendency to move towards suitable habitats. The parametric model can be connected to SSFs when long term space use is modelled as a function of environmental covariates. A unique property of this framework, compared to many multiscale modelling approaches, is its formulation in continuous time, which allows for the accommodation of irregular or incomplete movement data.

In a discrete-time formulation, Michelot et al. (2019a) proposed the MCMC step selection framework, an approach that introduces a novel model of fine-scale movement decisions of an animal that when fitted to data has the property that the model parameters can be used directly to make predictions of space use at the broad scale.

The key novelty of the Monte Carlo Markov chain (MCMC) step selection approach proposed by Michelot et al. (2020, 2019a) is to conceptualise animal movement as the movement of an MCMC sampler in parameter space (for an overview of MCMC methods see, for example, Bolstad (2009)). The use of this analogy enabled the design of an animal movement model that directly links the parameters of the step selection mechanism to the stationary utilisation distribution of the animal i.e. the distribution that would be obtained via an RSF analysis. This important advance allows selection coefficients to map directly to the utilisation distribution of the animal without requiring the individual based simulations typically associated with iSSF analysis (Signer et al., 2024). While the MCMC step selection approach obviates the need for expensive, individual-based movement model simulations and still provides movement steps that map directly

to the parameters of the long-term utilisation distribution, the original formulation does not obtain full posterior distributions for the model parameters but instead relies on numerical estimation of the Hessian matrix to formally quantify model parameter uncertainty. In general, failing to formally quantify uncertainty in animal movement models may lead to over-confidence in habitat selection parameters and inaccurate predictions of long term space use. As the number of measured movement steps change, model predictions will vary due to either underestimation or overestimation of the response resulting in an inability to identify data inadequacies, inaccurate utilisation distributions, compromised reliability of the estimates (Jansen et al., 2022; Rocchini et al., 2011), and a lack of transferability to the same species in areas not sampled (Wenger and Olden, 2012; Yates et al., 2018). However, as the approach proposed by Michelot et al. (2019a) is likelihood-based it is therefore amenable to Bayesian inference either using sampling-based approaches or approximate inference techniques. Further, since there is a direct link between the parameters of the step selection mechanism and the stationary utilisation distribution of the animal, uncertainty may be propagated directly from the fine-scale model to the long term predictions.

In this work, we propose the use of variational inference (VI) (Blei et al., 2017) to formally quantify the uncertainty in step selection models in an efficient and flexible way, thus enabling predictions of an animal's long term space use that accounts for uncertainty. Instead of directly sampling the parameters of interest from the posterior distribution using Markov chain Monte Carlo methods, as is common in standard Bayesian practice, VI turns the inference process into an optimisation problem. While approximate, this method offers notable improvements in speed and computational efficiency, as it involves dividing the data into batches (mini-batching), and optimising parameters and any hyperparameters associated with the model via stochastic gradient descent using

the Kullback-Leibler divergence (Ranganath et al., 2014) between the posterior distribution and its approximating distribution as the objective function of the optimiser. If the true posterior distribution belongs to the same family as the approximating distribution, VI returns the exact posterior, whereas if the approximating family does not contain the true posterior then we obtain only an approximate posterior. In what follows we select the multivariate normal as the approximating family of distributions, however arbitrary distributions may be employed within the same framework.

To showcase how VI can be used to estimate habitat selection and movement parameters jointly, we simulate synthetic movement data based on known parameters and investigate the performance of our method in recovering those movement parameters. We demonstrate the flexibility of the VI method by testing its capability with four different combinations of movement parameters (positive-positive coefficients, positive-negative coefficients, negative-negative coefficients, and negative-positive coefficients). To illustrate its scalability and robustness, we evaluate the computation speed and accuracy of recovering movement parameters based on datasets of up to 1 million observations. We demonstrate that VI is an effective tool for processing large scale animal movement datasets and further show how even relatively large datasets give rise to high uncertainty in the coefficients of habitat selection models. Finally, we present a resource selection analysis case study using GPS locations from a fisher (*Martes pennanti*), incorporating three environmental covariates (two continuous and one categorical) analysed through VI.

3.2 Background

3.2.1 Resource selection functions (RSFs)

RSFs are mathematical functions that evaluate the ratio of used habitats by an animal in relation to available habitats in order to estimate the probability an animal will occupy a specific location (Johnson, 1980; Manly et al., 2007). RSFs are usually fitted to animal location data using logistic regression so that the landscape features of locations that are visited by an animal (such as vegetation) are compared to what is available as a means of estimating selection or avoidance. Used locations are telemetry locations or points where animals are observed, and available locations are typically sampled randomly within the domain available to the animals and are fixed over time. This may be within the individual or population home range, where the home range is usually defined as the area traversed by an animal over a particular period of time (such as its lifetime or during the specific period of time the observations were made) or study area (geographical area where the animals are found). If a resource occurs more frequently in the used locations than the available locations, the resource is preferentially selected by the animal (selection), but if a resource is used less frequently than its availability, it means the resource is avoided by the animal (avoidance). Finally, if a resource is used randomly (such that the used resources is equal to available resources), it indicates neither selection nor avoidance (Manly et al., 2007). Mathematically, RSFs are typically written as

$$w(\mathbf{c}) = \exp(\beta_1 c_1 + \beta_2 c_2 \cdots + \beta_p c_p) \tag{3.1}$$

where $w(\mathbf{c})$ is proportional to the probability that a location with covariate value $\mathbf{c}$ is used, $\beta_1 \cdots \beta_p$ are the parameters (coefficients) to be estimated associated with the vector

c of predictor variables (or environmental covariates) $c_1 \cdots c_p$. The selection coefficients $\boldsymbol{\beta}$ can be estimated using maximum likelihood or Bayesian methods. Since the RSF is proportional to the probability of use, the long term average probability of an animal being found in a specific location within a study domain, termed the utilisation distribution, can be found by rescaling eqn. 3.1 with a normalising constant,

$$\pi(\mathbf{x}) = \frac{\exp\left(\beta_1 c_1(\mathbf{x}) + \beta_2 c_2(\mathbf{x}) + ... + \beta_n c_n(\mathbf{x})\right)}{\int_{\Omega} \exp\left(\beta_1 c_1(\mathbf{z}) + \beta_2 c_2(\mathbf{z}) + ... + \beta_n c_n(\mathbf{z})\right) d\mathbf{z}} \tag{3.2}$$

where $\boldsymbol{\Omega}$ denotes the study region and $c(\mathbf{z})$ and $c(\mathbf{x})$ associate the spatial locations to the corresponding covariate values. The denominator in eqn. 3.2 normalise the utilisation distribution to ensure that it defines a valid probability distribution for spatial location $\mathbf{x}$.

In addition, the selection coefficients $\boldsymbol{\beta}$ obtained from the RSFs model (eqn. 3.1) fitted using weighted logistic regression is equal to that of the intensity function (eqn. 3.3) of an Inhomogeneous Poisson Point Process model (IPP), when the number of available points sampled in RSFs is assigned an infinite weight and is large enough to ensure stability of the coefficient estimates (Fieberg et al., 2021; Fithian and Hastie, 2012; Warton and Shepherd, 2010), hence, making a link between IPP and RSFs in addressing applied ecological questions on attractiveness and repulsiveness of resources by the animal. The intensity function $\boldsymbol{\lambda}(\mathbf{s})$ can be modelled as a log-linear function of spatial covariates and is expressed mathematically as:

$$\boldsymbol{\lambda}(\mathbf{s}) = \exp(\beta_0 + \beta_1 c_1(s) \cdots + \beta_p c_p(s)) \tag{3.3}$$

where $\beta_1 \cdots \beta_p$ are the parameters (coefficients) to be estimated associated with the vector **c** of the spatial predictor variables $c_1 \cdots c_p$ at location **s**. β_0 in the eqn. 3.3 above

is the intercept and is ecologically not meaningful, but rather is used to determine the log density of locations within a spatial domain (area) that is small and homogeneous around $\mathbf{s}$ when all $c_i(\mathbf{s})(i = 1, \cdots, p)$ are zero.

When performing an RSF analysis, there are some assumptions inherent in the modelling framework that must be considered. The most significant assumptions are (Manly et al., 2007; Millspaugh et al., 1998): (i) the observations of the same individual are independent of each other, so there is no temporal auto-correlation in the location, (ii) all available habitats are equally accessible or available to animals all the time, (iii) sampling of the available points is random and independent,(iv) the selection of resource by an individual at one location is independent of any other location previous or consecutive location that an individual visited, and (v) resources must be heterogeneously distributed across the landscape to allow the detection of selection; without variation, preferences or avoidance cannot be inferred. Notably, RSFs are heavily influenced by the definition of availability (Paton and Matthiopoulos, 2016) since as the extent of the available habitat changes the strength of selection or avoidance will also change (Beyer et al., 2010). To avoid changes in habitat selection parameters when the extent of available area changes, Alston et al. (2023) and Matthiopoulos et al. (2023) have proposed objective methods that guarantee stable estimates regardless of changes in the extent of available habitat. For migratory animals and those which have a large home range, the assumption that all locations are equally available all the time is violated due to physical separation (Manly et al., 2007).

To overcome the issues associated with RSFs alternative methods have been developed, in the form of step selection functions (Forester et al., 2009), that account for spatial and temporal auto-correlation of telemetry locations and model resource availability so that it varies within individual animals in space and over time.

3.2.2 Step selection functions (SSFs)

SSFs are a mechanistic movement model derived from locally biased correlated random walks, where an animal's movement decisions are influenced by the local environmental conditions in its immediate vicinity. Mathematically, SSFs are functions that evaluate the effects of environmental features on the movement decisions of animals. SSFs provide a more biologically plausible and realistic comparison between what is used and what is available by incorporating movement characteristics of the animal, so available locations are constrained to be within range of a typical movement step and habitat selection is conditioned on the movement (Forester et al., 2009; Thurfjell et al., 2014). SSFs extend RSFs by combining a resource-independent movement kernel, which describes the animal's movement characteristics in the absence of resource selection, with a habitat selection function, which describes how the animal selects resources when not constrained by movement. SSFs account for factors such as distance and turning angle of the animal's movements, and they relax the assumptions of independence commonly associated with RSFs. This combination allows used and available locations to share the same starting point but different end points due to variations in distance and turning angle.

Available locations for an animal at a particular time step are generated by sampling from a distribution of turning angles (heading between two sequential locations) and step lengths (distance between two consecutive locations) derived from the movement data. Popular choices for the distribution of the step lengths are the Gamma or Weibull distributions, while turn angles are typically sampled from von Mises or wrapped Cauchy distributions. Alternatively, steps may be sampled directly from the empirical distribution given by the movement data (Fortin et al., 2005). SSFs are fitted using standard conditional logistic regression with a matched case-control design, where cases

are the observed steps and controls are the randomly sampled steps. In an SSF model, the probability of an animal moving to a location $\mathbf{x}_{t+1}$ given it is at location $\mathbf{x}_t$ is given by

$$p(\mathbf{x}_{t+1}|\mathbf{x}_t) = \frac{\phi(\mathbf{x}_{t+1}|\mathbf{x}_t)w\{c(\mathbf{x}_{t+1})\}}{\int_\Omega \phi(\mathbf{z}|\mathbf{x}_t)w\{c(\mathbf{z})\}\,d\mathbf{z}}, \tag{3.4}$$

where $\phi(\mathbf{x}_{t+1}|\mathbf{x}_t)$ is the resource-independent movement kernel (probability of moving from $\mathbf{x}$ to $\mathbf{x}_{t+1}$ in a homogeneous landscape), and $w\{c(\mathbf{x}_{t+1})\}$ is the attractiveness of the resources which typically takes an exponential form,

$$w\{c\left(\mathbf{x}_{t+1}\right)\} = \exp\left[\beta_1 c_1(\mathbf{x}_{t+1}) + \beta_2 c_2(\mathbf{x}_{t+1}) \cdots + \beta_p c_p(\mathbf{x}_{t+1})\right]. \tag{3.5}$$

The denominator is a normalising constant to ensure that eqn. 3.4 is a valid probability distribution with respect to $\mathbf{x}_{t+1}$.

Despite being widely used in resource selection modelling, step selection models produce biased estimates of the resource selection coefficients (Forester et al., 2009; Thurfjell et al., 2014) due to the separation of the movement and resource selection processes (Avgar et al., 2016). To rectify the biased estimates produced by SSFs, including movement characteristics when fitting the model (step length, or natural logarithm of step length or turning angle) is required (Fieberg et al., 2021; Forester et al., 2009). Similarly, the SSF framework does not take into account how the characteristics of an animal's movement may change when in a particular environment, thus making it difficult to address biological questions such as, does the animal move faster or slower when in a particular habitat type, or does the animal's movement tend to be more or less persistent when moving in a particular environment. To address such questions and enable unbiased estimation of resource selection coefficients, Avgar et al. (2016) extended SSFs and proposed a new class of resource selection models termed integrated step selection functions (iSSFs)

that simultaneously estimate movement and resource selection parameters as the product of two independent kernels (a selection-free movement kernel and a movement-free selection function). The probability density function in location $\mathbf{x}$ at time t is given by

$$f(\mathbf{x}_t|\mathbf{x}_{t-2},\mathbf{x}_{t-1};\boldsymbol{\beta},\boldsymbol{\theta}) = \frac{\phi(\mathbf{x}_{t-2},\mathbf{x}_{t-1},\mathbf{x}_t;\boldsymbol{\theta})w\{c(\mathbf{x}_t);\boldsymbol{\beta}\}}{\int_\Omega \phi(\mathbf{x}_{t-2},\mathbf{x}_{t-1},\mathbf{z};\boldsymbol{\theta})w\{c(\mathbf{z});\boldsymbol{\beta}\}d\mathbf{z}}, \tag{3.6}$$

where $\phi(\mathbf{x}_{t-2},\mathbf{x}_{t-1},\mathbf{x}_t;\boldsymbol{\theta})$ is a selection-free movement kernel, and $w\{c(\mathbf{x}_t);\boldsymbol{\beta}\}$ is a movement-free selection function. The denominator in eqn. 3.6 is a normalising constant.

3.2.3 Selection coefficients and the utilisation distribution

While SSFs enable the understanding of how animals select resources at a fine spatio-temporal scale, the resource selection parameters estimated by standard SSFs and iSSFs do not directly translate to predictions of the long term space use by animals (Barnett and Moorcroft, 2008; Moorcroft and Barnett, 2008; Signer et al., 2017). This means that the coefficients of eqn. 3.1 are not the same coefficients as eqn. 3.5 despite their apparent similarity. Therefore, in order for SSFs coefficients to converge to the long term space use, numerical simulations of fitted SSF models are required in order to estimate the utilisation distribution (Potts et al., 2014; Signer et al., 2017, 2024), which are generally challenging to implement and computationally expensive. Ecologically, however, the selection of resources at the broad scale is the consequence of movement processes that occur or are observed at the fine scale, hence a model fit to these fine scale processes should in principle to be able to directly estimate long-term space use (Michelot et al., 2019a).

To solve this issue, Michelot et al. (2020, 2019a) make use of the fundamental properties of Markov chain Monte Carlo (MCMC) methods and propose the MCMC step selection

model. MCMC methods are a widely used technique in Bayesian inference for sampling from posterior distributions. The transition probabilities of an MCMC sampler are specified so that the long term stationary distribution of the chain will converge to the target posterior distribution. Michelot et al. (2019a) design a step selection model that proposes an analogy between the movement of an individual animal and the steps of an MCMC sampler so that the fine-scale step selection rules are guaranteed to converge to the parameters of the underlying long term utilisation distribution. To implement this, they developed a rejection-free MCMC sampler which they termed the local Gibbs sampler. A key characteristic of the local Gibbs sampler is the introduction of an intermediate random step into the standard step selection mechanism so that eqn. 3.4 becomes,

$$p(\mathbf{x}_{t+1}|\mathbf{x}_t) = w\{c(\mathbf{x}_{t+1})\} \int_{\boldsymbol{\mu}\in\Omega} \frac{\phi(\mathbf{x}_{t+1}|\boldsymbol{\mu})\phi(\boldsymbol{\mu}|\mathbf{x}_t)}{\int_{\mathbf{z}\in\Omega} w\{c(\mathbf{z})\}\phi(\mathbf{z}|\boldsymbol{\mu})\, d\mathbf{z}} d\boldsymbol{\mu} \tag{3.7}$$

where $\boldsymbol{\mu}$ is the random intermediate step, and $\phi(\mathbf{x}_{t+1}|\boldsymbol{\mu})$ is a symmetric, resource-independent movement kernel. The intermediate random step is not intended to model animal movement behaviour but is a technical requirement that enables the construction of a valid transition kernel.

Due to the design of the movement step the long term stationary utilisation distribution and the fine-scale resource selection process are unified and the coefficients of $w\{c(\mathbf{x}_{t+1})\}$ in eqn. 3.7 may be used directly to make predictions of the long term space use of an animal. This is illustrated in Fig 3.1 which shows the theoretical utilisation distribution and the simulated utilisation distribution of an animal following the movement process described in eqn. 3.7.

The MCMC step selection approach provides valuable insight into the movement and resource selection studies, and is able to provide uncertainty quantification by estimat-

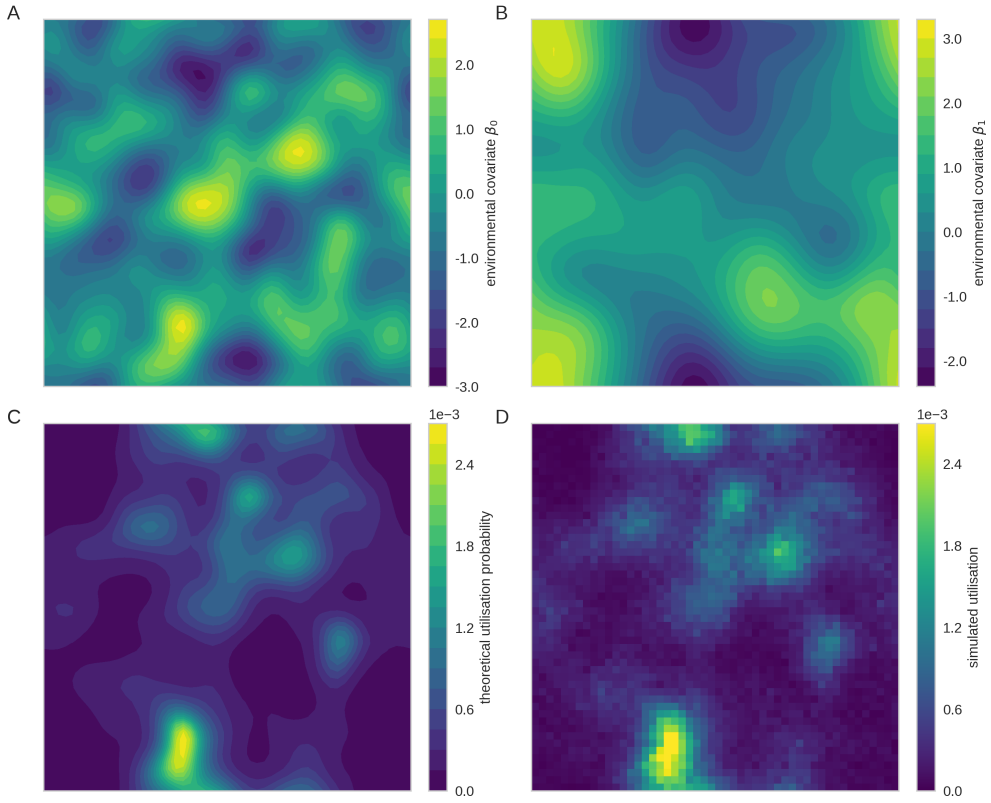

Figure 3.1: Simulated and theoretical utilisation distributions for an MCMC step selection model. A) and B) show random covariates generated using Gaussian random fields that represent an attractive and a repelling environmental covariate, C) theoretical utilisation distribution of animals using movement parameters $\beta_0 = 0.5$, $\beta_1 = -0.8$ and the distribution function given by eqn. 3.2, D) simulated utilisation distribution of an individual using the MCMC step selection movement model using parameters $\beta_0 = 0.5$, $\beta_1 = -0.8$.

ing the variance of the maximum likelihood estimate using the Hessian matrix. This paper aims to take this process one step further by proposing variational inference (VI) to quantify uncertainty in step selection models within a Bayesian framework, enabling more robust inference for conservation planning and environmental prediction and supporting a more precise understanding of long-term space use by animal populations.

3.3 Methods

3.3.1 Synthetic data generation

In order to test the accuracy of our proposed approach and to demonstrate the amount of uncertainty inherent in step selection models, we simulate synthetic movement data based on known parameters and investigate the performance of our method in recovering those movement parameters and associated posterior distributions. We therefore simulate animal movement in a 2-dimensional environment in which we assume movement steps are selected following the process proposed by Michelot et al. (2019a) whereby an intermediate random step is chosen, followed by a resource dependent step.

We simulated environmental covariate layers as Gaussian random fields with periodic boundary conditions by sampling from a 2-dimensional Gaussian process with a periodic covariance function (Rasmussen, 2006) to mimic the landscape feature of a real case study (see Fig. 3.1A and B for example spatial fields) while removing any effects of the boundary. The covariance kernel used was therefore,

$$k(x,x') = \alpha^2 \exp\left(-\frac{2}{\ell^2}\sin^2\left(\frac{\pi}{p}|x-x'|\right)\right) \tag{3.8}$$

where α^2 is the amplitude and a value of 1 was used for both fields, p is the size of the domain, ℓ is the length scale parameter where for spatial field A, $\ell = 0.4$ and for spatial field B, $\ell = 0.9$. The spatial fields were restricted to a domain of $\mathbf{\Omega} = [0,50]$.

Subsequently, we simulated synthetic movement data representing the selection process of animals in a 2-dimensional geographical space, with multiple runs consisting of simulated movement tracks of 10,000, 100,000, and 1,000,000 steps each. Each movement track started at the center of the domain and periodic boundaries were employed to

Table 3.1: Movement parameters used for simulation of synthetic movement locations data. Note that there is no significant difference between the use of positive and negative coefficients in the table below and identical distributions would be obtained if both the coefficient and the covariate were multiplied by minus one (-1)

Coefficients	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$
Positive-positive	1.2	1.8
Negative-negative	-1.5	-1.8
Positive-negative	0.5.	-0.8
Negative-positive	-1.5	1.8

.

ensure the animal remained within the domain. At each time step t=1,2,...n, an intermediate step $\boldsymbol{\mu}_t$ was generated as a random offset from the starting location $\boldsymbol{\mu}_t \sim \mathcal{N}\left(\mathbf{x}_t, \frac{\sigma^2}{2}\right)$. From this intermediate location, a set of 100 potential random locations were generated by sampling from the 2-d normal distribution $\mathcal{N}\left(\boldsymbol{\mu}_t, \frac{\sigma^2}{2}\right)$. For each potential point, the environmental covariates at the location were calculated, and the next point was selected with probability proportional to the standard resource selection function of Eq. 3.1. The chosen point was then the location of the animal at time $t+1$ and became the starting point for the next step. The resource independent movement parameter σ was set to 1 in all simulations whereas the resource selection coefficients that determined the response to each of the covariate fields, denoted by $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, were either specified as a positive value indicating selection of resources or a negative value indicating the avoidance of resources (see Table 3.1 for the values employed in the simulations).

3.3.2 Model likelihood

Similar to any movement and resource selection studies, the framework presented here describes a step selection model with parameters that specify the resource-independent movement process of the animal and the parameters of the target utilisation distribution, denoted $\boldsymbol{\beta}$ with β_i defining the response to the i^{th} covariate, which describe how animals' select resources (avoidance or selection) on a landscape (habitat selection parameters). The full parameter set, which we denote $\boldsymbol{\theta}$, is estimated simultaneously from

environmental layers and movement data using variational inference in a manner which ensures that the parameters inferred from fine-scale movement data define the long-term utilisation distribution of the animal.

Variational inference is a form of approximate Bayesian inference and as such approximates the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$ where $\mathbf{X}$ represents the sequence of animal locations $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$. In standard Bayesian inference this posterior distribution is found by applying Bayes' rule,

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{L(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{X})} \tag{3.9}$$

where $L(\mathbf{X}|\boldsymbol{\theta})$ is the likelihood of the data given the parameters, $p(\boldsymbol{\theta})$ is the prior distribution of the parameters, and $p(\mathbf{X})$ is the marginal likelihood of the data. In general, the marginal likelihood is intractable and so inference typically proceeds using sampling methods or through the use of an approximation. In either scenario it is necessary to calculate the likelihood function $L(\mathbf{X}|\boldsymbol{\theta})$.

Since the likelihood factorises it can be written as,

$$L(\mathbf{X}|\boldsymbol{\theta}) = \prod_{t=1}^{T-1} p(\mathbf{x}_{t+1}|\mathbf{x}_t, \boldsymbol{\theta}) \tag{3.10}$$

and the probability of each step is given by eqn. 3.7, i.e.

$$p(\mathbf{x}_{t+1}|\mathbf{x}_t, \boldsymbol{\theta}) = \int_{\boldsymbol{\mu}\in\Omega} \frac{w\{c(\mathbf{x}_{t+1})\}\phi(\mathbf{x}_{t+1}|\boldsymbol{\mu})}{\int_{\mathbf{z}\in\Omega} w\{c(\mathbf{z})\}\phi(\mathbf{z}|\boldsymbol{\mu})\, d\mathbf{z}} \phi(\boldsymbol{\mu}|\mathbf{x}_t) d\boldsymbol{\mu}. \tag{3.11}$$

For the resource independent movement kernel we use a 2-dimensional normal distribution so that $\phi(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{y}, \frac{\sigma^2}{2})$ where $\mathbf{y}$ represents the current steps (current locations) and $\mathbf{x}$ the potential end steps, so that $\boldsymbol{\theta} = \{\beta_1, \beta_2, \cdots, \beta_n, \sigma\}$. Since the movement ker-

nel is a Gaussian we employ Gauss-Hermite quadrature to approximate the integrals of eqn. 3.11. We use 3 Gauss-Hermite points so that the integral is exact if the resource field can be approximated as a 5 degree polynomial within the length scale of a movement step which leads to 9 Gauss-Hermite points and associated function evaluations in 2-dimensions. Since typically the environmental covariates will vary slowly with respect to the length scale of an average movement step we expect the error introduced by this approximation to be negligible.

While a Laplace approximation could potentially reduce computational cost, it is less appropriate for the type of integral appearing in eqn. 3.11, where the integrand comprising the product of the movement kernel and resource selection function is not guaranteed to be sharply peaked or well-approximated by a local Gaussian expansion. In particular, heterogeneous or multi-modal landscapes may give rise to multiple regions of elevated probability mass, violating the unimodality and local quadratic assumptions underlying Laplace's method. As a result, this could lead to underestimation of the integral and biased likelihood contributions. Our implementation uses fixed-point Gauss-Hermite quadrature, which, while not adaptive, provides a controlled and accurate approximation given the assumed local smoothness of the covariate fields relative to movement scale.

Given eqn. 3.11, along with a set of telemetry locations, and a set of covariate grids for each environmental layer, the log likelihood function may be computed directly. We employ the machine learning library TensorFlow (Abadi et al., 2016) to perform these calculations since it facilitates parallel computation of the likelihood using GPU-accelerated hardware. Note that calculating the nested integrals of eqn. 3.11 requires the computation of the inner integral for each of the Gauss-Hermite points of the outer integral and is the limiting step in terms of memory when parallelising the computation.

3.3.3 Variational inference

Due to the volumes of data associated with modern movement ecology studies and the complexity of the likelihood computation for the step selection model, it is likely to be infeasible to sample from the posterior distribution using Markov chain Monte Carlo methods in most cases. Instead, we propose the use of variational inference (Blei et al., 2017) to obtain approximate posterior distributions of the model parameters. In variational inference, the posterior distribution is approximated by a variational distribution q_λ that is restricted to belong to a family of distributions parameterised by $\boldsymbol{\lambda}$. The variational parameters $\boldsymbol{\lambda}$ are then optimised so that the difference between the true posterior and the approximating posterior is as small as possible given the distribution family. In this way the sampling associated with MCMC methods is replaced with an optimisation process which is not only more efficient in most cases, but also enables the use of stochastic gradient descent techniques developed in the domain of deep learning such as automatic differentiation and adaptive optimisation schemes.

In order to minimise the difference between the true posterior and the approximating posterior, it is necessary to define a distance metric between the two distributions. As is standard in variational inference, we use the Kullback-Leibler divergence (Kullback and Leibler, 1951) as the distance metric. The Kullback-Leibler (KL) divergence between the true posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$ and the variational distribution $q_\lambda(\boldsymbol{\theta})$ is given by,

$$KL[q_\lambda(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{X})] = \int q_\lambda(\boldsymbol{\theta}) \log\left(\frac{q_\lambda(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{X})}\right) d\boldsymbol{\theta}. \tag{3.12}$$

An obvious issue with the KL divergence is that it depends on the unknown posterior

distribution $p(\boldsymbol{\theta}|\mathbf{X})$ which is the quantity we are trying to approximate. However, as

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{X})} \tag{3.13}$$

we can rewrite the KL divergence as

$$\log p(\mathbf{X}) - \int q_{\lambda}(\boldsymbol{\theta}) \log p\left(\mathbf{X}|\boldsymbol{\theta}\right) d\boldsymbol{\theta} + \int q_{\lambda}(\boldsymbol{\theta}) \log \left(\frac{q_{\lambda}(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} \right) d\boldsymbol{\theta}. \tag{3.14}$$

Since $\log p(\mathbf{X})$ does not depend on $q_{\lambda}(\boldsymbol{\theta})$ minimising the KL divergence between the variational distribution and the true posterior is equivalent to maximising,

$$\int q_{\lambda}(\boldsymbol{\theta}) \log p\left(\mathbf{X}|\boldsymbol{\theta}\right) d\boldsymbol{\theta} - KL\left[q_{\lambda}(\boldsymbol{\theta})||p(\boldsymbol{\theta})\right] \tag{3.15}$$

a quantity that is known as the evidence lower bound (ELBO). It can therefore be seen that maximising the ELBO represents a trade-off between maximising the expected log-likelihood of the data under the variational distribution and minimising the KL divergence between the variational distribution and the prior distribution. Optimisation is achieved using stochastic gradient descent which enables the data to be processed in batches. The KL divergence term may be calculated in closed form for many parametric distributions, however the expected log-likelihood is typically intractable. Since we employ a multivariate Gaussian distribution as the family for the variational distribution we again make use of Gauss-Hermite quadrature to approximate the expectation term of eqn. 3.15. All numerical computation was undertaken using TensorFlow (Abadi et al., 2016) and TensorFlow Probability (Dillon et al., 2017). We used diffuse priors to estimate selection and avoidance parameters, reflecting the fact that these coefficients are unbounded and can take on both positive and negative values. This choice implies no prior assumption of preference or avoidance for any covariate, while also avoiding

undue constraints on ecologically plausible effect sizes. By allowing the coefficients to vary broadly, the priors support a wide range of potential selection strengths and let the data specifically, the covariate values observed in animal movements drive posterior inference. The priors used in the model are specified as follows:

$$\begin{aligned} \beta_1 &\sim N(0,10), \\ \beta_2 &\sim N(0,10), \end{aligned}$$

3.4 Results

To evaluate the performance of our approach we simulated movement datasets of varying sizes and attempted to infer the movement and resource selection parameters that were employed in the simulations. We employed a multivariate normal distribution as the variational posterior and used an optimisation scheme to maximise the evidence lower bound described above. In order to assess model convergence we examine how the optimised posterior distribution changes between subsequent epochs and cease model training when the change in the distribution falls below a threshold value.

In Fig. 3.2 we show the posterior distributions of the recovered resource selection parameters for each of the four combinations of parameters used in the simulations. The true values of the parameters are indicated by the vertical dashed line. It can be seen that the recovered parameters are approximately to the true (known) values and that the uncertainty in the parameters decreases as the number of observations increases. While there is not an exact match between the recovered posterior mean values and the true values it can be seen that the true values are contained within the 95% credible intervals

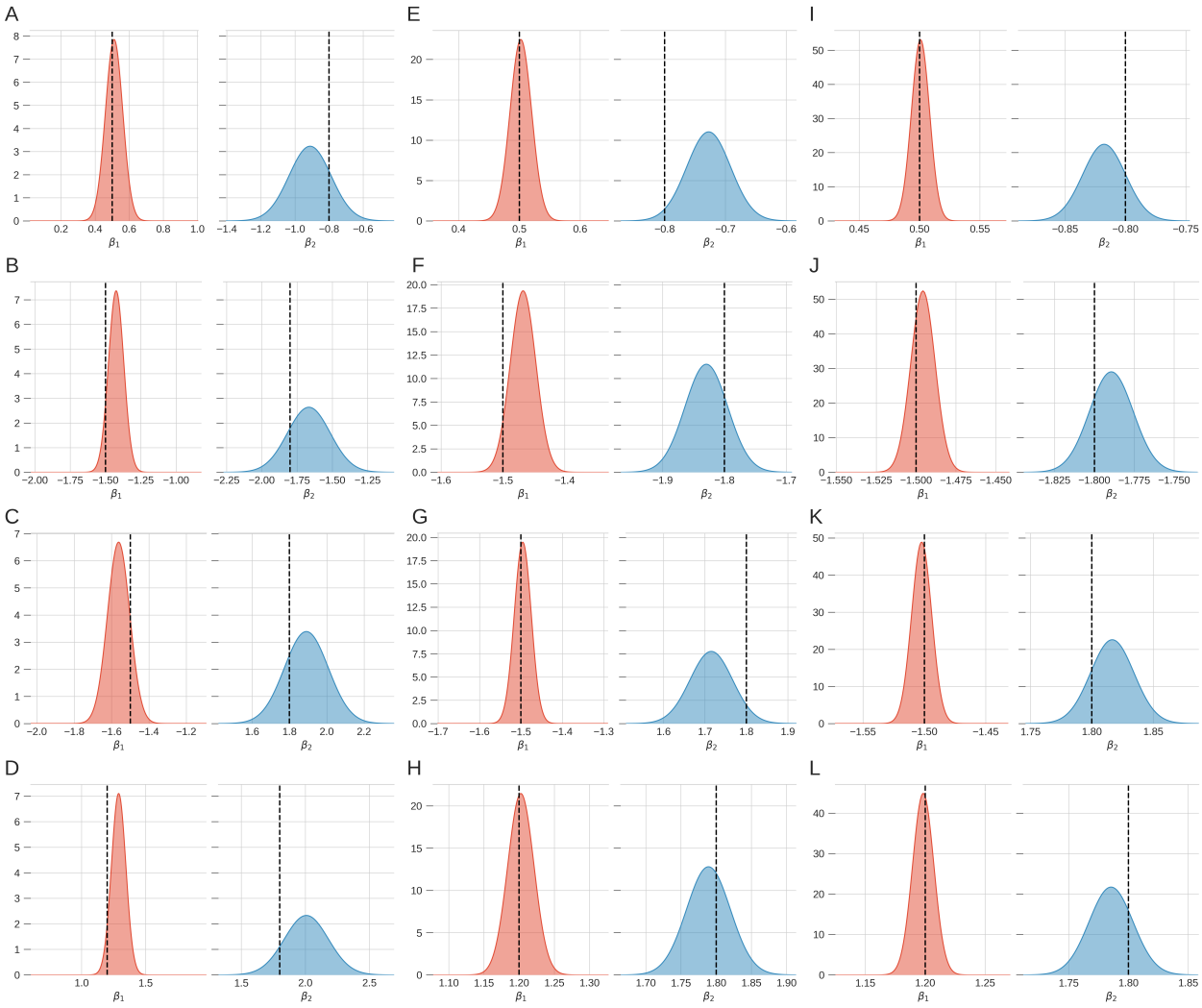

Figure 3.2: Posterior probability distribution of recovered movement parameters using VI from the simulated data. Left column, middle column and right column represent 10,000, 100,000, and 1,000,000 observations, respectively. Note, the scale of the axes are changing from left to right to account for the reduction in uncertainty as the number of observations increases. The movement parameters values for first row A,E, and I are $\beta_1 = 0.5$ and $\beta_2 = -0.8$; second row B,F, and J is $\beta_1 = -1.5$ and $\beta_2 = -1.8$; third row C, G, and K is $\beta_1 = -1.5$ and $\beta_2 = 1.8$, and Last row D, H, and L is $\beta_1 = 1.2$ and $\beta_2 = 1.8$. The vertical dashed line (black in colour) indicates the true values.

of the optimized posterior distributions (Table S1).

To further explore the uncertainty quantification of the recovered resource selection parameters, we computed the posterior Z-scores of the true simulation values (β_1, β_2) for 10 independent runs of synthetic data generation and subsequent variational inference.

This gave 240 scores corresponding to each of the 10 runs for four combinations of the two resource selection parameters and three observation sizes. We then compared the quantiles of the posterior z-scores with a standard normal distribution. The QQ-plot of the Z-scores is shown in Fig. 3.3. It can be seen that the Z-scores are approximately normally distributed meaning that the posterior distributions obtained via our framework are appropriately representing the uncertainty in the resource selection coefficient estimates.

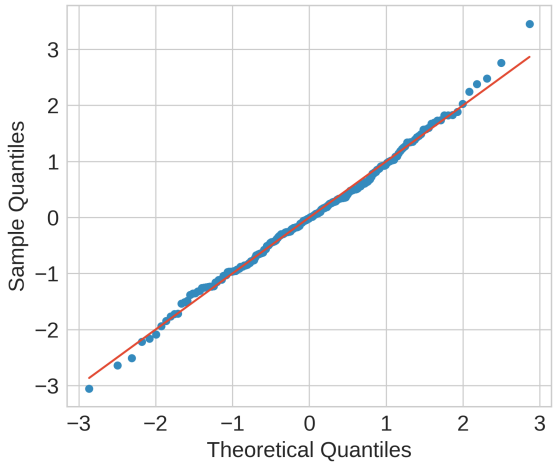

Figure 3.3: QQ-plot for 240 Z-scores of recovered resource selection parameters (β_1, β_2) from 10 runs for each combination of resource selection parameter (see Table 3.1) with varying observation sizes of 10,000, 100,000, and 1,000,000.

While the analysis of posterior Z-scores of the true values provide a useful metric for assessing the uncertainty in the recovered parameters for our simulation study, this approach requires the true values to be known. We therefore also include a comparison of the results of VI on the smaller movement dataset with an MCMC sampling approach on the same data. For 10,000 observations we were able to run a Hamiltonian Monte Carlo sampler for a single simulated dataset for each of the 4 parameter combinations within a reasonable time scale. Results from this analysis are shown in the supplementary material (Fig. S1 and Table S2) from which it can be seen that the results from HMC

are an almost exact match to the results obtained with VI.

Finally, to emphasise the importance of accounting for uncertainty in estimates of selection coefficients when scaling to predictions of an animal's utilisation distribution, we performed simulations of an animal responding to a single environmental covariate. We ran 10 independent simulations for dataset sizes of 10,000, 100,000, and 1,000,000 observations with the environment modelled as a simple sine function that varied in one axis only for ease of visualisation. A plot of the environment is shown in Fig. 3.4A. We then used the VI approach to obtain the maximum a posteriori probability (MAP) estimate, defined as,

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \mathbf{X}) = \arg\max_{\boldsymbol{\theta}} \left[p(\mathbf{X} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \right], \tag{3.16}$$

for the selection coefficient for each of the independent simulations. Since the predicted utilisation distribution can be obtained directly from the selection coefficient we then computed the predicted distribution for each of the simulations and these are plotted in Fig. 3.4B-D along with the true distribution. The results show that for 10,000 observations there is large variation in the utilisation distribution derived from each of the simulations even though all used the same parameters and initial conditions. This is consistent with the observed levels of uncertainty when considering the posterior distribution obtained from the analysis of a single data set of the same size. As the size of the dataset increases the variation between the different simulation results decreases, and for 1 million observations all predictions give a consistent picture of animal space use.

This highlights that uncertainty in estimated space use patterns is strongly influenced by the number of observations. However, in practice, the informativeness of a dataset also depends on its temporal coverage and resolution. A large number of closely spaced ob-

servations over a short time window (e.g., 1 million locations over 24 hours) may reduce uncertainty in fine-scale estimates but may fail to capture broader temporal patterns in movement behaviour or habitat use. Conversely, a sparser dataset collected over a longer period (e.g., 10,000 locations over a year) may better reflect long-term or seasonal dynamics but with greater uncertainty in short-term utilisation patterns. Therefore, both the number of observations and temporal extent of data collection influence the overall uncertainty and should be balanced according to the ecological processes or specific questions of interest.

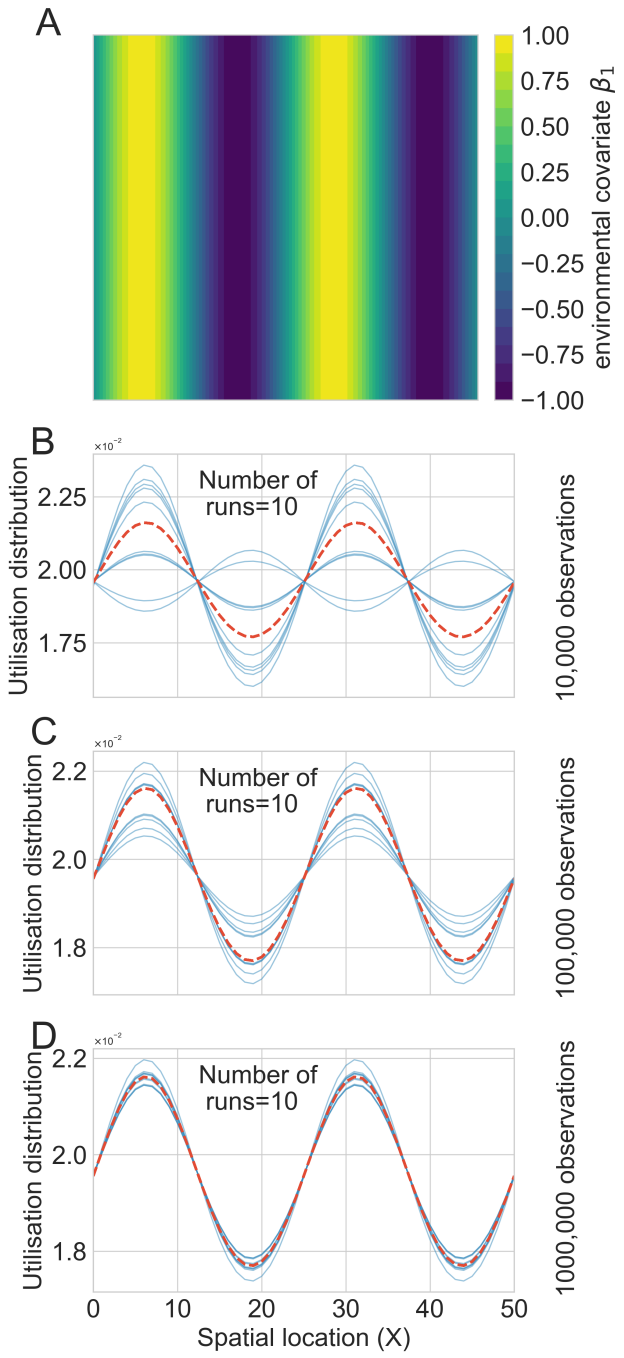

Figure 3.4: A) Simulated map of environmental covariate used to infer utilisation distribution, B) Estimated utilisation distribution using 10,000 observations across 10 runs, C) Estimated utilisation distribution using 100,000 observations across 10 runs, D) Estimated utilisation distribution using 1000,000 observations across 10 runs. Horizontal dashed line (red in colour) represent true utilisation distribution, and blue in colour is the estimated across 10 runs. The movement parameter used during the simulation was $\beta_1 = 0.1$.

3.4.1 Case study

To showcase the applicability of the VI approach on a real animal tracking dataset, we performed an analysis with GPS data obtained on a fisher tracked near Albany, New York, USA (Signer et al., 2019). Location data were collected between December, 2010 and January, 2011, and consisted of 3,004 locations. Further details on data collection can be found in LaPoint et al. (2013) and Fieberg et al. (2021). The environmental covariates used to estimate fisher's resource selection were elevation, population density, and a categorical land use variable with three categories: forest, grass and wet which are shown in Fig. 3.5 (again further details can be found in (Fieberg et al., 2021)). We used our VI framework to estimate habitat selection parameters and inferred posterior probability distributions which are presented in Fig. 3.6. For continuous covariates, our analysis reveals that fisher tends to select areas with relatively high elevation and low human population density, although the 95% credible interval for the effect of human population density contains zero. For the categorical variables, fisher avoids grass areas compared to forest, but we observed almost no attraction to, or avoidance of, wet areas compared to forest.

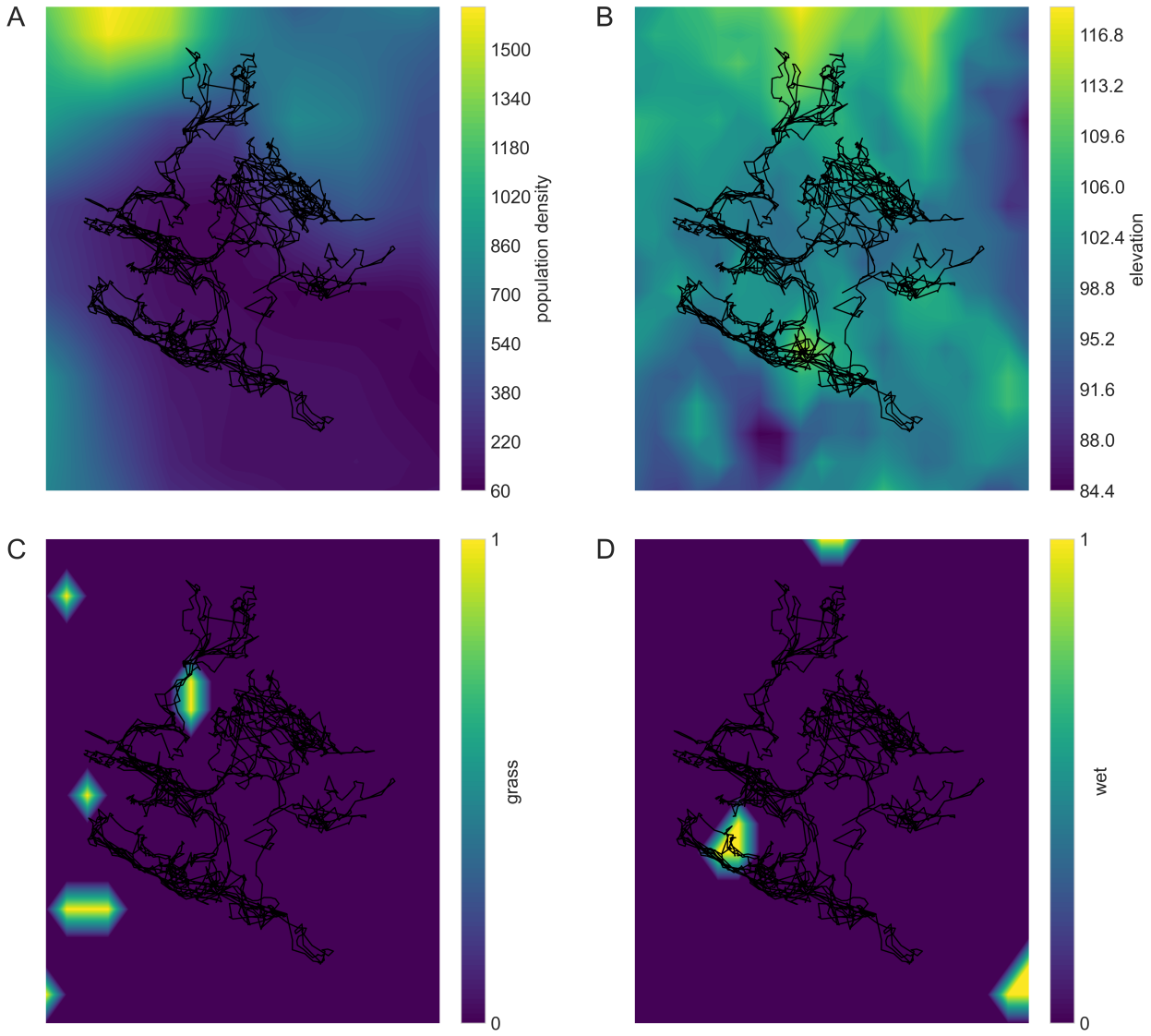

Figure 3.5: Covariate maps for the fisher analysis. A) Population density, B) Elevation, C) Grass area, and D) Wet area. The black lines represent the fisher's movement track. The covariate layers for population density, elevation, grass, and wet have spatial resolutions of 659 meters, 44 meters, 220 meters, and 220 meters, respectively

.

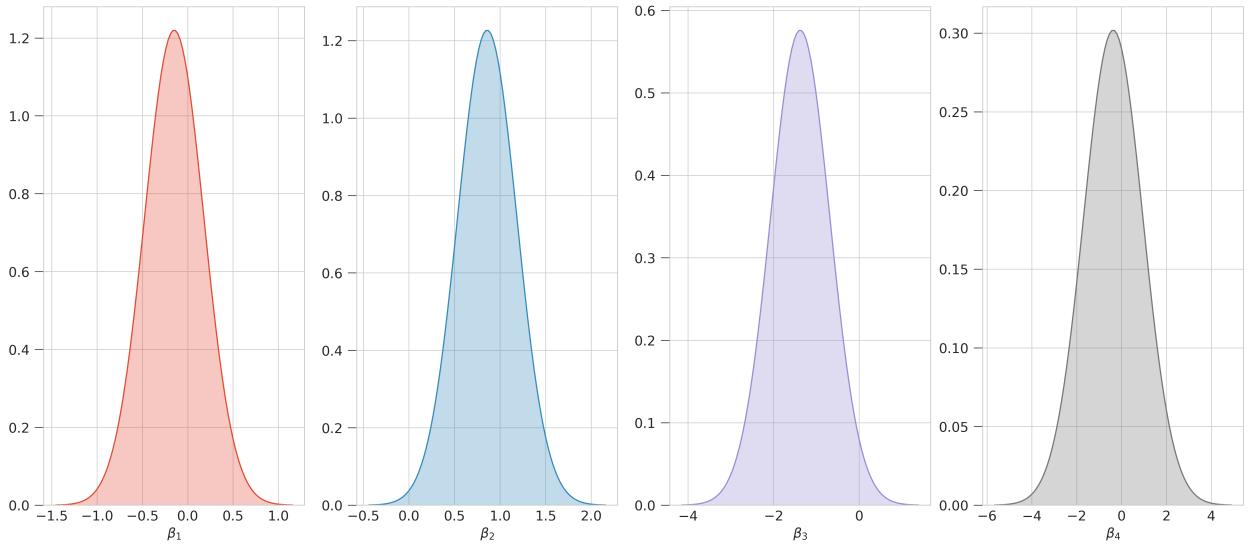

Figure 3.6: Posterior probability distribution of inferred habitat selection parameters using VI from fisher's locations data. First from left is the human population density (β_1) with a posterior mean of -0.15 and posterior standard deviation of 0.33, second from left is the elevation (β_2) with a posterior mean of 0.86 and posterior standard deviation of 0.33, third from left is the grass (β_3) with a posterior mean of -1.38 and posterior standard deviation of 0.69, fourth from left is the wet (β_4) with a posterior mean of -0.37 and posterior standard deviation of 1.32. Forest is the reference category for land use.

3.5 Discussion

We have presented an efficient Bayesian inference scheme for scalable step selection models that provides approximate posterior distributions of resource selection coefficients. This approach reveals that there are substantial levels of uncertainty in estimates of resource selection coefficients even when using datasets typically considered large for movement ecology studies, such as circa 10,000 observations. When scaling selection estimates to predictions about long term space use, this uncertainty can significantly impact the accuracy and utility of these predictions. In the context of animal move-

ment data, such uncertainty may arise from limited sampling duration, GPS location error, missing data due to device malfunction, or biased coverage across individuals, time periods, or spatial extents. Likewise, environmental covariates may introduce uncertainty due to coarse spatial or temporal resolution, measurement error, or missing values. Addressing these sources of uncertainty and developing methods to quantify them is essential for producing robust and ecologically meaningful space use predictions, an approach implemented in this study through the use of variational inference to quantify uncertainty in resource selection and avoidance parameters.

There are several key advantages to the approach we propose here. Firstly, it is fast and scalable making it suitable for handling large datasets. Notably, the time taken to optimise the variational distribution in a single gradient step does not scale linearly with the number of observations due to the use of mini-batching, meaning that significant computational savings can be achieved (Zhang et al., 2018). This is in contrast to MCMC sampling approaches which use the entire dataset to evaluate the likelihood in each step of the sampler. In addition, it relies on gradient-based optimisation, where the parameters of the variational distribution and any associated hyper-parameters within the model are optimised with stochastic gradient descent. This enables the use of techniques such as automatic differentiation and adaptive optimisation schemes that have received significant attention in the past years due to their importance for training deep neural networks. Secondly, it offers explicit posterior probability distributions of the estimated habitat selection parameters, allowing for uncertainty quantification. By combining the method of Michelot et al. (2019a) with variational inference, we present a comprehensive framework for analysing animal movement data that is both computationally efficient and provides a detailed understanding of the uncertainty in the estimates. This enables the uncertainty in resource selection coefficients to be propagated to estimates of long

term space use in an efficient and rigorous way. While alternative techniques, such as Potts et al. (2014); Signer et al. (2017, 2024), for estimating utilisation distributions from step selection coefficients do exist, they can be challenging to implement and computationally expensive, especially when performing multiple simulations to account for parameter uncertainty.

Currently, standard practices in multiscale modelling frameworks for animal movement frequently rely on methods such as maximum likelihood estimation (MLE) to estimate movement, resource avoidance, and selection parameters. However, MLE-based approaches have been shown to perform poorly when data is limited and complex models are used, which is a common scenario in movement ecology (Ferguson, 1982; Kéry and Schaub, 2011). Our study advances existing approaches by offering improved computational efficiency and approximate posterior distributions of model parameters that may be employed to create predictions of space use that incorporate systemic uncertainty.

Furthermore, the proposal to use VI in multiscale step selection models is particularly timely, as the findings of previous studies have shown that Bayesian, machine learning, and deep learning approaches outperform conventional statistical inference in species distribution modelling, particularly in terms of generalisability to novel ecosystems (Aldossari et al., 2022) and predictive accuracy of suitable habitats (Noda et al., 2024).

Moreover, the differences in predictions of utilisation distribution between observations of varying sizes (Fig. 3.4B-D) explicitly highlight the importance of propagating the uncertainty in parameter estimates into estimates of utilisation distributions. In a conservation context, this creates the risk of failing to pinpoint correctly animal utilisation hotspots in protected areas, misallocation of sparse and valuable resources that need to be channelled into the protection of animals, and leading wildlife managers to make incorrect decisions due to overconfidence in predictions and failure to properly account

for uncertainty (Jansen et al., 2022; Rocchini et al., 2011).

The results from the case study in Section **3.4.1** indicate that fisher selects areas with high elevation and avoids grass areas compared to forests. These findings are consistent with the results of a previous study by Fieberg et al. (2021), which employed integrated SSFs to infer habitat selection parameters using fisher's location data. In that study, available steps were generated by sampling step lengths from a gamma distribution and turning angles from a von Mises distribution. In contrast to the SSF approach, the parameters inferred from the variational inference method can directly be used to estimate fisher's utilisation distribution without the need for simulation, offering formal uncertainty quantification in an animal's habitat selection parameters and the resulting utilisation distribution.

The development of efficient inference methods for animal movement data is of critical importance for conservation management and planning due to the increasing availability of GPS-tagged animal telemetry data and the growth in the size of movement datasets (Joo et al., 2020; Nathan et al., 2022). Processing these datasets within a reasonable time scale is essential if the information is to be used in making informed conservation management decisions such as where to allow infrastructure development in protected areas while ensuring minimal impact to migratory, nomadic, dispersing, and sedentary animals. Our approach offers such a possibility by estimating the degree of uncertainty around the predictions of how animals use or avoid certain landscape features, such as accessing forage while avoiding human infrastructure. The application of this approach is a useful compromise in modeling animal resource selection and movement, quantifying uncertainty, and minimising computational cost.

In this study, a symmetric, resource-independent movement kernel was specified using a Gaussian distribution. This assumption offers significant computational and analytical

advantages due to its smooth and symmetric form and effectively captures movement patterns where short steps are more frequent than long ones, which is consistent with many animal telemetry datasets. However, this simplification may not fully capture the movement of animals in real-world, particularly in cases where individuals take long distance steps more frequently than the Gaussian tail would allow. This can lead to a misrepresentation of movement behaviour, although it is unlikely to substantially bias habitat selection parameters (Michelot et al., 2020). The discrepancy primarily arises from the inability of the Normal kernel to represent heavy-tailed movement distributions and directional biases, thereby underestimating the probability of large movements. While the Normal kernel remains useful for reliable inference of habitat selection and avoidance parameters, alternative approaches such as incorporating a random availability radius modelled using a gamma distribution may better capture variation in step lengths and movement speed (Michelot et al., 2020), thus improving the ecological realism of the movement process.

In summary, we have described a Bayesian framework that is able to accurately infer resource selection and movement coefficients from synthetic movement data. By evaluating the performance of our framework in a range of simulated scenarios we have shown how it is able to consistently recover different combinations of selection and movement parameters and quantify their uncertainty in an efficient and scalable manner. Our results highlight the importance of obtaining posterior distributions for model parameters rather than simply point estimates especially when making predictions that span multiple scales, from fine scale decision-making to the long-term use of space by animal populations.

Chapter 4

Revealing the effects of anthropogenic structures on the spatial distribution of migratory wildebeest

Note:

The content of this chapter is currently under review for publication in the Journal of Applied Ecology.

Abstract

The increasing interaction between wildlife and humans, both within and outside protected areas, highlights the importance of understanding how migratory animals respond to anthropogenic disturbance. To effectively safeguard migratory populations, we must understand their habitat use, particularly in response to the expanding presence of human-made structures in their environments. In this work, we employed a multiscale step selection model within a Bayesian framework to explore the impact of human-made structures on the movement patterns and habitat preferences of migratory wildebeest in the Serengeti. Our findings reveal that wildebeest tend to avoid areas near these structures, even in the core of the protected area where tourist infrastructure is the most prevalent. Although buildings do not entirely exclude wildebeest, they do reduce the amount of time wildebeest spend in their vicinity. Individuals weigh multiple trade-offs in deciding whether to remain or move during migration, and if animals forego access to key resources in the areas around buildings, this could lead to reduced fitness and demographic consequences that may not be immediately apparent. We further find that increasing numbers of co-located buildings have a diminishing rather than a compounding effect on the spatial distribution of wildebeest, meaning that clustering buildings away from key grazing areas could be a beneficial strategy. Synthesis and Applications: In light of these findings, we recommend careful regulation and spatial planning of infrastructure development within ecosystems that considers the nuanced effects human-made structures can have on the behaviour and habitat use of migratory animals.

4.1 Introduction

Protected areas worldwide are experiencing increases in anthropogenic disturbance, driven by expanding human activities including tourism, commercial development, and settlement within boundaries (Dirzo et al., 2014; Jones et al., 2018). The impacts of disturbance on wildlife populations and implications for the sustainable conservation of biodiversity within protected areas is unknown. Leveraging nature capital for tourism has emerged as a prominent strategy within the field of conservation management in recent years, particularly in protected areas in Africa (Lindsey et al., 2020). This has led to a significant proliferation of hard infrastructure development, including road networks and tourist accommodations such as campsites, lodges, and hotels within these once pristine areas (Larsen et al., 2020; Tverijonaite et al., 2018). Of particular concern is how wildlife respond to structural alterations in their habitats, with documented effects ranging from behavioural changes such as reduced movement (Doherty et al., 2021; Stabach et al., 2022), shielding effects (Berger, 2007) and noise avoidance (Zanette et al., 2023), to physiological responses including elevated stress hormones (Creel et al., 2002), and ultimately to population-level impacts through altered species interactions (Shannon et al., 2017) and reduced reproductive success (Phillips and Alldredge, 2000). Beyond these direct effects, animals may abandon areas where vital habitat features are fragmented, where forage quality and quantity decline, or where the perceived risk outweighs the value of the resource following infrastructure development. Understanding these responses at fine spatial and temporal scales is therefore crucial for wildlife managers and policy makers to evaluate and improve conservation planning strategies.

Migratory species are particularly vulnerable to anthropogenic disturbances because they require unimpeded access to seasonally available resources that are spatially sepa-

rated across large landscapes (Kauffman et al., 2021). Migratory animals often have to trade-off resource acquisition while avoiding perceived risks as they navigate through dynamic and heterogeneous ecosystems (Hopcraft et al., 2014, 2010). These trade-offs can have consequences for their individual fitness as well as the ecological community, given their influence on shaping ecosystem structure, functions, and processes (Dobson, 2009; Wilcove and Wikelski, 2008). Moreover, the concurrent effects of global environmental change and the increase in human disturbance pose dual threats to migratory animals by limiting the availability of suitable habitats in both space and time (Stabach et al., 2022; Wilcove and Wikelski, 2008), intensifying competition for limited resources in less productive areas (Stabach et al., 2022), and disrupting behavioural and migratory patterns (Aikens et al., 2022; Larsen et al., 2020; Paun et al., 2022; Veldhuis et al., 2019). To protect migratory populations over the long-term, it is essential to accurately identify the threats to their viability. For example, assessing the uncertainty in direct and indirect anthropogenic impacts on animal behaviour can help conserve species and strengthen ecosystem resilience.

Understanding the response of animals to anthropogenic pressure requires an efficient modelling framework that is sufficiently flexible to be able to model fine-scale movement behaviours as well as being able to quantify uncertainty in predictions about how individual decisions scale to produce long-term shifts in space use at the population-level (Masolele et al., 2024; Michelot et al., 2020, 2019a). Many studies that model movement decisions as a response to anthropogenic disturbances focus on features such as roads (Prokopenko et al., 2017; Scrafford et al., 2018; Singh et al., 2024) and fences (Robb et al., 2022) using step selection functions (SSF) and general and generalised linear models by estimating the change of movement metrics (such as home range, movement distance, and speed) (Aikens et al., 2022; Doherty et al., 2021; Mendgen et al., 2023; Tucker

et al., 2023), and fences using barrier behaviour analysis (Xu et al., 2021). However, how these effects of anthropogenic disturbances on individual animal movement decisions translate into animal's space use at the population level remains largely unknown. To accurately predict these processes and movement decisions at the population level, multiscale models are required that are capable of linking fine-scale animal movement decisions with predictions of animal space use at the broad scale (i.e. the long-term utilisation distribution) (Michelot et al., 2020). This translation from individual movement to population-level space use is achieved by modelling movement as a stochastic process, where both short-term step selection and long-term utilisation arise from the same underlying habitat selection mechanism. In other words, repeated application of local movement rules leads to emergent long-term space-use patterns. For multiple individuals, movement data are discretised into steps, and the overall likelihood of observed tracks is calculated as the product of the likelihoods of individual steps (Michelot et al., 2019a).

The Greater Mara-Serengeti ecosystem is currently experiencing rapid transformation as a result of landscape modifications driven by infrastructure development (Larsen et al., 2020). Examining the response of migratory herbivores to these landscape changes is often overlooked since responses are difficult to detect and are not a primary area of focus for wildlife managers when compared to other drivers of biodiversity decline such as poaching. However, these nonlethal effects on animal populations could potentially lead to large-scale impacts on the migratory system, as many of these species operate at their physiological limits and rely on large and extensive undisturbed landscapes to maintain viable and self-sustaining populations (Hopcraft et al., 2014). The Serengeti wildebeest (*Connochaetes taurinus*) serves as a prime example of such species, which in addition to being an iconic migratory species, also serves as an indicator species in the

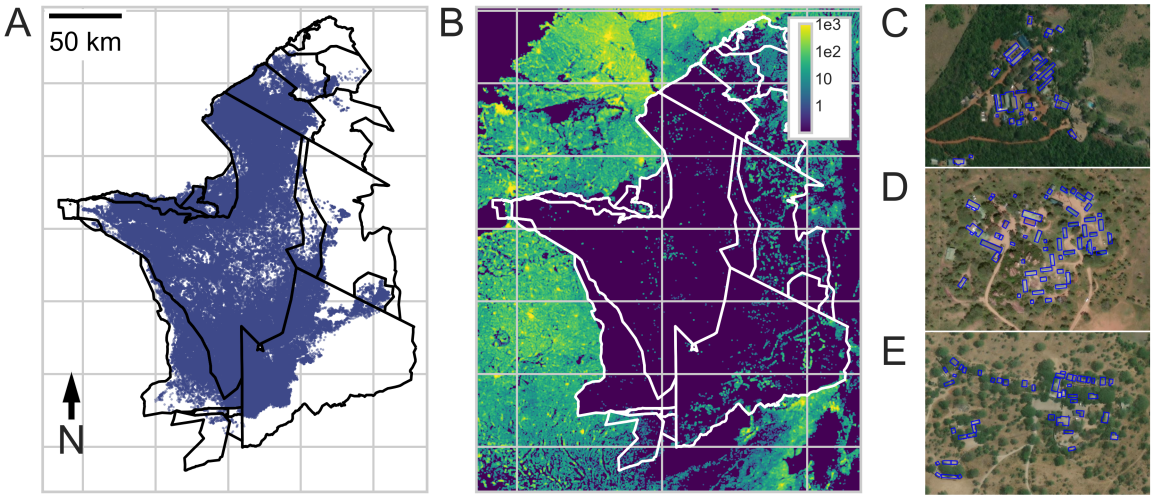

Figure 4.1: A) A map of the Serengeti ecosystem with wildebeest GPS locations shown as blue points, black line indicates the boundary of the Serengeti ecosystem and associated protected areas, B) Building density per square kilometer within the Serengeti ecosystem and within a buffer zone of approximately 20 Kilometers. Note, the colormap employs a log scale. (C-E) Satellite images illustrating selected locations within distinct regions of the ecosystem along the wildebeest migration route (C) Maasai Mara National Reserve in Kenya, (D) Seronera in the Serengeti National Park and (E) Ndutu within the Ngorongoro Conservation Area Authority. Blue lines in the satellite imagery show the outlines of buildings.

ecosystem (Mduma et al., 1999; Torney et al., 2018), and are sensitive to habitat changes resulting from human-induced disturbance (Kavwele et al., 2022).

In this work, we explore how movement decisions of wildebeest are influenced by human-made structures using a step selection model (Michelot et al., 2019a). This approach allows us to explore how the presence of buildings affect long-term space use by migratory wildebeest in the Serengeti. We quantify the behavioural response of wildebeest to buildings as a function of distance and determine if there are diminishing or compounding effects of multiple co-located human-made structures. Our findings aim to guide spatial planning of infrastructure development within ecosystems by accounting for the nuanced effects of human-made structures on the behaviour and habitat use of migratory animals.

4.2 Methods

4.2.1 Empirical data collection

GPS collars were deployed on 57 migratory wildebeest (*Connochaetes taurinus*) in the Serengeti National Park, Tanzania, between January 2019 and September 2023. Because the acquisition of locations varied among collared migratory wildebeest, we filtered all data to retain only locations where there was between one hour and less than 24 hours between successive fixes. The resulting movement data set consisted of 143,268 locations collected by 57 collared wildebeest. Fig 4.1A shows a map of the Serengeti ecosystem along with the recorded wildebeest movement data. We obtained the large-scale open data set that contains the outlines of buildings derived from high-resolution satellite imagery accessible at open buildings managed by (Sirko et al., 2021). We used version 2 of the open buildings dataset which was created in August 2022 on imagery cover-

ing 3.91×10^7 km^2 of Africa, South and South-East Asia. Building footprint detections from the open buildings dataset with a confidence score below 70% were discarded. These detections are derived from satellite imagery using deep learning models, and the confidence score reflects the model's estimated likelihood that a detected footprint corresponds to an actual building. Fig 4.1B shows a map of buildings per square kilometer for the Serengeti ecosystem and within a buffer zone of approximately 20 kilometers.

4.2.2 Model inference

In order to assess the effects of anthropogenic structures on the movement and long-term space use of wildebeest in the landscape, we included a distance to buildings covariate within a step selection model (Michelot et al., 2020, 2019a). Specifically, we calculated the distance of a potential location an animal could select to the 10 nearest buildings and then transformed these distances into a set of covariates that are used within the step selection model. In the model, the likelihood of a wildebeest moving to a location $\mathbf{x}_{t+1}$ at time $t+1$ given it is at location $\mathbf{x}_t$ at time t is given by

$$p(\mathbf{x}_{t+1}|\mathbf{x}_t) = \int_{\boldsymbol{\mu}\in\Omega} \frac{w\{\mathbf{c}(\mathbf{x}_{t+1})\}\phi(\mathbf{x}_{t+1}|\boldsymbol{\mu})}{\int_{\mathbf{z}\in\Omega} w\{\mathbf{c}(\mathbf{z})\}\phi(\mathbf{z}|\boldsymbol{\mu})\, d\mathbf{z}} \phi(\boldsymbol{\mu}|\mathbf{x}_t) d\boldsymbol{\mu}. \tag{4.1}$$

where $\boldsymbol{\mu}$ is a random intermediate step, $\phi(\mathbf{x}_{t+1}|\boldsymbol{\mu})$ is a symmetric, resource-independent movement kernel, and $w\{\mathbf{c}(\mathbf{x}_{t+1})\}$ is the selection function that models preference (attraction or repulsion) for the environmental covariates $\mathbf{c}(\mathbf{x})$. Implicit in eqn. 4.1 is a dependence on a parameter vector $\boldsymbol{\theta}$ which consists of the selection parameters and a resource independent movement parameter. The intermediate random step used in the likelihood calculation is not intended to model animal movement behaviour but is a technical requirement that enables the model to be fitted to fine-scale movement data and then used to make predictions of long-term space use (the interested reader should

refer to (Michelot et al., 2020, 2019a) for details).

The selection function $w\{c(\mathbf{x}_{t+1})\}$ takes an exponential form and is defined as

$$\exp\left[\beta_1 c_1(\mathbf{x}_{t+1}) + \beta_2 c_2(\mathbf{x}_{t+1}) \cdots + \beta_p c_p(\mathbf{x}_{t+1})\right]. \tag{4.2}$$

where β_i is the selection coefficient for environmental covariate c_i. Since any potential wildebeest response will decrease with increasing distance away from the buildings, we define the covariates based on a monotonically decreasing function so that

$$c_i(\mathbf{x}) = \left(1 + \exp\left[\alpha\Big(\|\mathbf{x} - \mathbf{b}_i\| - \gamma\Big)\right]\right)^{-1} \tag{4.3}$$

where α, γ are model parameters that control the shape of the function, $\|\cdot\|$ is the Euclidean norm, and $\mathbf{b}_i$ is a vector representing the location of the i^{th} nearest building to $\mathbf{x}$. Adjusting the parameters α and γ will change the shape of the response function. For example, a larger α will result in a steeper decline in the response as the distance increases, while a larger γ will shift the point at which the response is at half its maximum further away from the building.

The selection coefficient β_i then scales the response to the covariate value c_i and may be positive, meaning that animals are more likely to select locations near to buildings, or negative, indicating an avoidance response. However, in all scenarios the response becomes weaker as the distance to the building increases due to spatial attenuation. We include within the covariate fields potential effects of up to 10 nearest buildings based on their ranked distances (e.g., first nearest, second nearest, up to the tenth nearest) on wildebeest selection. Instead of allowing the selection coefficients to be independent of one another, we impose the constraint that there is a baseline response to the nearest building ($\beta \equiv \beta_1$) and subsequent buildings induce a response that are scaled versions

of this baseline, i.e.

$$\beta_i = \omega^{i-1}\beta_1 \qquad \forall i \in \{2,..,p\} \tag{4.4}$$

where ω is a model parameter. This enables us to encode the response to an arbitrary number of buildings with two parameters, while still addressing whether subsequent buildings have a diminishing effect ($\omega < 1$), a compounding effect ($\omega > 1$), or if the effect depends only on the Euclidean distance to buildings ($\omega = 1$). The final selection function is therefore

$$w(\mathbf{x}) = \exp\left(\sum_{i=0}^{N-1} \omega^i \beta \left(1 + \exp\left[\alpha\Big(\|\mathbf{x} - \mathbf{b}_i\| - \gamma\Big)\right]\right)^{-1}\right) \tag{4.5}$$

The response to buildings is therefore determined by four parameters in our model, the response to the closest building β_1(baseline selection coefficient value), hereafter referred to as β, the relative effect of additional buildings ω, where $\omega < 1$ indicates a diminishing effect and $\omega > 1$ indicates a compounding effect, the steepness of any threshold response to buildings α, and the inflection distance of the response γ. N is the number of buildings considered which we set at 10 in our analysis. We also experimented with both smaller ($N < 10$) and larger ($N > 10$) numbers of nearest buildings and observed that the relationship of interest remained consistent across these thresholds. Therefore, 10 nearest buildings were chosen to minimise unnecessary computational expense while maintaining the reliability of the results.

Given the wildebeest movement dataset $\mathcal{D}$ and the building locations data, we are able to compute the log likelihood function given in eqn. 4.1. In order to infer model parameters we employ variational Bayesian inference (Blei et al., 2017) following the methodology described in (Masolele et al., 2024). The idea behind variational inference is first to pro-

pose a family of variational distributions with which to approximate an unknown posterior, followed by optimising the parameters of the proposed distribution to minimise the distance (defined using the Kullback–Leibler divergence) between the approximate distribution and the true posterior. For our case, we employ an independent Gaussian distribution as the family for the variational distribution for each parameter. We then optimise the parameters of the Gaussian distributions and a hyperparameter associated with the resource independent movement process using a stochastic gradient descent algorithm (Hoffman et al., 2013). This approach enabled wildebeest telemetry locations to be processed in batches, which reduces memory requirements and increases computational efficiency. Three parameters (α, γ, ω) are constrained to be positive by applying an exponential transform and diffuse priors are employed as follows,

$$\begin{aligned} \beta &\sim N(0,10), \\ \ln\alpha &\sim N(0,10), \\ \ln\gamma &\sim N(0,10), \\ \ln\omega &\sim N(0,10), \end{aligned} \tag{4.6}$$

This choice of priors reflects no prior assumption of preference or avoidance for buildings, while avoiding overly restrictive constraints on ecologically plausible effect sizes. It allows the distances to buildings calculated at the end steps of the migratory wildebeest movement steps to drive the inference.

The parameters were optimised using stochastic gradient descent with a learning rate of 0.1 and a batch size of 1024, which means that 1024 movement steps were passed to the optimiser at each iteration, with the dataset shuffled at the end of each epoch. To assess convergence the change in the variational posteriors after each epoch of training was

monitored and the algorithm was halted once the posteriors had converged. All analyses were performed in Python using TensorFlow (Abadi et al., 2016) and TensorFlow Probability (Dillon et al., 2017).

4.3 Results

4.3.1 Displacement effects of buildings in the Serengeti ecosystem

We observe that the presence of anthropogenic structures has a significantly negative effect on space use by wildebeest. Given the posterior distribution for the selection coefficient β it is highly probably that this parameter is less than zero, with posterior mean $\hat{\beta} = -0.059$ (95% CI: $[-0.076, -0.041]$; $P(\beta > 0|\mathcal{D}) = 2.13e^{-11}$). In Fig. 4.2A we show the posterior distribution for the log selection function in the presence of a single building, while in Fig. 4.2B the expected reduction in space use across the ecosystem is shown, conditional on the location of buildings in the 2022 open buildings dataset. In the most affected areas, we predict a reduction of approximately 22% in wildebeest use. Furthermore, there is a substantial reduction in wildebeest use at the centre of the park compared to the south-east and areas to north of the centre of the park due to the relatively high number of buildings in this area as shown in Fig. 4.1B. During this period, we observed that wildebeest moved on average approximately 630 meters/hour.

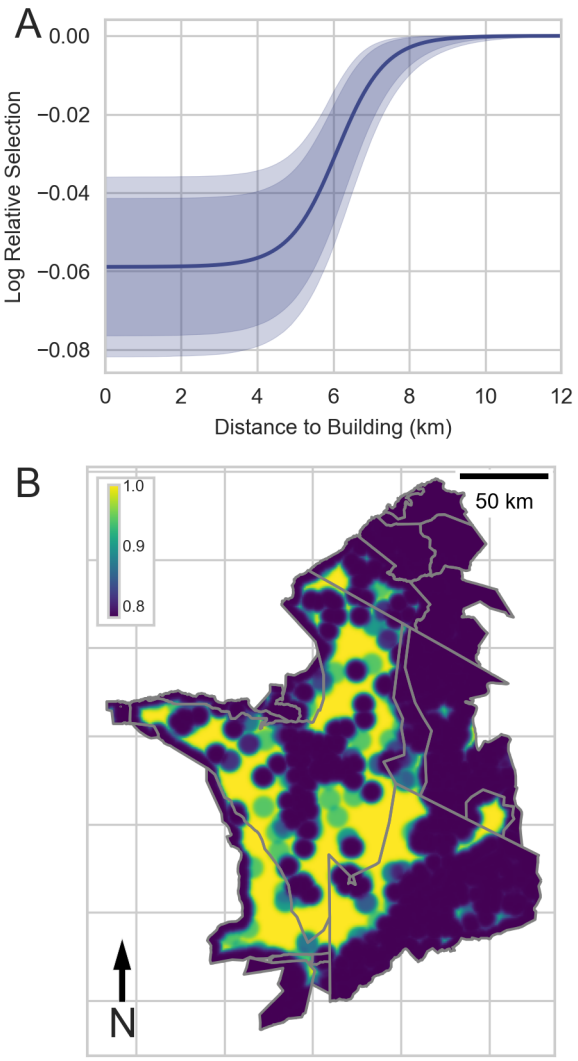

Figure 4.2: A) Log relative step selection strength as a function of distance to a single building. Blue line indicates the inferred posterior mean and shaded gray regions (dark to light) represent 95%, and 99% credible intervals respectively. B) Predicted reduction in space use due to the response to buildings. Colors indicate expected reduction in use from no reduction (yellow) to around 22% reduction (dark blue). Gray lines indicate the boundary of the Serengeti ecosystem and associated protected areas.

4.3.2 Nonlinear decay and threshold effects

We include the parameters γ and α in the model to quantify the spatial extent of the influence of the buildings on wildebeest and the manner in which this influence decays with increasing distance. When α is small the functional form of eqn. **??** results in a linear relationship between the buildings and wildebeest response. In contrast, for larger values of α ($\alpha \gg 1$), the response is nonlinear and there is a threshold response whereby buildings closer than γ elicit a strong response while buildings at a greater distance than this threshold have a far weaker effect. Our findings reveal that the response of the wildebeest to buildings decays in a nonlinear manner, with areas in close proximity to buildings being less likely to be selected but the effect diminishes rapidly once the distance from the buildings reaches a threshold. This pattern is illustrated in Fig. 4.2A, where the response is initially constant until the distance to the building reaches approximately 4 kilometers before transitioning to zero (no response) at a distance greater than 8 kilometers. The nonlinear response of the wildebeest to buildings is controlled by the parameter α, which has a posterior mean of $\hat{\alpha} = 1.617$ (95% CI: $[1.262, 1.972]$). Furthermore, the effect of a building reaches half of its maximum strength at a distance of approximately 6.1 kilometers from the building, indicating that the transition from a negative selection response to a neutral response is centreed at this distance ($\hat{\gamma} = 6.092$; 95% CI: $[6.011, 6.173]$).

4.3.3 Interacting effects of multiple buildings

Next we examine the cumulative impact of multiple buildings on wildebeest behaviour by analyzing their combined response to multiple structures at different distances. Our model incorporates the ranked distances to buildings (including from nearest to tenth nearest) and allows their combined influence on wildebeest selection and space use pat-

terns to depend on the Euclidean distance along with the rank order, i.e. a building at a distance of 5 kilometers to a potential location will have a different effect depending on whether it is the nearest building to the location or the 10th nearest. In this way clusters of buildings may have a compounding or diminishing effect. The nature of the effect of multiple buildings is controlled by the parameter ω in our model which defines the relative influence of each additional building in the vicinity of a potential location. If $\omega = 1$ then every building essentially has an independent effect that depends only its Euclidean distance to a location. If $\omega > 1$ then the effect is compounding and two buildings close to one another will have more than double the effect on the environment than a single isolated building, whereas if $\omega < 1$ the effect diminishes.

By allowing the effect of multiple buildings to be modelled in this way we find that there is a strong avoidance response to the nearest building that diminishes with each additional building in proximity (Fig. 4.3A). We assess the strength of this diminishing effect by examining the posterior distribution of the weight decay parameter ω which has a posterior mean of $\hat{\omega} = 0.776$ (95% CI: $[0.667, 0.884]$). Since $\omega < 1$ with high probability ($P(\omega \geq 1|\mathcal{D}) = 2.32e^{-5}$; Fig. 4.3B), this indicates that each subsequent building has a weaker effect than the previous one. These findings suggest that clustering buildings together may help minimise their overall spatial impact on wildebeest movement patterns and habitat use compared to dispersing buildings across the landscape.

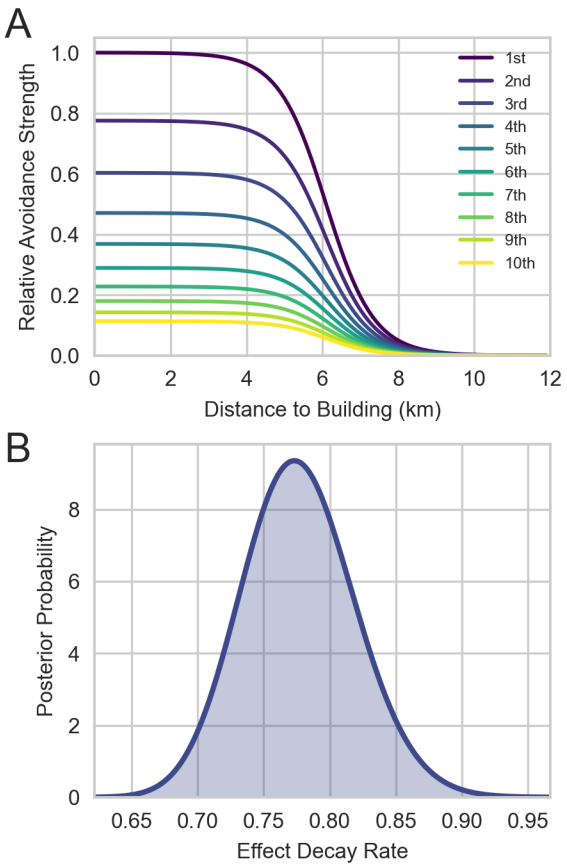

Figure 4.3: A) The relative avoidance strength of multiple co-located buildings (Distance to nearest ten (10) buildings) on wildebeest selection pattern, B) Posterior probability distribution of inferred decay rate of the effect of building, with a posterior mean of 0.776 and posterior standard deviation of 0.055.

4.4 Discussion

This study quantifies how anthropogenic structures influence wildebeest movement decisions in the Serengeti ecosystem using a novel multiscale analysis approach. By examining the cumulative impacts of multiple co-located buildings on the behaviour of wildebeest, we reveal patterns of avoidance that have not been previously documented.

Our findings demonstrate that wildebeest typically avoid areas near buildings, though this avoidance does not result in complete exclusion from these areas. Instead, the analysis reveals a reduction in space use of approximately 22% in the most affected regions.

The densest building areas within the region are in the central Maasai Mara National Reserve and the Serengeti National Park as compared to other protected areas in the ecosystem where human settlement is restricted (such as the Maswa Game Reserve, Kijereshi Game Reserve, Ikorongo-Grumeti Game Reserve and Pololeti Game Reserve). In the Serengeti National Park, the densest building areas are in the central Seronera region and to the north and north-west in the Kogatende region (Fig. 4.1B). Approximately 80% of all habitat areas in the Greater Serengeti-Mara ecosystem are located within 6.1 kilometers of a building (the distance at which we observe the response to buildings is approximately half of its maximum). This means that a relatively small area of the ecosystem is unaffected by anthropogenic disturbances. In the core protected area of the ecosystem, the Serengeti National Park, where wildebeest spend most of the year, approximately 65% of all habitat areas are within 6.1 kilometers of buildings, leaving only 35% of the habitat relatively unaffected.

Despite the high numbers of anthropogenic structures in the ecosystem, we find that wildebeest do not completely abandon areas near buildings but instead reduce their usage, suggesting they weigh localised, nonlethal risks against surrounding levels of resources. Their response aligns with current theory surrounding the 'landscape of fear' concept that describes how animals integrate risk information across multiple spatiotemporal scales (Gaynor et al., 2019; Palmer et al., 2023; Tablado and Jenni, 2017). Given that the edges of the Serengeti ecosystem are facing increasing pressure from human activities (Veldhuis et al., 2019), the continued use of regions in the vicinity of buildings may indicate a scarcity of viable alternative areas, rather than a display of tolerance or site fi-

delity. Although the response of wildebeest does not exclude them from these areas, the impact of buildings may limit their access to resources and could have increased energetic costs. Furthermore, these changes may have cumulative effects, potentially leading to long-term consequences for the viability and sustainability of the population (Smith et al., 2021).

Clustering buildings so they are close to each other has a diminishing rather than a compounding effect on, meaning that the largest impact on wildebeest occupancy occurs with the placement of the first building in an area, which has a larger impact on wildebeest behaviour than subsequent buildings. This is somewhat surprising since we may expect wildebeest to show stronger responses to areas where the building density is highest, assuming the density of buildings is correlated with the degree of human disturbance such as noise or lights at night. A possible explanation for the diminishing effects of co-located buildings could be that wildebeest react to the presence or absence of humans, rather than on the level of human activity.

There are several potential mechanisms that could explain why wildebeest avoid areas near buildings. These areas are likely to experience high levels of traffic disturbance and vehicle movement, creating suboptimal conditions for resting and foraging. This aligns with studies of other migratory ungulates such as a study on elk (Prokopenko et al., 2017) which found that these ungulates avoided areas near human-made infrastructure such as linear features, potentially due to relatively high disturbance. The presence of humans is known to illicit strong avoidance responses in wildlife (Stiegler et al., 2024; Zanette et al., 2023) and buildings may be considered a proxy for human presence (Potapov et al., 2014) with the response mediated by anthropogenic noise (Harding et al., 2019; Shannon et al., 2016), light pollution (de Wilde and de Souza, 2022), or memory of their locations (Bracis and Mueller, 2017; Verzuh et al., 2024). Another potential

explanation could be that humans select building sites that are associated with naturally risky features in the landscapes that wildebeest may inherently avoid such as rocky outcrops (kopjes) or permanent waterholes, which attract predators such as lions (Hopcraft et al., 2005; Valeix et al., 2009). Without a controlled experimental intervention, we cannot rule out the presence of a confounding variable that is positively correlated with the presence of buildings but negatively correlated with wildebeest selection. However, the strong avoidance response we observe suggest that human presence is a key driver of the observed patterns.

The parametric distance function employed in eqn. 4.5 reflects a biologically informed assumption that the influence of buildings on wildebeest movement decreases smoothly with distance. This formulation captures the ecological expectation that animals are more likely to respond to nearby anthropogenic features than to those farther away. By allowing the effect of buildings to diminish gradually with distance, the model avoids the unrealistic implication of persistent long-range influences. The parameters α and γ control the shape and scale of this decay, ensuring that the function remains monotonically decreasing and biologically interpretable. In particular, γ governs the approximate distance at which behavioural responses begin to transition, acting as a soft threshold rather than enforcing a sudden cut-off distance. This allows the model to capture a gradual shift in selection or avoidance behaviour, which is more consistent with observed patterns in animal movement than models assuming abrupt or binary responses without attenuation. Overall, this formulation provides a flexible and ecologically grounded way to represent how migratory wildebeest interact with built environments during movement decisions.

To ensure that our model provided a good fit to the data when using variational inference, we monitored the evidence lower bound (ELBO), which serves as a proxy for

the model's log marginal likelihood. Specifically, we monitored ELBO at each training epoch, where one epoch corresponded to a complete iteration through the entire migratory wildebeest movement dataset. To avoid overfitting and reduce unnecessary computation, training was halted early if the ELBO failed to improve over five consecutive epochs. This type of convergence monitoring is commonly used in variational inference frameworks and has been applied in ecological studies, such as the inference of spatially varying migratory wildebeest movement characteristics in the Serengeti National Park by Paun et al. (2022).

In summary, this study represents the first attempt to assess the potential nonlinear effects of human-made structures on the spatial distribution of wildebeest, while taking into account interacting effects of multiple co-located buildings. Our findings suggest that priority should be given to conservation efforts in high-use areas such as key foraging grounds and migratory routes. Limiting the expansion of building structures in these critical habitats and clustering buildings away from these areas will mitigate the impacts on migratory wildlife. Taken together, our results suggest that developing infrastructure in the ecosystem and surrounding areas that minimises the effects on migratory wildebeest requires careful planning, knowledge of wildebeest behaviour, and information on the history of their migratory route. Future research should incorporate additional characteristics of anthropogenic structures, such as the size of the building, the associated levels of traffic, and the nature of the structure, such as tented camps versus permanent structures, thus providing more accurate information for relevant conservation and management efforts.

Chapter 5

Identifying the migration routes of Serengeti wildebeest with hierarchical sparse Gaussian processes

Note:

This chapter has been prepared as a manuscript which I aim to submit to the journal Movement Ecology.

Abstract

Understanding animal migration patterns through the identification of migratory routes is essential for advancing our knowledge of animal movement behaviour, habitat connectivity, delineating migratory corridors, and understanding how animals interact with the landscape and respond to dynamic resource distributions, all of which are critical for effective conservation planning and for uncovering the mechanisms underlying migration. However, modelling animal movement at scale presents significant challenges due to the high volume of fine-scale, high-accuracy tracking data, the cyclic and often stochastic nature of migratory behaviour, and the inherent complexity of spatiotemporal dynamics. In this study, we introduce a novel hierarchical sparse Gaussian process (HSGP) framework for identifying population-level animal migration routes using large-scale datasets of GPS locations. Our method leverages the flexibility of Gaussian processes for non-parametric modelling of animal migration, while incorporating sparsity and hierarchical structure to achieve computational efficiency and capture both individual movement dynamics and the autocorrelation inherent in GPS tracking data. We apply this framework to a decade of movement data from migratory wildebeest within the Serengeti ecosystem, one of the worlds most iconic and ecologically significant migratory systems. Our results reveal that at the onset of the dry season, wildebeest transit through the western Serengeti toward the northern regions, whereas during the short rainy season, typically around November, they begin moving southward toward the nutrient rich short grass plains of the Serengeti, although their migratory routes show increased variability during this period. These findings demonstrate the utility of HSGP in ecological movement studies and provide actionable insights for tracking migration, preserving critical migratory pathways and guiding conservation planning in dynamic, large-scale landscapes.

5.1 Introduction

Conserving animal populations, such as migratory animals, requires a detailed understanding of the underlying drivers of observed spatial distributions. Identifying the timing of the temporal pattern of animal movement would enable an understanding of the set of ecological interactions in which they are involved. For example, due to their wide range patterns, migratory animals often interact with large numbers of other animals in various ways; such interactions can range from consumption, mutualism, disease spread, and competition (Bauer and Hoye, 2014). Therefore, in the current era with ecosystems experiencing changing climates, increased human population growth, increased infrastructure development, and an assortment of other destabilising effects, it is important to understand how migratory species may shift their migration routes over time in the system in which they occur. This information is important for understanding niche partitioning, migratory species ecology and evolution, and how these systems in which they occur will endure into the future given the role they play in shaping ecosystem structure and functions, and biological processes.

Nomadic, dispersal, and periodic migration are typical examples of macroscale and mesoscale movement phase processes exhibited by migratory animals and are long-range movement patterns in the sense that they occur at a large (spatial and temporal) scale (Mueller and Fagan, 2008; Nathan et al., 2008). For migratory species, both macroscale and mesoscale movement phase processes, such as periodic migration, are primarily motivated by the need to track and select temporary but high-quality forages (Fryxell and Sinclair, 1988). These long distance movements make the animal incur migration cost in terms of energy, time, and exposure to risks. Poor quality or low quantity forage and depleting fat reserves significantly impact an animals ability to meet these

cost requirements: hardly digestible forages reduce fat gain, low fat reserves increase foraging time or weaken the animal if not able to locate quality forage to replenish the lost reserve, and risks such as predation and poaching cost animal life or injuries. Therefore, through learning and memory (short-term and long-term reference memory), migratory animals are expected to have acquired a navigational capacity that allows them to select routes that minimise risk and energetic costs while maximising resource acquisition under the different environmental conditions they encounter. Identifying these migration routes is of critical importance, as it helps predict responses to environmental changes and directs conservation efforts focused on protecting vital corridors from habitat fragmentation. Furthermore, understanding these routes allows for the identification of key pathways, enabling their protection from development activities such as tourism infrastructure, thereby ensuring the continued survival of migratory species and biodiversity conservation.

Identifying migration routes of animals requires a modelling framework that can capture the cyclic or periodic patterns that an animal exhibits between different areas of the landscape due to the change in selection patterns driven by the change in environmental resources. To address this, several approaches have been developed that use animal telemetry data. For instance, a multilevel Gaussian process model in continuous-time has been used to identify animal migration routes and activity patterns (Torney et al., 2021), while likelihood-based approaches have been used to estimate range-shift parameters such as transition durations and site fidelity (Gurarie et al., 2017; Patin et al., 2020), and migration parameters such as distance and timing (Bunnefeld et al., 2011; Gurarie et al., 2019). Although the multilevel Gaussian process framework proposed by (Torney et al., 2021) is capable of detecting multiscale patterns and trends in movement trajectory data such as periodic activity patterns and altered migratory routes, the original

formulation relies on the adaptive Metropolis adjusted Langevin algorithm (MALA) to sample directly from the posterior distribution of latent functions and parameters. In general, this method may suffer from poor mixing and slow convergence of Markov chains as the size of the animal movement observations increases, particularly when inferring high-dimensional parameter spaces that include kernel hyperparameters. This limits their scalability and practical application in large-scale and long-term ecological studies in movement ecology. Thus, to accurately infer population-level migration routes of animals, it is essential to use multiscale models that provide scalable inference, maintain flexibility, and ensure computational efficiency. This is particularly important when modelling animal movement data, which are inherently complex and often exhibit substantial individual-level variation in movement patterns.

In this study, we use a hierarchical sparse Gaussian process (HSGP) to model and identify the mean migration routes of animals at the population level. Sparse Gaussian processes (SGP) are an efficient approximation of standard Gaussian processes (GPs) that uses a subset of observations, known as inducing points, to approximate the true posterior distribution of parameters and hyperparameters (Snelson and Ghahramani, 2005). This reduction enhances computational efficiency without significantly compromising model accuracy as fewer observations are involved in the modelling process. The model parameters, including the locations of the inducing points and the covariance function hyperparameters, are optimised using variational inference, resulting in a tractable approximation to the full GPs. HSGP extends SGP by incorporating a hierarchical structure into the observational framework, enabling the capture of complex dependencies and multiscale patterns within the data while maintaining the computational advantages of sparse approximations. This hierarchical framework facilitates the capture of both the population-level migration route and accounting for individual movement dy-

namics while also quantifying uncertainties at multiple levels. Finally, we demonstrate the utility of the HSGP in identifying the mean migration routes at the population-level of the Serengeti migratory wildebeest (*Connochaetes taurinus*). Building on this, we refine predictions of areas where wildebeest are likely to spend the most time during critical life-history stages such as calving, weaning, rutting, and migration, by integrating the inferred mean migration route, representing long-term memory effects, with fine-scale wildebeest space use patterns estimated using a buildings dataset described in (Masolele et al., 2025).

5.2 Methods

To identify the expected migration route of Serengeti wildebeest, we analysed a large-scale dataset of GPS locations using a hierarchical sparse Gaussian processes (HSGP). Wildebeest movement data was collected in the Serengeti National Park, Tanzania, between 2013 and 2023. In this period, a total of 63 wildebeest were fitted with GPS collars (Followit, formerly 'Televilt,' GSM or Iridium transmitters with GPS location) programmed to record locations at intervals ranging from 6 to 24 hours, resulting in 204,129 location observations.

The annual wildebeest migration covers, and in fact defines, the range of the Greater Serengeti ecosystem. Herds move south in early November away from their dry season refuge in Kenya, towards the southern short grass plains of Tanzania, where they spend the wet season (December to May) in search of high-quality forage. During the dry season (August to November) herds move towards the northern woodlands of the Serengeti National Park and on to the Maasai Mara National Reserve, where they remain until the short rains begin and the cycle begins again (Torney et al., 2018). The migration route is not fixed and can vary from year to year as the herds follow gradients of rainfall and

nutrients (Holdo et al., 2009), as well as being influenced by site fidelity (Morrison et al., 2021) and social information (Berdahl et al., 2018; Torney et al., 2018).

In order to infer a mean migration route, we employ Gaussian process modelling (Rasmussen, 2006) which allows us to decompose the observed data into a periodic component and a stochastic movement component by combining a periodic kernel corresponding to the long-term average herd migration route, and an Ornstein-Uhlenbeck (OU) process kernel corresponding to individual movement patterns. The periodic kernel captures the cyclic nature of the migration route, while the OU process kernel captures the stochastic movement patterns of individual wildebeest. This combination allows us to model the mean migration route while accounting for individual variability in movement patterns and the autocorrelation inherent in GPS location data.

Gaussian processes (GPs) are non-parametric models that can be used to model complex relationships between inputs and outputs by defining a mean value and a covariance structure, with the fundamental assumption that any finite collection of function values follows a multivariate normal distribution. This property allows Gaussian processes to provide not only predictions but also uncertainty estimates, making them particularly suitable for modelling complex spatial and temporal patterns in animal movement data (Torney et al., 2021).

5.2.1 Hierarchical Gaussian process model

As in standard GP regression, we begin by placing a GP prior on a latent function $g(\mathbf{t})$ that models the true location of wildebeest across N time points $\mathbf{t}$, and then assume that observations $y(\mathbf{t})$ are generated from the latent function with additive measurement noise, i.e.

$$g(\mathbf{t}) \sim \mathcal{GP}\left(\mathbf{0}, K(\mathbf{t}, \mathbf{t}')\right).$$

Note, the Gaussian process is defined over a continuous domain however the data is observed at discrete observation points.

where $g(\mathbf{t})$ denotes the collection of latent locations at all observation times $\mathbf{t} = [t_1, \ldots, t_N]^\top$. At each observation time t_i, we then observe a noisy measurement of the true location,

$$y_i = g_i + \epsilon_i,$$

where $y_i = y(t_i)$ and $g_i = g(t_i)$ denote the observed two-dimensional location and the true location, respectively, at time t_i. Both y_i and g_i are elements of $\mathbb{R}^2$ representing two-dimensional spatial coordinates. The measurement noise ϵ_i is assumed to be independent and identically distributed (i.i.d.) two-dimensional isotropic Gaussian white noise,

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2_{\text{obs}} \mathbf{I}_2)$$

where σ^2_{obs} is the observation noise variance. Note, to improve clarity, throughout this work we use non-bold symbols for single two-dimensional vectors such as y_i and g_i, and bold symbols (e.g. $\mathbf{y}$) for collections of such vectors where $\mathbf{y} = y(\mathbf{t})$ is the stacked $2N$-dimensional vector of N observations.

In our framework, we assume that the covariance kernel, $K(\mathbf{t}, \mathbf{t}')$ is composed of two components, a periodic kernel that captures the long-term average migration route, and a stochastic movement kernel that captures the individual movement dynamics. The periodic kernel is defined as an annual periodic function, while the stochastic movement kernel defines an OU process, a mean reverting random walk that has been widely used to model animal movement patterns (Calabrese et al., 2016; Dunn and Gipson, 1977; Torney et al., 2021). For a single individual, the covariance function is therefore defined

as

$$K(\mathbf{t}, \mathbf{t}') = K_m(\mathbf{t}, \mathbf{t}') + K_{OU}(\mathbf{t}, \mathbf{t}') \tag{5.1}$$

where $K_m(\mathbf{t}, \mathbf{t}')$ is the migration kernel and $K_{OU}(\mathbf{t}, \mathbf{t}')$ is the individual movement kernel. The migration kernel is defined as,

$$K_m(\mathbf{t}, \mathbf{t}') = \sigma_m^2 \exp\left(\frac{-2\sin^2(\pi|\mathbf{t} - \mathbf{t}'|/P)}{\ell_m^2}\right) \tag{5.2}$$

where σ_m^2 is the amplitude, P is the period (365 days), and ℓ_m is the length scale (which defines how quickly the migratory wildebeest location can change over time within each migration cycle) controls the smoothness of the modelled migration route. The movement kernel is defined as the Matern 1/2 kernel,

$$K_{OU}(\mathbf{t}, \mathbf{t}') = \sigma_{OU}^2 \exp\left(-\frac{|\mathbf{t} - \mathbf{t}'|}{\ell_{OU}}\right) \tag{5.3}$$

where σ_{OU}^2 is the amplitude and ℓ_{OU} is the length scale, which is the covariance function associated with the OU process (Torney et al., 2021).

Since the average migration location is a population level process, we assume that the migration kernel is shared across all individuals whereas the OU process kernel on the other hand, models individual movement around this mean. This hierarchical structure allows us to capture both the population-level migration route while also accounting for individual-level variability in movement. For multiple individuals, the pair wise covariance function between any two observations is defined as

$$\text{Cov}\left(y_i^p, y_j^q\right) = K_m(t_i, t_j) + \delta_{\text{ind}} K_{OU}(t_i, t_j) \tag{5.4}$$

where y_i^p indicates the location observation is from individual p at time t_i, and δ_{ind} is

an indicator function that is 1 if individual p and individual q are the same individual ($p = q$) and 0 otherwise.

Primarily, we are interested in learning about the mean migration route of the wildebeest population and not the true location of a single individual. To that end, we may reformulate the model in a hierarchical structure as follows,

$$\begin{aligned} \mathbf{y} &\sim \mathcal{N}\left(\mathbf{g}, \sigma^2_{\text{obs}}\mathbf{I}\right), \\ \mathbf{g} &\sim \mathcal{GP}\left(\mathbf{f}, \delta_{\text{ind}} K_{OU}(\mathbf{t}, \mathbf{t}')\right), \\ \mathbf{f} &\sim \mathcal{GP}\left(\mathbf{0}, K_m(\mathbf{t}, \mathbf{t}')\right). \end{aligned} \tag{5.5}$$

where we have introduced a separate GP for $\mathbf{f}$ that denotes the mean migration route at all time points in $\mathbf{t}$, $\mathbf{g}$ remains the latent true location of wildebeest, now conditional on $\mathbf{f}$, and $\mathbf{I}$ denotes the $2N \times 2N$ identity matrix, corresponding to the vectorised form of the N two-dimensional observations. Given the model defined in eqn. 5.5, we focus on obtaining the posterior distribution of the mean migration location given the telemetry observations, $p(\mathbf{f}|\mathbf{y})$. Inference for GP models involves optimising the parameters of the covariance kernels, followed by analytically calculating the posterior distributions by exploiting the conjugate structure of the model. In small data scenarios, the former may be solved by directly maximising the log marginal likelihood,

$$\log p(\mathbf{y}) = \log \iint p(\mathbf{y} \mid \mathbf{g})\, p(\mathbf{g} \mid \mathbf{f})\, p(\mathbf{f})\, d\mathbf{g}\, d\mathbf{f}. \tag{5.6}$$

However, for large-scale datasets this is impractical since, although tractable, eqn. 5.6 involves inverting the full covariance matrix, an operation that scales cubically with dataset size.

5.2.2 Sparse variational inference for Gaussian processes

To enable inference for large telemetry datasets, such as the wildebeest migration data, we employ a sparse approximation (Titsias, 2009) to full GP regression by introducing a set of inducing points $\mathbf{z}$ and defining $\mathbf{f}_z$ as the unknown function values at these points. The key approximation introduced in this approach is that the latent function $\mathbf{f}^*$ at any test inputs and the latent function at the training locations $\mathbf{f}$ are conditionally independent given $\mathbf{f}_z$,

$$p(\mathbf{f}^* \mid \mathbf{f}, \mathbf{f}_z) = p(\mathbf{f}^* \mid \mathbf{f}_z). \tag{5.7}$$

This allows us to introduce a variational distribution $q(\mathbf{f}, \mathbf{f}_z)$ that approximates the posterior distribution of the latent function values at the training locations and the inducing points. The variational distribution is defined as

$$q(\mathbf{f}, \mathbf{f}_z) = p(\mathbf{f} \mid \mathbf{f}_z)\phi(\mathbf{f}_z), \tag{5.8}$$

where $\phi(\mathbf{f}_z)$ is a Gaussian distribution with mean $\boldsymbol{\mu}_z$ and covariance $\boldsymbol{\Sigma}_z$. To determine the optimal values of the variational parameters $\boldsymbol{\mu}_z$ and $\boldsymbol{\Sigma}_z$, we maximise a lower bound on the marginal log-likelihood that is obtained via Jensen's inequality (Saul et al., 2016)

$$\begin{aligned}
\log p(\mathbf{y}) &= \log \iint p(\mathbf{y} \mid \mathbf{f})\, p(\mathbf{f}, \mathbf{f}_z)\, d\mathbf{f}\, d\mathbf{f}_z \\
&= \log \iint p(\mathbf{y} \mid \mathbf{f})\, p(\mathbf{f}, \mathbf{f}_z) \frac{q(\mathbf{f}, \mathbf{f}_z)}{q(\mathbf{f}, \mathbf{f}_z)}\, d\mathbf{f}\, d\mathbf{f}_z \\
&\geq \iint q(\mathbf{f}, \mathbf{f}_z) \log \left(\frac{p(\mathbf{y} \mid \mathbf{f})\, p(\mathbf{f}, \mathbf{f}_z)}{q(\mathbf{f}, \mathbf{f}_z)} \right) d\mathbf{f} d\mathbf{f}_z \\
&\geq \int \log \left(p(\mathbf{y} \mid \mathbf{f}) \right) q(\mathbf{f}) d\mathbf{f} - \mathrm{KL}\left(\phi(\mathbf{f}_z) \| p(\mathbf{f}_z) \right)
\end{aligned} \tag{5.9}$$

where $q(\mathbf{f}) = \int q(\mathbf{f}, \mathbf{f}_z) d\mathbf{f}_z$, KL is the Kullback-Leibler divergence, and

$$p(\mathbf{y} \mid \mathbf{f}) = \int p(\mathbf{y} \mid \mathbf{g}) p(\mathbf{g} \mid \mathbf{f}) d\mathbf{g}. \tag{5.10}$$

The final term of eqn. 5.9 is known as the evidence lower bound (or ELBO) and maximising this bound results in optimal variational parameters i.e. parameters that minimise the KL divergence between the approximate posterior and the true posterior. In practice, the ELBO also depends on the hyperparameters associated with the covariance kernels and the observation process, therefore, we may jointly optimise the variational parameters and the model hyperparameters by maximising the ELBO with respect to all of them.

The inducing point approach proposed in (Titsias, 2009) was later extended to enable stochastic variational inference (SVI) for GP regression (Hensman et al., 2013). This approach allows the model to be trained on arbitrarily large datasets by optimising the ELBO with respect to both the variational parameters and the kernel hyperparameters using mini-batch stochastic gradients. The key advance presented in Hensman et al. (2013) was to maintain an explicit representation of the variational distribution which acts as a set of global parameters, enabling efficient and scalable variational inference. We follow the approach of (Hensman et al., 2013) to perform variational inference for our hierarchical sparse Gaussian process model; however, we modify the approach to account for the dependence between observations from the same individual.

5.2.3 Evidence lower bound

We next derive the ELBO for our hierarchical sparse Gaussian process model given that the observations depend on the latent true location $\mathbf{g}$ which in turn depends on the mean migration route $\mathbf{f}$ and show that the lower bound can be analytically computed

therefore avoiding the need for Monte Carlo sampling (Ranganath et al., 2014) to estimate the expectation of the log likelihood term in eqn. 5.9. Beginning with the ELBO from eqn. 5.9,

$$\text{ELBO} = \int \log\left(p(\mathbf{y}|\mathbf{f})\right) q(\mathbf{f}) d\mathbf{f} - \text{KL}(\phi(\mathbf{f}_z)||p(\mathbf{f}_z)). \tag{5.11}$$

The second term of this equation is the KL divergence between the variational distribution of the inducing points and the prior distribution of the inducing points which is a Gaussian distribution and can be calculated in closed form. The first term involves the expected value of $\log p(\mathbf{y}|\mathbf{f})$ under the variational distribution. Given the model defined in eqn. 5.5,

$$p(\mathbf{y}|\mathbf{f}) = (2\pi)^{-N/2}|\mathbf{\Sigma}|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{f})^T\mathbf{\Sigma}^{-1}(\mathbf{y}-\mathbf{f})\right) \tag{5.12}$$

where N is the number of observations and we have introduced the notation $\mathbf{\Sigma} = \delta_{\text{ind}}K_{OU}(\mathbf{t},\mathbf{t}') + \sigma^2_{\text{obs}}\mathbf{I}$ for clarity. Since the observations of a single individual are conditionally independent given the latent function values, we can write this probability as a product of the probabilities for each individual, i.e.

$$p(\mathbf{y}|\mathbf{f}) = \prod_{j=1}^{M}(2\pi)^{-N_j/2}|\mathbf{\Sigma}_j|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{y}_j-\mathbf{f}_j)^T\mathbf{\Sigma}_j^{-1}(\mathbf{y}_j-\mathbf{f}_j)\right) \tag{5.13}$$

where M is the number of individuals, $\mathbf{y}_j$ is the vector of observations for individual j, $\mathbf{f}_j$ is the vector of latent function values for individual j, and $\mathbf{\Sigma}_j = K_{OU}(\mathbf{t}_j,\mathbf{t}_j') + \sigma^2_{\text{obs}}\mathbf{I}$ is the covariance matrix for individual j. Taking the logarithm of this probability gives us the log likelihood of the observations,

$$\log(p(\mathbf{y}|\mathbf{f})) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{M}\left(\log|\mathbf{\Sigma}_j| + (\mathbf{y}_j-\mathbf{f}_j)^T\mathbf{\Sigma}_j^{-1}(\mathbf{y}_j-\mathbf{f}_j)\right). \tag{5.14}$$

We may now substitute this expression into the first term of the ELBO in eqn. 5.11 to obtain

$$\int q(\mathbf{f}) \log\left(p(\mathbf{y}|\mathbf{f})\right) d\mathbf{f} = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{M}\log|\boldsymbol{\Sigma}_j| \\ - \frac{1}{2}\sum_{j=1}^{M}\left(\int q(\mathbf{f}_j)(\mathbf{y}_j - \mathbf{f}_j)^T\boldsymbol{\Sigma}_j^{-1}(\mathbf{y}_j - \mathbf{f}_j)d\mathbf{f}_j\right). \tag{5.15}$$

Since $q(\mathbf{f}_j)$ is a Gaussian distribution, we can write the remaining integral as

$$\int q(\mathbf{f}_j)(\mathbf{y}_j - \mathbf{f}_j)^T\boldsymbol{\Sigma}_j^{-1}(\mathbf{y}_j - \mathbf{f}_j)d\mathbf{f}_j \\ = \int \mathcal{N}(\mathbf{f}_j|\boldsymbol{\mu}_j, \boldsymbol{C}_j)(\mathbf{y}_j - \mathbf{f}_j)^T\boldsymbol{\Sigma}_j^{-1}(\mathbf{y}_j - \mathbf{f}_j)d\mathbf{f}_j. \tag{5.16}$$

where $\boldsymbol{\mu}_j$ and $\boldsymbol{C}_j$ are the mean and covariance of the probability distribution $q(\mathbf{f}_j)$ which can be computed analytically by marginalising eqn. 5.8 over the inducing points $\mathbf{z}$. By introducing a substitution for $\mathbf{f}_j - \mathbf{y}_j$ we can view this integral as the expectation of a quadratic form, which has a known solution (Kendrick, 1981) and is given by

$$(\mathbf{y}_j - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\mathbf{y}_j - \boldsymbol{\mu}_j) + \mathrm{Tr}(\boldsymbol{\Sigma}_j^{-1}\boldsymbol{C}_j). \tag{5.17}$$

All combined, the expected log likelihood term in the ELBO can be written as

$$-\frac{1}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{M}\left(\log|\boldsymbol{\Sigma}_j| + (\mathbf{y}_j - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\mathbf{y}_j - \boldsymbol{\mu}_j) + \mathrm{Tr}(\boldsymbol{\Sigma}_j^{-1}\boldsymbol{C}_j)\right), \tag{5.18}$$

which in combination with the KL divergence term in eqn. 5.11 gives us a tractable expression for the ELBO that can be optimised with respect to the variational parameters and the kernel hyperparameters. It should be noted that it is only possible to compute the expected log likelihood term in eqn. 5.11 because the lower levels of the hierarchical

model only affect the mean location of the observations and not the parameters of the movement kernel. This is a key difference between our model and the original hierarchical GP model proposed in (Torney et al., 2021) which does not allow for analytical computation of the expected log likelihood term.

5.2.4 Numerical implementation

To maximise the expected log likelihood and infer the parameters of the migration and movement kernel, we employed variational inference using stochastic optimisation. Because of large wildebeest movement observations, we used a trajectory segmentation technique, where we divided the individual animal trajectories into multiple and computationally manageable segments, so that each segment has 1024 movement observations. Additionally, we used 300 inducing points ($\mathbf{z}$) that were evenly spaced across the year (assuming 1 year =365 days) and provided an optimal balance between accuracy and reducing computational costs. All the analysis were performed in Python programming language using TensorFlow (Abadi et al., 2016) and TensorFlow Probability (Dillon et al., 2017).

5.2.5 Integrating local environmental features

Understanding the mechanisms underlying animal space use during specific days of the year or key life history events often requires more than simple consideration of broad scale movement patterns. In many cases animals respond to finer scale environmental features such as the presence of buildings or watering holes, and to dynamic features such as vegetation growth.

Similarly, fine scale features alone do not adequately explain animal movement patterns or habitat selection. To improve predictions of where animals are likely to spend most of

their time during critical life-history stages such as calving, weaning, rutting, or migration, it is essential to incorporate the role of long-term spatial memory. Animals often retain information about past experiences and may preferentially return to familiar locations, even when those areas do not offer the highest immediate resource quality. In order to better estimate the spatial distribution of Serengeti wildebeest we employ the inferred population mean migration routes as a proxy for the effects of long-term memory on fine scale movement decisions. Firstly, we note that for any time t we may calculate a probability density function for the location of a wildebeest at that time of the year $p_t(x)$. The density function is a multivariate normal with mean and covariance calculated from eqn. 5.8 combined with the stationary covariance of the OU process (Gardiner, 2009). Due to the periodicity of the mean migration route we also note that $p_t(x) = p_{t+365}(x)$.

In the absence of any environmental responses beyond what is encoded in the mean migration location, the probability density function may be reformulated within a step selection model (Thurfjell et al., 2014) framework by assuming the covariate field on which movement steps are based is given as $\log p_t(x)$, the selection coefficient is 1, and the movement model follows the local Gibbs sampler proposed in (Michelot et al., 2019a) which ensures that the step selection rules map directly to the parameters of the long term utilisation distibution. Specifically, the selection function is defined as

$$w(x,t) = \exp\left(\log(p_t(x)) + \sum_i \beta_i c_i(x,t)\right) \tag{5.19}$$

where β_i and c_i are the selection coefficients and covariate values respectively. As described in (Michelot et al., 2019a), if movement steps are assumed to follow the local Gibbs algorithm, the utilisation distribution is given by,

$$\pi(x,t) = \frac{1}{Z}\exp\left(\log(p_t(x)) + \sum_i \beta_i c_i(x,t)\right) \tag{5.20}$$

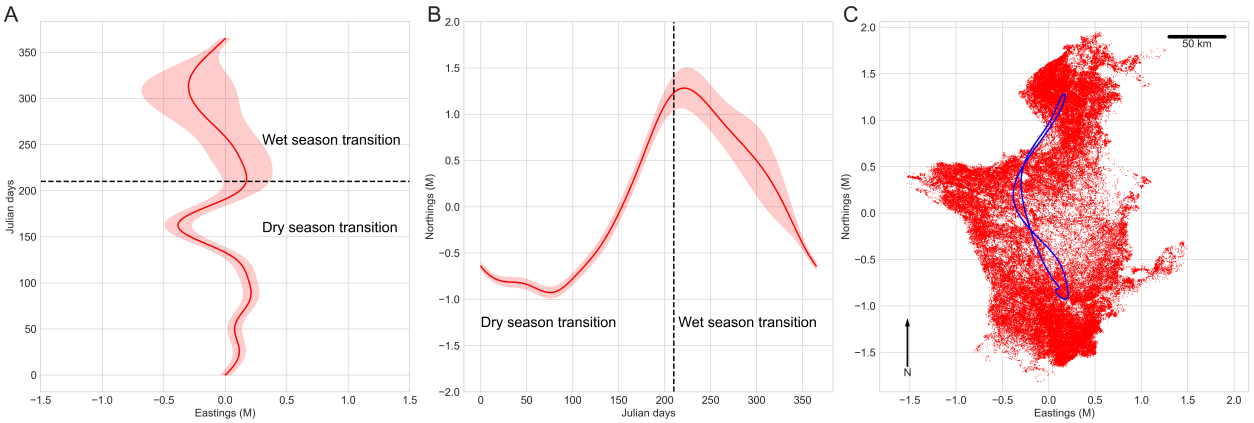

Figure 5.1: Wildebeest migration route in Serengeti, A) Inferred population-level posterior mean migration route in easting and shaded red regions represent 95% credible interval, B) Inferred population-level posterior mean migration route in northing and and shaded red regions represent 95% credible interval, C) Wildebeest locations (red points) and inferred population-level posterior mean migration route (blue line).

where Z is a normalising constant. In the absence of any covariate fields (or if $\beta_i = 0$ for all i) we recover the long term predicted space use $\pi(x,t) = p_t(x)$.

To demonstrate how predictions of wildebeest space use may be iteratively refined, we analyse movement decisions of wildebeest in the context of human made structures within the national park, using a buildings dataset described in (Masolele et al., 2025).

5.3 Results

The results reveal that at the onset of the dry season, wildebeest migrate through the western Serengeti toward the northern regions in search for forage and water (Fig. 5.1A). During the short rainy season, typically around November, they begin moving southward toward the nutrient rich short grass plains of the Serengeti, although their migratory routes show increased variability during this period (Fig. 5.1B).

Furthermore, wildebeest have lower uncertainty in their memory-informed space use during the calving period, as shown with the contour plot in Fig. 5.2A, but the un-

certainty increases as they transition to another life event such as rutting (Fig. 5.2B), weaning (Fig. 5.2C), and becomes much higher during their migration routes back to the southern part of the ecosystem from the North (Fig. 5.2D).

5.4 Discussion

Gaussian processes have long been a mainstream machine learning technique for analysing time series data in spatial statistics (Williams and Rasmussen, 2006). However, their application to modelling animal movement data has emerged only recently (e.g., Cobb et al., 2018; Torney et al., 2021). In this study, we have demonstrated how the HSGP framework can effectively model movement observations of migratory animals to infer the mean population-level migration route. To account for uncertainty in the dynamics of migration routes, we introduced the use of variational inference, a fast and computationally efficient approximate Bayesian inference that has also been employed in modelling animal observations to capture spatially varying movement characteristics (Paun et al., 2022), and to formally quantify uncertainty in multiscale step selection models (Masolele et al., 2024). Our results suggest that the HSGP approach could also be useful for identifying shifts or changes in the migratory route and activity periods of animals using GPS data collected from multiple individuals while reducing computational complexity.

Models of animal migration routes often depict them as circuit-like flows between habitat patches, using circuit theory to represent connectivity (McRae et al., 2008). While these models are useful for understanding broad-scale connectivity, they tend to oversimplify migration routes as closed loops between seasonal ranges (e.g., Sawyer et al., 2009). This simplification limits the ability to capture fine-scale spatiotemporal movement patterns of animals. This study demonstrates that the wildebeest migration route is not a circuit, as it is typically presented, but rather a to-and-fro migration with greater variability

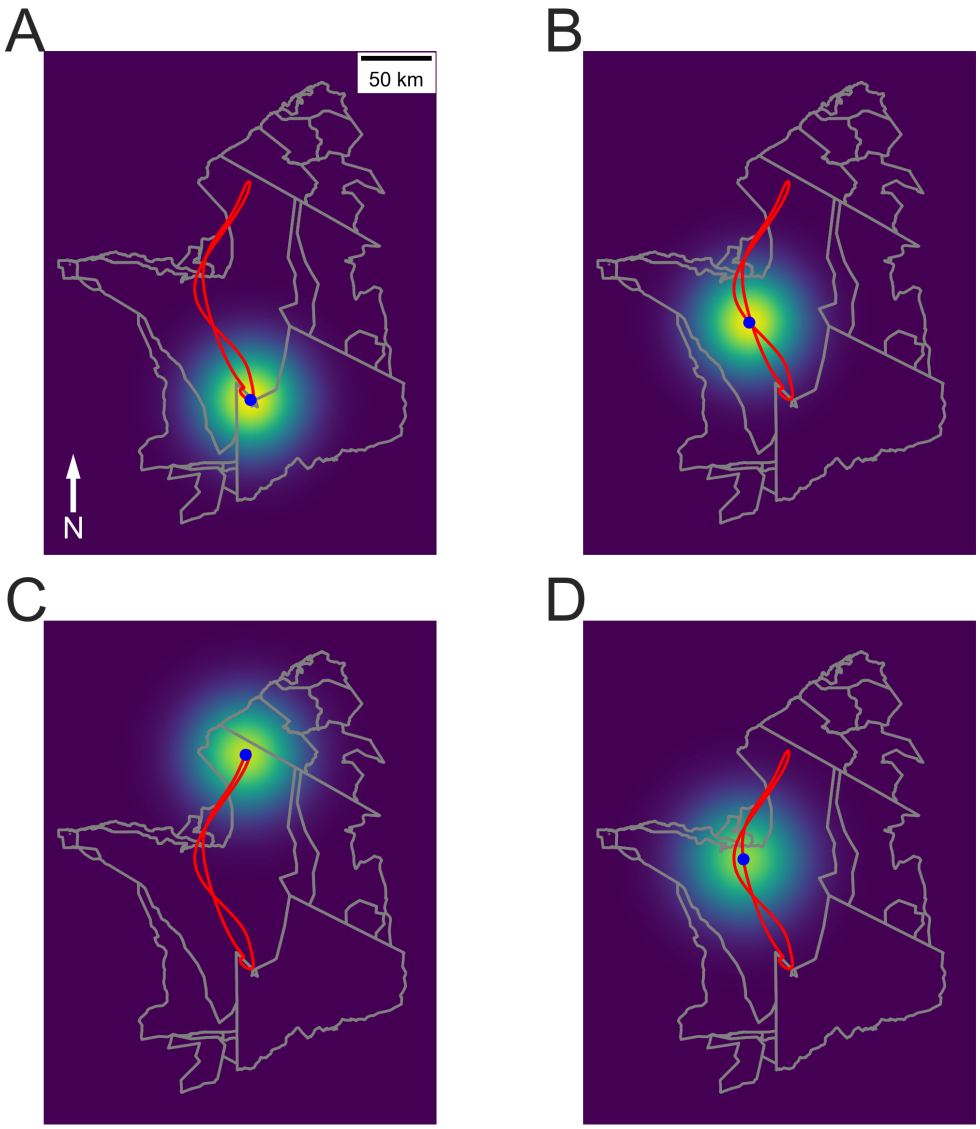

Figure 5.2: Inference of migratory wildebeest memory-informed space use in some of the key life events during the course of their annual migration, A) Calving, B) Rutting, C) Weaning, D) Migrating to the southern part of the ecosystem. The blue dot represent the inferred posterior mean population-level migration route in that location. The background regions (from dark purple (low values) to bright yellow (high values)) represent wildebeest long term space use in that particular event. The red line represent the inferred population-level posterior mean migration route and Gray line represent the boundary of the Greater Serengeti-Mara ecosystem.

in the southward movements. This shift in perspective advances our understanding of wildebeest movement behaviour, spatial variability, and deviations due to environmental or anthropogenic factors, and provides a more accurate representation of the Serengeti migration. The inferred mean population-level migration route indicates that wildebeest transit through the western Serengeti at the onset of the dry season (Fig. 5.1A). During the short rainy season, they begin moving southward toward the nutrient-rich short grass plains of the Serengeti, although their migration routes exhibit greater variability during this period (Fig. 5.1B). These patterns align with seasonal changes in resource distribution and landscape heterogeneity in the ecosystem. During the onset of the dry season, the western Serengeti typically begins to experience a decrease in water availability and a decline in forage quality due to reduced rainfall. In such conditions, wildebeest are likely to minimise their residence time in this region to avoid nutritional stress and dehydration. The observed route through the western Serengeti may represent opportunistic foraging periods and salt licking for essential minerals such as sodium (Buchanan, 2020) as herds move towards the more reliable water and forage resources of the northern Serengeti.

In contrast, the short rainy season triggers localised forage growth and temporary water availability (Boone et al., 2006), encouraging extensive search and capitalisation of newly available resources. However, due to the spatial unpredictability of rainfall during this period, migration routes become more variable as herds adjust their movement in response to dynamic, patchy resources, leading to increased uncertainty. The observed variation in the duration of southward movement may also be influenced by social cues and collective decision making, especially under uncertain conditions. Wildebeest migrations are known to be partially guided by collective decision making and social cues (Torney et al., 2018), and such behaviours may become more prominent when the

distribution of the resource is unpredictable. The increased variability during short rains may thus reflect both environmental heterogeneity and socially mediated exploration.

The lower uncertainty in the wildebeest memory-informed space use during the calving period (Fig. 5.2A) suggests that migratory wildebeest employ a mobile aggregated distribution pattern at this time of the year by forming large aggregations that move in the same direction over a large area based on where they are able to locate and exploit short, highly rich in protein and nutritious forages. Furthermore, the level of uncertainty increases as they move to the rutting period (Fig. 5.2B) because of the changing environmental conditions such as localised high-quality forages and wildebeest apt to disperse and form clusters (groups) consisting of small herds and large herds.

Through the use of HSGP we have shown that it is possible to combine both migration kernel, and a movement kernel to encode into the model periodic structures and the stochastic nature of animal movement behaviour, respectively, thereby introducing multiscale processes into the model. This method further allows learning of the arbitrary model structures to infer the mean route of the animal migration without limiting the model to a particular functional form. Additionally, our non-parametric approach can be extended to incorporate fine-scale movement decisions and categorical variables such as whether an animal is inside or outside a protected area when the goal of the inference is to assess spatial use patterns and potential boundary-crossing behaviour.

The inferred population-level posterior mean migration route in Fig. 5.1C indicates that, at some point in the north-west of the ecosystem, the route passes through areas with relatively few wildebeest observations. This arises primarily due to increased uncertainty in data-sparse regions and the use of stationary GPs, which assume constant smoothness and correlation structure across space. Such assumptions limit the model's ability to adapt to heterogeneous observation densities. One possible way to address this limi-

tation is to adopt a non-stationary GPs framework, where the covariance structure varies spatially. This would allow the model to be more flexible and responsive in areas with dense data, while remaining more conservative in regions with limited observational support.

However, a key limitation of our model lies in the assumptions made to enable efficient inference. Specifically, VI in this framework is tractable primarily because it is the mean function of the higher-level GPs that is allowed to vary while the kernel covariance parameters remain fixed or stationary. This simplification allows closed-form expectations during optimisation, making VI computationally efficient (Titsias, 2009). If instead the model employed a non-stationary kernel, where the covariance structure changes over input space, then the expectation term in the variational objective would no longer have an analytic solution. In such cases, computationally more demanding methods such as the Metropolis-Adjusted Langevin Algorithm (MALA) or variational inference augmented with Monte Carlo approximations of the expectation would be employed (Heinonen et al., 2016), both of which increase inference complexity and runtime.

Herein, we have described hierarchical sparse Gaussian processes (HSGP) that offers computationally flexible scalable inference and is able to formally quantify uncertainty around the inferred population mean migration routes of animals from telemetry data in an efficient manner. Our results highlight the importance of using HSGP in animal movement modelling especially when inferring animal migration routes or periodic patterns that are multiscale processes, include multiple individuals, and multiple covariance kernels either through multiplication or addition to capture complex patterns such as annual or diurnal patterns exhibited by animal populations. Furthermore, the seasonal differences observed in the uncertainty around the migratory route of wilde-

beest can be attributed to the interplay of rainfall-driven resource dynamics, the costs of movement, and the use of social information. These findings underscore the sensitivity of wildebeest migration in a highly variable environment and highlight the importance of conserving heterogeneous landscapes that support this iconic migratory system.

Chapter 6

Predicted impact of anthropogenic structures on the Serengeti migratory wildebeest population

Note:

This chapter has been prepared as a manuscript which I aim to submit to the journal of Ecological Applications.

Abstract

Given the rapid increase of anthropogenic structures in previously pristine environments, understanding how migratory animals navigate these altered landscapes and where they predominantly spend their time is paramount to devising evidence-based conservation interventions. In this study, we employ a Bayesian multiscale step selection model developed for migratory wildebeest, together with an anthropogenic-structure simulation model, to investigate how different conservation spatial planning strategies for allocating new buildings both individually and in combination may influence the long-term spatial use of migratory wildebeest in the Greater Serengeti–Mara ecosystem. The simulation model assumes that new buildings are added either in proximity to existing structures, following a preferential attachment mechanism, or randomly in previously undeveloped areas. Allocation near existing structures is governed by a preferential attachment exponent, which determines the strength of the tendency for new buildings to cluster around existing development. Our simulation results indicate that the impact on wildebeest space use is greater when new buildings are added to previously undeveloped areas or away from existing infrastructure. Furthermore, even a modest increase in infrastructure, such as a 10% addition to the existing buildings, results in a measurable change in wildebeest space use and access to key grazing habitats, likely because factors such as human presence, road networks, vehicle traffic, noise, and light pollution may repel animals from these areas. Given the critical role of wildebeest in maintaining the Serengeti ecosystem's structure, function, and key processes such as nutrient cycling and storage, we recommend restricting or ceasing new construction in the core regions of the ecosystem where wildebeest concentrate on key resources, in order to preserve their long-term population viability. Furthermore, our results highlight the broader need to assess the long-term impacts of human activity on the habitat use of

migratory species and their potential to displace animals from key resources.

6.1 Introduction

Migratory animals are currently facing an existential crisis due to a combination of multifaceted anthropogenic activities including climate change. Alterations in habitat and environmental conditions are resulting in shifts in spatial distribution, migration timing, and animal abundance. As a consequence, many species are experiencing population declines or risk of extinction (Kauffman et al., 2021), primarily due to their inability to adapt to or cope with the rapid changes that occur in their ecosystems. To effectively protect these species, it is crucial to understand and mitigate the impacts of human-induced environmental change. However, to predict future distribution patterns, such as where animals are likely to move or spend most of the time, it is essential to investigate how they respond to local environmental change and identify critical threats in high-risk areas using robust computational models grounded on current conditions.

Recent advances in technology have facilitated the development of fast, powerful, and more flexible mechanistic models in the field of computational ecology (de Koning et al., 2023; Thiele and Grimm, 2015). These models are increasingly becoming indispensable tools in ecological modelling, allowing applied ecologists, conservationists, and policy makers to continuously reassess and refine their conservation strategies and ecological management approaches (Schuwirth et al., 2019). However, the complex ecological interactions inherent in many ecosystems, coupled with rapid global changes in protected areas driven by anthropogenic activities (Riva et al., 2023), highlight the need for dynamic computational models that can combine data, domain knowledge and continuous alignment with real-world ecological systems (de Koning et al., 2023) to create a more mechanistic understanding. These models must maintain consistency over time to support long-term ecological assessments, which is essential for producing reliable

predictions that guide evidence-based management and policy interventions.

Ecosystems, by their nature, are characterised by complex and dynamic ecological interactions (Mouquet et al., 2015) and are increasingly exposed to a range of internal and external stressors, including human pressures, climate change, habitat loss and degradation, among others (Riva et al., 2023). As these stressors intensify, there is a growing need to anticipate potential future outcomes to develop effective management and conservation strategies that can mitigate risks or prevent further damage to these natural systems. A widely used approach in ecology for this purpose is scenario modelling, which enables ecologists to examine the potential impacts of environmental change on species distributions, the spread of invasive species, and the outcomes of various management actions (Bennett et al., 2003; Cumming, 2007). Scenario analysis facilitates the exploration of different hypothetical quantitative estimates of future population dynamics or distribution patterns of animals and plants by considering various environmental covariates, ecological assumptions about how the system works, management plans or strategies, and sources of uncertainties. This approach enables simulation of potential outcomes grounded on current ecological patterns, principles, and assumptions that reflect prevailing conditions. By considering different states, such as baseline, current and projected future conditions, it is possible to assess changes in spatial distribution and identify context-specific management interventions for each biodiversity scenario that may reduce risk or enhance resilience. The ability to simulate and compare such ecological scenarios offers valuable insight to wildlife managers and policymakers, enabling informed evidence-based decision making in the face of ecological uncertainty.

In ecology, simulation models have been used to understand and provide probabilistic predictions of ecological outcomes under varying scenarios. These models, informed by initial conditions and parameters, help explore ecosystem behaviour, population dynam-

ics (Colomer and Margalida, 2025), species distributions over time (Willis et al., 2009), the risk of species extinction (Mashayekhi et al., 2014; Ovaskainen and Meerson, 2010; Schleuning et al., 2016) and predicting the impact of human-made infrastructure, such as roads, on animal migration (Holdo et al., 2011). Additionally, they simulate species behaviour at multiple scales, such as foraging patterns, plant competition, or community dynamics, and offer insight into long-term trends in ecosystem processes, including human-environment interactions (Railsback and Grimm, 2019). This approach provides valuable insights on the inherent uncertainty of ecological systems and aids in predicting future outcomes based on the implementation of specific management actions.

In many large wildlife-rich ecosystems around the world, expanding human settlement, tourism, and commercial development have increasingly fragmented wildlife habitats, limiting access to key foraging areas and disrupting traditional migratory routes (Bolger et al., 2008; Harris et al., 2009; Kauffman et al., 2021; Liu et al., 2024). Although tourism provides essential revenue for conservation efforts and supports local livelihoods (Larsen et al., 2020), it can also degrade natural habitats when not managed properly. This dual role highlights the need to balance the economic and conservation benefits of tourism with its potential ecological costs. To prevent conservation interventions from inadvertently compromising the ecosystems they aim to safeguard, it is critical to identify the areas most affected by anthropogenic disturbances associated with tourism infrastructure, including lodges, campsites, and hotels. Quantifying animal space use is fundamental to inform sustainable tourism strategies that support both biodiversity conservation and ecosystem resilience. These analyses facilitate the identification of high- and low-use wildlife areas, thereby guiding the development of targeted, cost-effective mitigation measures that maintain ecological connectivity and minimise the risks associated with human pressures or barriers that impede animal movement. Crucially, effective

mitigation requires movement models that accurately reflect current animal behaviour, supported by high-resolution data collected at fine spatial and temporal scales. Such models are essential for anticipating the ecological consequences of planned and unplanned development and ensuring continued wildlife access to key resources in the ecosystem.

In the Serengeti ecosystem, wildebeest have been observed to reduce the time they spend in areas near human-made structures such as buildings (Masolele et al., 2025). However, as the landscape continues to be modified by the addition of new anthropogenic structures, it becomes increasingly important to predict future changes in wildebeest space use based on observed behavioural responses. In this study, we evaluate the impact of anthropogenic disturbances on the spatial distribution of migratory wildebeest in the Serengeti ecosystem, which is undergoing rapid environmental changes due to the development of human-made structures such as buildings. Specifically, we investigate how the long-term space use of migratory wildebeest may be altered as a result of the addition of new buildings in the ecosystem. We use an existing multiscale step selection model developed for migratory wildebeest (Masolele et al., 2025), together with a simulation model that simulates the addition of new buildings in areas with existing infrastructure and in previously undeveloped regions of the landscape, considering three scenarios of 10%, 50%, and 100% increase of existing buildings in the landscape. Finally, we evaluate the shift in wildebeest space use in future scenarios relative to their current patterns. This approach demonstrates how animal movement models can inform infrastructure development in anthropogenically altered landscapes and provide valuable predictions that can help mitigate or reduce associated risks.

6.2 Methods

6.2.1 Study area description

The Greater Serengeti–Mara Ecosystem is a transboundary conservation landscape located in East Africa, spanning the border between Tanzania and Kenya (approximately $33°30'–35°30'E$ and $1°15'–3°30'S$), and covering an area of about 37,516 km^2. At its core is the Serengeti National Park, which is surrounded by a mosaic of other protected areas, including the Ngorongoro Conservation Area, Masai Mara National Reserve, and the Maswa, Grumeti, Ikorongo, Kijereshi, and Pololeti Game Reserves as shown in Fig. 6.1A. In addition, the ecosystem includes a network of community-managed wildlife conservancies in Kenya and wildlife management areas in Tanzania. Each of these protected areas operates under distinct management objectives tailored to meet their conservation priorities.

We obtained version 2 of the location data (Fig. 6.1B) of existing buildings from the large-scale open data set that contains the outlines of buildings derived from high-resolution satellite imagery accessible at open buildings managed by (Sirko et al., 2021). The wildebeest movement dataset consisted of 143,268 GPS locations from 57 collared migratory wildebeest collected in the Serengeti ecosystem, between January 2019 and September 2023 (Fig. 6.1A).

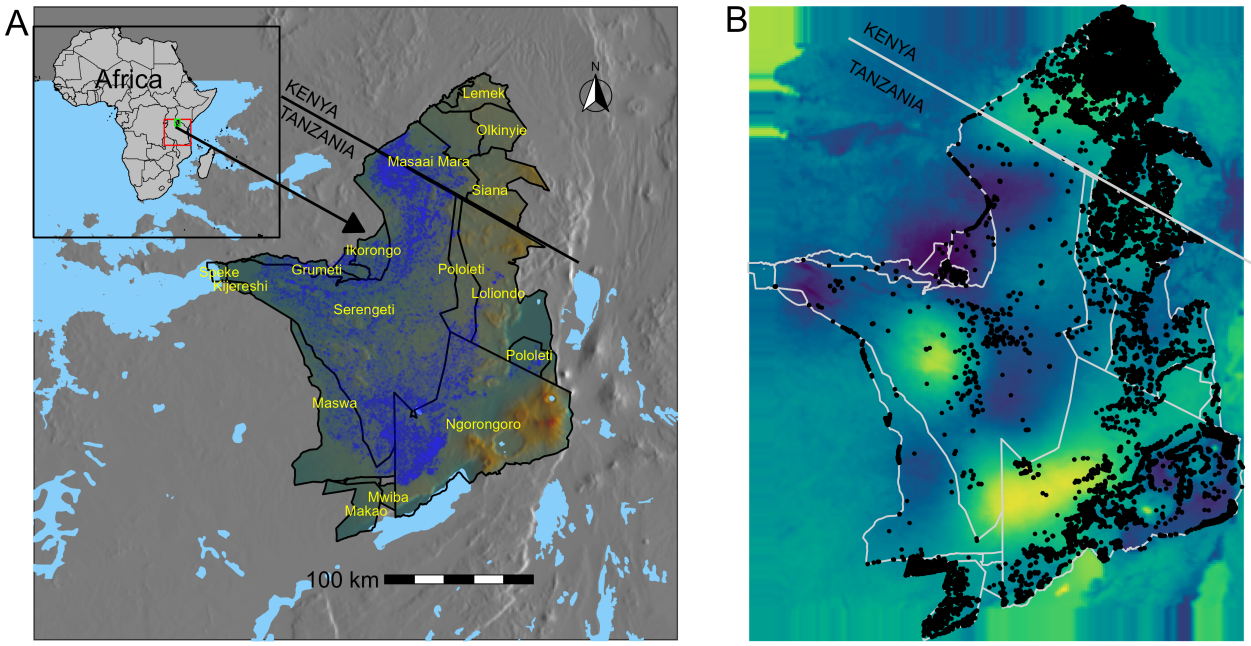

Figure 6.1: A) A map of the Serengeti ecosystem with wildebeest GPS locations shown as blue points, black line indicates the boundary of the Serengeti ecosystem and associated protected areas, B) Building distribution within the Serengeti ecosystem shown as black points, grey lines indicates the boundary of the Serengeti ecosystem and associated protected areas. The background regions from dark purple (low values) to bright yellow (high values) represent grass nitrogen concentration in the ecosystem.

6.2.2 Model scenarios

Our modelling objective is to investigate how the long-term spatial use patterns of migratory wildebeest may be affected by the addition of new buildings in the ecosystem. To achieve this, we simulate the addition of new buildings in both areas with existing infrastructure and previously undeveloped regions. We examine three development scenarios, representing the increase in 10%, 50%, and 100% of the number of existing buildings in the landscape.

The current observed spatial distribution of human-made structures within the ecosystem reveals a high-level of clustering in the locations of buildings. This indicates that new buildings are often constructed near existing infrastructure rather than located in previously undeveloped areas. Clustered patterns suggest that the selection of the building site is influenced by logistical considerations, such as proximity to road networks or favourable wildlife viewing locations such as proximity to water provisioning points. In contrast, isolated buildings imply that some developments may prioritise remoteness and access to relatively undisturbed environments. To capture both dynamics, we implement a hybrid spatial anthropogenic structure simulation model: with probability δ, locations are selected via random placement to simulate preferences for isolation or previously undeveloped regions; with probability $1-\delta$, new structures follow a non-linear preferential attachment rule (Barabási and Albert, 1999; Kunegis et al., 2013), favouring locations near existing buildings (described below). Specifically, a new building is allocated using preferential attachment if the proposal is less than a threshold $(1-\delta)$, where $\delta \in [0,1]$; otherwise, the new building is allocated randomly. This approach allows new buildings to be allocated to previously undeveloped regions and results in a decrease in clustering as the value of δ increases from 0 to 1. Additionally, when δ is between 0 and 1, the method implements a hybrid allocation strategy that combines the random placement of new buildings with the clustering, allowing a continuous and automated transition between these two spatial distribution patterns of human-made structures within the ecosystem. Furthermore, given that the Serengeti ecosystem comprises multiple protected areas, each with different management regime and objectives and varying numbers of existing buildings, the allocation of new buildings in all three scenarios is proportional to the current number of structures within each protected area.

To define the attachment rule, we assign each existing building a probability of receiv-

ing a new neighbouring building. This probability is proportional to the number of buildings already located within a 6 kilometres radius, a range within which habitat is significantly influenced by anthropogenic disturbance (Masolele et al., 2025). To account for varying spatial development patterns in areas with existing structures, ranging from random distribution (no preference for the number of existing neighbouring structures), to linear preferential attachment (new buildings are more likely to appear near areas with many existing structures), to clustering (new buildings preferentially attach near large, dense clusters), we introduce a preferential attachment exponent, denoted by α, as described in Eq. 6.1. This exponent modulates the likelihood that a new building will be placed adjacent to a given existing building based on the current number of structures. Specifically:

- When $\alpha = 0$, new buildings are placed uniformly at random, with no preference for the number of neighbouring structures.
- When $\alpha = 1$, the probability of a new building attaching to an existing one is directly proportional to the number of neighbouring structures .
- When $\alpha = 2$, new buildings are strongly biased toward attaching near existing structures with many neighbours, resulting in enhanced clustering.

The probability $P(i)$ that a new building connects to an existing building i is given by,

$$P(i) = \frac{k_i^{\alpha}}{\sum_{j=1}^{M} k_j^{\alpha}} \tag{6.1}$$

where k_i is the number of neighbouring buildings of building i within a radius of 6 kilometres, an area within which habitat is significantly influenced by anthropogenic disturbance, and where the effect of a building on wildebeest behavioural response reaches

half of its maximum strength (Masolele et al., 2025). $\sum_{j=1}^{M} k_j^{\alpha}$ represents the sum of the degrees of all existing buildings, that is, the number of connections each building has, each raised to the power of α (this normalises the probability across all existing buildings to sum to one), M is the total number of existing buildings.

New buildings are placed uniformly within a radius of 6 km from the selected neighbouring building i, hence the location of the new building is given by,

$$\begin{aligned} r &= 6\sqrt{u}, \\ x' &= x + r\cos\phi, \\ y' &= y + r\sin\phi \end{aligned} \tag{6.2}$$

where (x', y') is the location of the new building, x and y is the location of the existing building to which the new building is attached and is chosen stochastically with its probability given by eqn. 6.1. u is a uniformly distributed random number between 0 and 1, r is the radius within which the new building is added and ϕ is a uniformly distributed random number between 0 and 2π.

6.2.3 Wildebeest space use estimation

For each simulation scenario, we estimate wildebeest space use using the covariate field defined on the basis of proximity of the combined existing and new simulated buildings in the ecosystem in eqn. 6.2. The estimated wildebeest space use is given by:

$$PSPU = \frac{1}{Z}\exp\left(\sum_{i=0}^{N-1} \omega^i \beta \left(1 + \exp\left[\lambda\left(\|\mathbf{b}_i\| - \gamma\right)\right]\right)^{-1}\right) \tag{6.3}$$

where β is the value of the selection coefficient of wildebeest response to buildings, N

is the number of nearest buildings considered that we set at 10 in our analysis to reduce unnecessary computational expense while preserving the stability of the results as demonstrated by Masolele et al. (2025), ω is a model parameter that represents the diminishing effects of the nearest buildings, γ and λ quantify the spatial extent of the influence of the buildings on wildebeest, $\|\mathbf{b}_i\|$ is the covariate field of the Euclidean distance to the i^{th} nearest building. To create this covariate, we generated a 1 km resolution grid covering the full spatial extent of all potential wildebeest locations and calculated the distance to the i^{th} nearest building for each point on the grid, and Z is a normalising constant. All model parameters used were inferred from the model in (Masolele et al., 2025) and presented in the appendix in Table S3.

Subsequently, we calculated the change in wildebeest space use ratio ($SPUR$) as the difference between the predicted space use from simulated buildings data ($PSPU$) and the space use obtained using the current building locations ($CSPU$) divided by the space use obtained using the current building locations:

$$SPUR = \frac{PSPU - CSPU}{CSPU} \times 100. \tag{6.4}$$

6.2.4 Quantifying change in wildebeest space use

We assessed how increasing the number of buildings would lead to a change from the current observed wildebeest space use by comparing the predicted space use from simulated buildings data to the space use of the real spatial data of buildings currently present in the ecosystem. For this purpose, we used the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) as the distance metric to compare between each simulated space use under various scenarios and the space use obtained using the current

building locations. Then the KL divergence was calculated as,

$$KL[Q(\boldsymbol{\theta})||P(\boldsymbol{\theta})] = \int Q(\boldsymbol{\theta}) \log\left(\frac{Q(\boldsymbol{\theta})}{P(\boldsymbol{\theta})}\right) d\boldsymbol{\theta}. \tag{6.5}$$

Where $P(\boldsymbol{\theta})$ is the predicted space use using the real building data and $Q(\boldsymbol{\theta})$ is the predicted space use from the simulated building data. If the simulated space use produces an identical distribution of usage as the observed space use, the KL distance will be equal to 0, but if they are not identical, the KL will be greater than 0 $(KL > 0)$.

6.3 Results

Comparison of the new building allocation strategy within the ecosystem across all three scenarios (10%, 50%, and 100%) demonstrates clear effects on the space use of migratory wildebeest, though the impact varies between the scenarios. Even a small increase in the number of buildings, for example, by 10% of existing buildings within the ecosystem is predicted to negatively affect the migratory wildebeest by partially displacing their spatial distribution and altering habitat use patterns in all values of parameter α in the preferential attachment mechanism (Top rows in Fig. 6.2A-E, Fig. 6.5A-E, Fig. 6.8A-E). Notably, when the allocation of new buildings is completely randomly distributed across the ecosystem regardless of the presence of prior structures (i.e., when the parameter $\delta = 1$), we observe a marked negative change in wildebeest space use in almost all areas of the ecosystem especially under large development scenarios (e.g., 50% and 100% increase in the number of existing buildings), as shown in Fig. 6.2J and O, Fig. 6.5J and O, and Fig. 6.8J and O.

Likewise, the simulation predicts that when the new buildings are randomly allocated to areas with existing structure (i.e., when the parameter $\alpha = 0$), meaning that both small

and large clusters have the same chance of the new building being assigned next to it, there will be a substantial change in wildebeest space use (Fig. 6.2), and the magnitude of the change increases when there is a large increase in the number of new buildings, that is, 50% and 100% increase of existing buildings as shown in Fig. 6.2F-I and K-N.

Figure 6.2: Predicted change in wildebeest space use due to the response to increasing buildings from the simulation. Colors indicate expected change in use from positive change (blue), no change (grey), to negative change (red). Top row A-E, middle row F-J and bottom row K-O represent 10%, 50%, and 100% increase of existing buildings, respectively. Note, the space use is changing from left to right due to decreasing of buildings clustering with δ values of 0, 0.25, 0.5, 0.75, and 1. The preferential attachment parameter used during the simulation of buildings was $\boldsymbol{\alpha} = 0$. Gray lines indicate the boundary of the Serengeti ecosystem and associated protected areas.

Figure 6.3: Predicted 95% lower credible intervals for the wildebeest space use due to the response to increasing buildings from the simulation. Colors indicate low space use (dark blue) to high space use (yellow). Top row A-E, middle row F-J and bottom row K-O represent 10%, 50%, and 100% increase of existing buildings, respectively. Note, the space use is changing from left to right due to decreasing of buildings clustering with δ values of 0, 0.25, 0.5, 0.75, and 1. The preferential attachment parameter used during the simulation of buildings was $\boldsymbol{\alpha} = 0$. Gray lines indicate the boundary of the Serengeti ecosystem and associated protected areas.

Figure 6.4: Predicted 95% upper credible intervals for the wildebeest space use due to the response to increasing buildings from the simulation. Colors indicate low space use (dark blue) to high space use (yellow). Top row A-E, middle row F-J and bottom row K-O represent 10%, 50%, and 100% increase of existing buildings, respectively. Note, the space use is changing from left to right due to decreasing of buildings clustering with δ values of 0, 0.25, 0.5, 0.75, and 1. The preferential attachment parameter used during the simulation of buildings was $\boldsymbol{\alpha} = 0$. Gray lines indicate the boundary of the Serengeti ecosystem and associated protected areas.

Furthermore, the simulation predicts a small change in wildebeest space use when new buildings are added to areas with existing structures but proportionate to the number

of structures already existing (i.e., when the parameter $\alpha = 1$) compared to random allocation regardless of the existing number of structures (i.e., when the parameter $\alpha = 0$), as shown in Fig. 6.5A, F, K. This means that small clusters receive fewer new allocations of buildings and large clusters receive more allocation of new buildings. In contrast, when random allocation and preferential attachment strategies are used at the same time, we observe a significant change in wildebeest space use regardless of the number of new buildings added as shown in Fig. 6.5 B-D, G-I, and L-N. But the changes are far more pronounced when the allocation is totally random (that is, when the parameter $\delta = 1$), meaning both areas with existing structures and wilderness regions (undeveloped regions) have the same probability of new buildings being assigned in those locations.

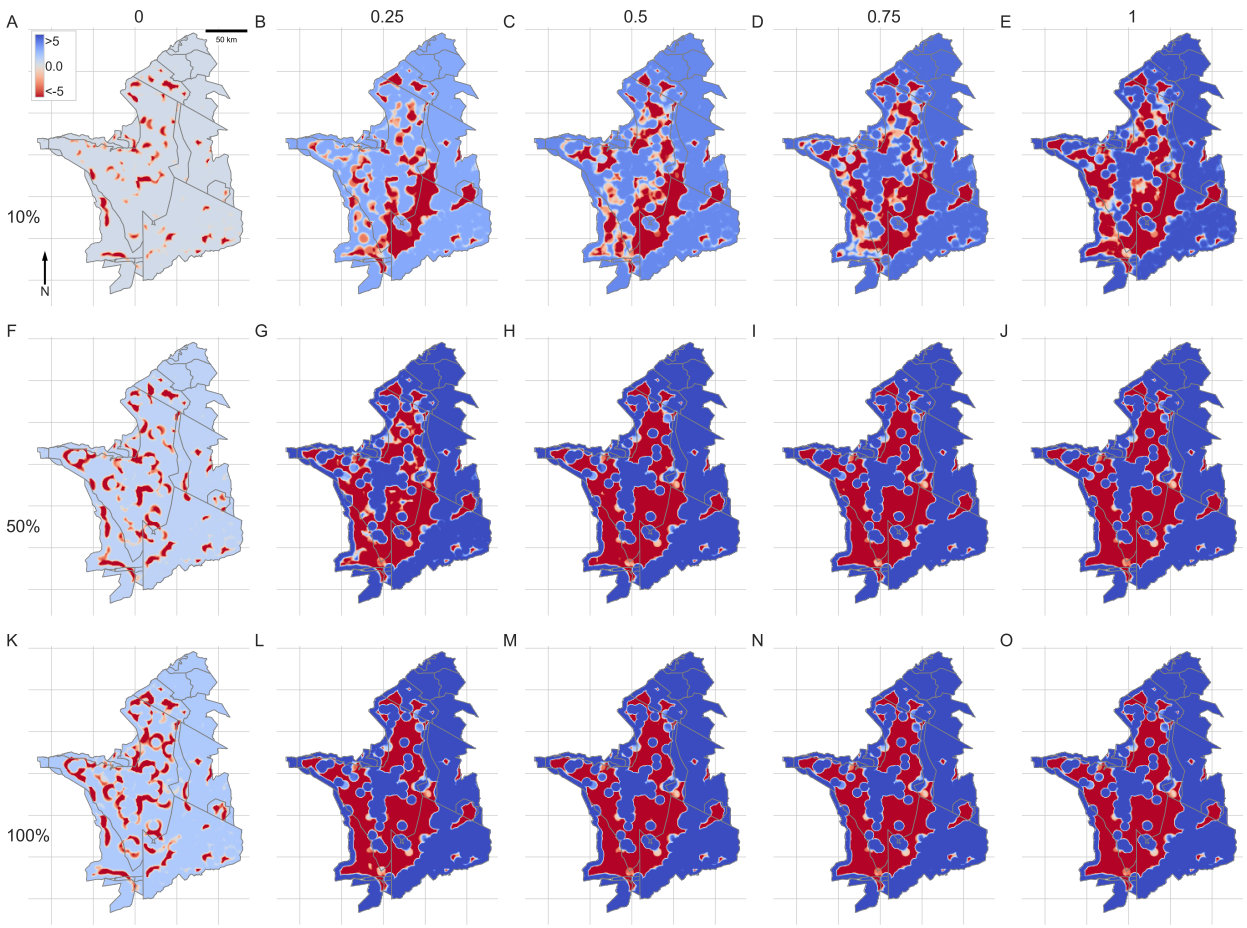

Figure 6.5: Predicted change in wildebeest space use due to the response to increasing buildings from the simulation. Colors indicate expected change in use from positive change (blue), no change (grey), to negative change (red). Top row A-E, middle row F-J and bottom row K-O represent 10%, 50%, and 100% increase of existing buildings, respectively. Note, the space use is changing from left to right due to decreasing of buildings clustering with δ values of 0, 0.25, 0.5, 0.75, and 1. The preferential attachment parameter used during the simulation of buildings was $\boldsymbol{\alpha} = 1$. Gray lines indicate the boundary of the Serengeti ecosystem and associated protected areas.

Figure 6.6: Predicted 95% lower credible intervals for the wildebeest space use due to the response to increasing buildings from the simulation. Colors indicate low space use (dark blue) to high space use (yellow). Top row A-E, middle row F-J and bottom row K-O represent 10%, 50%, and 100% increase of existing buildings, respectively. Note, the space use is changing from left to right due to decreasing of buildings clustering with δ values of 0, 0.25, 0.5, 0.75, and 1. The preferential attachment parameter used during the simulation of buildings was $\boldsymbol{\alpha} = 1$. Gray lines indicate the boundary of the Serengeti ecosystem and associated protected areas.

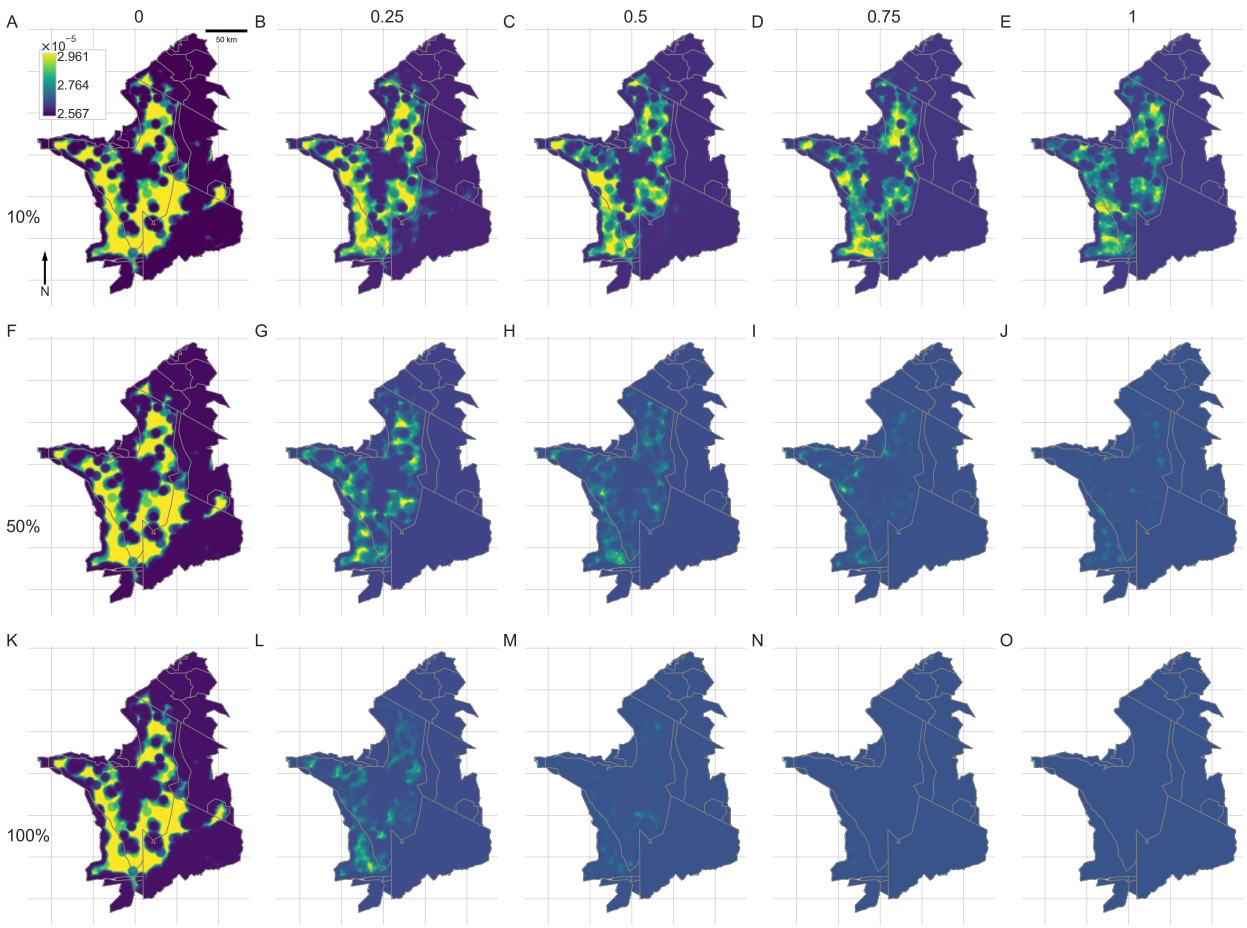

Figure 6.7: Predicted 95% upper credible intervals for the wildebeest space use due to the response to increasing buildings from the simulation. Colors indicate low space use (dark blue) to high space use (yellow). Top row A-E, middle row F-J and bottom row K-O represent 10%, 50%, and 100% increase of existing buildings, respectively. Note, the space use is changing from left to right due to decreasing of buildings clustering with δ values of 0, 0.25, 0.5, 0.75, and 1. The preferential attachment parameter used during the simulation of buildings was $\boldsymbol{\alpha} = 1$. Gray lines indicate the boundary of the Serengeti ecosystem and associated protected areas.

In addition, when the new buildings are clustered (i.e., when the parameter $\alpha = 2$), there is a relatively small change in wildebeest space use compared to when they are allocated

randomly (Fig. 6.8A, F, K). Furthermore, we observed that when a mixed strategy (for example, combining clustering and random allocation) is used for a large increase in the number of new buildings (50% and 100% increases of existing buildings in Fig. 6.8G-I and Fig. 6.8L-N, respectively), there is an observable change in wildebeest space use, even when the number of new buildings added randomly is relatively small compared to the clustered ones.

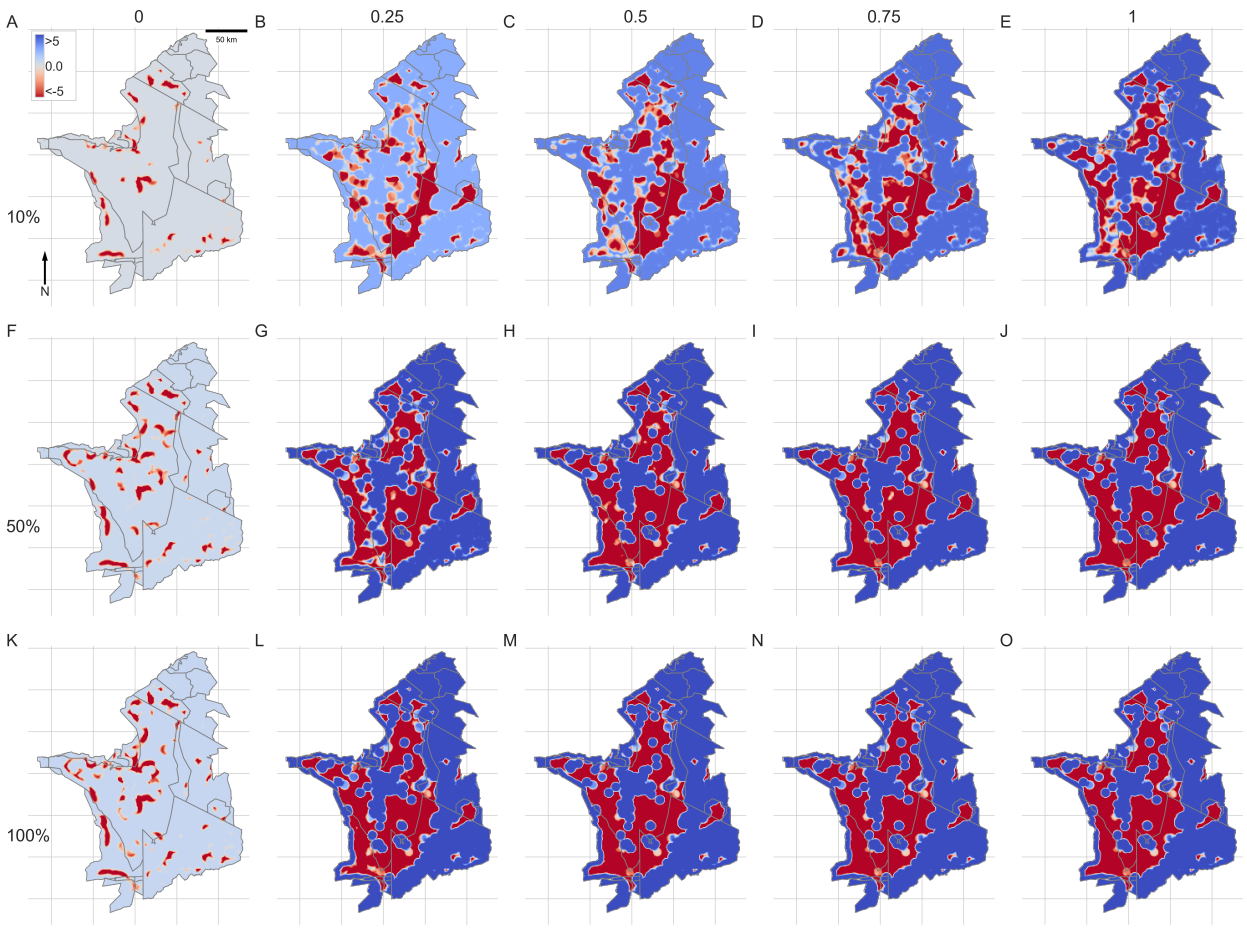

Figure 6.8: Predicted change in wildebeest space use due to the response to increasing buildings from the simulation. Colors indicate expected change in use from positive change (blue), no change (grey), to negative change (red). Top row A-E, middle row F-J and bottom row K-O represent 10%, 50%, and 100% increase of existing buildings, respectively. Note, the space use is changing from left to right due to decreasing of buildings clustering with δ values of 0, 0.25, 0.5, 0.75, and 1. The preferential attachment parameter used during the simulation of buildings was $\boldsymbol{\alpha} = 2$. Gray lines indicate the boundary of the Serengeti ecosystem and associated protected areas.

Figure 6.9: Predicted 95% lower credible intervals for the wildebeest space use due to the response to increasing buildings from the simulation. Colors indicate low space use (dark blue) to high space use (yellow). Top row A-E, middle row F-J and bottom row K-O represent 10%, 50%, and 100% increase of existing buildings, respectively. Note, the space use is changing from left to right due to decreasing of buildings clustering with δ values of 0, 0.25, 0.5, 0.75, and 1. The preferential attachment parameter used during the simulation of buildings was $\boldsymbol{\alpha} = 2$. Gray lines indicate the boundary of the Serengeti ecosystem and associated protected areas.

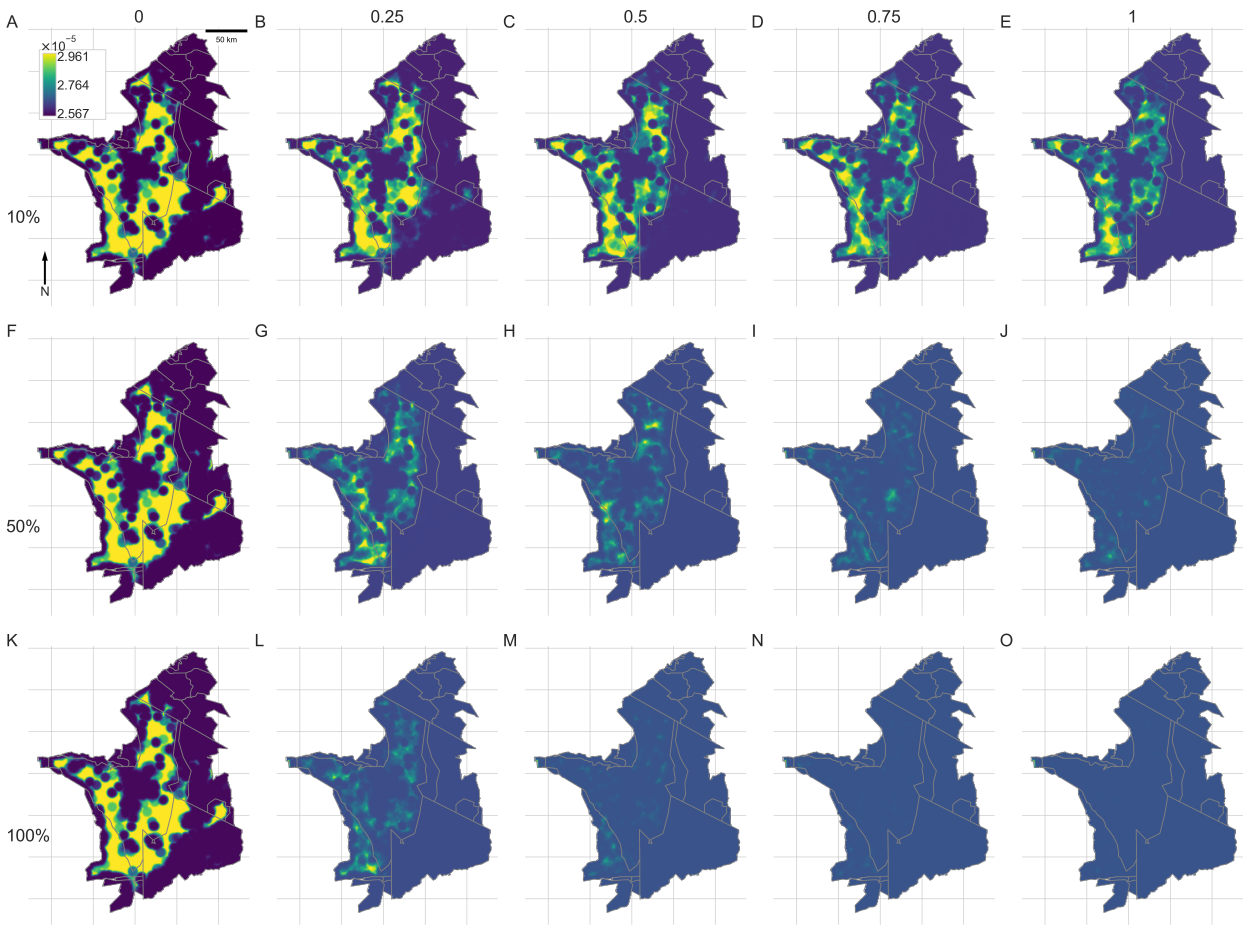

Figure 6.10: Predicted 95% upper credible intervals for the wildebeest space use due to the response to increasing buildings from the simulation. Colors indicate low space use (dark blue) to high space use (yellow). Top row A-E, middle row F-J and bottom row K-O represent 10%, 50%, and 100% increase of existing buildings, respectively. Note, the space use is changing from left to right due to decreasing of buildings clustering with δ values of 0, 0.25, 0.5, 0.75, and 1. The preferential attachment parameter used during the simulation of buildings was $\boldsymbol{\alpha} = 2$. Gray lines indicate the boundary of the Serengeti ecosystem and associated protected areas.

6.3.1 Predicted change in wildebeest space use

To demonstrate how an increase in the number of buildings within the ecosystem changes wildebeest space use, we performed simulations that incorporate new buildings in three scenarios: a 10%, 50%, and 100% increase in the number of existing buildings. We used the parameter α that governed the clustering of the new buildings, with its range spanning from 0 to 2, while the parameter δ (ranging from 0 to 1) represented the decrease in clustering. For each value of $\alpha \in \{0, 0.2, 0.4, 0.6, 0.8, ..., 2\}$, we iterated over a second set of parameters $\delta \in \{0, 0.1, 0.2, 0.3, 0.4, ..., 1\}$ during the building simulation across all scenarios. This resulted in a decrease in the clustering of buildings as the value of δ increased from 0 to 1. We then estimated space use based on the simulated buildings, as described in Eqn. 6.3. To assess the difference between the simulated and observed wildebeest space use, we used the Kullback-Leibler divergence (Eqn. 6.5), with the results plotted in Fig. 6.11A-C. The results indicate that when the new buildings are clustered, there is a small change in wildebeest space use compared to when they are not clustered (Fig. 6.11A, B, C). In contrast, we observe a large shift in wildebeest space use when there is a decrease in clustering of new buildings. This large change is observed even for a small reduction in clustering when a large number of buildings are added to the ecosystem, for example, 50% and 100% of existing buildings in Fig. 6.11B-C.

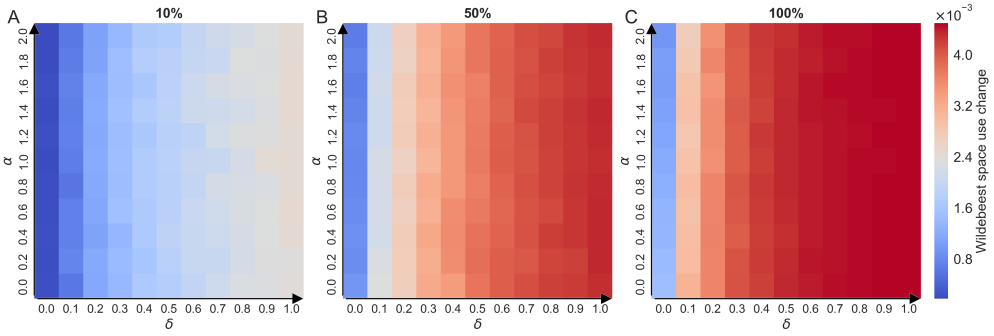

Figure 6.11: Predicted change in wildebeest space use resulting from increasing the number of new buildings from the current observed usage in the Serengeti ecosystem. A) Number of new added buildings is 10% of existing buildings, B) Number of new added buildings is 50% of existing buildings, C) Number of new added buildings is 100% of existing buildings.

6.4 Discussion

We have developed a framework that predicts the impact of buildings on wildebeest space use change. The framework allows for the simultaneous application of multiple building allocation strategies (random placement or clustered), where one strategy can be increased while the other is decreased in terms of the placement of new buildings within the ecosystem. We have ensured that our simulation of additional buildings aligns with real-world practices in building allocation in the ecosystem, thereby facilitating a direct link to how these structures may impact wildebeest space use. We achieve it by combining a simple non-linear preferential attachment rule (governed by α parameter) with a modifier parameter (δ) that decreases the clustering of buildings over time to understand how wildebeest space use will be changed by new additions of buildings in the ecosystem. Our simulation model predicts change in wildebeest space use as the number of new buildings are placed in the ecosystem; however, these changes are

more pronounced when buildings are added haphazardly. While the framework presented here does not directly address wildebeest migration or incorporate demographic changes, buildings may pose a potential risk if they are placed in areas with key resources essential for survival during migration. If such areas are abandoned by the animals, this could affect their fitness and have broader demographic consequences.

Our simulation model presents one possible scenario as far as the effects of the buildings on wildebeest space use are concerned in the absence of other environmental factors such as rainfall. However, there are reasons to believe that as the number of buildings increases, people, roads, vehicle traffic, noise, and light pollution follow, eventually making the development of new buildings a de facto disturbance and fragmenting habitats that could reduce the use of wildebeest in those areas. The fragmentation of habitats that is likely to result from building activities has the potential to dissect the ecosystem habitat into separate habitats, which could lead to the loss of functional heterogeneity and a decrease in vegetation productivity (Hobbs et al., 2008), leading to inadequate support for wildebeest in these areas. In worst-case scenarios, this could affect the fitness of individuals if these areas where new buildings are built are the refuge of last resort that provide key resources for wildebeest during times of need, such as times of dietary stress in the course of their migration.

While clustering buildings does not completely eliminate the impact of buildings on wildebeest space use, it significantly reduces the negative change compared to random and linear preferential attachment building allocation strategies. This reduction likely results from the diminishing effects, in which the first structure introduced into an undeveloped area elicits the strongest behavioural response from wildebeest. Subsequent buildings in already developed zones or areas contribute comparatively little additional displacement, as many associated stressors of human disturbance are already

present (Masolele et al., 2025). From a conservation management point of view, this finding underscores the importance of directing new infrastructure into already impacted areas, as this may help buffer wildebeest from further habitat fragmentation and behavioural displacement. Implementing conservation spatial planning policies that promote clustered development can help limit the expansion of the human footprint into intact habitats, thereby reducing habitat fragmentation and preserving functional space for migratory species like wildebeest. Such spatial planning approaches are essential for balancing development needs with the long-term conservation of wide-ranging wildlife.

Like all models that model hypothetical scenarios, the framework presented here inevitably has inherent limitations. For example, although we allow for new buildings to be allocated in both areas with existing structures and previously undeveloped regions in our simulations, we assume that wildebeest have avoidance patterns toward buildings, multiple co-located buildings have diminishing effects, and that only the number of buildings varies. However, it is clear that a more precise identification and separation of the potential correlates of buildings is crucial to derive more refined predictions of future wildebeest space use in the presence of additional buildings. For now, in some cases, the simulations extrapolate beyond the observed range of building densities, and the assumption is that increasing the number of buildings within a cluster does not fundamentally alter wildebeest movement patterns. This assumption is informed by recent telemetry data on wildebeest movements, in conjunction with spatial data on building distributions, used to model their responses to anthropogenic structures (Masolele et al., 2025). Future refinements may include the development of a dynamic virtual model driven by the integration of multisource data within the simulation framework. Additionally, employing a digital twin approach could facilitate real-time updates, thereby enhancing the models ability to adapt to changes in the distribution and expansion of

anthropogenic structures within the ecosystem.

We have provided a simulation framework to assess the impact of buildings on wildebeest space use under various scenarios that involve the addition of new buildings to the Greater Serengeti-Mara ecosystem. The framework accommodates the use of multiple strategies simultaneously (random placement or clustered), and each strategy independently, allowing for the exploration of both individual and combined effects of building allocation strategies. We assess the impacts of these new building additions using the Kullback-Leibler divergence, enabling a comprehensive evaluation of changes in animal space use. We believe that our simulation framework provides insight for wildlife managers and policymakers in minimising the adverse effects of new anthropogenic structures when making decisions regarding their addition in protected areas.

Chapter 7

Conclusions

Statistical methods have become indispensable in ecological modelling, enhancing generalisability, predictive capabilities, and the scalability of inferences across spatial scales. These approaches offer critical insights into animal responses to anthropogenic pressures in the Anthropocene, making them especially relevant for contemporary ecological applications. This thesis contributes to the field by introducing novel, scalable statistical methods for analysing animal movement and migration, integrating both simulation studies and modelling real-world animal telemetry data, while maintaining high computational efficiency throughout. The methodological contributions that I have given are found in Chapter 3 and include its applications to real telemetry data of animals in Chapters 4 and 5. Chapter 6 focuses on simulations.

In Chapter 2, I highlighted the significance of existing modelling techniques and their contribution to our comprehension of animal movement at various scales such as fine-scale and broad-scale, and identified research gaps that need to be filled in this area.

In Chapter 3, I proposed a novel approximate Bayesian inference method to quantify un-

certainty in a multiscale step selection model, which links fine-scale animal movement decisions to broad-scale space use. Specifically, this approach allows the parameters inferred from a model fitted to fine-scale data to be used directly to estimate population-level space use (Michelot et al., 2020, 2019a). The method is based on variational inference, which employs a simple distribution, known as the variational distribution, to approximate the true posterior probability distribution (Blei et al., 2017) through optimization techniques such as stochastic gradient descent (Hoffman et al., 2013). This approach enables the division of telemetry data into mini-batches during inference, facilitating parallel computation, reducing memory requirements, and enhancing computational efficiency. Unlike standard Bayesian sampling methods, which require evaluating the model likelihood at each step of the sampler using the whole dataset, thereby significantly increasing computational complexity, this method obviates the need to sample directly from the posterior distribution. Furthermore, by iterating data batches multiple times through the optimization algorithm, variational inference effectively learns the selection and movement parameters from movement observations and the associated covariate field. This powerful and versatile modelling framework provides reliable parameter estimates within a reasonable timeframe, making it a viable alternative to standard Bayesian sampling-based approaches, which can be computationally prohibitive. As a result, this method has great potential for applications in movement ecology.

The novel multiscale step selection framework presented in Chapter 4 addresses key challenges in assessing the effects of anthropogenic disturbances on animal habitat selection, including identifying thresholds at which their impact diminishes, capturing nonlinear responses, and evaluating the interacting effects of multiple co-located human-made structures. The primary objective was to examine whether the presence of a single human-made structure, such as a building, elicits a stronger response in wildebeest com-

pared to multiple structures in close proximity, thereby assessing whether the combined effects of such structures are compounding or diminishing. In the literature, traditional resource selection functions (RSFs) (Manly et al., 2007) and standard step selection functions (SSFs) (Forester et al., 2009; Fortin et al., 2005; Thurfjell et al., 2014) are commonly used for this purpose. While these models are intuitive to apply, they have limitations in terms of scaling properties. For example, the RSF parameters represent broad-scale selection, whereas the SSF parameters capture fine-scale movement decisions, making it difficult to translate individual animal movement responses to anthropogenic disturbances to population-level space use. To overcome these limitations, the proposed flexible multiscale model enables the identification of thresholds, nonlinear responses, and interactions effects of multiple co-located anthropogenic structures on animal habitat preferences. Additionally, it allows fine-scale movement effects to be directly propagated to broad-scale space use patterns. This model was applied to quantify the impact of buildings on the movement and spatial distribution of migratory wildebeest in the Greater Serengeti-Mara ecosystem. The analysis revealed a reduction in wildebeest use near buildings, with diminishing effects observed when buildings were clustered. Future research could enhance the current model by incorporating additional covariates that were not previously included. This would improve our understanding of how various abiotic, biotic, and anthropogenic factors influence wildebeest movement decisions and spatial use patterns.

In Chapter 5, I demonstrated the application of hierarchical sparse Gaussian processes to model the population-level mean migration routes of the Serengeti wildebeest. By using these routes as a proxy for long-term spatial memory and integrating them with space use patterns derived from local environmental features, I demonstrated how this approach improves the prediction of animal habitat use during ecologically and demo-

graphically critical periods. This integrative approach underscores the importance of accounting for both environmental features and long-term spatial memory in movement models when predicting the spatial distribution of animals over time and space.

In Chapter 6, I demonstrated how the addition of new buildings in the Serengeti ecosystem would impact wildebeest space use. To assess this, I introduced a novel simulation approach that modelled building expansion under various scenarios (10%, 50%, and 100% increases in existing buildings), while considering different allocation strategies, including random, linear, and clustering in areas with existing structures, and completely random placement regardless of prior structures. The simulation framework was based on concepts from non-linear preferential attachment (Kunegis et al., 2013), augmented with an accept-reject mechanism. This approach allowed for the realistic simulation of new buildings in both previously developed and undeveloped areas while simultaneously quantifying the changes in wildebeest space use due to additional infrastructure. To evaluate these changes, Kullback-Leibler divergence was used to quantify the shift in estimated space use relative to the observed space use inferred from real telemetry data. The simulation results indicated that if new buildings must be added, the strategy that minimises the impact on wildebeest space use is clustering buildings together in areas away from key grazing sites. However, it is important to note that the current simulation framework does not explicitly model wildebeest migratory dynamics or account for demographic processes. As a result, the model provides only a partial understanding of how additional development may influence wildebeest behavioural responses. To fully assess the ecological consequences of expanding human-made structures within the ecosystem, future models should integrate migratory patterns and demographic parameters. This will help determine whether increased infrastructure imposes additional barriers to migration or disrupts population viability over time.

7.1 Future work

The computational modelling techniques developed in this study have a broad range of potential applications and can be used effectively in various ecosystems to study animal movement and migration using telemetry data. The core concept underlying these methods is the ability to handle large-scale datasets of animal movement while improving computational efficiency. The Bayesian multiscale step selection framework developed in this study is based on a discrete-time approach. As a result, there remains a gap in our mechanistic understanding of how multiscale inference operates in continuous time within a Bayesian paradigm. A promising direction for future research would be to extend this framework to a continuous-time setting, which could provide deeper insights into animal movement and habitat selection in the Anthropocene. Additionally, while the current method relies solely on telemetry data from GPS-tagged individuals, future work could explore the integration of multiscale inference with traditional resource selection functions (RSFs), which are based on spatial survey (location) data. Combining these approaches would improve estimates and predictions of animal space use at the population level by incorporating information from areas visited by untagged conspecifics. This integration would enhance the robustness of habitat preference modelling and contribute to a more comprehensive understanding of animal movement. Given the increasing availability of fine-scale movement observations, leveraging this information is crucial for advancing movement ecology research.

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: A system for large-scale machine learning. arXiv:1605.08695.

Aikens, E.O., Wyckoff, T.B., Sawyer, H., Kauffman, M.J., 2022. Industrial energy development decouples ungulate migration from the green wave. Nature Ecology & Evolution 6, 1733–1741.

Aldossari, S., Husmeier, D., Matthiopoulos, J., 2022. Transferable species distribution modelling: Comparative performance of generalised functional response models. Ecological Informatics 71, 101803.

Alston, J.M., Fleming, C.H., Kays, R., Streicher, J.P., Downs, C.T., Ramesh, T., Reineking, B., Calabrese, J.M., 2023. Mitigating pseudoreplication and bias in resource selection functions with autocorrelation-informed weighting. Methods in Ecology and Evolution 14, 643–654.

Arce Guillen, R., Lindgren, F., Muff, S., Glass, T.W., Breed, G.A., Schlägel, U.E., 2023. Accounting for unobserved spatial variation in step selection analyses of animal movement via spatial random effects. Methods in Ecology and Evolution 14, 2639–2653.

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

Barabási, A.L., Albert, R., 1999. Emergence of scaling in random networks. science 286, 509–512.

Barnett, A.H., Moorcroft, P.R., 2008. Analytic steady-state space use patterns and rapid computations in mechanistic home range analysis. Journal of mathematical biology 57, 139–159.

Bauer, S., Hoye, B.J., 2014. Migratory animals couple biodiversity and ecosystem functioning worldwide. Science 344, 1242552.

Beaudrot, L., Palmer, M.S., Anderson, T.M., Packer, C., 2020. Mixed-species groups of Serengeti grazers: a test of the stress gradient hypothesis. Ecology 101, e03163.

Benhamou, S., 2014. Of scales and stationarity in animal movements. Ecology letters 17, 261–272.

Bennett, E.M., Carpenter, S.R., Peterson, G.D., Cumming, G.S., Zurek, M., Pingali, P., 2003. Why global scenarios need ecology. Frontiers in Ecology and the Environment 1, 322–329.

Berdahl, A.M., Kao, A.B., Flack, A., Westley, P.A., Codling, E.A., Couzin, I.D., Dell, A.I., Biro, D., 2018. Collective animal navigation and migratory culture: from theoretical models to empirical evidence. Philosophical Transactions of the Royal Society B: Biological Sciences 373, 20170009.

Berger, J., 2007. Fear, human shields and the redistribution of prey and predators in protected areas. Biology letters 3, 620–623.

Beyer, H.L., Haydon, D.T., Morales, J.M., Frair, J.L., Hebblewhite, M., Mitchell, M., Matthiopoulos, J., 2010. The interpretation of habitat preference metrics under use–availability designs. Philosophical Transactions of the Royal Society B: Biological Sciences 365, 2245–2254.

Blei, D.M., Kucukelbir, A., McAuliffe, J.D., 2017. Variational inference: A review for statisticians. Journal of the American Statistical Association 112, 859–877.

Bolger, D.T., Newmark, W.D., Morrison, T.A., Doak, D.F., 2008. The need for integrative approaches to understand and conserve migratory ungulates. Ecology letters 11, 63–77.

Bolstad, W.M., 2009. Understanding computational Bayesian statistics. volume 644. John Wiley & Sons.

Boone, R.B., Thirgood, S.J., Hopcraft, J.G.C., 2006. Serengeti wildebeest migratory patterns modeled from rainfall and new vegetation growth. Ecology 87, 1987–1994.

Bracis, C., Mueller, T., 2017. Memory, not just perception, plays an important role in terrestrial mammalian migration. Proceedings of the Royal Society B: Biological Sciences 284, 20170449.

Brooks, S., Gelman, A., Jones, G., Meng, X.L., 2011. Handbook of markov chain monte carlo. CRC press.

Buchanan, C., 2020. Long-term physiological trends and their drivers: linking hair hormone concentrations with telemetry data in GPS-collared Serengeti wildebeest. Ph.D. thesis. University of Glasgow.

Bunnefeld, N., Börger, L., van Moorter, B., Rolandsen, C.M., Dettki, H., Solberg, E.J., Ericsson, G., 2011. A model-driven approach to quantify migration patterns: individual, regional and yearly differences. Journal of Animal Ecology 80, 466–476.

Calabrese, J.M., Fleming, C.H., Gurarie, E., 2016. ctmm: An R package for analyzing animal relocation data as a continuous-time stochastic process. Methods in Ecology and Evolution 7, 1124–1132.

Caravaggi, A., Banks, P.B., Burton, A.C., Finlay, C.M., Haswell, P.M., Hayward, M.W., Rowcliffe, M.J., Wood, M.D., 2017. A review of camera trapping for conservation behaviour research. Remote Sensing in Ecology and Conservation 3, 109–122.

Clontz, L.M., Pepin, K.M., VerCauteren, K.C., Beasley, J.C., 2021. Behavioral state resource selection in invasive wild pigs in the Southeastern United States. Scientific reports 11, 6924.

Cobb, A.D., Everett, R., Markham, A., Roberts, S.J., 2018. Identifying sources and sinks in the presence of multiple agents with gaussian process vector calculus, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1254–1262.

Colomer, M., Margalida, A., 2025. Demographic effects of sanitary policies on European vulture population dynamics: A retrospective modeling approach. Ecological Applications 35, e3093.

Creel, S., Fox, J.E., Hardy, A., Sands, J., Garrott, B., Peterson, R.O., 2002. Snowmobile activity and glucocorticoid stress responses in wolves and elk. Conservation biology 16, 809–814.

Cumming, G.S., 2007. Global biodiversity scenarios and landscape ecology. Landscape ecology 22, 671–685.

DeCesare, N.J., Hebblewhite, M., Schmiegelow, F., Hervieux, D., McDermid, G.J., Neufeld, L., Bradley, M., Whittington, J., Smith, K.G., Morgantini, L.E., et al., 2012. Transcending scale dependence in identifying habitat with resource selection functions. Ecological Applications 22, 1068–1083.

Dejeante, R., Valeix, M., Chamaillé-Jammes, S., 2024. Time-varying habitat selection anal-

ysis: A model and applications for studying diel, seasonal, and post-release changes. Ecology 105, e4233.

Dillon, J.V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., Patton, B., Alemi, A., Hoffman, M., Saurous, R.A., 2017. Tensorflow distributions. arXiv preprint arXiv:1711.10604 .

Dirzo, R., Young, H.S., Galetti, M., Ceballos, G., Isaac, N.J., Collen, B., 2014. Defaunation in the Anthropocene. science 345, 401–406.

Dobson, A., 2009. Food-web structure and ecosystem services: insights from the Serengeti. Philosophical Transactions of the Royal Society B: Biological Sciences 364, 1665–1682.

Doherty, T.S., Hays, G.C., Driscoll, D.A., 2021. Human disturbance causes widespread disruption of animal movement. Nature Ecology & Evolution 5, 513–519.

Dunn, J.E., Gipson, P.S., 1977. Analysis of radio telemetry data in studies of home range. Biometrics , 85–101.

Fagan, W.F., Calabrese, J.M., 2014. The correlated random walk and the rise of movement ecology. Bulletin of the Ecological Society of America 95, 204–206.

Fahrmeir, L., 1992. Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. Journal of the American Statistical Association 87, 501–509.

Ferguson, T.S., 1982. An inconsistent maximum likelihood estimate. Journal of the American Statistical Association 77, 831–834.

Fieberg, J., Matthiopoulos, J., Hebblewhite, M., Boyce, M.S., Frair, J.L., 2010. Correlation and studies of habitat selection: problem, red herring or opportunity? Philosophical Transactions of the Royal Society B: Biological Sciences 365, 2233–2244.

Fieberg, J., Signer, J., Smith, B., Avgar, T., 2021. A 'How to'guide for interpreting parameters in habitat-selection analyses. Journal of Animal Ecology 90, 1027–1043.

Fieberg, J.R., Vitense, K., Johnson, D.H., 2020. Resampling-based methods for biologists. PeerJ 8, e9089.

Fithian, W., Hastie, T., 2012. Statistical models for presence-only data: finite-sample equivalence and addressing observer bias. Ann Appl Stat .

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

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

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

Friedman, M., 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the american statistical association 32, 675–701.

Fryxell, J.M., Sinclair, A., 1988. Causes and consequences of migration by large herbivores. Trends in ecology & evolution 3, 237–241.

Gardiner, C.W., 2009. Stochastic Methods: A Handbook for the Natural and Social Sciences. 4th ed., Springer.

Gaynor, K.M., Brown, J.S., Middleton, A.D., Power, M.E., Brashares, J.S., 2019. Landscapes of fear: spatial patterns of risk perception and response. Trends in ecology & evolution 34, 355–368.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian data analysis. Texts in Statistical Science. 3rd ed., Chapman & Hall/CRC, 2000 Corporate Blvd N.W., Boca Raton, FL 33431, USA.

Gilks, W.R., Richardson, S., Spiegelhalter, D., 1995. Markov chain Monte Carlo in practice. CRC press.

Gurarie, E., Cagnacci, F., Peters, W., Fleming, C.H., Calabrese, J.M., Mueller, T., Fagan, W.F., 2017. A framework for modelling range shifts and migrations: asking when, whither, whether and will it return. Journal of Animal Ecology 86, 943–959.

Gurarie, E., Hebblewhite, M., Joly, K., Kelly, A.P., Adamczewski, J., Davidson, S.C., Davison, T., Gunn, A., Suitor, M.J., Fagan, W.F., et al., 2019. Tactical departures and strategic arrivals: Divergent effects of climate and weather on caribou spring migrations. Ecosphere 10, e02971.

Harding, H.R., Gordon, T.A., Eastcott, E., Simpson, S.D., Radford, A.N., 2019. Causes and consequences of intraspecific variation in animal responses to anthropogenic noise. Behavioral Ecology 30, 1501–1511.

Harris, G., Thirgood, S., Hopcraft, J.G.C., Cromsigt, J.P., Berger, J., 2009. Global decline in aggregated migrations of large terrestrial mammals. Endangered Species Research 7, 55–76.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57, 97–108.

Hawkes, C., 2009. Linking movement behaviour, dispersal and population processes: is individual variation a key? Journal of Animal Ecology 78, 894–906.

Haydon, D.T., Morales, J.M., Yott, A., Jenkins, D.A., Rosatte, R., Fryxell, J.M., 2008. Socially informed random walks: incorporating group dynamics into models of population spread and growth. Proceedings of the Royal Society B: Biological Sciences 275, 1101–1109.

Heinonen, M., Mannerström, H., Rousu, J., Kaski, S., Lähdesmäki, H., 2016. Non-stationary gaussian process regression with hamiltonian monte carlo, in: Artificial Intelligence and Statistics, PMLR. pp. 732–740.

Hensman, J., Fusi, N., Lawrence, N.D., 2013. Gaussian processes for big data. arXiv preprint arXiv:1309.6835 .

Hobbs, N.T., Galvin, K.A., Stokes, C.J., Lackett, J.M., Ash, A.J., Boone, R.B., Reid, R.S., Thornton, P.K., 2008. Fragmentation of rangelands: Implications for humans, animals, and landscapes. Global environmental change 18, 776–785.

Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J., 2013. Stochastic variational inference. Journal of Machine Learning Research .

Hofmann, D.D., Cozzi, G., Fieberg, J., 2024. Methods for implementing integrated step-selection functions with incomplete data. Movement Ecology 12, 37.

Holdo, R.M., Fryxell, J.M., Sinclair, A.R., Dobson, A., Holt, R.D., 2011. Predicted impact of barriers to migration on the Serengeti wildebeest population. PloS one 6, e16370.

Holdo, R.M., Holt, R.D., Fryxell, J.M., 2009. Opposing rainfall and plant nutritional gradients best explain the wildebeest migration in the Serengeti. The American Naturalist 173, 431–445.

Holdo, R.M., Roach, R.R., 2013. Inferring animal population distributions from individual tracking data: theoretical insights and potential pitfalls. Journal of Animal Ecology 82, 175–181.

Holloway, P., Miller, J.A., 2014. Uncertainty analysis of step-selection functions: The effect of model parameters on inferences about the relationship between animal movement and the environment, in: Geographic Information Science: 8th International Conference, GIScience 2014, Vienna, Austria, September 24-26, 2014. Proceedings 8, Springer. pp. 48–63.

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

Hopcraft, J.G.C., Anderson, T.M., Pérez-Vila, S., Mayemba, E., Olff, H., 2012. Body size and the division of niche space: food and predation differentially shape the distribution of Serengeti grazers. Journal of Animal Ecology 81, 201–213.

Hopcraft, J.G.C., Morales, J.M., Beyer, H., Borner, M., Mwangomo, E., Sinclair, A., Olff, H., Haydon, D.T., 2014. Competition, predation, and migration: individual choice patterns of Serengeti migrants captured by hierarchical models. Ecological Monographs 84, 355–372.

Hopcraft, J.G.C., Olff, H., Sinclair, A., 2010. Herbivores, resources and risks: alternating regulation along primary environmental gradients in savannas. Trends in ecology & evolution 25, 119–128.

Hopcraft, J.G.C., Sinclair, A., Packer, C., 2005. Planning for success: Serengeti lions seek prey accessibility rather than abundance. Journal of Animal Ecology 74, 559–566.

Jansen, J., Woolley, S.N., Dunstan, P.K., Foster, S.D., Hill, N.A., Haward, M., Johnson, C.R., 2022. Stop ignoring map uncertainty in biodiversity science and conservation policy. Nature Ecology & Evolution 6, 828–829.

Johnson, D.H., 1980. The comparison of usage and availability measurements for evaluating resource preference. Ecology 61, 65–71.

Jones, K.R., Venter, O., Fuller, R.A., Allan, J.R., Maxwell, S.L., Negret, P.J., Watson, J.E., 2018. One-third of global protected land is under intense human pressure. Science 360, 788–791.

Joo, R., Picardi, S., Boone, M.E., Clay, T.A., Patrick, S.C., Romero-Romero, V.S., Basille, M., 2020. A decade of movement ecology. arXiv preprint arXiv:2006.00110 .

Kareiva, P., Shigesada, N., 1983. Analyzing insect movement as a correlated random walk. Oecologia 56, 234–238.

Kauffman, M.J., Cagnacci, F., Chamaillé-Jammes, S., Hebblewhite, M., Hopcraft, J.G.C., Merkle, J.A., Mueller, T., Mysterud, A., Peters, W., Roettger, C., et al., 2021. Mapping out a future for ungulate migrations. Science 372, 566–569.

Kavwele, C.M., Torney, C.J., Morrison, T.A., Fulford, S., Masolele, M.M., Masoy, J., Hopcraft, J.G.C., 2022. Non-local effects of human activity on the spatial distribution of migratory wildlife in Serengeti National Park, Tanzania. Ecological Solutions and Evidence 3, e12159.

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

Kendrick, D.A., 1981. Stochastic Control for Econometric Models. McGraw-Hill Inc., US.

Kéry, M., Schaub, M., 2011. Bayesian population analysis using WinBUGS: a hierarchical perspective. Academic Press.

Kihwele, E., Veldhuis, M., Loishooki, A., Hongoa, J., Hopcraft, J., Olff, H., Wolanski, E., 2021. Upstream land-use negatively affects river flow dynamics in the Serengeti National Park. Ecohydrology & Hydrobiology 21, 1–12.

Klappstein, N.J., Thomas, L., Michelot, T., 2023. Flexible hidden Markov models for behaviour-dependent habitat selection. Movement Ecology 11, 30.

de Koning, K., Broekhuijsen, J., Kühn, I., Ovaskainen, O., Taubert, F., Endresen, D., Schigel, D., Grimm, V., 2023. Digital twins: dynamic model-data fusion for ecology. Trends in ecology & evolution 38, 916–926.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. The Annals of Mathematical Statistics 22, 79–86.

Kunegis, J., Blattner, M., Moser, C., 2013. Preferential attachment in online networks: Measurement and explanations, in: Proceedings of the 5th annual ACM web science conference, pp. 205–214.

Langrock, R., King, R., Matthiopoulos, J., Thomas, L., Fortin, D., Morales, J.M., 2012. Flexible and practical modeling of animal telemetry data: hidden Markov models and extensions. Ecology 93, 2336–2342.

LaPoint, S., Gallery, P., Wikelski, M., Kays, R., 2013. Animal behavior, cost-based corridor models, and real corridors. Landscape ecology 28, 1615–1630.

Larsen, F., Hopcraft, J.G.C., Hanley, N., Hongoa, J.R., Hynes, S., Loibooki, M., Mafuru, G., Needham, K., Shirima, F., Morrison, T.A., 2020. Wildebeest migration drives tourism demand in the Serengeti. Biological Conservation 248, 108688.

Lindsey, P., Allan, J., Brehony, P., Dickman, A., Robson, A., Begg, C., Bhammar, H., Blanken, L., Breuer, T., Fitzgerald, K., et al., 2020. Conserving Africa's wildlife and wildlands through the COVID-19 crisis and beyond. Nature ecology & evolution 4, 1300–1310.

Liu, X., Lin, L., Sinding, M.H.S., Bertola, L.D., Hanghøj, K., Quinn, L., Garcia-Erill, G., Rasmussen, M.S., Schubert, M., Pečnerová, P., et al., 2024. Introgression and disruption of migration routes have shaped the genetic integrity of wildebeest populations. Nature Communications 15, 2921.

Lyamuya, R.D., Munisi, E.J., Hariohay, K.M., Masenga, E.H., Bukombe, J.K., Mwakalebe, G.G., Mdaki, M.L., Nkwabi, A.K., Fyumagwa, R.D., 2022. Patterns of mammalian roadkill in the Serengeti ecosystem, northern Tanzania. International Journal of Biodiversity and Conservation .

Majaliwa, M., Hughey, L., Stabach, J., Songer, M., Whyle, K., Alhashmi, A., Al Remeithi, M., Pusey, R., Chaibo, H., Ngari Walsoumon, A., et al., 2022. Experience and social factors influence movement and habitat selection in scimitar-horned oryx (Oryx dammah) reintroduced into Chad. Movement Ecology 10, 47.

Manly, B., McDonald, L., Thomas, D.L., McDonald, T.L., Erickson, W.P., 2007. Resource selection by animals: statistical design and analysis for field studies. Springer Science & Business Media.

Mashayekhi, M., MacPherson, B., Gras, R., 2014. A machine learning approach to investigate the reasons behind species extinction. Ecological informatics 20, 58–66.

Masolele, M.M., Hopcraft, J.G.C., Faust, C.L., Torney, C.J., 2025. Revealing the effects of anthropogenic structures on the spatial distribution of migratory wildebeest. Manuscript in review.

Masolele, M.M., Hopcraft, J.G.C., Torney, C.J., 2024. Efficient approximate Bayesian inference for quantifying uncertainty in multiscale animal movement models. Ecological Informatics 84, 102853.

Matthiopoulos, J., 2003. The use of space by animals as a function of accessibility and preference. Ecological Modelling 159, 239–268.

Matthiopoulos, J., Fieberg, J.R., Aarts, G., 2020. Species-Habitat Associations: Spatial data, predictive models, and ecological insights.

Matthiopoulos, J., Fieberg, J.R., Aarts, G., 2023. Species-Habitat Associations: Spatial data, predictive models, and ecological insights. University of Minnesota Libraries Publishing.

McClintock, B.T., Johnson, D.S., Hooten, M.B., Ver Hoef, J.M., Morales, J.M., 2014. When to be discrete: the importance of time formulation in understanding animal movement. Movement ecology 2, 1–14.

McNaughton, S., 1985. Ecology of a grazing ecosystem: the Serengeti. Ecological monographs 55, 259–294.

McRae, B.H., Dickson, B.G., Keitt, T.H., Shah, V.B., 2008. Using circuit theory to model connectivity in ecology, evolution, and conservation. Ecology 89, 2712–2724.

Mduma, S.A., Sinclair, A., Hilborn, R., 1999. Food regulates the Serengeti wildebeest: a 40-year record. Journal of Animal Ecology 68, 1101–1122.

Mendgen, P., Dejid, N., Olson, K., Buuveibaatar, B., Calabrese, J.M., Chimeddorj, B., Dalannast, M., Fagan, W.F., Leimgruber, P., Müller, T., 2023. Nomadic ungulate movements under threat: Declining mobility of Mongolian gazelles in the Eastern Steppe of Mongolia. Biological Conservation 286, 110271.

Michelot, T., Blackwell, P.G., Chamaillé-Jammes, S., Matthiopoulos, J., 2020. Inference in MCMC step selection models. Biometrics 76, 438–447.

Michelot, T., Blackwell, P.G., Matthiopoulos, J., 2019a. Linking resource selection and step selection models for habitat preferences in animals. Ecology 100, e02452.

Michelot, T., Gloaguen, P., Blackwell, P.G., Étienne, M.P., 2019b. The Langevin diffusion as a continuous-time model of animal movement and habitat selection. Methods in ecology and evolution 10, 1894–1907.

Michelot, T., Klappstein, N.J., Potts, J.R., Fieberg, J., 2024. Understanding step selection analysis through numerical integration. Methods in Ecology and Evolution 15, 24–35.

Millspaugh, J.J., Skalski, J.R., Kernohan, B.J., Raedeke, K.J., Brundige, G.C., Cooper, A.B., 1998. Some comments on spatial independence in studies of resource selection. Wildlife Society Bulletin , 232–236.

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

Morales, J.M., Haydon, D.T., Frair, J., Holsinger, K.E., Fryxell, J.M., 2004. Extracting more out of relocation data: building movement models as mixtures of random walks. Ecology 85, 2436–2445.

Morrison, T.A., Merkle, J.A., Hopcraft, J.G.C., Aikens, E.O., Beck, J.L., Boone, R.B.,

Courtemanch, A.B., Dwinnell, S.P., Fairbanks, W.S., Griffith, B., et al., 2021. Drivers of site fidelity in ungulates. Journal of animal ecology 90, 955–966.

Mouquet, N., Lagadeuc, Y., Devictor, V., Doyen, L., Duputié, A., Eveillard, D., Faure, D., Garnier, E., Gimenez, O., Huneman, P., et al., 2015. Predictive ecology in a changing world. Journal of applied ecology 52, 1293–1310.

Mueller, T., Fagan, W.F., 2008. Search and navigation in dynamic environments–from individual behaviors to population distributions. Oikos 117, 654–664.

Murphy, K.P., 2012. Machine learning: a probabilistic perspective. MIT press.

Nathan, R., Getz, W.M., Revilla, E., Holyoak, M., Kadmon, R., Saltz, D., Smouse, P.E., 2008. A movement ecology paradigm for unifying organismal movement research. Proceedings of the National Academy of Sciences 105, 19052–19059.

Nathan, R., Monk, C.T., Arlinghaus, R., Adam, T., Alós, J., Assaf, M., Baktoft, H., Beardsworth, C.E., Bertram, M.G., Bijleveld, A.I., et al., 2022. Big-data approaches lead to an increased understanding of the ecology of animal movement. Science 375, eabg1780.

Neal, R.M., 2012. MCMC using Hamiltonian dynamics. arXiv preprint arXiv:1206.1901 .

Neu, C.W., Byers, C.R., Peek, J.M., 1974. A technique for analysis of utilization-availability data. The Journal of Wildlife Management , 541–545.

Nickisch, H., Rasmussen, C.E., et al., 2008. Approximations for binary Gaussian process classification. Journal of Machine Learning Research 9, 2035–2078.

Noda, R., Mechenich, M.F., Saarinen, J., Vehtari, A., Žliobaitė, I., 2024. Predicting habitat suitability for Asian elephants in non-analog ecosystems with Bayesian models. Ecological Informatics , 102658.

Noonan, M.J., Tucker, M.A., Fleming, C.H., Akre, T.S., Alberts, S.C., Ali, A.H., Altmann, J., Antunes, P.C., Belant, J.L., Beyer, D., et al., 2019. A comprehensive analysis of autocorrelation and bias in home range estimation. Ecological Monographs 89, e01344.

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

Ovaskainen, O., Meerson, B., 2010. Stochastic models of population extinction. Trends in ecology & evolution 25, 643–652.

Palmer, M.S., Gaynor, K.M., Abraham, J.O., Pringle, R.M., 2023. The role of humans in dynamic landscapes of fear. Trends in Ecology & Evolution 38.

Patin, R., Etienne, M.P., Lebarbier, E., Chamaillé-Jammes, S., Benhamou, S., 2020. Identifying stationary phases in multivariate time series for highlighting behavioural modes and home range settlements. Journal of Animal Ecology 89, 44–56.

Paton, R.S., Matthiopoulos, J., 2016. Defining the scale of habitat availability for models of habitat selection. Ecology 97, 1113–1122.

Paun, I., Husmeier, D., Hopcraft, J.G.C., Masolele, M.M., Torney, C.J., 2022. Inferring spatially varying animal movement characteristics using a hierarchical continuous-time velocity model. Ecology Letters 25, 2726–2738.

Pettorelli, N., Lobora, A., Msuha, M., Foley, C., Durant, S., 2010. Carnivore biodiversity in Tanzania: revealing the distribution patterns of secretive mammals using camera traps. Animal Conservation 13, 131–139.

Phillips, G.E., Alldredge, A.W., 2000. Reproductive success of elk following disturbance by humans during calving season. The Journal of Wildlife Management 64, 521–530.

Picardi, S., Coates, P., Kolar, J., O'Neil, S., Mathews, S., Dahlgren, D., 2022. Behavioural state-dependent habitat selection and implications for animal translocations. Journal of Applied Ecology 59, 624–635.

Piironen, A., Piironen, J., Laaksonen, T., 2022. Predicting spatio-temporal distributions of migratory populations using Gaussian process modelling. Journal of Applied Ecology 59, 1146–1156.

Pohle, J., Signer, J., Eccard, J.A., Dammhahn, M., Schlägel, U.E., 2024. How to account for behavioral states in step-selection analysis: a model comparison. PeerJ 12, e16509.

Potapov, E., Bedford, A., Bryntesson, F., Cooper, S., 2014. White-tailed deer (Odocoileus virginianus) suburban habitat use along disturbance gradients. The American Midland Naturalist 171, 128–138.

Potts, J.R., Bastille-Rousseau, G., Murray, D.L., Schaefer, J.A., 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, 253–262.

Prokopenko, C.M., Boyce, M.S., Avgar, T., 2017. Characterizing wildlife behavioural responses to roads using integrated step selection analysis. Journal of Applied Ecology 54, 470–479.

Quade, D., 1979. Using weighted rankings in the analysis of complete blocks with additive block effects. Journal of the American Statistical Association 74, 680–683.

Railsback, S.F., Grimm, V., 2019. Agent-based and individual-based modeling: a practical introduction. Princeton university press.

Ranganath, R., Gerrish, S., Blei, D., 2014. Black box variational inference, in: Artificial intelligence and statistics, PMLR. pp. 814–822.

Rasmussen, C.E., 2006. Gaussian processes for machine learning. the MIT Press 1, 255–259.

Rieber, C.J., Hefley, T.J., Haukos, D.A., 2024. Treed Gaussian processes for animal movement modeling. Ecology and Evolution 14, e11447.

Riva, F., Graco-Roza, C., Daskalova, G.N., Hudgins, E.J., Lewthwaite, J.M., Newman, E.A., Ryo, M., Mammola, S., 2023. Toward a cohesive understanding of ecological complexity. Science advances 9, eabq4207.

Robb, B.S., Merkle, J.A., Sawyer, H., Beck, J.L., Kauffman, M.J., 2022. Nowhere to run: semi-permeable barriers affect pronghorn space use. The Journal of Wildlife Management 86, e22212.

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jimenez-Valverde, A., Ricotta, C., Bacaro, G., Chiarucci, A., 2011. Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. Progress in Physical Geography 35, 211–226.

Roever, C.L., Beyer, H., Chase, M.J., Van Aarde, R.J., 2014. The pitfalls of ignoring behaviour when quantifying habitat selection. Diversity and Distributions 20, 322–333.

Rowcliffe, J.M., 2017. Key frontiers in camera trapping research. Remote Sensing in Ecology and Conservation 3, 107–108.

Saul, A.D., Hensman, J., Vehtari, A., Lawrence, N.D., 2016. Chained Gaussian processes, in: Artificial intelligence and statistics, PMLR. pp. 1431–1440.

Sawyer, H., Kauffman, M.J., Nielson, R.M., Horne, J.S., 2009. Identifying and prioritizing ungulate migration routes for landscape-level conservation. Ecological Applications 19, 2016–2025.

Schleuning, M., Fründ, J., Schweiger, O., Welk, E., Albrecht, J., Albrecht, M., Beil, M., Benadi, G., Blüthgen, N., Bruelheide, H., et al., 2016. Ecological networks are more sensitive to plant than to animal extinction under climate change. Nature communications 7, 13965.

Schuwirth, N., Borgwardt, F., Domisch, S., Friedrichs, M., Kattwinkel, M., Kneis, D., Kuemmerlen, M., Langhans, S.D., Martínez-López, J., Vermeiren, P., 2019. How to make ecological models useful for environmental management. Ecological Modelling 411, 108784.

Scrafford, M.A., Avgar, T., Heeres, R., Boyce, M.S., 2018. Roads elicit negative movement and habitat-selection responses by wolverines (Gulo gulo luscus). Behavioral Ecology 29, 534–542.

Shannon, G., Larson, C.L., Reed, S.E., Crooks, K.R., Angeloni, L.M., 2017. Ecological consequences of ecotourism for wildlife populations and communities, in: Blumstein, D., Geffroy, B., Samia, D., Bessa, E. (Eds.), Ecotourism's promise and peril: A biological evaluation. Springer International Publish, New York, NY. chapter 3, pp. 29–46.

Shannon, G., McKenna, M.F., Angeloni, L.M., Crooks, K.R., Fristrup, K.M., Brown, E., Warner, K.A., Nelson, M.D., White, C., Briggs, J., et al., 2016. A synthesis of two decades of research documenting the effects of noise on wildlife. Biological Reviews 91, 982–1005.

Signer, J., Fieberg, J., Avgar, T., 2017. Estimating utilization distributions from fitted step-selection functions. Ecosphere 8, e01771.

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

Signer, J., Fieberg, J., Reineking, B., Schlägel, U., Smith, B., Balkenhol, N., Avgar, T., 2024. Simulating animal space use from fitted integrated Step-Selection Functions (iSSF). Methods in Ecology and Evolution 15, 43–50.

Sinclair, A.R., Hopcraft, J.G.C., Mduma, S., Galvin, K., J, G.S., 2008. Historical and future changes to the Serengeti ecosystem. Serengeti III: Human Impacts on Ecosystem Dynamics , 7.

Singh, N.J., Etienne, M., Spong, G., Ecke, F., Hörnfeldt, B., 2024. Linear infrastructure and associated wildlife accidents create an ecological trap for an apex predator and scavenger. Science of the Total Environment 955, 176934.

Sirko, W., Kashubin, S., Ritter, M., Annkah, A., Bouchareb, Y.S.E., Dauphin, Y., Keysers, D., Neumann, M., Cisse, M., Quinn, J., 2021. Continental-scale building detection from high resolution satellite imagery. arXiv preprint arXiv:2107.12283 1.

Smith, J.A., Gaynor, K.M., Suraci, J.P., 2021. Mismatch between risk and response may amplify lethal and non-lethal effects of humans on wild animal populations. Frontiers in Ecology and Evolution 9, 604973.

Smouse, P.E., Focardi, S., Moorcroft, P.R., Kie, J.G., Forester, J.D., Morales, J.M., 2010. Stochastic modelling of animal movement. Philosophical Transactions of the Royal Society B: Biological Sciences 365, 2201–2211.

Snelson, E., Ghahramani, Z., 2005. Sparse Gaussian processes using pseudo-inputs. Advances in neural information processing systems 18.

Snoek, J., Larochelle, H., Adams, R.P., 2012. Practical bayesian optimization of machine learning algorithms. Advances in neural information processing systems 25.

Stabach, J.A., Hughey, L.F., Crego, R.D., Fleming, C.H., Hopcraft, J.G.C., Leimgruber, P., Morrison, T.A., Ogutu, J.O., Reid, R.S., Worden, J.S., et al., 2022. Increasing anthropogenic disturbance restricts wildebeest movement across East African grazing systems. Frontiers in Ecology and Evolution 10, 846171.

Stiegler, J., Gallagher, C.A., Hering, R., Müller, T., Tucker, M., Apollonio, M., Arnold, J., Barker, N.A., Barthel, L., Bassano, B., et al., 2024. Mammals show faster recovery from capture and tagging in human-disturbed landscapes. Nature communications 15, 8079.

Tablado, Z., Jenni, L., 2017. Determinants of uncertainty in wildlife responses to human disturbance. Biological Reviews 92, 216–233.

Thiele, J.C., Grimm, V., 2015. Replicating and breaking models: good for you and good for ecology. Oikos 124, 691–696.

Thurfjell, H., Ciuti, S., Boyce, M.S., 2014. Applications of step-selection functions in ecology and conservation. Movement ecology 2, 1–12.

Titsias, M., 2009. Variational learning of inducing variables in sparse Gaussian processes, in: Artificial intelligence and statistics, PMLR. pp. 567–574.

Torney, C.J., Hopcraft, J.G.C., Morrison, T.A., Couzin, I.D., Levin, S.A., 2018. From single steps to mass migration: the problem of scale in the movement ecology of the Serengeti wildebeest. Philosophical Transactions of the Royal Society B: Biological Sciences 373, 20170012.

Torney, C.J., Morales, J.M., Husmeier, D., 2021. A hierarchical machine learning framework for the analysis of large scale animal movement data. Movement ecology 9, 1–11.

Tucker, M.A., Schipper, A.M., Adams, T.S., Attias, N., Avgar, T., Babic, N.L., Barker, K.J., Bastille-Rousseau, G., Behr, D.M., Belant, J.L., et al., 2023. Behavioral responses of terrestrial mammals to COVID-19 lockdowns. Science 380, 1059–1064.

Tverijonaite, E., Ólafsdóttir, R., Thorsteinsson, T., 2018. Accessibility of protected areas and visitor behaviour: A case study from Iceland. Journal of outdoor recreation and tourism 24, 1–10.

Valeix, M., Loveridge, A., Chamaillé-Jammes, S., Davidson, Z., Murindagomo, F., Fritz, H., Macdonald, D., 2009. Behavioral adjustments of African herbivores to predation risk by lions: spatiotemporal variations influence habitat use. Ecology 90, 23–30.

Veldhuis, M.P., Ritchie, M.E., Ogutu, J.O., Morrison, T.A., Beale, C.M., Estes, A.B., Mwakilema, W., Ojwang, G.O., Parr, C.L., Probert, J., et al., 2019. Cross-boundary human impacts compromise the Serengeti-Mara ecosystem. Science 363, 1424–1428.

Verzuh, T.L., Heuer, K., Merkle, J.A., 2024. Leveraging how animals learn in conservation science: Behavioral responses of reintroduced bison to management interventions. Conservation Science and Practice 6, e13240.

Wang, P., Mihaylova, L., Chakraborty, R., Munir, S., Mayfield, M., Alam, K., Khokhar, M.F., Zheng, Z., Jiang, C., Fang, H., 2021. A Gaussian process method with uncertainty quantification for air quality monitoring. Atmosphere 12, 1344.

Warton, D.I., Shepherd, L.C., 2010. Poisson point process models solve the" pseudo-absence problem" for presence-only data in ecology. The Annals of Applied Statistics , 1383–1402.

Wenger, S.J., Olden, J.D., 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Methods in Ecology and Evolution 3, 260–267.

Wilcove, D.S., Wikelski, M., 2008. Going, going, gone: is animal migration disappearing. PLoS biology 6, e188.

de Wilde, P., de Souza, C.B., 2022. Interactions between buildings, building stakeholders and animals: A scoping review. Journal of Cleaner Production 367, 133055.

Williams, C.K., Rasmussen, C.E., 2006. Gaussian processes for machine learning. volume 2. MIT press Cambridge, MA.

Willis, S.G., Thomas, C.D., Hill, J.K., Collingham, Y.C., Telfer, M.G., Fox, R., Huntley, B., 2009. Dynamic distribution modelling: predicting the present from the past. Ecography 32, 5–12.

Xu, W., Dejid, N., Herrmann, V., Sawyer, H., Middleton, A.D., 2021. Barrier Behaviour Analysis (BaBA) reveals extensive effects of fencing on wide-ranging ungulates. Journal of Applied Ecology 58, 690–698.

Yates, K.L., Bouchet, P.J., Caley, M.J., Mengersen, K., Randin, C.F., Parnell, S., Fielding, A.H., Bamford, A.J., Ban, S., Barbosa, A.M., et al., 2018. Outstanding challenges in the transferability of ecological models. Trends in ecology & evolution 33, 790–802.

Zanette, L.Y., Frizzelle, N.R., Clinchy, M., Peel, M.J., Keller, C.B., Huebner, S.E., Packer, C., 2023. Fear of the human "super predator" pervades the South African savanna . Current biology 33, 4689–4696.

Zhang, C., Bütepage, J., Kjellström, H., Mandt, S., 2018. Advances in variational inference. IEEE transactions on pattern analysis and machine intelligence 41, 2008–2026.

Supplementary materials

S1 Variational Inference estimates

Table S1: Estimates and 95% credible intervals of resource selection parameters (β_1 and β_2) recovered from the simulated synthetic movement data using VI.

Coefficients combinations	No. observations	Parameter	True values	Estimates	Credible intervals
Positive-Negative	10,000	β_1	0.5	0.51	[0.41 , 0.61]
	100,000	β_1	0.5	0.50	[0.48, 0.54]
	1,000,000	β_1	0.5	0.50	[0.49, 0.52]
	10,000	β_2	-0.8	-0.91	[-1.15, -0.67]
	100,000	β_2	-0.8	-0.73	[-0.79, -0.66]
	1,000,000	β_2	-0.8	-0.82	[-0.85, -0.78]
Negative-Negative	10,000	β_1	-1.5	-1.42	[-1.53, -1.32]
	100,000	β_1	-1.5	-1.47	[-1.51, -1.43]
	1,000,000	β_1	-1.5	-1.49	[-1.51, -1.48]
	10,000	β_2	-1.8	-1.67	[-1.96, -1.37]
	100,000	β_2	-1.8	-1.83	[-1.89, -1.76]
	1,000,000	β_2	-1.8	-1.79	[-1.82, -1.76]
Negative-Positive	10000	β_1	-1.5	-1.56	[-1.68, -1.45]
	100,000	β_1	-1.5	-1.49	[-1.54, -1.46]
	1,000,000	β_1	-1.5	-1.50	[-1.52, -1.49]
	10,000	β_2	1.8	1.89	[1.66, 2.12]
	100,000	β_2	1.8	1.72	[1.61, 1.82]
	1,000,000	β_2	1.8	1.82	[1.78, 1.85]
Positive-Positive	10000	β_1	1.2	1.29	[1.18, 1.39]
	100,000	β_1	1.2	1.20	[1.17, 1.24]
	1,000,000	β_1	1.2	1.19	[1.18, 1.22]
	10,000	β_2	1.8	2.01	[1.67, 2.34]
	100,000	β_2	1.8	1.79	[1.73, 1.85]
	1,000,000	β_2	1.8	1.79	[1.75, 1.82]

S2 Hamiltonian Monte Carlo

We used Hamiltonian Monte Carlo (HMC) sampling to recover the parameters used for the simulation of 10,000 observations of synthetic movement data in a 2-dimension geographical space. The movement parameters used for the simulation were positive indicating selection of resources, and negative indicating the avoidance of resources. The number of leapfrog steps, step size, and burn-in steps were 3, 0.1, and 1000, respectively and a step size adaptation algorithm was used during the burn-in phase. The sampler was then run for 10000 steps with 4 independent chains. Convergence and mixing was assessed by calculating effective sample sizes and potential scale reduction factors. Posterior distributions of β_1, and β_2 and summary statistics of recovered parameters are shown in Fig. S1 and Table S2, respectively. As we are using simulated data, this can be compared to the true values used to simulate the movement data. We observe a close agreement between the recovered values and the parameter values used for simulation. Though, is not an exact match, but the true values are contained within the 95% credible intervals of the recovered parameter values. We also note a very close match to the posterior distributions obtained with VI.

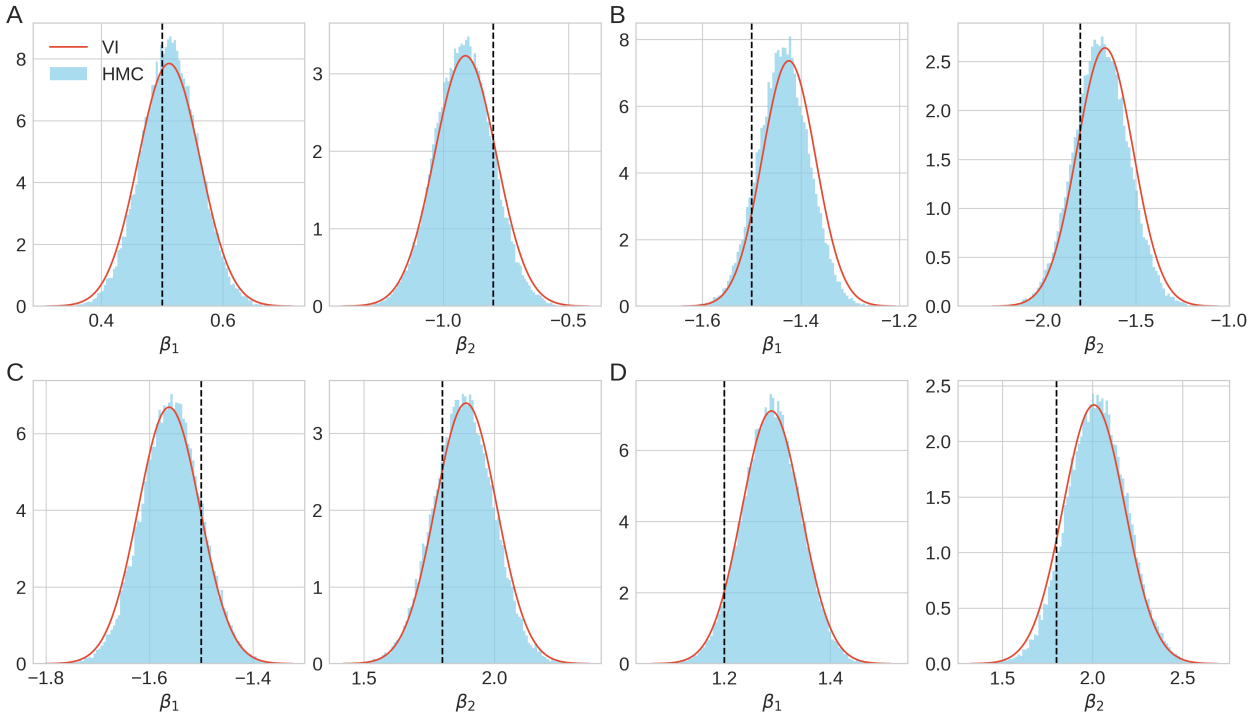

Figure S1: Posterior probability distribution of recovered movement parameters using Hamiltonian Monte Carlo sampling technique and Variational inference from the simulated data with 10,000 observations using a combination of movement parameter values (see Table 3.1) of A) β_1=0.5 and β_2=-0.8, B) β_1=-1.5 and β_2=-1.8, C) β_1=-1.5 and β_2=1.8, and D) β_1=1.2 and β_2=1.8. The vertical dashed line (black in colour) indicates the true values.

Table S2: Estimates and 95% credible intervals of resource selection parameters (β_1 and β_2) recovered from the simulated synthetic movement data using HMC.

Coefficients combinations	No. observations	Parameter	True value	Estimates	Credible interval
Positive-Negative	10,000	β_1	0.5	0.51	[0.42, 0.61]
	10,000	β_2	-0.8	-0.92	[-1.15, -0.69]
Negative-Negative	10,000	β_1	-1.5	-1.44	[-1.54, -1.33]
	10000	β_2	-1.8	-1.69	[-1.98, -1.40]
Negative-Positive	10,000	β_1	-1.5	-1.56	[-1.67, -1.44]
	10,000	β_2	1.8	1.88	[1.65, 2.11]
Positive-Positive	10,000	β_1	1.2	1.29	[1.18, 1.39]
	10,000	β_2	1.8	2.03	[1.69, 2.35]

S3 Buildings simulation

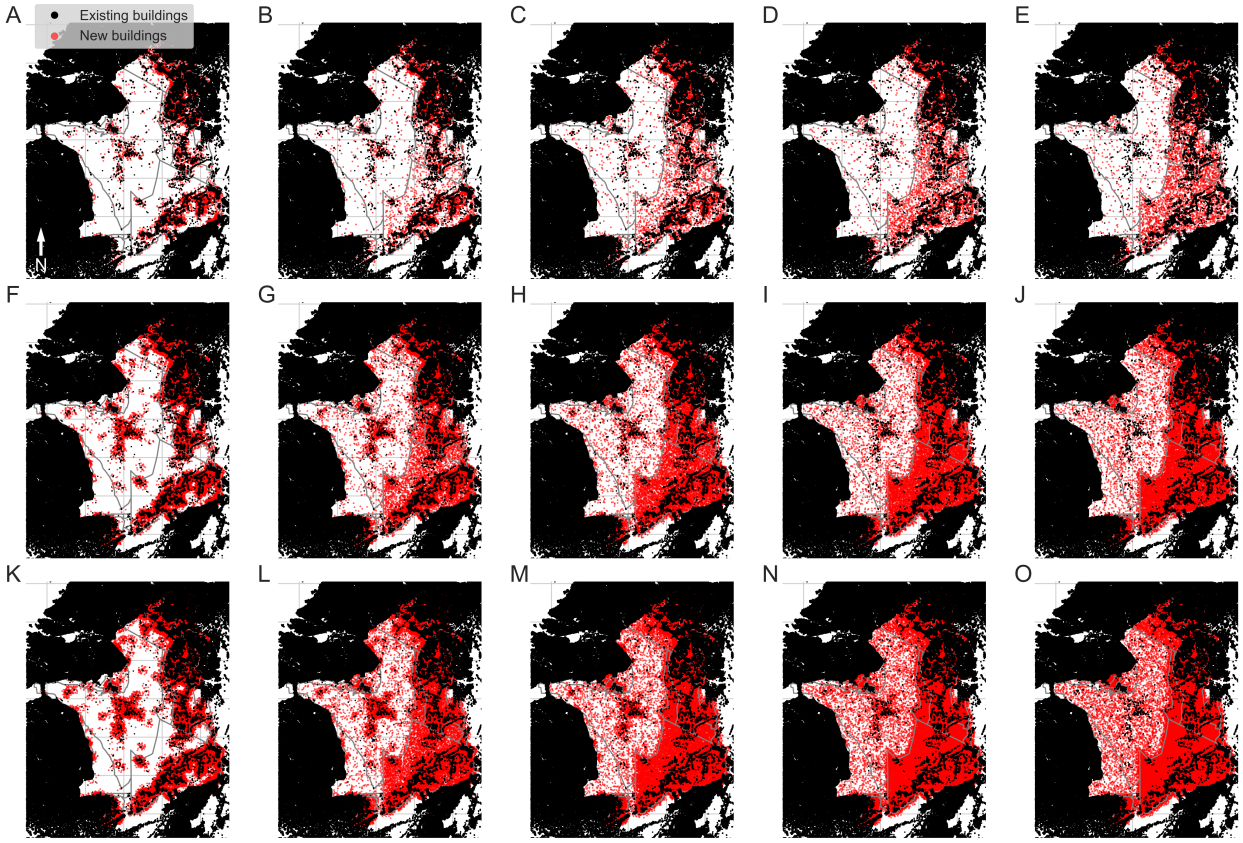

Figure S2: Simulated buildings in the Greater Serengeti-Mara Ecosystem. Top row A-E, middle row F-J and bottom row K-O represent 10%, 50%, and 100% increase of existing buildings, respectively. Note, the distribution is changing from left to right due to decreasing of buildings clustering with δ values of 0, 0.25, 0.5, 0.75, and 1. The preferential attachment parameter used during the simulation of buildings was $\boldsymbol{\alpha} = 0$. Black dots indicate the locations of existing buildings and red dots indicate the simulated locations of new additional buildings in the ecosystem. Gray lines indicate the boundary of the Serengeti ecosystem and associated protected areas.

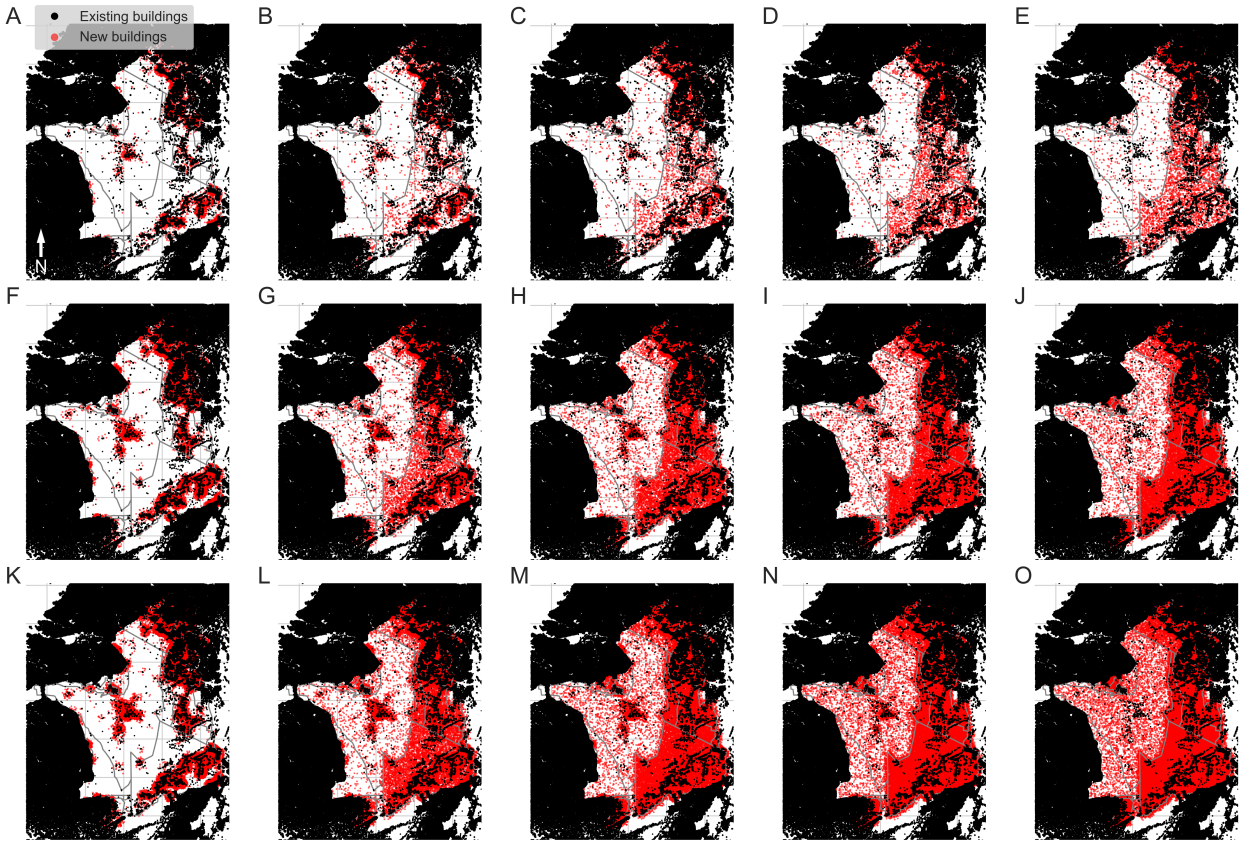

Figure S3: Simulated buildings in the Greater Serengeti-Mara Ecosystem. Top row A-E, middle row F-J and bottom row K-O represent 10%, 50%, and 100% increase of existing buildings, respectively. Note, the distribution is changing from left to right due to decreasing of buildings clustering with δ values of 0, 0.25, 0.5, 0.75, and 1. The preferential attachment parameter used during the simulation of buildings was $\boldsymbol{\alpha} = 1$. Black dots indicate the locations of existing buildings and red dots indicate the simulated locations of new additional buildings in the ecosystem. Gray lines indicate the boundary of the Serengeti ecosystem and associated protected areas.

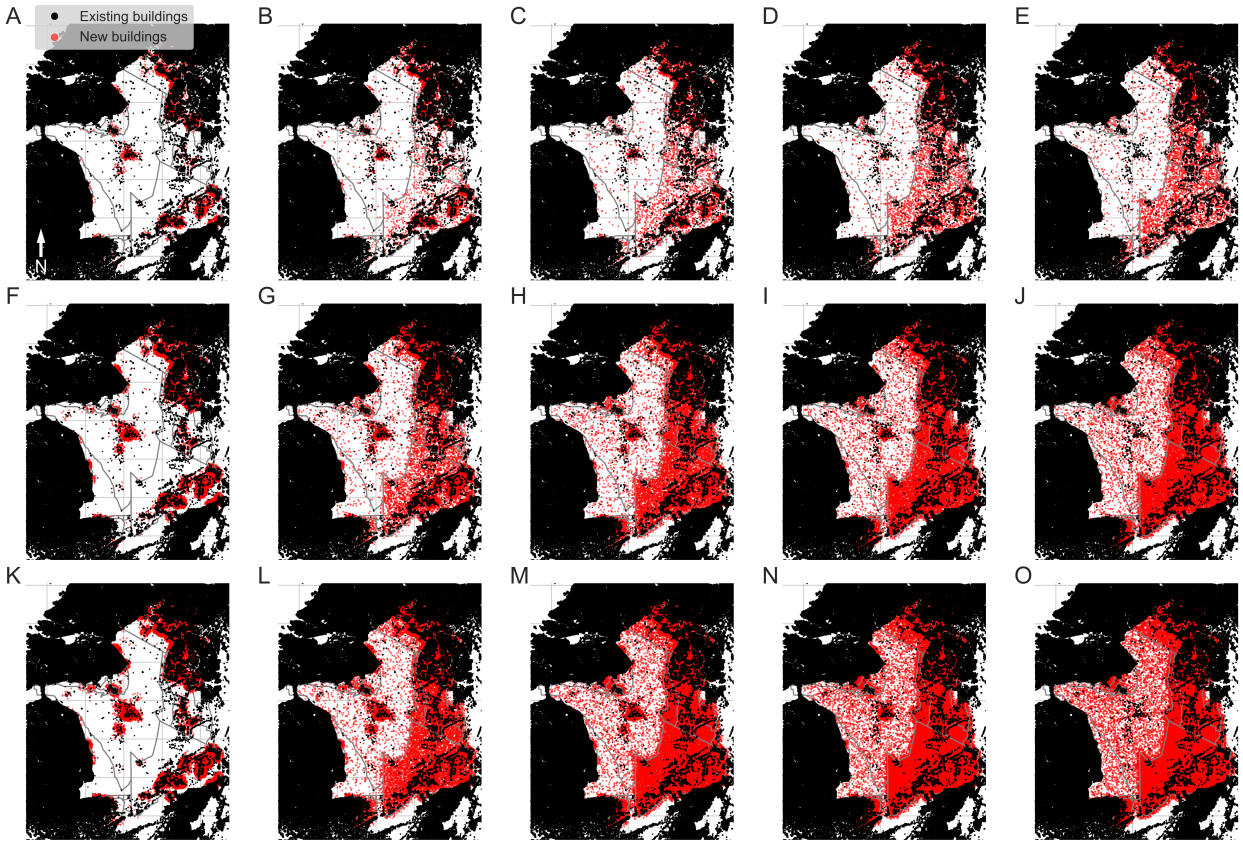

Figure S4: Simulated buildings in the Greater Serengeti-Mara Ecosystem. Top row A-E, middle row F-J and bottom row K-O represent 10%, 50%, and 100% increase of existing buildings, respectively. Note, the distribution is changing from left to right due to decreasing of buildings clustering with δ values of 0, 0.25, 0.5, 0.75, and 1. The preferential attachment parameter used during the simulation of buildings was $\boldsymbol{\alpha} = 2$. Black dots indicate the locations of existing buildings and red dots indicate the simulated locations of new additional buildings in the ecosystem. Gray lines indicate the boundary of the Serengeti ecosystem and associated protected areas.

S4 Parameters inferred from multiscale step selection Model

Table S3: Estimates and 95% credible intervals of parameters used for the estimation of simulated wildebeest space use. β is the coefficient value of wildebeest selection, ω is a parameter that indicates the diminishing effects of the subsequent buildings, λ and γ quantify the spatial extent of the influence of the buildings on wildebeest.

Parameters	Estimates	Credible interval
$\hat{\beta}$	-0.059	$[-0.076, -0.041]$
$\hat{\gamma}$	6.092	$[6.011, 6.173]$
$\hat{\lambda}$	1.617	$[1.262, 1.972]$
$\hat{\omega}$	0.776	$[0.667, 0.884]$