Species-Habitat associations

Spatial Data • Predictive Models • Ecological Insights

bookdown

Jason Matthiopoulos • John Fieberg • Geert Aarts

2

Suggested Citation:

Matthiopoulos, Jason; Fieberg, John; Aarts, Geert. (2020). Species-Habitat Associations: Spatial data, predictive models, and ecological insights. University of Minnesota Libraries Publishing. Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/217469.

Related Works:

A copy of the book, which we plan to continuously update (with new versions in the future) can be accessed in gitbook format at: https://bookdown.org/jfieberg/SHABook/.

Cover photograph:

Guanacos, a camelid native to South America, grazing in Torres del Paine National Park in the Pantagonia region of Chile. ©Gary R. Jensen, www.GaryRobertPhotography.com.

License:

Contents

About the Authors

Jason Matthiopoulos

Jason Matthiopoulos is Professor of Spatial and Population Ecology at the University of Glasgow, Scotland. He has a joint degree in mathematics and zoology, a PhD in population ecology, qualifications in statistics, and almost 30 years' experience in teaching mathematics, statistics and computing to biology students at all levels. His research interests revolve around population dynamics and spatial ecology and he has published widely on taxa as diverse as invertebrates, fish, birds and mammals. He lives with his wife and son in suburban Glasgow and enjoys listening loudly to music that most people hate and playing the bass poorly along with music most people like.

John Fieberg

John Fieberg is an Associate Professor of Quantitative Ecology and McKnight Presidential Fellow in the Department of Fisheries, Wildlife, and Conservation Biology at the University of Minnesota where he teaches undergraduate and graduate-level courses in statistics. Prior to joining the faculty at the University of Minnesota, he worked for ten years as a research statistician with the Minnesota Department of Natural Resources and two and a half years as a Biometrician with the Northwest Indian Fisheries Commission. In these positions, he had the pleasure of collaborating with many highly engaged fisheries, wildlife, and conservation biologists on a wide range of applied problems. These experiences have shaped his approach to problem solving and helped develop his ability to communicate difficult concepts to diverse audiences. When not at work, he enjoys playing ultimate frisbee with his wife, watching his girls play hockey, playing guitar with his son (a drummer), and traveling and camping with his family.

Geert Aarts

Geert Aarts is a researcher at Wageningen Marine Research and the Department of Wildlife Ecology and Conservation of Wageningen University, and he is a postdoctoral researcher at the Royal Netherlands Institute for Sea Research. His main field of interest is quantitative marine ecology and wildlife conservation, with a special interest in studying marine mammals. When not at work, he enjoys being outdoors in nature and to (trail)run, cycle, canoe or swim. He lives with his family on the Wadden Sea island Texel, located in the Netherlands.

Acknowledgments

Abbreviations

Term	Abbreviation
Generalised Additive Model	GAM
Generalised Linear Mixed effects Model	GLMM
Generalised Linear Model	GLM
Geographic Information System	GIS
Homogeneous Poisson Process	HPP
Ideal-Free Distribution	IFD
Inhomogenenious Poisson Process	IPP
Kernel Density Estimator	KDE
Maximum Entropy	MaxEnt
Principal Components Analysis	PCA
Resource-Selection Function	RSF
Species-Habitat Association	SHA
Species Distribution Model	SDM
Step-Selection Function	SSF

Glossary

Accuracy: The lack of bias in an estimator of a model **parameter** or model prediction.

Allee effects: A discounting of population growth rates in small populations. It is a form of density dependence (often known as depensatory density dependence) which populations can overcome when they benefit from a critical size. For instance, in small (or low-density) populations, Allee effects may originate from the inability of individuals to encounter suitable breeding mates, or to form efficient hunting packs.

Antagonistic resources: Two resources that, when taken together, offset the benefits of each other.

Available locations or **Background locations**: Locations in geographic or environmental space that are assumed to be accessible to individuals under study. These are locations that moving animals or dispersing plants can reach but may not eventually be found in.

Camera trap: A stationary, autonomous camera, triggered by movement.

Climate envelope models: A model of the distribution of a species with particular applications to climate change. Model predictions are based on a species' tolerances to different climate covariates (e.g. temperature or precipitation). Tolerances are represented as deterministic or probabilistic ranges (or, envelopes).

Colonization credit: The opposite of **extinction debt**. The number of species or members of a species that can reestablish in a region, and will do so given enough time.

Complementary resources: Two resources that, when taken together, have a more beneficial effect (per resource) than when taken individually.

Condition: Environmental variable (such as ambient temperature, humidity, salinity or pressure) that influences an organism's functioning. Extreme values of conditions (e.g. very high or very low temperatures) are usually detrimental to an organism, so responses to conditions are often modeled with non-linear (unimodal) functions.

Demography: The study of birth, death, immigration and emigration rates at the level of populations.

Demographic sorting: The process of habitat-driven changes in the vital rates (i.e. survival, growth and reproduction) that lead to differences in species density between habitats.

Density, species: The observed or expected number of individuals per unit of area. See also **intensity function**.

Density estimation method: A statistical method for estimating the number organisms per unit of area in a landscape, often relying on smoothing or spatial interpolation between observations.

Depletion: The reduction of resource density caused by the action of an organism (e.g. through consumption or occupation).

Design matrix: A matrix containing the values of explanatory variables for a set of observations. Each row represents a different observation and the columns contain the specific values of the explanatory variables associated with that observation. The design matrix provides a convenient shorthand for describing and implementing regression models.

Deterministic variable: A variable whose exact value can be predicted from a mathematical model (to a predetermined degree of precision) without the need for it to be measured experimentally.

Dirichlet tessellation: Divides a two-dimensional geographic space into a set of geometric shapes or tiles with no overlap or gaps. In a spatial point pattern, the Dirichlet tile associated with a particular point is the region of space that is closer to that point than to any other point. The Dirichlet tessellation is also known as the Voronoi or Thiessen tessellation.

Dispersal: The movement of members of a species (or their propagules) away from their birth site.

Distance sampling: Methodology for estimating density or population abundance using point or line-transect sampling. The distances of individuals to observers is used parameterize detection functions, which describe how the probability of detection depends on distance from the observer.

Extinction debt: The opposite of **colonization credit**. The number of species that are doomed to (local) extinction due to environmental change, but have yet to become extinct. In relation to species distributions, it implies that SHA-models predict zero density for a region, but remnants of the species are still present.

Ecosystem engineers: A species that significantly modifies its environment, thereby shaping the distribution and abundance of other species.

Empirical model or **Phenomenological model** A mathematical model whose behavior emulates the behavior of a physical phenomenon, but whose mathematical structure has not been derived from first principles. All mathematical models are, at some level, empirical, particularly in ecology, because the reductionist approach of specifying every aspect of a model from first principles leads to unnecessarily complicated models. (Contrast with **Mechanistic model**). Classic examples of empirical models are statistical **regression** models, but many of the time-honored physical models (e.g. Newton's law of gravitation) are also empirical.

Environmental Space (E-space): The space whose dimensions are environmental variables (compare with **Geographical space**).

Environmental variable: A measurable environmental trait that, for a particular focal species, could represent a **resource**, a **condition** or **risk**. In SHA models, the term is synonymous with **explanatory variable** or **covariate**.

Equilibrium assumption: The assumption that sufficient time has elapsed for a species to reach a stable spatial distribution. This assumption is vulnerable not just because of transient dynamics, but also due to persistent, long-term instabilities in species-habitat associations.

Fitness, partial: The **fitness** of a population living in an environment made up entirely of a single **habitat** (a set of resources, conditions and risks).

Fitness, individual: From a demographic viewpoint, the log of the combined effect of life-long survival and reproductive success of an individual. Equivalently, from an evolutionary perspective, the log of the individual's contribution to the gene pool of the next generation (note that this omits the effects of inclusive fitness).

Fitness, population: The average **fitness of individuals** in a population. In low densities, the population fitness afforded by the environment is equal to the population's **intrinsic growth rate** (excluding the possibility of **Allee effects**).

Fitness integration: The process whereby organisms accumulate resources, reduce risk and mitigate environmental conditions in space or time. As a result, viable populations may be observed in landscapes where no single point in space or time is characterized by positive fitness.

Functional response: In SHA models, the idea that the response of an organism to a particular environmental variable depends critically on the values of other environmental variables. This definition is slightly different to the more often encountered use of the terr in trophic ecology, where a functional response describes the **intake rate** of a food type by a single consumer, in response to the abundance of that (as well as other) food types.

Gaussian random field: A stochastic process (or collection of random variables in time or space) that have a multivariate Normal (Gaussian) distribution characterized by a mean function and covariance function.

Geographical Space (G-space): The space defined by physical dimensions (e.g. latitude, longitude and altitude/depth).

Habitat: A point in **Environmental space** (defined by a set of resources, risks and conditions).

Habitat availability: The relative proportion of habitats making up the region of **G-space** that is accessible to an organism or population.

Habitat context: The habitat composition within a neighborhood of a particular habitat. This is influenced by the autocorrelation within and cross-correlations between the environmental variables in **G-space**.

Habitat selection: The process whereby an organisms uses a habitat disproportionately more (or less) than that habitat's availability. Together with **Demographic sorting** it shapes species distributions.

Habitat usage: The proportion of an individual's time or the proportion of a population associated with a single unit (e.g. m^2) of a particular habitat.

Hessian: Matrix of second derivatives of the log-likelihood with respect to model **parameters**. The inverse of the Hessian provides an estimate of the variance-covariance matrix of parameters when estimated using **maximum Likelihood**.

Homogeneous Poisson point Process (HPP): A model for locations or events in geographic space. The number of events in a fixed region, G, is assumed to be Poisson-distributed with rate λ. In an HPP, the rate is assumed to be constant across space and time.

Ideal-Free Distribution (IFD): A conceptual model of density-dependent resource exploitation in a patchy environment. Individuals are aware of the quality of each and every patch (i.e. individuals are *ideal*) and *free* to enter any patch. As a result, they settle in the most profitable patch. In the special case where the quality of a patch decreases linearly with the density of animals in the patch, the equilibrium density of animals will be proportional to the intrinsic habitat quality of all occupied patches.

Indiscriminate: Hypothetical organisms that do not appear to be selecting one habitat over another. Equivalently, organisms that use **G-space** uniformly randomly.

Inhomogeneous Poisson point Process (IPP): A model for locations or events in geographic space where the expected density of points depends on local spatial predictors through a spatially-varying **intensity function**. The number of events in a fixed region, G, is assumed to be Poisson distributed with mean given by the average intensity function over the region.

Intake rate: The uptake of food or other resources per unit of time.

Integrated data models: Models fit simultaneously to multiple data sets, often with the goal of increasing spatial coverage, precision, and accuracy of estimators of species distributions by addressing issues related to sampling bias or imperfect detection.

Intensity function: The **Intensity surface** described as a function of spatially-varying environmental variables.

Intensity surface: The expected spatial **density** of individuals or observations.

Intrinsic population growth rate: The population growth rate without any density dependence.

Landscape of fear: Quantifies an organism's perception of risk in space and often has an influence on its distribution.

Likelihood: A statistical expression that describes the data generating mechanism in terms of one or more parameters.

Logistic population model: A population model that formulates the rate of population growth as a decreasing function of population density. It therefore incorporates conspecific crowding effects acting through biological processes such as competition or cannibalism.

Marginal Value Theorem: Theorem that states an animal should leave its current habitat patch when that patch's quality (measured in terms of energy gain per unit time) falls below the average quality of other habitat patches available to the animal.

MaxEnt or **Maximum entropy**: A use-availability method (or software) for modeling species distributions. It uses entropy as a model fitting criterion, as opposed to **likelihood**.

Maximum Likelihood: A method for estimating parameters in a statistical model. Parameter values are chosen to make the likelihood of the data as large as possible.

Mechanistic model: A model whose mathematical form is derived, to some extent, from physical or biological first principles. For example, in dispersal studies, deriving a diffusion kernel from a model of Brownian motion adds mechanistic content to the process of dispersal. In mechanistic models, the participating parameters often have a clear physical interpretation (contrast with **Empirical model**).

Monte-Carlo integration: A simulation-based method for approximating an integral when a closed-form analytical solution may not exist.

Niche, fundamental: The set of points in **E-space** (**habitats**) that allow the intrinsic growth rate of a population to be positive.

Niche, realized: The collection of all habitats where a species has been found.

Niche, Grinnellian: See **fundamental niche**, but with the inclusion of other influential species (e.g. prey, predators) as resource or risk dimensions of the environmental space.

Niche, Eltonian: See **fundamental niche**, but with the inclusion of bi-directional relationship between the focal species and other environmental variables and non-focal species, and thereby acknowledging that a species also influences the availability and behavior of properties of its environment.

Null model, general: A mathematical construct, derived from a set of baseline assumptions. Often it is a simplification of a physical process that forms the basis for further model development, as the baseline assumptions become increasingly relaxed.

Null model, SHA: The equilibrium density of a species at habitat **x** being proportional to the intrinsic growth rate at low density in habitat **x** (also see **Pseudo-equilibrium assumption** and **Ideal-free distribution**).

Numerical integration: A formal set of rules applied to an integral that allows us to approximate its value when a closed-form analytical solution is not available.

Occupancy: The presence of (at least one member of) a species within a predefined spatial region and time window.

Occurrence distribution: Statistical distribution describing the position of an animal during a specific observation time window.

Offset: A predictor variable with regression coefficient fixed at the value 1, often used to account for varying (but known) levels of observation effort in time or space.

Opportunistic sampling: Data collected without a formal sampling design. An example is a web site that allows citizen scientists to report any observed locations of a species of interest.

Parameter: A numerical quantity that participates in a model and (together with the model's mathematical structure and any initial conditions), shapes the model's behavior. Parameters are not observed directly and are usually assumed not to change. Instead, they are estimated via statistical inference methods. So, parameters should not be confused with variables, or covariates.

Phenotype: The composite observable characteristics or traits of an organism.

Probabilistic survey: A survey that samples a fixed region and uses a randomized design for allocating survey effort.

Precision: The level of uncertainty in a **parameter** estimate or model prediction.

Point Process Events: (or points, locations): observations of animals in space.

Population closure: An assumption often made when estimating abundance or occupancy that the sampled population is not changing during the observation period (i.e., there are no births, deaths, or immigration or emigration).

Predictor function: A function on the log scale, describing the effect of environmental variables on habitat selection. It is equivalent to the link-scale predictor in GLMs and GAMs.

Principal component analysis (PCA): An axis-rotation technique for multivariate data that results in a new set of variable definitions. These new variables are orthogonal and each accounts for a decreasing amount of variation in the original data (from maximum to minimum). PCA is used to eliminate collinearity between candidate covariates and is sometimes used to achieve dimension-reduction, by only retaining those variables that account for most of the variation in the original data.

Profile method, SDM: In the species distribution literature, this is a name occasionally used to describe a model that predicts **habitat usage** while ignoring variations in **habitat availability**.

Process component: The process component of a SHA model deals with the biological phenomena of interest. Its parameters should be interpretable in terms of organism responses to their environment, their density or their conspecifics.

Pseudo-equilibrium assumption: The assumption that the density of species at any point in a landscape responds to the measured habitat covariates with no delay, and that this is a reflection of the species fitness at that point.

Quadrature points: Points within the domain of integration at which a function is evaluated when approximating an integral numerically. See **numerical integration**.

Quadrature weights: Weights given to different points when evaluating an integral numerically. See **numerical integration**.

Random variable or **Variate**: A measurable quantity whose value is unknown before a measurement is taken (contrast with **Deterministic variable**).

Range, Semivariogram: The distance (in time or space) over which observations are correlated.

Range distribution: The asymptotic (or equilibrium) distribution that results from assuming animals move consistently in a range-restricted manner.

Raster: A rectangular grid containing data, with an associated spatial extent and resolution or cell size.

Regression methods: Statistical models that relate the mean value of one (response or dependent) variable to the values of several other (explanatory or independent) variables.

Resource: A substance, object or place required by an organism for growth, maintenance and reproduction, and whose quantities may be reduced by the organism. In contrast to **conditions** and **risks**, resources are assumed to always have a positive effect on fitness in the context of a SHA model.

Resource (standing stock) density: The density of a resource at a specific point in space and time. This value is most influential on the **intrinsic growth rate** or **population fitness**.

Resource productivity: The change in resource density per unit of area and time. This variable is most influential on the equilibrium density of a species.

Resource selection function: A weighting function that describes the relative likelihood of selecting a location as a function of its environmental characteristics; equivalent to the intensity function of a Inhomogeneous Poisson point-process model with the intercept removed.

Ricker model: A particular version of the **logistic population model** which has the density-dependent part inside an exponential function.

Risk: Environmental variable that has a negative relationship with fitness by lowering the actual or perceived chances of individual survival or reproduction.

Robustness, model: The desirable combination of **accurate** and **precise** predictions from a model even when its assumptions are violated to some extent.

Semivariogram: A function that measures the degree of statistical dependence between observations as a function of their distance (in space or time). It displays expected variability between two points as a function of their distance, so it increases from low values to a high asymptote (the **sill**, representing baseline variability between distant, independent points).

Sill, Semivariogram: The asymptote of a semivariogram; describes the spatial or temporal variance for locations that are far enough to be statistically independent.

Spatial point-process model: A model for the data-generating process associated with locations or events in geographic space.

Substitutable resources: Two resources that contribute similarly to an individual's fitness, such that an organism can replace one lost unit of one resource by a fixed number of units of the other resource.

Species distribution model (SDM): A model that captures variations in species density as a function of spatial coordinates or environmental covariates.

Species-habitat association (SHA): A model that connects aspects of habitat (resources, conditions, risks) to particular observations about a species (recorded at the individual, group or population levels). Most often, the observations relate to species densities (see **Species distribution model**), but can also relate to population fitness or *changes* in species density. The main difference between a SHA and SDM is that in SHA model we acknowledge and account for species distributions not being at equilibrium.

Step-selection function: A weighting function that describes the relative likelihood of selecting a location as a function of its environmental characteristics; used in models that characterize time-specific habitat availability using a model of animal movement.

Synoptic SHA model: A model of space use that incorporates both home range and resource-selection processes.

Telemetry: A collection of measurements (animal locations and other physiological readings) obtained remotely.

Thinned point-process model: A point-process model that assumes only a subset of locations or events in space are observed. The "thinning" process provides a way to model observation biases. Specifically, the full set of locations are "thinned" based on model mechanisms, resulting in the data set that is available for analysis.

Transferability of model predictions: The ability to use a model to make accurate and precise predictions in new time and place.

Use-availability scheme: A **SHA model** that combines information on **habitat usage** together with **habitat availability** in order to quantify **habitat selection** and predict usage in a **transferable** way.

Utilization distribution: Or **usage**. A spatial probability distribution describing the expected time of an individual or the expected density of a population at a particular place.

Weighted distribution theory: A framework for modeling distributions of observed random variables that are influenced by various forms of selection biases.

Zero-niche paradox: The ability of animals to survive in environments where no single point affords positive fitness, via complementary use of resources from different locations (also see **Fitness integration**). An equivalent idea can be considered for plants, whereby integration of fitness happens across instants in time (e.g. across seasons).

Notation

Notation	Term	Explanation
$f_a(x)$	Distribution of available habitat in environmental space	describes risks, recources, and conditions at available locations in environmental space
$f_a(s)$	Distribution of available habitat in geographic space	describes risks, resources, and conditions at available locations in geographic space
$f_u(x)$	Distribution of used habitat in environmental space	describes risks, recources, and conditions at used locations in environmental space
$f_u(s)$	Distribution of used habitat in geographic space	describes risks, resources, and conditions at used locations in geographic space
$\lambda(s)$	Intensity function of an Inhomogenenious Poisson Process Model	Describes the expected number of points within an infinitely small area surrounding location s

Preface

0.1 A "live" project

Developments in the literature linking species to their environment are rapid and multifaceted (we might even say, splintered). The area of species distribution models was recently ranked as one of the top 5 research fronts in ecology and the environmental sciences by ISI's Essential Science Indicators (Renner & Warton, 2013). At the same time, the actors and audience in this area are a sufficiently focused group of scientists to be accessible as a coherent community. To enhance this coherence, for some time, we have felt the need for a synthetic approach in this area but, crucially, one that can remain freely-available and unfossilized. Our chosen publication model is therefore one of online distribution through non-proprietary electronic archives. This is a double-edged blade. On the one hand, you the reader, may benefit from an accessible and current monograph. On the other, some of the chapters that tie up several of the vital plotlines of our narrative may be missing in the early editions. So, please be patient in your reading. This story is unfolding as it is being written...

0.2 Audience

We envision this book will be of interest to:

- Graduate students and professionals looking for a clear introduction to the ecological and statistical underpinnings of Species-Habitat-Association (SHA) models.

- Practitioners with data and an interest in learning about animal movements or species distributions and looking for guidance (and R code!).

- Quantitative ecologists looking to contribute new methods addressing the limitations of the current incarnations of SHA models.

We have made only modest assumptions about the prior knowledge of our readers. Of our ecological readership we envisage a basic understanding of statistical inference (to the level of Generalized Linear or Additive Models) and of our statistical readership we assume an exposure to the key questions in spatial and population ecology.

We therefore hope to provide field ecologists and theoreticians with guidance on how to avoid pitfalls in statistical inference and biological interpretation, by describing the limitations and output of available frameworks for studying SHAs. Statistical models that link species distribution data with environmental variables have become so easy to fit and can produce such compelling maps, that it is easy to neglect to pause and consider what these maps mean. Interpretation of any model requires us to trace its form and function back to its fundamental assumptions and the physical meaning of its participating variables. This task is not easy in the case of phenomenological models (i.e. models that put more emphasis on form, rather than function), such as the vast majority of statistical models available to us today. Nevertheless, failing to meet this challenge can impede, or misdirect, efforts of conservation for threatened species and programs of elimination of pests and disease.

0.3 Objectives

Our overarching objective is to describe the state of the art in models that connect the spatial distribution of species to their environment, while incorporating as much biological mechanism as possible in these mathematical and statistical formalisms. More specifically, we aim to:

1. Highlight the importance of well-known, but often neglected, ecological concepts and their role in driving dispersal, movements, population dynamics and species distributions.

2. Synthesize and connect parallel modeling frameworks developed to infer the importance of biotic and abiotic variables on species distributions.

3. Motivate the development of new analytical methods that better capture ecological mechanisms, thereby improving the predictive abilities of SHA models.

SHAs need to go beyond describing the apparent correlations between observed density of organisms and the local conditions they happen to be observed in. As ecologists, we need to push the envelope of our existing correlational models to include cornerstone ecological concepts such as the fundamental niche of a species, the ideal free distribution, density dependence, resource depletion, population dynamics, landscapes of fear and numerous others. We have therefore aimed for an approach that does not shy away from the details of mathematical models but neither does it deprive the reader of the crucial biological motivations that lead to those models. We aspire to account for processes that span the hierarchy of ecological complexity, from the movement and behavior of individual organisms, through processes that regulate population size and distribution, all the way to community interactions between species. Above all, we aim to be synthetic, bringing together several of the apparently disconnected pieces of ecological wisdom and statistical technique in our field's literature. This unifying approach is the pre-eminent aspiration of our book and has been attempted along four different axes (taxonomy, scale, statistical and mathematical methodology), as we explain below.

0.4 Why is this book unique?

By drawing parallels and intersections between plants and animals, the material presented here tries to *unify our conceptual and methodological approach to SHAs across living taxa*. By doing so, we hope to generate some cross-fertilization between the disparate analytical approaches traditionally used to investigate the distribution and habitat preferences of sessile and mobile species. Indeed, we aim to extend movement-related models and habitat selection concepts that are usually associated solely with animals to plant movement and selection over longer time scales, between life-stages, between generations, or even across different morphs along the evolutionary timeline.

We also try to achieve *unification across scales*. Species distributions are dynamic and spatially structured, but the data we collect often cover a specific spatial region or time window. For example, tracking an animal moving within a well-defined home-range, may tell us something about the behavior and resource selection of that individual, but may not be informative about how that animal established its home range there. Alternatively, using museum records, we may look into the distant past, and capture global distributions, but these distributions may not help us understand fine-scale selection, and they may also not match the current distribution of a species following recent anthropogenic change. To address these issues, we need to consider hierarchical spatial models that allow for non-equilibrium dynamics.

This book also tries to achieve *convergence between existing statistical methodologies*. Our feeling is that the literature currently comprises many methods that are already well connected, a few methods that represent analytical dead-ends and several methods that merely appear isolated and are waiting to be linked to our main body of work via re-interpretation. Hence, rather than offer a mixed bag of quantitative recipes, we have selected available and emerging methods that we feel combine into a coherent and expandable framework.

Finally, we try to achieve *unification of modeling paradigms* by bringing together mathematically formulated mechanisms with the statistical machinery needed for extracting information from field data. For example, classic mechanistic models from the 1970s, such as the ideal free distribution (Fretwell & Lucas, 1969), have

connected environmental productivity with species distribution by considering the behavior of ideal individuals (Křivan, Cressman, & Schneider, 2008). In more recent years, flexibility in species-habitat association models has come from statistical approaches, like generalized additive models (Wood, 2006) using smooth functions of covariates, whose constraints are data-driven. We can gain much by replacing the extremely flexible smooth functions with pre-defined dependencies between the organism and the environment, e.g. by motivating the functional forms of models from biological first principles, or by including informative priors for some of our model parameters.

0.5 Why model species habitat associations?

There are many specific reasons why ecologists are interested in developing SHA models, but, in their essence, these can be split into three broad categories. Given a sample of spatial observations from a species (or a population, a social group, or even a single individual) together with environmental data from the same region in space and time period, we aim to quantify: 1. where organisms occur (spatial estimation); 2. why they occur there (inference); 3. where else they might occur (prediction); There is a clear ramp-up in difficulty in these questions. In its simplest form, spatial estimation is purely pattern-based whereas prediction (in space or time) arguably requires deeper insights into behavioral, energetic, demographic and community mechanisms.

0.5.1 Spatial estimation

Spatial estimation or 'map-making' could be achieved by means of density estimation methods, using only the spatial coordinates of observed locations (e.g., various smoothing approaches such as kernel or spline methods; Silverman, 1986). Species distribution maps can be valuable for conservation and population management purposes. For instance, by highlighting where certain (rare) species occur, they can assist in the designation of protected areas (Moilanen, Wilson, Kerrie, & Possingham, 2008). Maps can also be used to quantify the impact of anthropogenic activities on wildlife by estimating direct encounters (e.g., collisions of birds with wind turbines, wildlife-vehicle collisions along road networks, bycatch of seabirds or marine mammals in longlines and fishing nets), or sub-lethal effects (e.g. impact of military sonars on marine mammals, exclusion of foragers from valuable food sources, alteration of migration routes due to climate change). Potentially, maps can be used by ecologists as stepping-stones for further analysis. E.g., when studying predator-prey interactions, a previously estimated density map of prey could be used as an explanatory variable for the distribution of the predator.

0.5.2 Inference

For the second aim of understanding why certain organisms occur where we observe them, a relationship needs to be established between the distribution of organisms and relevant environmental variables that surround the organisms. For example, plant distribution modeling may be used to quantify the temperature or soil pH ranges within which the study species occur, or to investigate their tolerance for extreme events, like droughts or inundation (Sarker, Reeve, Thompson, Paul, & Matthiopoulos, 2016). This has led plant ecologists to think primarily in terms of physiological tolerances, and environmental envelopes (Pearson & Dawson, 2003). In contrast, animal ecologists use selection functions, such as Resource-Selection Functions (Boyce & McDonald, 1999) or Step-Selection Functions (Thurfjell, Ciuti, & Boyce, 2014) to quantify which combinations of environmental attributes animals select from a list of available options. These models provide insights into why and how animals migrate (McClintock et al., 2012; Börger et al., 2013), what drives the distribution of their breeding sites (e.g. in colonial marine predators) (Robinson et al., 2017), and possible symbiotic (or exclusion) effects between species (Ovaskainen & Abrego, 2020).

0.5.3 Prediction

The third aim, predicting species distributions in space and time (Elith & Graham, 2009), is probably the most challenging, and the most reliant on the successful completion of prior steps (i.e. a good model fit to observed density using sufficiently insightful environmental variables). Applications related to prediction are targeted at vital questions, such as species range expansion or contraction (Scheele, Foster, Banks, &

Lindenmayer, 2017), risk assessment for invasive species (Gallien, Douzet, Pratte, Zimmermann, & Thuiller, 2012), recovery after temporary displacement (Russell et al., 2016), and redistribution following permanent displacement (Street et al., 2015), as well as classic questions regarding impacts of habitat destruction or fragmentation due to human infrastructure development (Beyer et al., 2016). All of these apparently divergent questions are driven by a desire to have models that are transferable to novel environments in space or time (Yates et al., 2018) - i.e. we want our models to give robust predictions even when they are ripped out of the spatiotemporal context in which they were trained. For ecologists, obtaining robust predictions under-change is a primary objective. However, most available statistical methods focus on association rather than causal inference, and rely on environmental explanatory variables that are most readily available, rather than have a causal relationship with the species' ecology. Further, we tend to evaluate models based on goodness-of-fit rather than their predictive capacity (Fourcade, Besnard, & Secondi, 2018). As we illustrate throughout this book, the pragmatic resolution of such dilemmas for real-world applications centers around enhancing the mechanistic content of our statistical models.

Part I

Fundamental concepts and methods

Chapter 1

The ecology behind species-habitat-association models

1.1 Objectives

1. To define what we mean by habitat and its constituent environmental variables.

2. To distinguish between the effect of habitat on fitness and on species distribution.

3. To identify the behavioral and demographic drivers of SHAs.

4. To introduce an appropriate null model for species-habitat associations (SHAs) that encapsulates fundamental concepts such as the Ideal free distribution (IDF), the pseudo-equilibrium assumption and the niche.

5. To outline the biological mechanisms that can violate the null model for SHAs.

6. To synthesize the state of the art in the currently best developed methodology for quantifying SHAs using species distribution models (SDMs).

1.2 How do living beings "see" the world around them?

Individuals experience the world around them in different ways, but what ultimately matters for an organism's fitness and distribution is whether it has a positive, negative or ambivalent relationship with aspects of its environment. This trichotomy allows us to classify environmental variables into **resources**, **risks** and **conditions**. Although there is clearly some ambiguity in classifying environmental variables into these three categories, they will facilitate the model development outlined in Chapter 2. Furthermore, these three classes will provide us the opportunity to present different examples of the environmental variables driving species-habitat associations.

1.2.1 Resources

Resources are substances, objects or places in the environment required by an organism for normal growth, maintenance, and reproduction, and therefore resources generally have a positive effect on fitness. Examples of resources for plants are sunlight, carbon dioxide, minerals and water, and for animals are prey, water, and nesting or resting spaces. In overabundance, resources could potentially turn into risks and their effect on fitness may become negative. For example, some food sources consumed by animals contain poisonous substances that may become lethal when consumed in large quantities, or for plants, the photosynthetic machinery can be damaged by too much light. However, most often resources are limited because they are reduced by an organism and its conspecifics or heterospecifics, for example, by consumption or occupancy.

Resource limitation leads to inter- or intra-specific competition, which limits species distributions globally and locally. Resource limitation can also lead to a dynamic feedback between resources and consumers, because resources often have their own dynamics.

1.2.2   Risks

Risks are environmental variables that have a negative relationship with fitness by lowering the actual or perceived chances of individual survival or reproduction. Risks may shape distributions by influencing demographic rates (e.g. a predator can cause death and a toxic substance can cause abortion), but the perception of risk by organisms may be just as influential (Laundré, Hernández, & Ripple, 2010). For example, in the case of human encroachment (Ciuti et al., 2012), even as some environmentally hostile anthropogenic activities are waning, the mere presence of humans can maintain entrenched avoidance behaviors by animals and limit their access to valuable resources (the idea of 'landscapes of fear'; Brown, Laundre, & Gurung, 1999). It is common for risks, like predation pressure, to have a deciding influence on species distributions. For example, intense commercial whaling has led to the extinction of Gray whales in the North Atlantic. While those main threats have now disappeared, Gray whales remain virtually absent in the North Atlantic ocean, despite suitable local conditions (Monsarrat et al. (2015), Fig. 1.1). Also, environments that favor parasites may suppress the distribution of their hosts locally, resulting in the absence of the host species from areas that may otherwise be optimal for their survival (Giannini, Chapman, Saraiva, Alves-dos-Santos, & Biesmeijer, 2013). Despite the potential importance of risk layers in shaping species distributions, they are rarely included in species distribution models because actual and perceived risks tend to be dynamic processes that are difficult to quantify (Palmer, Fieberg, Swanson, Kosmala, & Packer, 2017).

Figure 1.1: Gray whales (*Eschrichtius robustus*) currently only reside in the Pacific Ocean, but used to live in the Atlantic. Their population became locally extinct in the Atlantic Ocean in the early 18th century, at least partly due to commercial whaling. Apart from the historic risks imposed by whaling, the Atlantic likely remains suitable habitat for gray whales (Monsarrat et al., 2015). Photo: Jeroen Hoekendijk.

1.2.3   Conditions

Conditions are environmental variables (such as ambient temperature, humidity, salinity or pressure) that surround the organism and influence its functioning. Crucially, conditions can have both positive and negative influences on fitness. For example, temperature regulates metabolism (particularly in cold-blooded organisms),

and for extreme high or low values (i.e. outside the thermal neutral 'Goldilocks zone'), organisms may not function properly and die. Although, like resources, conditions may be modified by the actions of organisms (see literature on ecosystem engineers; e.g. Odling-Smee, Laland, & Feldman, 2013), they are generally considered external drivers.

1.2.4 Interactions between resources, risks and conditions

All three classes of environmental variables interact, and this can lead to apparently counter-intuitive effects on species distributions. The *resource-risk* interaction is nicely captured by the phrase: 'Is it worth the risk?' (Brown & Kotler, 2004). Predation risks, in particular, may cause organisms to avoid risky habitats despite high resource density (McNamara & Houston, 1987). Sometimes interactions between risks and resources may lead to counter-intuitive effects on species-habitat associations. For example, while populations exposed to risk may have lower survival, those individuals that survive may experience less resource competition, and this may ultimately increase overall population fitness (Van Leeuwen, De Roos, & Persson, 2008). *Resource-condition* interactions relate to the effect of environmental conditions on the detection and intake of resources. For example, depth and sediment type may influence the accessibility of marine invertebrates to their predators (De Goeij & Honkoop, 2002; Van Gils, Edelaar, Escudero, & Piersma, 2004). An example of a *risk-condition* interaction is the effect of canopy cover reducing or enhancing the exposure to predation (Brown & Kotler, 2004). Finally, all three classes of variables may interact in counter-intuitive ways. For example, although warm-blooded marine mammals need to invest heavily in thermal regulation to avoid the risk of hypothermia, many species of seals and whales are particularly abundant in polar environments. It is believed that their warm-blooded nature allows them to outswim their cold-blooded prey giving them a competitive advantage in colder conditions over other top predators, like sharks (Grady et al., 2019).

1.3 What is a habitat?

Although it is unambiguous that habitat is somehow related to environmental variables, its precise definition in the contemporary ecological literature is rather vague (Dennis, 2012; Kirk et al., 2018). To clarify the concept, we need to answer three questions that can be traced back to ambiguous uses of the term in the literature: 1) Is habitat a place, or a set of circumstances? 2) Should habitat comprise just conditions, or should it also include resources and risks? 3) Is habitat species-specific, or species-independent? Such details of semantics are worth considering because different answers lead to different analytical decisions and interpretations of results.

First, let us consider the possibility of habitat as a region in space. Intuitively, classifying a region in space as a habitat manages to encompass the spatial context in which environmental variables (resources, conditions and risks) are arranged in relation to each other. For instance, from the perspective of an arboreal primate, a single tree is of no value, but a collection of trees in close proximity is home. Equivalently, for a small grazer the availability of short vegetation for food may need to be combined with tall vegetation for cover. These two types of vegetation may not coexist locally, but they must be found in close proximity to be of value to the grazer. Several authors have therefore defined habitat as a suite of resources, risks and environmental conditions, together with their spatial configuration in a landscape (Caughley & Sinclair, 1994; Weddell, 2002). Such a region-specific definition of habitat makes it easier to link habitat to species distribution data collected from the field.

However, a region-specific definition of habitat has several methodological disadvantages. First, it requires us to divide space into discrete parts *a-priori*, which can be challenging, particularly when multiple, nested spatial scales are relevant for different environmental variables. Second, the appropriate scale of spatial configuration also depends on the species' mobility and the spatial scales at which it perceives its surroundings. Finally, the environmental makeup of regions in space does not stay constant. It might change daily or seasonally, but also abruptly due to human intervention or natural drivers. So, if we define a habitat as a place, we either have to insist that its composition is static, or we have to allow the environmental properties of "habitat" to change. Overall therefore, although habitats can be mapped onto geographical space and although models should be equipped to understand spatial associations (e.g. in order to capture complementary use of neighboring habitats), the definition of habitat as a spatial region is pre-emptive before an analysis, and

severely constrictive for the predictions that can be generated after an analysis. Furthermore, and most importantly, a model that has been trained with data from a particular place cannot be transferred elsewhere if its predictions are specified only to that place.

We continue with the second question. Should a habitat comprise just conditions, or should it also include resources and risks? Traditionally, studies of the distribution of sessile organisms, like plants, primarily focused on conditions like temperature, humidity or soil pH. In contrast, studies of the distribution of moving animals originally used 'resource-selection functions' (Manly, Thomas, McDonald, & Erickson, 2002) to study food selection, but soon also allowed the inclusion of other environmental variables like conditions or risks. As described in the previous section, all three types of environmental variables - resources, risks and conditions - ultimately influence fitness and distribution, and complex interactions between these variables prevent us from treating them in isolation. Hence, particularly if we want to facilitate unification of frameworks across taxa, we need a definition of habitat that incorporates all three types environmental variables (Hall et al., 1997).

Regarding the final question, when considering the distribution of, for example, polar bears, is it helpful to think about "polar bear habitat" or "polar habitat"? Although many studies have a species-specific focus with the objective of classifying environments as suitable or unsuitable, there are two problems with assuming a species-specific definition of habitat. First, at the onset of a study, it is often not known which environmental properties are important to any focal species. We therefore need to be able to compare the suitability of different habitats as part of data analysis. Second, the inferred relationship of species with habitats is not binary (suitable or unsuitable), and models must allow for gradients of suitability. We therefore feel that a species-specific definition of habitat is prejudicial when setting up a data analysis and under-powered when used to interpret its results.

In summary, at the onset of the analysis, we define habitat as a point in environmental space, a space whose dimensions are environmental variables (conditions, resources and risks). Although we promote this strictly species-independent definition of habitat, much of this book is devoted to deriving species-specific habitat-association models. A habitat-association model captures the dependence between a species and environmental variables. Most current studies on SHAs take this approach by modeling the distribution of a species as a function of environmental covariates. Disconnecting the spatial observations from geographical space enables these studies to make predictions for regions in space or time that have not been sampled. Nevertheless, it is still necessary to map habitats onto geographical space and models of species-habitat associations must consider spatial context.

1.4   What is a species-habitat association?

In the broadest sense, a SHA connects aspects of habitat (resources, conditions, risks) to particular observations about a species (recorded at the individual, group or population levels). There are several different aspects of a species that can be associated causally with habitat and there are a multitude of interesting biological questions that can be informed by studying such associations. For the logical coherence of later chapters, we distinguish between three general categories of SHAs (Fig. 1.2).

First, there are associations between habitat and fitness. Here we define habitat-associated fitness as the contribution of each unit of habitat, $\mathbf{x}$, to the population's long-term log-rate of change (referred to as partial fitness in Matthiopoulos et al., 2015). Partial fitness can be interpreted as the fitness of a population living in an environment made up entirely of habitat type $\mathbf{x}$. The mechanisms underpinning the association between habitat and fitness are often sub-organismal, operating at the level of physiological capabilities, anatomy and life history strategies. We can often measure these fitness relationships with carefully constructed experiments or field observations. For example, fundamental biology of resource uptake or temperature tolerance in plants may be experimentally measured in the lab, metabolic costs can be gleaned from thermodynamic first principles, or for animals in the wild, accelerometers can be used to measure energetic costs or gains (Wilson et al., 2020). For example, changes in elephant seal body fat may be inferred from changes in sink-rate during drift dives, occurring in different habitats (Bailleul et al., 2007). Habitats that make a low contribution to fitness may ultimately lead to a loss of condition and subsequent demographic costs.

Second, we see associations between habitat and *changes* in species-density. Many organisms can mitigate
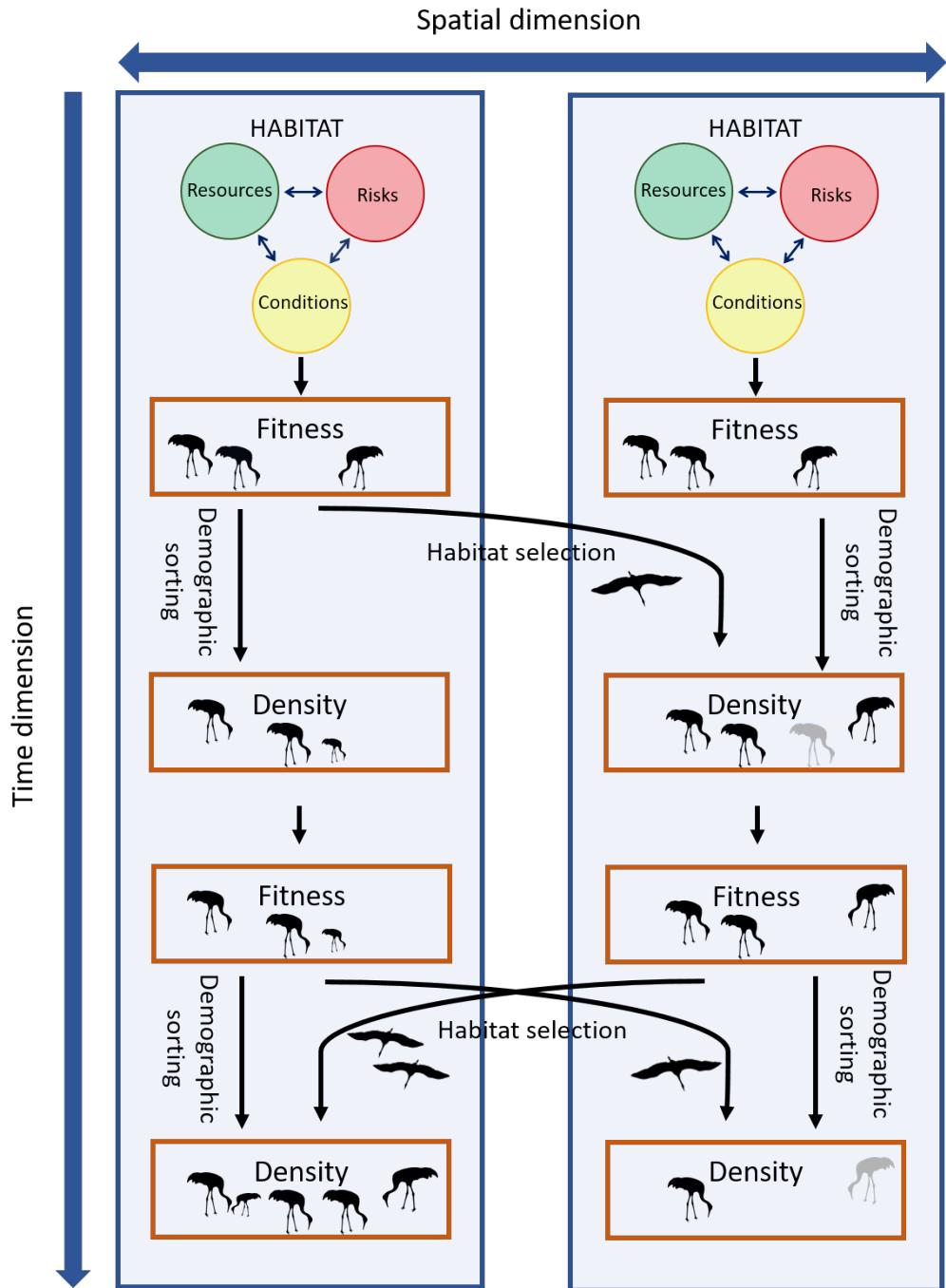
Figure 1.2: Fitness is the evolutionary currency determining an individual's ability to transfer its genes to future generations. Individuals may be able to improve their fitness via plastic adaptations in physiology or behavior, for example by relocating and selecting more suitable habitats elsewhere. However, eventually they will have to face the demographic consequences of the environment they are exposed to. Their vital rates; births (depicted by small juvenile cranes) and deaths (depicted by gray cranes), will vary between habitats, and together with the redistribution of individuals, this demographic sorting will drive the spatial distribution and abundance of species.

the consequences of fitness, for example through their patterns of growth, defense activation and even whole body-metamorphosis. However, when organisms living in specific habitats experience low fitness, they either have to relocate or face the demographic consequences, like mortality and loss of reproductive potential. These processes will manifest themselves as changes in species density. The dynamical processes of relocation involves movement of individual animals, groups, or entire populations. The patterns in these movement trajectories can inform us about how the animal responds to its environment. For example, movement trajectories that look localized and sinuous are likely to be occurring in regions of high resource availability, or at least regions that appear worth exploring for resources. For organisms that cannot relocate, their density will ultimately be affected by habitat-mediated survival and reproduction. These vital rates define the intrinsic growth rate of a population and relate to the fundamental niche of a species (Holt, 2009).

Third, there are associations between habitat and species-density. The existence of an organism somewhere is mediated by fitness and dynamics in a series of complex causal steps. The emerging species distribution patterns are arguably the easiest to observe directly, in the wild, and are therefore the easiest to construct empirical models for (collectively called Species Distribution Models (SDMs)). While spatial variations in density may not reflect mechanisms operating at the immediate short term (e.g. days, hours or minutes), the spatial density of a species reflects accumulated effects of processes operating over longer time scales (e.g. years or multiple generations) and may ultimately be more informative about the place and role of that species in the earth's biosphere.

In summary, phrased in more traditional terminology, associations with fitness define the fundamental niche of a species. The realized niche is represented by the eventual mapping of the distribution of a species in geographical space. In-between those two, we have a wealth of dynamics, represented by the associations between habitat and mechanisms of change. So, while species distribution is often implicitly assumed to be proportionally (or, at least, monotonically) linked with fitness, this is only the case in special circumstances.

1.5   What mechanisms drive habitat-mediated changes in species densities?

The underlying, and often unseen, forces shaping fitness manifest themselves in the distribution of a species via different dynamic mechanisms. Here, we make the important distinction between selective use of habitats at the individual level and demographic forces at the population level.

1.5.1   Behavioural selection

Basic patterns of behavior can be assigned to all organisms. For example, some plants may be triggered to release seeds only under certain conditions and propagule dispersal may be allowed to continue until suitable settlement conditions have been met (Grohmann, Hartmann, Kovalev, & Gorb, 2019; Seale et al., 2019). However, behavioral selection is most clearly seen in organisms that can perceive multiple stimuli from their environment and have control over their mobility. For some organisms, voluntary mobility may happen only once during a specific life-stage. For example, most bivalve species have a pelagic, long-ranging, larval stage, but once they settle, they may remain sessile for the rest of their lives. Most herbivores and their predators, however, may move great distances during their lives and continuously select habitats that fit their needs.

The process of voluntary movement comprises two different decision processes, each with different determinants and cognitive requirements. First, a decision must be made on whether to stay or leave, and second, a decision is needed on where to go. These decisions may happen simultaneously, sequentially or iteratively. Decisions on leaving are closely linked to patch departure rules or giving-up times (Charnov, 1976; Brown, 1988). For example, one well-known theoretical approach to optimal decisions for patchy resources is the marginal value theorem which posits that once an individual depletes resources locally to a point that its intake rate drops below the average long-term intake rate, it should leave the patch (Charnov, 1976). However, since the actual or perceived long-term intake rate may gradually change over time (e.g. due to environmental change or learning abilities) and because animals often cannot instantaneously determine what local intake rate is (e.g. due to the stochastic nature of resources), they have to learn (Ollason, 1980). Several models are possible for the process of learning, even ones using direct metaphors from statistical laws, such as Bayesian updating

Figure 1.3: Plant species, like the dandelion (*Taraxacum officinale*) can regulate their dispersal and soil attachment in response to environmental conditions, like moisture and soil structure (Grohmann et al., 2019; Seale et al., 2019). Photo credit: Half-Seeded Dandelion by Wonglijie – CC BY-SA 3.0.

of experience (Biernaskie, Walker, & Gegear, 2009). Learning takes time and may lead to sub-optimal decisions and a miss-match between the distribution of resources and species distributions, particularly if the environment is unpredictable (Kamil, Misthal, & Stephens, 1993; Riotte-Lambert & Matthiopoulos, 2019).

Having decided to leave, animals may then move randomly (e.g. following isotropic or correlated random walks) or exhibit directed movement based on their sensory perception, spatial memory or information from conspecifics (Fagan et al., 2013; Bracis & Mueller, 2017; Jesmer et al., 2018). While these well-recognized behavioral processes are of vital importance, they have only recently been explicitly included in SHA models (Oliveira-Santos, Forester, Piovezan, Tomas, & Fernandez, 2016; Merkle et al., 2019). For example, use-versus-availability analyses (see Chapter 2) compare used habitats to those habitats available within a next movement step (e.g. Step-Selection Functions) or within their home-range (e.g. Resource-Selection Functions). But, what about habitats beyond this range - do individuals perceive and avoid them? It is clear that realistic SHA models need to consider that individuals integrate information over much larger time intervals than the arbitrary scale of data collection. Cognitive processes can operate at several scales, some very small, others very large, and none of them need be similar or biologically relevant to the scale used in the data collection and subsequent analysis. If we are to understand SHAs, it is important to recognize the underlying mechanisms shaping individual movements and find a way to include these (or include emergent, accessibility constraints that are derived from these mechanisms) into our empirical models.

1.5.2 Demographic sorting

While the fine-scale distribution of organisms, such as animals, that can move in a voluntary and controlled fashion will be strongly shaped by behavioral selection processes, the distribution of all organisms will be shaped by demographic processes that are highly dependent on spatially heterogeneous environmental variables. Demographic sorting is particularly relevant for plants. Following an initial period of dispersal, which could be highly localized (e.g. via stolons) or wide-ranging (e.g. wind or animal dispersal), most plant propagules will become sessile. From that point onward, demographic processes determine whether a seedling will manage to germinate and survive. This combination of dispersal and habitat-driven demographic sorting can cause

intergenerational movement across the landscape (see Fig. 1.4).

Similar demographic processes shape mobile animal distributions, but the combined action of demography and mobility makes their relative contributions difficult to tease apart. For example, when we fix an observation device (e.g. a GPS logger) onto a mammal, our view of the individual's lifetime performance is truncated to the time-window of observation and all our inferences are conditional on the animal being alive and present at the original capture location. Demographic sorting has taken place prior to tagging, and clearly we can only study the behavioral selection of those individuals that managed to stay alive.
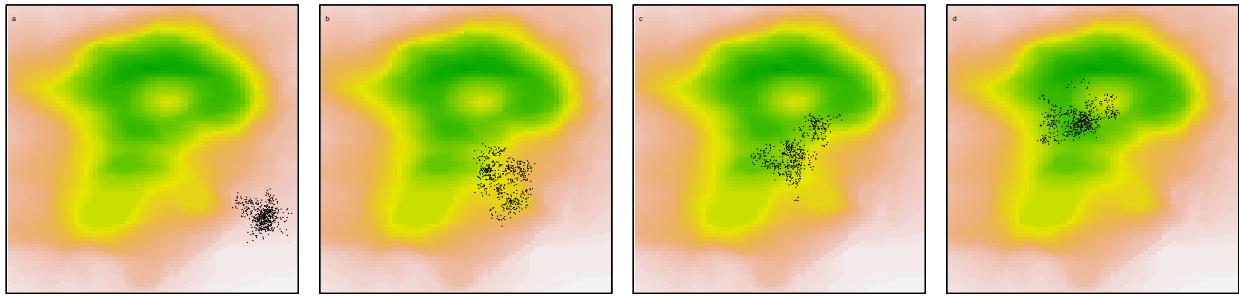

Figure 1.4: Inter-generational movement of members of a sessile species (e.g. a plant) progressing directionally through a combination of non-directional dispersal and demographic sorting. Suitable habitats are in shades of green, sub-optimal ones in yellow-to-brown. If we examine plant distributions over longer (multi-generational) time-scales, such that demographic selection is considered instantaneous, they display high mobility and dynamics. At these scales, patterns of selectivity may be reminiscent of animal redistribution.

1.6   When is species density a reliable reflection of habitat suitability?

When we observe a species at high density, we often make the implicit assumption that the local habitat must be good for the species, i.e. the intrinsic fitness contributed by the habitat ($F_{\mathbf{x}}$) is high, but this assumption is not necessarily valid. Below, we outline assumptions necessary for the distribution of a species to be proportional to fitness. Since the transition between fitness and species distribution can be driven by habitat-driven demographics and behavioral selection, we will formalize the link between fitness and distribution for these two avenues separately using a null model.

1.6.1   Why do we need a null model for SHA?

Each species will most likely be unique with respect to its relationship with its environment, and therefore it is tempting to develop SHA models in isolation. However, achieving some unification and contributing towards broader ecological theories is desirable, and this calls for a null model of species-habitat associations. Good null models can be useful for making sense of the world quantitatively, and they form the basis for further model development, as the initial assumptions become increasingly relaxed. These two properties of utility and expandability can stimulate scientific thinking and make null models good platforms for development. However, it is important to guard against over-interpreting null models in the face of data (this often happens when we inadvertently forget about their underlying assumptions).

A null model is a mathematical construct, derived from a set of assumptions that form the basis for discussion on a physical process. In the development of null models, mathematical convenience is prioritized over the realism of the assumptions, to ensure that the basic results can be derived analytically and exactly. For example, in physics, an ideal gas is a hypothetical form of matter that comprises dimensionless and massless molecules, that are unaffected by forces other than those resulting from perfectly elastic collisions between them. Despite these patently unrealistic assumptions, the null model of ideal gasses has led to invaluable and approximately applicable rules in thermodynamics. In economics, the null model of supply

and demand has been used successfully to understand price determination for goods and services. Its key simplifying assumption is that it considers omniscient and logical consumers who focus exclusively on a particular commodity, while keeping their behavior constant with regard to other goods. The model has been so successful among economists, despite these assumptions, that people are occasionally (and unfairly) surprised when its predictions diverge from reality.

Ecology has its fair share of null models. Examples include the logistic model in population dynamics (Verhulst, 1845; Pearl & Reed, 1920), the model of perfect mixing in predator-prey ecology and epidemiology (Law et al., 2008), the marginal value theorem in foraging ecology (Charnov, 1976), and the neutral theory of biodiversity (Hubbell, 2001). In the early stages of structuring this book, it occurred to us that for SHA, the research community has been implicitly using a null model, without specifying its mathematical formulation or explicitly stating its assumptions. This implicit null-model is essentially one that gives rise to a "pseudo-equilibrium assumption", meaning that species respond to measured habitat covariates with no delay, and that species density is proportional to species fitness. Yet, the underlying assumptions required for this implicit null-model are rarely if ever spelled out.

Species fitness, dynamics and population distribution, the three aspects of species that respond to habitat, are connected. Fitness affects dynamics, which affects distribution. In turn, distribution (through density dependent processes) affects habitat and fitness (Fig. 1.5). Below, we will formalize the link between fitness, dynamics and distribution, and we will try to connect our SHA null model with many of the already existing ecological null models. In the process, we will make sure that the null model has the two valuable properties of utility and expandability.

In Section (1.5) we have identified two mechanisms that determine SHA: Demographic sorting and behavioral selection. Below we will first develop a null model for each mechanism independently, and subsequently show how these two null models converge.
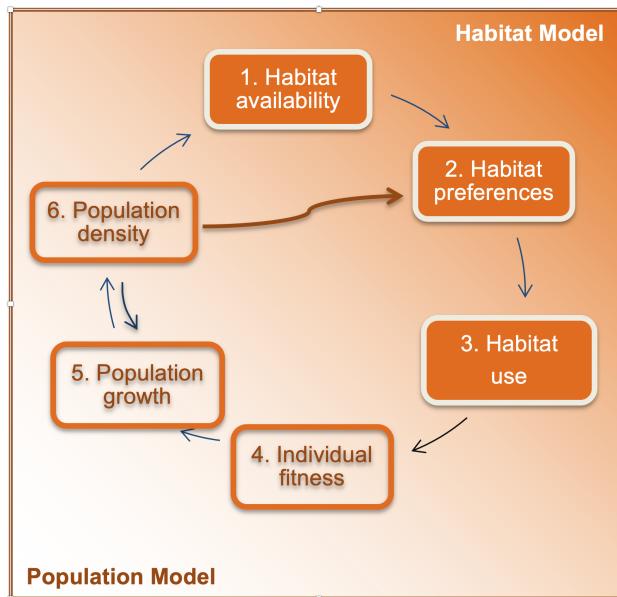
Figure 1.5: The chain highlighting how species fitness, dynamics and population distribution interact with habitat (modified from Matthiopoulos, Field, & MacLeod, 2019). A region in space is characterized by the availability and spatial configuration of different habitat types (step 1). Organisms may actively or passively select specific habitats (step 2) which may lead them to use habitats disproportionately to their availability. Habitat availability and habitat selection give rise to the observed spatial distribution (step 3). The exposure of individuals to different habitat types will influence their fitness (step 4), which determines the collective capability of a population to grow (step 5). Processes of population change determine current population density (step 6), which has the opportunity to alter habitat availability and feed back directly into habitat selection and population growth via density dependent and spatial crowding.

1.6.2   A null model for habitat-driven demographic sorting

Let us imagine an ideal life-history organism with the following properties:

1. It has a highly mobile dispersal stage, so that all of space is equally accessible to all of its propagules.
2. It has a completely sessile reproductive stage, so that local density results from local habitat quality.
3. Its intrinsic population growth is solely driven by habitat, and it is positive everywhere in the landscape.
4. Its local population growth decreases linearly with local density.
5. It has no other species to compete with.

For a given spatial location, $\mathbf{s}$, with a particular set of environmental characteristics, $\mathbf{x}$, we can model growth of local population density, $N_{\mathbf{s}}$, in discrete time as

$$N_{\mathbf{s},t+1} = N_{\mathbf{s},t} \exp(F(\mathbf{x}, N_{\mathbf{s},t})) \tag{1.1}$$

The function F has two convenient names: growth rate and fitness. In classic population ecology, the population's growth rate between two successive measurements is defined as $\log(N_{t+1}/N_t)$. The population growth rate is equivalently called the fitness of a population (Nur, 1987; Matthiopoulos et al., 2015). So, let us consider the local growth rate of a population as a trade-off between the intrinsic fitness value of the habitat (i.e. $F(\mathbf{x}, N_t \approx 0) = F_0(\mathbf{x})$) and the attrition b inflicted onto population growth (i.e. onto population fitness) by the addition of a single individual:

$$F(\mathbf{x}, N_{\mathbf{s},t}) = F_0(\mathbf{x}) - bN_{\mathbf{s},t} \tag{1.2}$$

From equations (1.2) and (1.1), we get the model

$$N_{\mathbf{s},t+1} = N_{\mathbf{s},t} \exp(F_0(\mathbf{x}) - bN_{\mathbf{s},t}) \tag{1.3}$$

We can now look at the two extremes of population density. At one extreme, when the population density is very small (i.e. $N_t \approx 0$), its growth rate is equal to the intrinsic growth rate $F_0(\mathbf{x})$. At the other extreme, when the species has sufficient time to reach an equilibrium local population density (i.e. $N_{\mathbf{s},t+1} = N_{\mathbf{s},t} = N_{\mathbf{s}}^*$):

$$N_{\mathbf{s}}^* = F_0(\mathbf{x})/b \tag{1.4}$$

In biological terms, attrition caused by density dependence has annulled the benefits of habitat, and the equilibrium distribution becomes proportional to the intrinsic growth rate. In Fig. 1.6 we try to visualize how, for a growing population of ideal life-history organisms, the equilibrium density in each cell eventually becomes proportional to $F_0(\mathbf{x})$.

Because our ideal organisms behave as a closed population within each location, the SHA null model can be linked to the null model used in single-species population dynamics, the logistic equation[1]

---

[1]Our eq. (1.3) is closest to a logistic form known as the Ricker model (Matthiopoulos, 2011):

$$N_{t+1} = N_t \exp\left(r\left(1 - \frac{N_t}{K}\right)\right).$$

By comparing the equation above with eq. (1.3), we note that the intrinsic growth rate in classical terminology is equal to habitat-specific fitness ($r = F_{\mathbf{x}}$) and the environment's carrying capacity is proportional to it ($K = F_0(\mathbf{x})/b$). If we replace $F_0(\mathbf{x})$ by r, our null model also implies a simple relationship between the two key parameters of the logistic null model

$$r \propto K$$
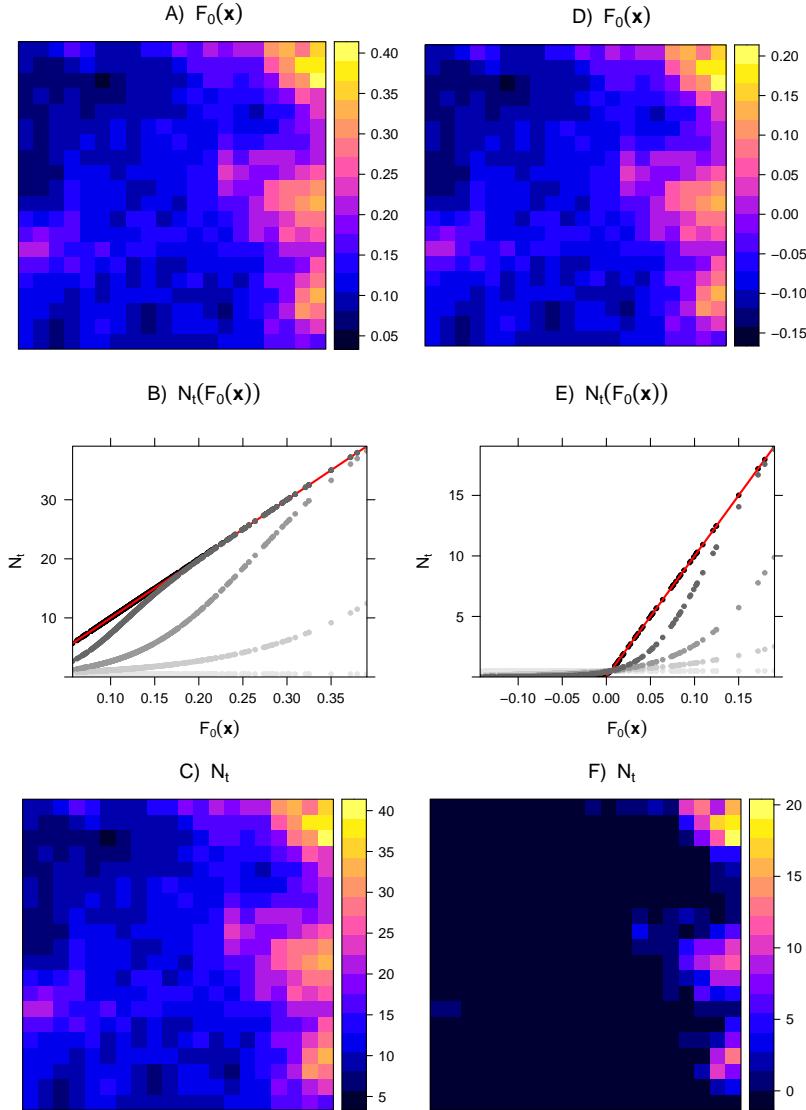
Figure 1.6: Simulation of an ideal life-history organism: We set out to visualize the relationship between fitness and distribution during the growth of a population from $N_{\mathbf{s},t=0} \approx 0$ to the local equilibrium density $N_{\mathbf{s}}^*$. The experienced fitness $F_{\mathbf{x}}$ at low population size is defined as a linear function of a covariate X representing food. We explore two scenarios, one where $F_{\mathbf{x}}$ is always positive across space (**A**) and one where $F_{\mathbf{x}}$ can also be negative in some areas of space (**D**), achieved by subtracting a constant value from the original non-negative $F_{\mathbf{x}}$. The population growth rate is defined as $N_{\mathbf{s},t+1} = N_{\mathbf{s},t} \exp(F_{\mathbf{x}} - bN_{\mathbf{s},t})$ where b=0.01. The simulation is run for 250 years. At the start of the simulation ($t = 0$), the local population size $N_{\mathbf{s}}$ is still independent of $F_{\mathbf{x}}$ (light gray points, in **B**). Ultimately, the relationship will equilibrate to a linear relationship between local carrying capacity $N_{\mathbf{s}}^*$ and $F_{\mathbf{x}}$ (black points in **B**), with slope coefficient of $1/b = 100$ (red line in **B**)). The same applies to the scenario where $F_{\mathbf{x}}$ can also be negative, but this will result in truncation of $N_{\mathbf{s}}^* = 0$ when $F_{\mathbf{x}} \leq 0$ (**E**). Spatial maps of $N_{\mathbf{s}}^*$ are shown in **C** and **F**. Note that when a landscape contains locations with negative fitness (i.e. $F_{\mathbf{x}} \leq 0$), this will result in the emergence of a patchy distribution of a species (**F**).

1.6.3   A null model for habitat selection

We build upon the work of Fretwell & Lucas (1969) who consider the distribution of an ideal free species to be a purely behavioral phenomenon, with no demographic sorting involved.

The assumptions underlying the ideal free organism are:

1. Individuals are aware of the current value of each patch[2] and settle in the patch most suitable to them (i.e. individuals are *ideal* foragers).
2. All individuals are *free* to enter any patch, and arriving individuals do not have a disadvantage compared to individuals that are already there, all individuals are genetically alike or of the same phenotype, and all patches are equally and instantaneously accessible to all members of the species.

Now let us consider the behavioral mechanism that shapes the distribution of ideal-free foragers. The first individual will select a single patch i with the highest baseline suitability B_i, ignoring all other, lower-quality patches. By making use of the patch, the patch baseline suitability B_i will be reduced to a lower suitability S_i. The next individual that comes along, will select a patch with the highest suitability S, which could be the same patch i or a different one. This process is repeated for all individuals in the population (see also Fig. 1.7). Eventually this process results in a landscape where all l occupied patches have equal suitability; $S_1 = S_2 = ... = S_l$. The formula quantifying how S_i depends on the baseline suitability B_i and species density N_i is:

$$S_i = B_i - f(N_i), \ i = 1, 2, ...N \tag{1.5}$$

Fretwell & Lucas (1969) treated habitats as discrete patches. For a given spatial location, $\mathbf{s}$, with a particular set of environmental characteristics, $\mathbf{x}$, the continuous formulation for eq. (1.5) is:

$$S(\mathbf{x}, N_{\mathbf{s},t}) = B_{\mathbf{x}} - f(N_{\mathbf{s},t}) \tag{1.6}$$

Since natural selection leads to the evolution of behavior, the perception of habitat suitability by well-adapted individuals should correspond to the fitness value of those habitats (see also assumption 1). In that case, the baseline suitability, B_i, corresponds to the intrinsic population growth rate at low population density ($F_{\mathbf{x}}$), and the emerging suitability in the presence of conspecifics, S_i, will be equivalent to $F(\mathbf{x}, N_t)$. In that case, eq. (1.6) becomes:

$$F(\mathbf{x}, N_{\mathbf{s},t}) = F_0(\mathbf{x}) - f(N_{\mathbf{s},t}) \tag{1.7}$$

Unlike us, Fretwell & Lucas (1969) were not explicit about the functional form or mechanism underlying $f(N_i)$, but two possible mechanisms that can lead to eq. (1.5) are interference competition and explorative or scramble competition (Křivan et al., 2008). Given these assumptions, Fretwell & Lucas (1969) showed that, for each positive total population size M, there will be a unique Ideal Free Distribution (IFD) (see also Fig. 1.7). Now, let us consider the special case where fitness declines linearly with local population density (i.e. $f(N_{\mathbf{s},t}) = bN_{\mathbf{s},t}$), and where the total population size M is at carrying capacity, such that, $F(\mathbf{x}, N_{\mathbf{s}}^*) = 0$ everywhere within the study region:

$$F(\mathbf{x}, N_{\mathbf{s}}^*) = K = F_0(\mathbf{x}) - bN_{\mathbf{s}}^* = 0 \tag{1.8}$$

which implies

$$N_{\mathbf{s}}^* = F_0(\mathbf{x})/b \tag{1.9}$$

---

[2]Fretwell & Lucas (1969) used the term habitat, rather than patch, and defined a habitat of a species as a portion of the surface of the earth where the species is able to colonize and live. We prefer to use the term patch to explicitly refer to a spatial unit.

In summary, when total population size of ideal free foragers is at carrying capacity, the relationship between density and habitat suitability or fitness (eq. (1.9)) is identical to that for the ideal life history organism (eq. (1.4)).

1.6.4 Is population size or population growth a better proxy for fitness?

Our identical null models (eq. (1.9) and eq. (1.4)) hint at the fact that equilibrium population size and intrinsic growth rate could both provide valuable clues about the underlying fitness contributed by habitats. Historically, the niche literature has focused on the ability of a population to grow (Holt, 2009) and the species distribution literature has focused on pseudo-equilibria (Guisan, Thuiller, & Zimmermann, 2017), but the intuitive notion is shared. Whether we are talking about fitness in niche space or habitat suitability in geographical space, we want to get a handle on which habitats are "good" and "bad" for an organism.

So, looking ahead at the types of data that we should be collecting and analyzing, does the null model offer any insights into whether data on intrinsic growth or equilibrium distribution is more useful? The short answer is that both have limitations and, ultimately, data on both may be required. A longer answer goes as follows: Data on intrinsic growth rates are difficult to collect. They require us to observe a local population at a small enough size that density dependence is not affecting growth. Growing populations, by definition stay small only temporarily. Catching a species at such a transient state at sufficiently many habitats to enable a SHA model to be fitted is quite unlikely. Certainly, longer-lasting states (e.g. close to a stable equilibrium) are more amenable to observation. However, as the null model suggests, landscapes at-equilibrium yield leveling inferences for the poorest of habitats, because local populations observed in negative fitness habitats are always extinct in the long-term, so we have no way of learning about different shades of "bad habitats" (See also Fig. 1.6F)

The choice between intrinsic rates and equilibrium sizes becomes more tangled when one of the limiting resources driving fitness is depletable. The overall abundance and detailed spatial distribution of the resource will be different when the consumer is sparse compared to when it is at its carrying capacity. Which of those resource distributions is more relevant to a SHA model?

Consider the following thought experiment, whereby the intrinsic population growth rate is a function of current food availability. At the earliest stages of consumer growth (and resource depletion), a correlation between the distribution of food and consumer growth rates would deliver an informative SHA. Food-rich locations would give faster growing consumer populations. However, it is rarely the case that populations are captured at such initial stages of development, and it is unlikely that the initial development happens in a spatially synchronized manner (as opposed to a local invasion). If, on the other hand, we wait for the equilibrium distribution to develop, depletion distorts our inferences. As predicted by the IFD, the equilibrium density of the consumer will yield no relationship with food density, but a strong positive correlation with rate of food replenishment. Hence, correlating consumer density with standing crop may give us the misleading impression that the consumer is indifferent or even avoids its key resource. Using resource productivity data as a covariate would, of course, dispel this myth but such data are almost never available in the field.

Hence, in the case of depletable resources, it makes sense to relate resource growth to equilibrium consumer distribution, and it also makes sense to relate initial resource distributions to consumer intrinsic growth rate, but it is not meaningful to correlate simultaneous snapshots of the distribution of the study species and its resource. Given that data on distributions are much easier to collect than data on rates of change, we have a fundamental problem of indeterminacy in correlational models.

1.7 When are the null model assumptions violated?

The two null models for demographic sorting and behavioral selection, both describes the proportional relationship between habitat-specific intrinsic fitness and the species' equilibrium density. There are two main reasons why this proportionality assumption is violated. First, the null models might be inappropriate. For example, the models assumed that an increase in species-density led to a proportional decrease in habitat quality and associated fitness. In many systems, changes in species density might have a non-linear effect on habitat-associated fitness. Second, the transition towards the equilibrium density takes time. In fact,
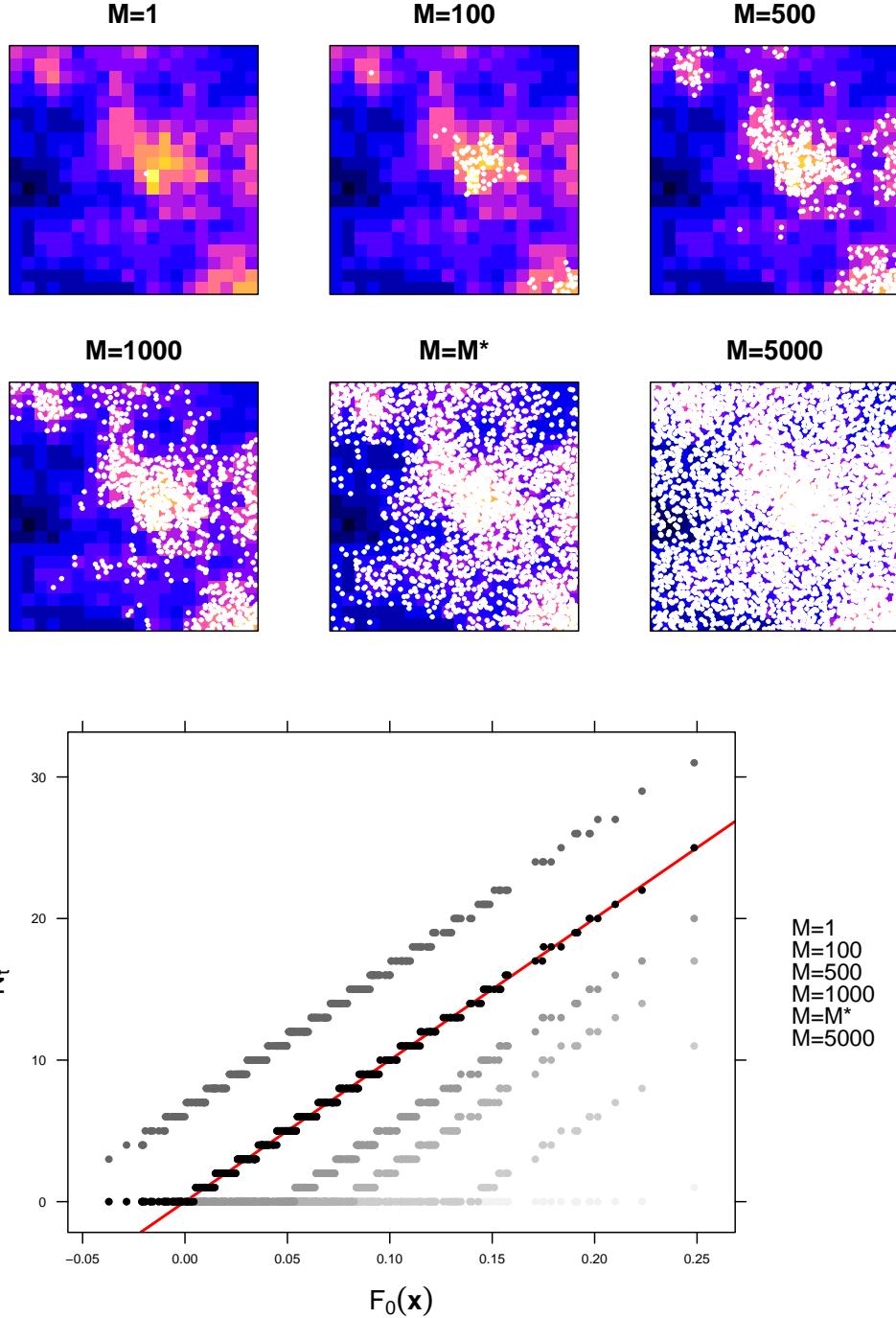
Figure 1.7: The relationship between the baseline suitability (B_i in Fretwell & Lucas, 1969) or intrinsic fitness $F_0(\mathbf{x})$, and local density ($N_\mathbf{s}$), shown for an increasing total population size (M) of ideal-free organisms (as proposed by Fretwell & Lucas, 1969). The top figures show habitat suitability represented by intrinsic fitness. The white dots represent the distribution of individual animals for different total population sizes (M). In the bottom figure, the different shades of gray represent the different population sizes ($M = 1, 50, 100, 500, 1000, 5000$), with light gray for $M = 1$ and dark gray for $M = 5000$. Black dots and red line show the relationship between $N_\mathbf{s}$ and $F_0(\mathbf{x})$ when the population is at carrying capacity, i.e. $F(\mathbf{x}, N_{\mathbf{s},t}) = 0$. Individuals always select the habitats with the highest current $F(\mathbf{x}, N_{\mathbf{s},t})$. When habitat quality diminishes as a linear function of local population density, this will result in a linear relationship between baseline habitat quality $F_0(\mathbf{x})$ and the observed species density $N_\mathbf{s}$ for all occupied locations, and $N_\mathbf{s} = 0$ for all other locations.

constraints in species demography and mobility, in conjunction with changing environments, might mean this equilibrium distribution is never reached. However, being able to consider a snapshot of a species distribution as a static reflection of underlying habitat suitability (the pseudo-equilibrium assumption) is certainly convenient and our null models allows us to think in these terms. Yet, it is important to remember just how restrictive a set of assumptions we needed to make for the pseudo-equilibrium assumption to hold. We take some time in this section to examine the ways in which the real world can violate the assumptions of our null models. These violations will prompt improvements to the basic SHA framework in later chapters of this book.

1.7.1 When the relation between species density and fitness is non-linear

Our null model assumed that species density responds continuously to fitness and that fitness is non-negative throughout the landscape. However, population density does not always scale linearly with fitness. For example, nesting seabirds establishing at suitable locations exclude their competitors and form equally spaced nesting arrangements. Almost undoubtedly, some nesting locations are more suitable than others, but as long as the suitability of a nesting site is above a nominal threshold, it will eventually be occupied. Once occupied, no additional breeding pair can settle on that location. Our observations at equilibrium will therefore only be able to capture a binary state of occupancy/vacancy, rather than a continuous measure of suitability.

The effect of species density on fitness can also be positive. For example, many species aggregate to create positive feedbacks between group members, to reduce predation risk, modify their the environment, and exchange information and services. These are also known as Allee effects (Allee, 1951; Stephens & Sutherland, 1999). For example, social species, like cetacean, operate in groups to improve the detection and exploitation of fish schools. As a result, even habitats with negative intrinsic fitness may generate positive fitness once the density of cooperating individuals increases (Fig. 1.8). As a result, the relationship between density and intrinsic fitness may be non-linear at low population densities.

Figure 1.8: Collective feeding by Humpback whales (*Megaptera novaeangliae*) in Alaska (USA). The null models for demographic sorting and behavioral selection assume that habitat quality declines linearly with the increase in species-density. For most social, cooperative organisms this linear relationship does not hold; fitness may increase as function of group size (the Allee effect), which results in a non-linear relationship between fitness and population densities. Photo: Jeroen Hoekendijk.

1.7.2   When species occur at places with negative fitness

Organisms are often encountered in poor-quality habitats. Most obviously, this can be a transient state; Organisms move at finite speed and may cross hostile or barren habitat while moving in-between good-quality habitats that are within reach (Matthiopoulos, Fieberg, Aarts, Barraquand, & Kendall, 2020). Organisms may also be pushed into suboptimal habitats more permanently, creating extinction debts: Temporal lags in the decline of a population may leave remnants in poor quality habitats and give misleading results if we incorrectly assume pseudo-equilibrium states.

Less intuitive, however, is that mobile organisms may actively choose to spend time in habitats of negative fitness. Indeed, viable animal populations may be observed in *landscapes where no single point is characterized by positive fitness*; the zero niche paradox. This possibility occurs because, while immobile organisms must match themselves to their environment, mobile organisms have the ability to match their environments to themselves (Begon, Harper, & Townsend, 1996). By moving around, mobile organisms can accumulate resources, reduce risk and mitigate environmental conditions to ensure fitness is positive overall. Sometimes, all necessary environmental conditions and resources are in relatively close proximity, which allows animals to live in a relatively compact home-range. However, for other species, like many migratory animals, the required resources/conditions might be transient, or separated by thousands of kilometers, requiring long-range movements. In general, whether a fitness integration is determined by the spacing between different habitats and the mobility of the animals that exploit them.

Fitness integration may also be possible for sessile organisms that may be able to exploit temporal (i.e. diurnal, or seasonal) transitions in the resources and conditions available at their fixed locations. For example, plants may regulate their tolerances so that they can withstand temporal separation between peaks of availability in the same, or complementary resources. The notion of temporal proximity of vital environmental circumstances is another manifestation of the problem of accessibility (introduced above in the context of space, and mobile organisms) that is not readily visible by simply mapping fitness in environmental space.

1.7.3   When species do not occur at places of positive fitness

Just as organisms may be found at places of negative fitness, they may be absent from areas of positive fitness. In conservation ecology this is known as a 'colonization credit' (Watts et al., 2020), a result of the fact that it may take time for individuals, populations and species to respond to changing environments. Environments can be highly dynamic. Mismatches between high habitat suitability and the abundance of a species may arise from imperfect knowledge of the landscape or limited relocation speed. A resource, like the massive release of eggs by spawning fish (Ims, 1990; Šmejkal et al., 2018), may suddenly become available and it may take some time before a forager can locate and exploit it. Sometimes, environmental changes are unpredictable and the ability of organisms to benefit from them is limited by their range of perception. The pressure to quickly and accurately assess the availability of resources has selected for sophisticated visual, auditory, gustatory, olfactory and somatosensory capabilities. However, the ultimate extensions to perception come from higher cognitive functions, such as memory and communication, which are extremely useful for exploiting both predictable (e.g. seasonal) or unpredictable habitats. For example, memory may allow an individual to relocate in anticipation of the appearance of resources, suitable conditions or the departure of predators. For example, many migratory birds start their journey in anticipation of high resource availability elsewhere and start breeding well before the peak in insect abundance.

Absence of a species from apparently suitable habitat may be caused by medium-term delays, covering the life span of individuals. Most organisms have distinct life stages, each potentially characterized by different mobility, resource requirements and sensitivity to environmental conditions. For example, jellyfish have a distinct and sessile polyp phase and a free-moving medusa phase. Insects, like butterflies, have well-defined larval and adult stages. Even the types of risk that individuals are vulnerable to can change as they metamorphose or grow. The habitat requirements of distinct life-stages may be dramatically different and, even, mutually exclusive. If the habitats needed for different life-stages are not sufficiently close to each other, then the organism will occur in none of them. This suggests that the mobility of different life stages must be understood in a habitat context. For example, many fish species have very specific spawning areas, from which free-floating fertilized eggs, or larvae, are transported by the currents towards nursery

areas. It also underlines that successful SHA models must consider how mobility constraints operate across an individual life history.

At longer time scales, delays can occur trans-generationally as a result of transient population dynamics, particularly if the landscape is changing continuously. Colonization and invasion processes may take many years even for the most advanced and mobile of animals (see Grey whale example, Fig. 1.1). Conceivably, trans-generational delays are the result of cultural behavior in animals with high levels of cognition (Matthiopoulos, Harwood, & Thomas, 2005). Grey seals had been absent from the Wadden Sea since the middle ages, despite there being suitable habitat, and the presence of nearby gray seals around the UK coastline. Once a breeding population established in the Wadden Sea, the population quickly grew (Brasseur et al., 2015).

Ecological dynamics beyond single-species growth becomes an even more severe complication to the null model notions about SHA. In natural systems, multiple species interact, either directly or indirectly. If the effects can be considered unidirectional, then the focal species can be regarded from a Hutchinsonian viewpoint (Hutchinson, 1957) by including all other influential species (e.g. prey, predators) as resource or risk dimensions of the environmental space. However, if interactions are fully bi-directional, then we need to take an Eltonian viewpoint (Soberón, 2007) to acknowledge that a species responds to the environment, but also influences the availability and behavior of properties of its environment. We have already discussed how consumer resource dynamics can complicate our inferences. Although we looked solely at the mechanism of depletion, predators may also impact the distribution of a mobile prey through indirect mechanisms such as landscapes of fear. Other mechanisms of environmental alteration without direct depletion include various forms of ecosystem engineering (Odling-Smee et al., 2013) and indirect community effects. For example, tree shadows reduce wind below the canopy and increase ambient humidity. Blue mussels (*Mytilus edulis*) attach themselves to each other and collectively create a structure that is less susceptible to environmental stress, like waves and currents, and can house a multitude of other organisms (Fig. 1.9). There are currently few attempts to fit dynamical, multi-species SHA models to field data (but see Ovaskainen & Abrego, 2020), a task that seems as intimidating as it is worthwhile.

Figure 1.9: The blue mussel (*Mytilus edulis*) is an important ecosystem engineer in coastal intertidal and subtidal ecosystems. Its association with its environment is bi-directional; Its presence can strongly alter the type of habitat, influencing the presence of other species. An Eltonian viewpoint is required to quantify their habitat association. Photo credit: Cornish Mussels by Mark A. Wilson – Public Domain.

Most of the above obstacles leading to an absence of the focal species from apparently suitable habitat are overcome by plastic (e.g. behavioral) responses of individuals or numerical responses of populations. Evolutionary responses may also occur at the species level and at potentially slower rates, although sometimes rapid selection occurs on time-scales similar to population dynamics (Bassar et al., 2010). Hence, a portion of the species may be able to live in a new region outside the current species range, but they may have yet to reach it. This discussion highlights the possibility that species tolerances are fluid, and hence that SHAs are non-stationary due to rapid evolution. Evolution would violate the assumptions of even the most sophisticated statistical SHA models, but is currently low on the wish-list of extensions.

1.8   The SHA models we wish for and the SHA models we have

Our proposed null model in section 1.6.2 may be new, but it spells out the implicit assumptions of nearly all SHA models in existence today. Their key desire is to assume that the study species is found at all suitable places and is absent from all unsuitable places (Guisan & Thuiller, 2005; Gallien et al., 2012). This is known as the pseudo-equilibrium assumption. But why the prefix "pseudo"? Because every ecologist is fully aware of the mechanisms that can violate our null model assumptions. For instance, every ecologist can recognize that species abundance may be capped above or below so that it is not consistently indicative of habitat suitability gradients. Species can survive indefinitely in areas of negative fitness by using habitats in a complementary way or by buffering against brief periods of adversity. Species may be observed temporarily in completely unsuitable areas due to extinction debts, but conversely, they may be absent from areas of positive fitness because of colonization credits.

The *raison d'etre* for the pseudo-equilibrium assumption (and, by implication, our null models) is that we wish to draw statistical inferences about fitness, suitability, viability and fundamental niches, but, instead, we have data on occurrence from museum records, line- or point-transect detections, individual telemetry tracking and spatial mark-recaptures. The type of SHA model that has so far been used to achieve the transition from distribution to fitness (and back again) is the Species Distribution Model (SDM). Its fitness inferences are valid only as a manifestation of our null model and the highly restrictive assumptions that it entails. Nevertheless, the general category of SDMs is currently our only route to modeling SHAs and must therefore be seen as a precious foundation on which to build. What kind of research program does this suggest for the future? The process of extending the SDM to become a fully-fledged SHA model will entail first, a recognition of the difference between fitness and distribution, second, an enumeration of the mechanisms that make them different (the mechanisms that violate the assumptions of the null model) and, third, the addition of modeling features in-between the state variables of fitness and distribution, to correct for these violations. We have taken the first and second of these steps in this introductory chapter. The third step is more of a voyage, and will take the rest of this book, and beyond, to complete.

To close this chapter, we present a comprehensive synthesis of the current versions of SDM that are encountered in the literature. Bear with us while we first pretend that there are only pure-forms of SDMs (section 1.9), to help us define the distinct flavors. We will then go on to present the results of cross-fertilization between them (section 1.10).

1.9   A puritan taxonomy of SDMs

At first sight, the diversity of methods available for converting spatial data to prediction maps can seem overwhelming. However, there is an emerging hierarchy in the methodological literature that considerably simplifies our effort to outline recommendations for best practice (Fig. 1.10). We can present this as a succession of four branchings, leading up to our preferred approach of inhomogeneous point process models

Spatial predictions can be generated by building mechanistic models of animal movement and demography from first principles and scaling them up to population distributions (Moorcroft, Lewis, & Crabtree, 1999, 2006; Moorcroft & Barnett, 2008; Moorcroft, 2012). Arguably, models with lots of biological detail (e.g. based on principles of physiological tolerance, movement behavior and social interactions) bring greater insights and predictive capability (Kearney & Porter, 2009; Hefley & Hooten, 2016). However, mechanistic modeling can
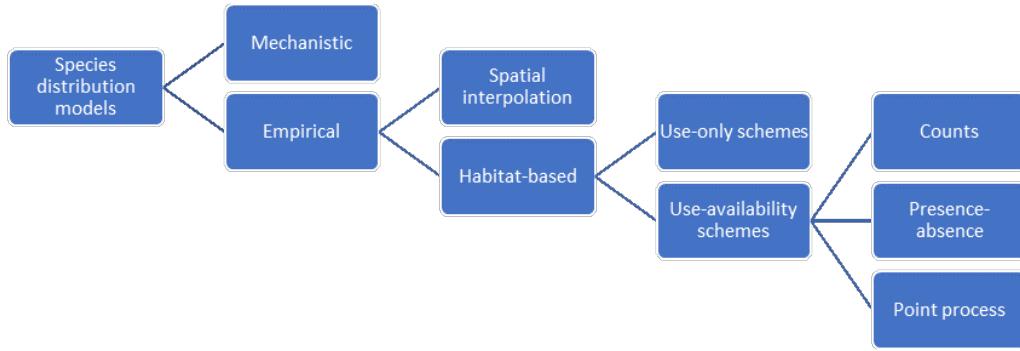
Figure 1.10: An overview of available methods for modeling species distributions leading up to our recommended approach.

be quite demanding technically and is generally more vulnerable to model misspecficification (models that disagree with the data) and parameter identifiability issues (situations where the data are not sufficient to estimate all model parameters). These problems can only be revealed if models are confronted with data. For example, the series of papers by Lewis and Moorcroft cited above, which form the state-of-the-art in mechanistic distribution modeling, require specific assumptions about animal movement and rely on sufficiently mathematical users who can formulate and manipulate partial differential equation models. Alternatively, we may choose to employ empirical, statistical models (Guisan & Zimmermann, 2000; Guisan & Thuiller, 2005; Guisan et al., 2017). *The deciding trade-off between mechanistic and empirical models is one of realism and predictive capacity against the ease of use and robustness to misspecification.* In this book, we build up from the foundations of the empirical model, but all of its extensions strive to increase its mechanistic content.

Within the class of empirical models (see review by Matthiopoulos & Aarts (2007)), we can distinguish between models that merely reconstruct the spatial density of a population (such as kernel smoothing, additive smoothing, or geostatistical methods) and regression methods that rely on habitat information as explanatory variables. Spatial density estimation methods rely on geographical proximity and the existence of spatial autocorrelation (Levin, 1992) to interpolate between observation points and map density in unobserved space, or alternatively, to smooth a finite data set of synoptic observations into a population-level expectation of usage. Density estimation methods focus on removing spurious variability from the predictions, but aim to stay as close as possible to the observations. Therefore, their ability to describe the available data is often better than that of habitat models (Bahn & Mcgill, 2007). Habitat models, on the other hand, are not by default spatial, since they are fitted in environmental (or niche) space (Pearman, Guisan, Broennimann, & Randin, 2008). Consequently, the greater ability of habitat models to interpolate and extrapolate spatially relies on the quality and relevance of their underpinning covariates. *The deciding trade-off between density estimation and habitat models is one of faithfulness to the particular distributional data collected and the ability to extend predictions beyond the spatial and temporal frame of data collection.* We aspire to generality, so we build on the habitat model as a foundation, but later return to spatially explicit extensions.

Within the class of habitat models, we distinguish between two main categories. The first, are known as profile methods and they argue that knowledge of where, in niche space, a species occurs is sufficient to understand its fundamental niche and map its current and future distribution. Broadly, this category includes methods such as climate envelope models and the use of multivariate statistical methods such as principal component analysis (PCA) for the analysis of presence-only methods (reviewed in Pearce & Boyce, 2006). The alternative class of use-availability schemes either contain representative information on the distribution of organisms (i.e. presence and absence), or they supplement presence data with availability data, allowing the models to contrast the habitat choices that organisms made, with the options that they had available to choose from. The broad area of use-availability schemes includes the vast literature on resource-selection functions (Boyce & McDonald, 1999; Morris, Proffitt, & Blackburn, 2016) and maximum entropy approaches (Elith et al., 2011; Merow, Smith, & Silander, 2013). *Profile methods have been critiqued extensively in the methodological literature* (see Pearce & Boyce, 2006 for a review), *and there is really no sound scientific reason for choosing to use a profile method.*

The final decision stage is mostly perceptual, relating to how space is conceptualized for the purposes of modeling the data. For example, space may be thought of as a regular grid (e.g. comprising squares, or other regular forms of tessellation, such as hexagons - see Grecian et al., 2016). In that case, the spatial data take the form of counts and are modeled by appropriate probability models such as the Poisson. Alternatively, and more realistically, space may be thought of as continuous and different spatial locations may be characterized by whether a species was present or absent. In that case, spatially referenced data take the form of zeroes and ones and the most appropriate probability model is Bernoulli (Aarts, MacKenzie, McConnell, Fedak, & Matthiopoulos, 2008; Gelfand & Shirota, 2019; Gelfand, 2020). Yet another approach within the continuous space framework is to imagine that observations of organisms appear at random locations almost like pin-lights that blink in and out at different time frames of observation. This framework, known as the Inhomogeneous Point Process (IPP) models the occurrence of events within a unit of time and space as originating from a smooth intensity surface, describing the instantaneous and infinitesimal rate of the Poisson process (Aarts, Fieberg, & Matthiopoulos, 2012; Fithian & Hastie, 2013; Renner et al., 2015; Fletcher et al., 2019; Miller, Pacifici, Sanderlin, & Reich, 2019). It is an elegant approach that makes an implicit comparison between use and availability, captures heterogeneities in the distribution of the population (e.g. due to environmental covariates) but, can with equal ease, use the intensity surface to represent heterogeneities in the distribution of spatial observation effort (so that, regions that receive no observation effort will have a zero intensity when modeling the data). *The deciding trade-off between count, presence-absence and point process models is in how the data are recorded and whether the user feels comfortable in conceptualizing infinitesimal quantities.* We consider the IPP to be the most general framework from which to view all other existing approaches and as the underlying generating process to all types of spatial data. We will introduce the IPP formally in Chapter 3.

1.10   Hybridisation of SDMs

The tapestry of different modeling approaches to species distributions is often perceived in as fragmented a way as we followed in the preceding section. This presentation often gives the impression that the only way to deal with the multiplicity of approaches is to compare their performance and choose the "best" (e.g. Oppel et al., 2012). This performance comparison is not a fruitful approach when we can strive for the best-of-several-worlds solution. Indeed, there is a methodological kinship between many of these approaches that is rarely apparent in the applied literature. Having so far organized the literature as a sequence of three strict dichotomies and a final trichotomy (Fig. 1.10), it is good to take a more synthetic and conciliatory view on the above decisions. It is, in fact, the case that a hybrid approach is possible that retains the best elements of all the approaches discussed above.

Specifically, if one starts from a purely empirical model, it is possible to move it towards a higher mechanistic content. In the simplest case, this can be done by carefully considering the biological relevance of the set of covariates that are offered to the model (Bell & Schlaepfer, 2016). It is also possible to construct more sophisticated covariates using mechanistic models to try and increase the explanatory power of empirical models (Kearney & Porter, 2009; Matthiopoulos et al., 2015). More recently, it is becoming possible to fit structurally complex models directly to data either by likelihood approaches, but most often, via Bayesian approaches. These developments have come mostly from the field of integrated population modeling (John R Fieberg et al., 2010; Matthiopoulos et al., 2014; Zipkin & Saunders, 2018; Yen et al., 2019), as well as data-integration; the main benefit to integrative frameworks is their capability to deal with nonlinearity, a feature characterizing most biologically realistic (i.e. mechanistic) models.

Also, the separation between spatial and habitat-based models can be made less strict. Geostatistical models can accept habitat covariates and habitat models can accept spatial autocorrelation structures (Dormann et al., 2007). A considerable advantage of these models is that they can separate effects caused by habitat covariates and dependencies in the data (but see Hodges & Reich, 2010). A perceived (but not entirely accurate) limitation of this approach is that predictions from hybrid spatial-habitat models are tied to the spatial extent of the data collection (i.e. the study area). However, inclusion of spatial autocorrelation structures can lead to a more accurate representation of the link between the species and environmental variables, which may also improve predictions outside the study area (see also Chapter 2).

The separation between use-only (or, profile) models and use-availability models may appear to be the most clear-cut of the branchings in Fig. 1.10. Profile methods assume that different habitats are equally available (Merow et al., 2014), and therefore, they interpret the habitat choices of organisms as purely the result of preference, not as a combination of preference and availability (Matthiopoulos, 2003a). Profile models also misappropriate the ecological term "niche" because they aspire to define a species' viable hypervolume in environmental space, and yet make no explicit connection between habitat data and population trends representing information on viability (Peterson et al., 2011; Matthiopoulos et al., 2015). Therefore, profile methods have some fundamental flaws from an ecological perspective. And yet, despite their limitations, their aspiration is worthwhile and connections to the use-availability model can be made. Using habitat models to make sense of population viability should be a key objective in our search for defining critical habitat and in driving conservation efforts. Recent publications (Matthiopoulos et al., 2015, 2019) have shown how this can be achieved in practice by using the more defensible option of use-availability methods as a platform to build upon.

Finally, the decision between different approaches to modeling space is driven by the available data, but it can be argued that the distinctions between approaches are not clear-cut because the IPP can be thought of as a data-generating process for all of them. Data types commonly used for SDMs are count, presence-absence and presence only (Hefley & Hooten, 2016). Count data can be divided into point counts (e.g. point or line transects) and quadrat counts (comprehensive count in an area) (Hefley & Hooten, 2016), although the distinctions between those two can be blurred. Presence-absence (or occupancy) data may either originate from count data that have been converted to binary form, or they may be the result of survey effort units that were terminated as soon as the species was detected once (Hefley & Hooten, 2016). Finally, presence-only data may include observations from known survey effort units (e.g. telemetry), or alternatively unknown effort surveys (such as museum records, or some citizen science programs). Several papers (Warton, Shepherd, & Others, 2010; Aarts et al., 2012; Fithian & Hastie, 2013; Hefley & Hooten, 2016) have shown that the separation between count, presence absence and point process models is not substantial. Indeed, all of these methods can be thought of and re-formulated under an Inhomogeneous Point Process Model. Furthermore, widely used spatial modeling packages such as MAXENT, can be thought of as point process models (Fithian & Hastie, 2013; Renner & Warton, 2013). Conversely, computational methods used for efficiently fitting point process models to data make use of spatial discretization, similar to grid-based methods, but using more efficient schemes tailored to the data (Lindgren & Rue, 2015).

1.11 Concluding remarks

For presentational reasons, we will introduce SDMs in two parts, which we will tentatively name process and observation models. The process model (which will form the subject of Chapter 2) encompasses the mathematical relationships between habitat variables and predicted usage or fitness. In these early discussions, the boundary between fitness and distribution will be blurred, because we will be operating under the pseudo-equilibrium assumption. In later chapters, as the assumptions of the null models become relaxed, the distinction between suitable and used habitats will become more detailed. By the same token, the distinction between an SDM and a SHA model will be non-existent, in these early chapters.

The second component of an SDM, the observation model (which will form the subject of Chapter 3) will include the statistical and conceptual machinery needed to fit the process model to distributional and environmental data. Therefore, in Chapter 2, we will discuss how to go from mathematical models of fitness and suitability to spatial maps of species distribution, and in Chapter 3, we will think of the data generating process in stochastic terms and discuss how to use spatial data to parameterize the underlying process model.

Chapter 2

Modelling Species-Habitat Associations

2.1 Objectives

The objectives of this chapter are to:

1. Explore the function and scope of the *process component* of a Species Habitat Association Model. How can we translate basic ecology into basic models?

2. Outline the two necessary ingredients of process models for species-habitat associations: the *availability* of habitats in the environment and their resulting *usage* by animals, or plants.

3. Present the Use-Availability (U-A) approach, a broad conceptual framework that encompasses all existing SHA process models and can be expanded to incorporate more biological realism.

4. Introduce the connection between U-A models fitted in *environmental* space (*E*-space) and the expected usage surfaces mapped in *geographical* space (*G*-space).

5. Explain how the structure of the process model can describe responses to resources, conditions and risks, capture trade-offs and complementarities between resources.

6. Suggest future extensions to process models that are not currently implemented in published studies, but are anticipated by theoretical models.

2.2 You are here

This chapter focuses on how the biological questions and challenges from Chapter 1 can be formalized into expandable mathematical frameworks. To link up with the biological motivation of SHAs, the underlying frameworks seek to address questions about individual selection (a behavioral process) and state (a physiological and life history process). They also aspire to be relevant to predictions about population viability and demographic functions. These biological motives are best satisfied in environmental (or niche) space. However, this chapter also aims to link forward to Chapter 3, by translating the predictions of environmental models into observable, geographical space. The first three chapters together, achieve the transition from questions-to-data and back again.

2.3 Formalising the association between habitat and species

The most common image conjured up when considering species-habitat associations is of several spatial layers of environmental variables being fed into a statistical model, that produces (often, with a considerable degree

of automation) a corresponding species layer (Fig. 2.1). This output layer is thought of as representing some notion of habitat suitability, or expected usage.

Within the SHA community, this kind of representation (a glorious sandwich of geographical layers achieved by a statistical model) has immense appeal for five reasons. First, geographical space is intuitive. It is, after all, the natural space that we inhabit, we perceive, and have built intuition in. Second, animals and plants give us evidence of their preferences for particular habitats by coinciding spatially with those habitats. Third, statistical models are simple, quantitative and sufficiently removed from biological mechanism. This makes them objective (a good thing) and, at the same time, almost immune to biological criticism (a really bad thing). Fourth, color maps of habitat suitability are intrinsically beguiling and often difficult to dispute, especially when the quantity they depict is not defined precisely. Finally, the geographical viewpoint is partly a historical relic from earlier times when studies did not use statistical models, but just visually compared different GIS layers. All these sources of appeal (to which we must also admit our own vulnerability), have made the SDM workflow appear understandable and, even, trivial.

There is little doubt that we need to (somehow) funnel the influence of multiple environmental variables into a single map representing habitat suitability, so the general notion of Fig. 2.1 is correct, but its implementation into models conceals two pitfalls. As the users and developers of these methods, we need to consider exactly how the input layers (let us call them $x_1, x_2, x_3, ...$) are combined within the model's black box, and we also need to define, precisely, the meaning of the model's output.
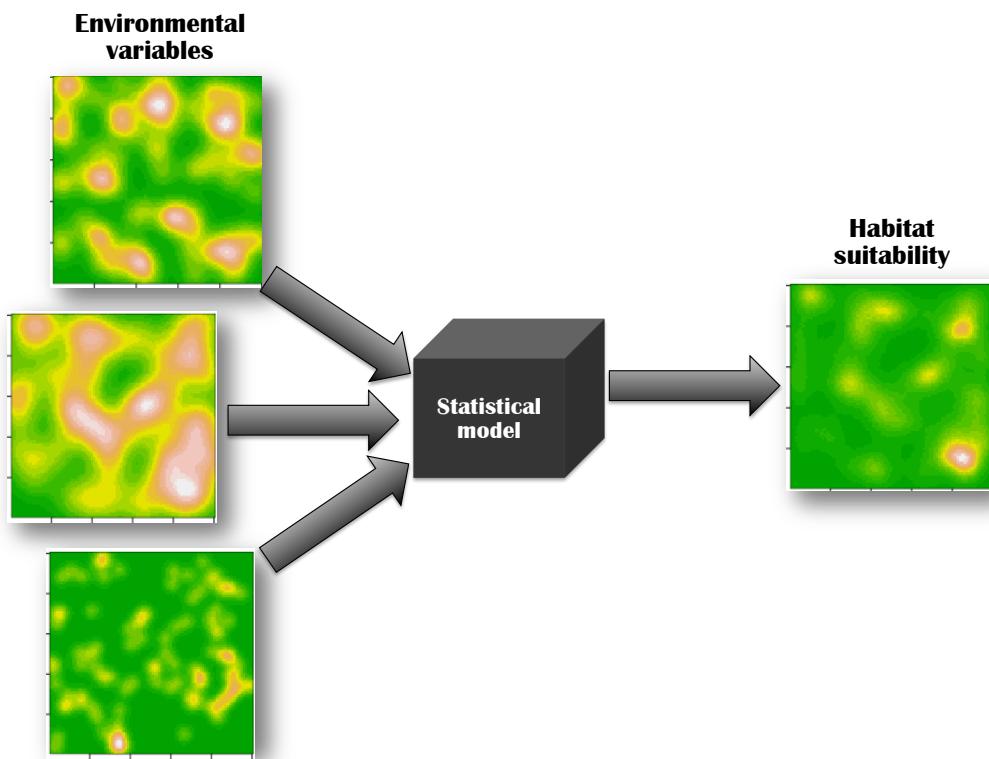
Figure 2.1: Species-habitat association models are usually imagined in geographical space as a superposition (a combination) of environmental layers, into some measure of habitat suitability, via a statistical model. The statistical model is often treated as a black box, and the emphasis is placed on the relative contribution of the input layers to the resulting map.

The exact mathematical translation of Fig. 2.1 is the addition of multiple layers into a single, resultant map.

Then, each cell of the output layer would be calculated by the corresponding values $(x_1, x_2, x_3, ...)$ of the cells of the input layers

$$\sum_{k=1}^{K} x_i = x_1 + x_2 + ... + x_K \tag{2.1}$$

Of course, not all inputs are equally important for the resultant (however we choose to define this resultant, biologically), so we should allow for the possibility of adjusting the contribution of the different layers according to some coefficients α. Something like

$$\sum_{k=1}^{K} \alpha_k x_k = \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_K x_K \tag{2.2}$$

If we impose no constraints on the sign and value of these coefficients and input layers, this kind of expression can take any values between $-\infty$ to $+\infty$, which is not ideal, if we want to use it to describe things like expected usage (a non-negative quantity). So, we may need to constrain this expression in an appropriate way, depending on the definition of the output layer.

More challenges like these soon present themselves. For example, there is nothing to say that such additive combinations of the inputs are realistic. As we saw in Chapter 1, environmental variables can act in ways that are highly non-linear. Trade-offs, synergies, complementarities between variables are the biological rule, not the exception in how the environment affects the distribution of species. Also, when asking population-level questions, how can we account for the fact that individual animals and plants can behave in different ways in different contexts and during different life-history stages? How can we account for population composition, density-dependent processes and evolutionary pressures?

If we think biologically about these problems (as we always should), then suddenly, the simplistic, phenomenological approach of Fig. 2.1, seems very far away from the questions we are trying to answer with it. How do we attempt to move from pattern-driven (phenomenological) models, to process-based (mechanistic) ones?

The process component of a SHA model is ultimately intended to encapsulate all of the ecological truths that can be feasibly extracted from the available data via the process of model fitting (see Chapter 3). Ideally therefore, the parameters built into the process model should be biologically interpretable. In theory, a fully-detailed process model should be able to inform us about the behavioral decisions mediating habitat selection, the energetic and nutritional trade-offs and synergies of resources on animals and plants, the effect of conspecifics and heterospecifics on the expected distribution of usage, etc. However, the degree to which process parameters are amenable to interpretation depends on how mechanistically-motivated a process model is. Traditionally, research in this area has tended to think of SHA process models as purely phenomenological (e.g. Generalized Linear Models - GLMs), and has attempted to catch glimpses of biological causality by observing the direction and magnitude of correlations between usage and its suspected covariates. Have a look at eq. (2.2), which already looks like the linear predictor of a GLM. The parameters α_i are a far-cry from basic biological parameters with clear physical interpretation (such as an individual's basal metabolic rate or the handling time needed to consume its prey).

However, the underlying process framework need not be constrained solely to providing such indirect inferences. For many years, we have had access to process models with high levels of biological detail (Nisbet & Gurney, 2004; Otto & Day, 2011; Murray, 2013), and recently, we have been acquiring the statistical ability to confront them with field data (Newman et al., 2014; Hooten, Johnson, Mcclintock, & Morales, 2017; Kery & Royle, 2020). A good strategy, when approaching the empiricism-mechanism dilemma, is to start with an empirical model that can be incrementally (and strategically) expanded with mechanistic detail. Therefore, in this chapter, we take the broadest possible view of process models in SHAs, aiming to embed simple phenomenological models in their true context and, also, to signpost the routes that can be followed to flesh-out these skeleton empirical models with more biological mechanism.

2.4 The meaning of the SHA model output

So, how should we interpret the output of the SHA process model? There are a great many options to choose from, unfortunately, and not all of them are correct for particular process models. For example, for most currently available models, the interpretation of the output as a measure of habitat suitability is more wishful thinking than truth. As we explained in Chapter 1, organisms are frequently found at places that are not suitable for them (e.g. either because they have to transit through them, or because they overflow from nearby suitable places). Conversely, organisms are frequently not found at places that may be suitable for them (e.g. either because they have not yet extended their range to include these places, or because they need complementary access to other habitats, that happen not to be nearby). Various flavors of "niche-related" models (Hirzel, Hausser, Chessel, Perrin, & Jul, 2002; Hirzel & Le Lay, 2008; Broennimann et al., 2012; Robertson; Caithness, & Villet, 2012) assume that population density is proportional to fitness. This presumption will influence the biological conclusions to an unknown extent (Peterson et al., 2011). In Chapter 1, we introduced a null model for SHAs that preserved the correspondence between distribution and fitness, precisely to allow us to discuss and develop the main ideas of SHAs without worrying about these additional biological complications. We will return to these problems (and proposed solutions) repeatedly in this book, but for now, we need a working definition of the model output.

The final output of a SHA process model represents expected spatial usage by an individual or a population

Here, we use the term *expectation* in its statistical sense (as a population mean or long-term average). *Spatial usage* may be defined as a relative measure, a proportion attributed to each cell in geographical space, or a probability density associated with each set of spatial coordinates. Alternatively, in the case of entire populations, *usage* may be defined as an absolute amount, integrating to the total population size, over the whole of space.

2.5 The process model, as a black box: Inputs and outputs

We begin by considering the inputs and outputs of a SHA model, particularly focusing on their constraints. In general, the process model can be thought of as a function h that takes input values from a domain $\mathbb{X}$ and gives output values in a range $\mathbb{U}$.

$$U = h(\mathbf{x}) \qquad\qquad h : \mathbb{X} \to \mathbb{U} \qquad\qquad (2.3)$$

The domain, range and structure of the mathematical model can vary in their definition and complexity. If we consider the process model as a black box operating on its input(s), to produce its output(s), the possible definitions of the domain and range can be quite informative.

For example, the domain $\mathbb{X}$ can be thought of environmentally, spatially and temporally. A spatial domain $\mathbb{G}$, or *G-space* may be defined in terms of latitude and longitude on a map (with the possible addition of altitude, or sea-depth as a third spatial dimension). Any subset of *G*-space is possible (e.g. resulting from the boundary of a particular study region). An environmental domain $\mathbb{E}$, or *E-space* may be defined in terms of continuous or discrete environmental variables. Examples of continuous variables include temperature, humidity, food abundance, predator abundance etc. Examples of discrete variables include pre-designed classifications of the environment into habitat classes (as is often done for the purposes of preparing color maps by Geographic Information Systems). A temporal domain $\mathbb{T}$, or *T-space* is defined as any subset of the real numbers signifying a time interval. Most SHA models are defined in *E*-space (so that, $\mathbb{X} = \mathbb{E}$). All purely spatial models (e.g. most geo-statistical or trend-fitting models) are defined in *G*-space (so that, $\mathbb{X} = \mathbb{G}$).

Combinations of those spaces are, of course, possible and there are some examples (Hedley & Buckland, 2004; Augustin, Trenkel, Wood, & Lorance, 2013) of habitat models that account for residual spatial autocorrelation by using spatial dimensions (so that, $\mathbb{X} = \mathbb{E} \cup \mathbb{G}$). Such models effectively fold the spatial location into the definition of "habitat" (making each location a unique habitat, and each habitat a unique location). Such geographical extensions of the process model are made mostly as heuristic tools to increase the variability

Table 2.1: Examples of the range of SHA models

Domain ($\mathbb{X}$)	Examples
$\mathbb{E}$	Species association with food and water
$\mathbb{G}$	Description of a spatial trend in species density
$\mathbb{E} \cup \mathbb{G}$	A species invasion gradient interacting with habitat variables
$\mathbb{E} \cup \mathbb{T}$	Seasonality in habitat responses
$\mathbb{G} \cup \mathbb{T}$	A dynamic gradient in space
$\mathbb{E} \cup \mathbb{G} \cup \mathbb{T}$	Seasonal responses to habitat on a multi-annual invasion gradient

explained by SHA models and, on some occasions, they can motivate ideas about the shape of missing covariates.

The temporal dimension is occasionally employed to augment the domain ($\mathbb{X} = \mathbb{E} \cup \mathbb{T}$), leading to the construction of dynamical SHA models (Augustin et al., 2013; Hooten, Hanks, Johnson, & Alldredge, 2014).

Table 2.1, shows some increasingly fanciful possibilities of process model domains, but for much of this book, we will consider SHA models fitted exclusively in *E*-space. Although we may occasionally consider spatial proximity between habitats, we will not usually employ the specific location within the process model.

The process model associates a value of absolute or relative usage, with each point in the model's domain (see Section 2.4). Therefore, for the vast majority of SHA models, the range $\mathbb{U}$ of values is assumed to be one-dimensional. Indeed, with no loss of generality, we can think of the output (the relative usage of any given habitat $\mathbf{x}$) as a proportion of an individual's time or the portion of a whole population, so that so that, $\mathbb{U} = [0, 1]$. Conceivably, it is possible to extend our process model's range to multiple dimensions. Such models could, for example, simultaneously look at the associations of multiple species with habitat (Guisan & Zimmermann, 2000; Ovaskainen, Hottola, & Shtonen, 2010; Wisz et al., 2013; Ovaskainen & Abrego, 2020).

2.6   Usage in Geographical and Environmental spaces: Model transferability

Let us consider the example of two environmental variables that act as resources for a particular species. In Figs 2.2a,b we show two such examples plotted in *G*-space. We assume that the organism can aggregate its usage in regions of space that offer a good combination of both resources (according to some unknown preference function, expressed by the organism through some unknown selection process). This behavior gives rise to a utilization distribution shown in Fig. 2.2c. The colors in that figure represent the frequency of use of a particular location $\mathbf{s}$ in geographical space (dark greens for low usage and light browns for high).

However, most of our SHA models are not designed to be explicitly spatial, for one very good reason to do with *model transferability*. A model that is formulated to produce predictions about particular points in space $\mathbf{s}$ is only directly usable for prediction if we are interested specifically in those points[1]. In contrast, a model that anticipates usage of a particular habitat $\mathbf{x}$ has the potential to give us useful predictions for times and places other than those for which we have data. Achieving transferability in predictions is the holy grail of SHA models (Randin et al., 2006; Tuanmu et al., 2011; Wenger & Olden, 2012; Sequeira, Bouchet, Yates, Mengersen, & Caley, 2018; Yates et al., 2018; Qiao et al., 2019), and a core theme of this book. Although this highly desirable property cannot be achieved simply by thinking and modeling in *E*-spaces, doing so is a prerequisite for transferable SHA models.

---

[1]Hence, a model that uses longitude and latitude as fixed or random effects is tied to the geographical region in which it was fitted. This may be a good idea when controlling for missing covariates, but in order to predict in other regions the spatial terms must be dropped. Such an approach (i.e. explicitly spatial during the model fitting stage, but non-spatial during the prediction stage) could increase the robustness of predictions by accounting for residual spatial autocorrelation (although it may inadvertently lead to the removal of covariates that are important Hodges & Reich, 2010).

It therefore helps to be able to think of biological processes in *E*-spaces, i.e. spaces whose dimensions are environmental variables, rather than geographical coordinates. In principle, they are more challenging to visualize because they can very easily become high-dimensional (no ecological system exists in more than three geographical dimensions, but most systems are usually influenced by a myriad of environmental variables). However, thinking about *E*-spaces in one or two dimensions (i.e. in terms of one or two environmental variables) builds intuition that is readily transferable to more complex, multidimensional *E*-spaces. Plotting the observed usage of *G*-space (Fig. 2.2c) in *E*-space, requires us to identify all G-cells that are characterized by similar combinations of resources, and to sum the proportions of usage they contain. We will use the symbol *f* to denote frequencies plotted in this way in *E*-space. Since each cell in this plot represents the frequency of usage of a habitat $\mathbf{x}$, we will denote this quantity by $f_u(\mathbf{x})$.

The result is shown in Fig. 2.2d, plotted in a different color scheme, to emphasize the transition from G- to *E*-space. Cells close to the origin of this plot represent habitats that are low in both resources, whereas habitats towards the top right corner represent utopian habitats with an embarrassment of riches. In that map, cold blue colors represent lack of usage, whereas warmer colors (towards yellow) show us habitats that are used a lot. As might be expected, the habitats at the bottom-left of *E*-space are not used very much, but, perhaps less intuitively, the utopian habitats at the top right are not used at all. What is going on?

2.7 Lifting the black box lid: Habitat usage and habitat availability

Well, the answer becomes easy to guess if, instead of usage, we focus on visualizing habitat availability in *E*-space (Fig. 2.2e). This new quantity, let us call it $f_a(\mathbf{x})$, represents the frequency with which habitats of similar characteristics appear in the environment at-large. Notice the dark blue colors at the top-right of that plot: Utopian habitats do not *physically* exist, which might explain why they appear not to be used in Fig. 2.2d. The pattern of apparent habitat preference, is better revealed in Fig. 2.2f, where we have attempted to *control* for the effects of habitat availability by plotting the ratio of usage over availability $f_u(\mathbf{x})/f_a(\mathbf{x})$. The white areas of this plot are essentially non-existent habitat, so the picture is not strictly complete, however, the colored areas reveal a very simple tendency in habitat preferences of "more-is-better", as the color gradient goes diagonally from low (bottom-left) to high (top-right).

Quantifying the availability of a habitat is an essential part of any attempt to quantify preference for that habitat (Johnson, 1980). It is generally impossible to express total usage of a habitat purely as a function of preference (Matthiopoulos, 2003a) because doing so would imply that organisms can use habitats that do not physically exist. If this point seems self-evident, then perhaps it is valuable to highlight that there are several methods that make just this assumption (Hirzel et al., 2002; Broennimann et al., 2012; Robertson; et al., 2012).

Fig. 2.2 also reveals another common truth about G- and *E*-spaces. In general, objects in *G*-space are more likely to have multiple peaks and troughs, since there may be several different areas in *G*-space with similar properties. Indeed, the whole part of the ecological literature that deals with patchy environments and metapopulation occupancy patterns, is a direct result of that fact. However, *E*-space gradients in habitat availability, usage and especially preference, are comparatively simpler constructs, that are more easily captured by monotonic, or unimodal mathematical models (Matthiopoulos et al., 2020). If you cast your mind back to our classification of environmental variables from Chapter 1, in theory, the behavior of habitat preference along any single dimension of *E*-space can be modeled as an increasing (resource), decreasing (risk), or unimodal (condition) function.

With all of the above in the background, it appears that to understand species-habitat associations, we need to consider habitat availability, because what an organism will be observed using depends on what it *needs/wants* and what it can *find/access*. We want to ensure that observed usage will be zero both for undesirable and for unavailable habitats. Multiplication is a good trick here[2]:

$$\text{What I use} = (\text{What I want}) \times (\text{What I can find}) \qquad (2.4)$$

---

[2]If you prefer to think in terms of probability, this expression can be written as $P(Use, Available) = P(Use|Available)P(Available)$. The conditional probability in this expression, represents habitat selection.
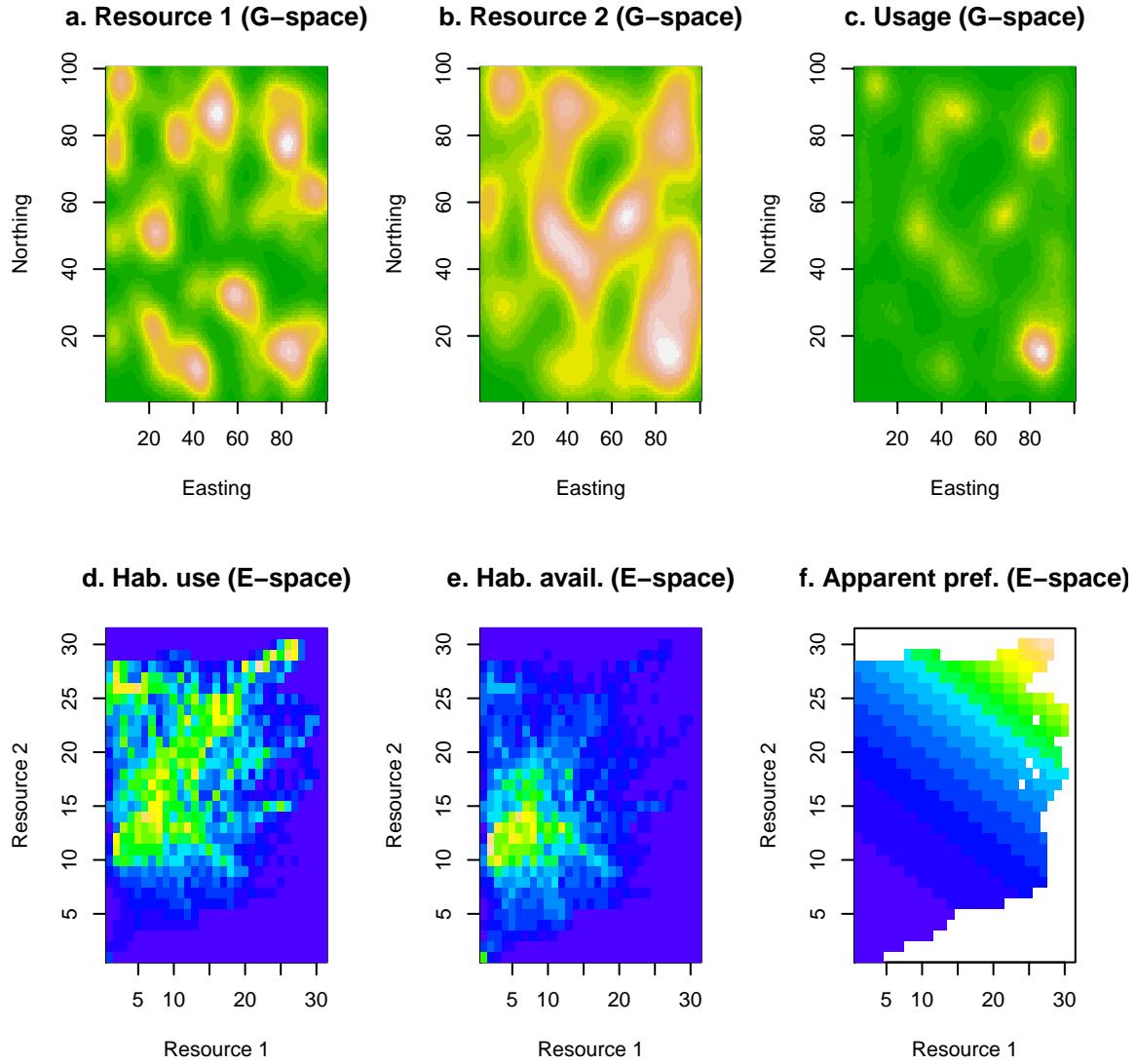
Figure 2.2: An example of use and availability of habitats in geographical and environmental space. We consider two geographical layers (a and b) representing the geographical distribution of two resources. On those, we superimpose the utilisation distribution (c) of a hypothetical animal. We can recast this information into environmental space. Usage of habitats in E-space (d) is, essentially, a histogram of the frequency with which different resource combinations are used by the animal. Similarly, the availability of habitats in E-space (e) is the frequency with which different combinations of resource abundances occur in the environment. A measure of apparent habitat preference (f) can be obtained by dividing habitat use (d) by habitat availability (e). Here, white space indicates division by zero (i.e. non-existent habitats).

The underlying biology of these loosely used words is considerable. The choices of organisms are not necessarily fully expressed by them within the frames of observation, or in a way that is always aligned with their fitness. Similarly, the availability of habitats to organisms relies on a complex interplay between landscape structuring, organisms' cognition and mobility (Matthiopoulos, 2003b; Matthiopoulos et al., 2020). Be that as it may, if we pretend for the moment that we have some intuitive understanding of these concepts, we can plausibly write the expected usage of a habitat in terms of some function (h) expressing habitat preference

$$f_u(\mathbf{x}) = h(\ )f_a(\mathbf{x}) \tag{2.5}$$

The function $h()$ is the core of the process model and will ultimately be required to absorb all the intricate biological mechanisms described in Chapter 1. So, how do we start giving it a form? Biologically, the simplest version of the function h is constant, implying that an organism will use a habitat *in proportion* to its availability.

$$f_u(\mathbf{x}) = hf_a(\mathbf{x}) \tag{2.6}$$

Indeed, if usage and availability are expressed as *density functions* (such that, $\int_{\mathbb{E}} f_u(\mathbf{x})d\mathbf{x} = 1$ and $\int_{\mathbb{E}} f_a(\mathbf{x})d\mathbf{x} = 1$), then the proportionality relation implies that usage is *equal* to habitat availability (i.e. $h = 1$). The biological scenarios that justify this kind of model are restricted to *indiscriminate* organisms, species that roam the entire landscape without preferentially selecting to move to, or even linger in, some habitats over others. Arguably, from the point of view of species-habitat association, this model is somewhat boring, because it implies no association at all. It is nevertheless, an appropriate null model for the process of selection, because it corresponds to uniform usage of G-space.

Deviations from this null model, are thought of as manifesting *disproportionate usage* of habitats compared to their availability. This is the point where we can start talking about *habitat selection*. Mathematically, we can achieve this behavior by writing h as a function of habitat.

$$f_u(\mathbf{x}) = h(\mathbf{x})f_a(\mathbf{x}) \tag{2.7}$$

Regarding the properties of $h(\mathbf{x})$, we know that it needs to be non-negative because it makes no sense to define a multiplier for availability that may lead to negative usage. We also know that it must preserve the unit-sum properties of usage because no individual or population can exist for more (or less) than 100% of their time in this world (i.e. $\int_{\mathbb{E}} f_u(\mathbf{x})d\mathbf{x} = 1$). Therefore, we need:

$$h(\mathbf{x}) \geq 0 \quad , \quad \int_{\mathbb{E}} h(\mathbf{x})f_a(\mathbf{x})d\mathbf{x} = 1 \tag{2.8}$$

In this first model of disproportionate usage, we can rewrite eq. (2.7) as

$$h(\mathbf{x}) = \frac{f_u(\mathbf{x})}{f_a(\mathbf{x})} \tag{2.9}$$

which gives us an intuitive model for the *habitat preference function*, as the ratio of a habitat's usage over its availability. If you need to visualize this, then just look at the example of Fig. 2.2f, which shows exactly this ratio across E-space.

2.8 Thinking inside the box: Mathematical formulations

Now, let us try and put some more details into our definition of the process model. We will do this incrementally, leading up to a correct expression for $h(\mathbf{x})$. Let us imagine that we want to write habitat preference as some simple linear combination $g(\mathbf{x})$ of the characteristics of a habitat. This follows the superposition rationale from

Section 2.3, and, in particular, eq. (2.2). In general, for a vector of n environmental variables $\mathbf{x} = (x_1, ..., x_K)$ this could be

$$g(\mathbf{x}) = \sum_{k=1}^{K} a_k x_k \tag{2.10}$$

We will call $g(\mathbf{x})$ the *predictor function*, a name that echoes the *linear predictor* functions of GLMs, but still allows a potential relaxation of linearity, in more elaborate models. For the simple example of habitat preference with two resources, seen in Fig. 2.2, the predictor function was

$$g(\mathbf{x}) = 0.04x_1 + 0.08x_2 \tag{2.11}$$

In this example, we had positive coefficients in response to both environmental variables but, in many cases (e.g. environmental risks), we might like to express avoidance for increasing values of an environmental variable. In these cases, the coefficients of the model may need to take negative values, and ultimately yield negative values overall for the preference function. This is expressly forbidden by eq. (2.8) which requires a non-negative valued function. There are many ways that an unbounded expression can be constrained to be non-negative (e.g. by taking its absolute value or raising it to an even power). The SHA literature customarily uses the exponential function. We will denote this transformed predictor function by $w(\mathbf{x})$

$$w(\mathbf{x}) = exp(g(\mathbf{x})) \tag{2.12}$$

We will call $w(\mathbf{x})$ the resource (or habitat) *selection function*, a well-established name for this function (Boyce & McDonald, 1999; Boyce, Vernier, Nielsen, & Schmiegelow, 2002). Although this ticks the positivity requirement for eq. (2.8), we still need to ensure that its second (unit-sum) requirement is observed. For example, as the values of x_1 and x_2 become larger across a landscape, there is nothing to stop eq. (2.12) from integrating to more than one. We can put a cap to that value by normalizing, as follows:

$$h(\mathbf{x}) = \frac{w(\mathbf{x})}{\int_{\mathbb{E}} w(\mathbf{x}) f_a(\mathbf{x}) d\mathbf{x}} \tag{2.13}$$

Alternative ways of writing the function h are

$$h(\mathbf{x}) \propto w(\mathbf{x}) \quad or \quad h(\mathbf{x}) = k^{-1} w(\mathbf{x}) \tag{2.14}$$

where

$$k = \int_{\mathbb{E}} w(\mathbf{x}) f_a(\mathbf{x}) d\mathbf{x} \tag{2.15}$$

The proportionality constant k^{-1} is needed if we want to eventually attach some numbers to the above symbols. Considering the value of k in detail is relatively unimportant for following the main thread of the story here, but it *does* hide a valuable story of its own. Calculating k from eq. (2.15) requires us to consider the relative response of organisms to *all* relevant habitats. This means that, ultimately, any quantification of the usage, or preference of a particular habitat, cannot be achieved without consideration of the usages and preferences of *all* accessible habitats in the environment of an organism. In other words, it doesn't matter how large the value of the selection function $w(\mathbf{x})$ is for a particular habitat $\mathbf{x}$, what matters is how much *larger* that value is compared to *all other habitats.*

We have so far presented three useful functions. The preference function $h(\mathbf{x})$, the selection function, $w(\mathbf{x})$, and the predictor function $g(\mathbf{x})$. These are nested within each-other as follows:

$$h(w(g(\mathbf{x}))) = \frac{exp(g(\mathbf{x}))}{\int_{\mathbb{E}} exp(g(\mathbf{x})) f_a(\mathbf{x}) d\mathbf{x}} \tag{2.16}$$

Bringing eq. (2.13) together with eq. (2.7) gives a complete expression for habitat use.

$$f_u(\mathbf{x}) = \frac{w(\mathbf{x}) f_a(\mathbf{x})}{\int_{\mathbb{E}} w(\mathbf{z}) f_a(\mathbf{z}) d\mathbf{z}} \tag{2.17}$$

Later, in section 3.5, we will see how this expression is employed by weighted distribution theory (Lele, 2009) to fit these models to data.

2.9 Spatial intensity from habitat preference

The development and interpretation of SHA models requires some familiarization with G- and *E*-spaces. We need to be able to move from physical to niche space and back again. In section 2.6, we made the forward transition (G- to *E*-space). In this section, we discuss how to go back. To visualize our predictions in *G*-space, we need a notion of usage per unit area of habitat. Let us consider the example of a species, observed in a completely synoptic way (i.e. all occurrences of every individual in the population are recorded). A total of 1000 observations are made. A total of 120 observations are recorded in a particular type of habitat. Taking a look at the map of the study area, we measure that the total area occupied by this habitat is $3000 m^2$. Therefore, the average number of observations (from our total sample of 1000) to be found in each m^2 of that habitat will be $\frac{120}{3000} = 0.04$. In general, therefore

$$\text{Usage per unit area of habitat } \mathbf{x} = \frac{\text{Total usage of habitat } \mathbf{x}}{\text{Total area of habitat } \mathbf{x}} \tag{2.18}$$

The proportion of usage of this population to be found in any one m^2 belonging to that particular habitat will be $\frac{0.04}{1000} = 4 \times 10^{-5}$. Generating this kind of (per-unit-area) calculation for every habitat type can allow us to populate each cell in *G*-space with values, leading to a map of expected usage by an individual, or a population.

Assuming that we have obtained functions for habitat use and availability in *E*-space, can we construct a per-unit-area calculation for plotting *G*-space maps? The trick is to look at the definition of habitat preference from eq. (2.9):

$$h(\mathbf{x}) = \frac{f_u(\mathbf{x})}{f_a(\mathbf{x})}$$

We can expand each of the three quantities in this expression. From eq. (2.14), we have

$$h(\mathbf{x}) = k^{-1} w(\mathbf{x})$$

From the definitions of $f_u(\mathbf{x})$ and $f_a(\mathbf{x})$ as probability functions, we can write:

$$f_u(\mathbf{x}) = \frac{\text{Total usage of habitat } \mathbf{x}}{\text{Total usage of space}} \tag{2.19}$$

$$f_a(\mathbf{x}) = \frac{\text{Total area of habitat } \mathbf{x}}{\text{Total area of space}} \tag{2.20}$$

Since k, the total usage of space and the total area of space are constants across some subset of geographical space (say, our study region), we can rewrite eq. (2.9) as

$$w(\mathbf{x}) \propto \frac{\text{Total usage of habitat } \mathbf{x}}{\text{Total area of habitat } \mathbf{x}} \tag{2.21}$$

Therefore, $w(\mathbf{x})$, as given in eq. (2.12), is proportional to the use of a cell of type $\mathbf{x}$ in G-space. We will use $\lambda(\mathbf{s})$ to denote the expected usage at a geographic location $\mathbf{s}$. This gives us the fundamental relationship between E-space and G-space.

$$\lambda(\mathbf{s}) \propto w(\mathbf{x}(\mathbf{s})) \tag{2.22}$$

The expected usage of a geographical location ($\mathbf{s}$) is proportional to the value of the selection function associate with the habitat characteristics at that location ($\mathbf{x}(\mathbf{s})$). Eq. (2.22) also delivers a simple algorithm for generating spatial maps of expected usage from a SHA model. Here are the steps:

1. Consider any spatial location $\mathbf{s}$ in G-space
2. Take the vector of habitat characteristics $\mathbf{x}(\mathbf{s})$ at the coordinates of that location
3. Express the value of the resource-selection function $w(\mathbf{x}(\mathbf{s}))$
4. Assign that value to the location $\mathbf{s}$
5. Repeat for all locations in G-space
6. Normalize all the values so that they add up to 1 (i.e. 100% of usage) for the whole of space

For example, consider the maps of two resources (say x_1 and x_2) in G-space. Let us assume that these maps are matrices[3] and have the same dimensions. Let us further assume that the RSF for this species is given, simply, by:

$$w(\mathbf{x}) = e^{0.04x_1 + 0.08x_2} \tag{2.23}$$

The R-code to plot a map of `lambda` would be:

```
w<-exp(0.04*x1+0.08*x2)
lambda<-w/sum(w)
image(lambda)
```

There is one feature of G-space that can wreak havoc on spatial predictions. We will call this *accessibility* for short and will come to define it more clearly in section 3.11.2. Consider the following biological examples where the transition from E- to G-space may give unrealistic maps:

1. We are interested in the distribution of usage of a particular pack of wolves who constrain their use of space to their home range, a subset of the landscape. Hence, not all of space is accessible to them and if we apply the above approach indiscriminately on all of the landscape, we will be spreading the pack's fixed amount of usage rather thinly, hence overestimating the usage of space outside their territory, and underestimating it inside.

2. We are interested in the distribution of invasive plant species across the landscape. We may have a good process model for the plant's habitat preferences, but if we do not combine this with current dispersal information, the spatial map will suffer. If, for example, only half of the landscape has become invaded by the plant, showing a map of complete accessibility of the landscape is misrepresenting reality.

2.10 Thinking outside the box: Biological mechanism in process models

Increase in the complexity of the preference function h can either come from first principles, or the function can acquire added "wiggliness" under the instruction of the data (Wood, 2006). In this section, we look at some immediate extensions that can enhance the biological realism and interpretation of SHA models. We

---

[3]You could think of them as rasters if you prefer to use GIS functionality in R
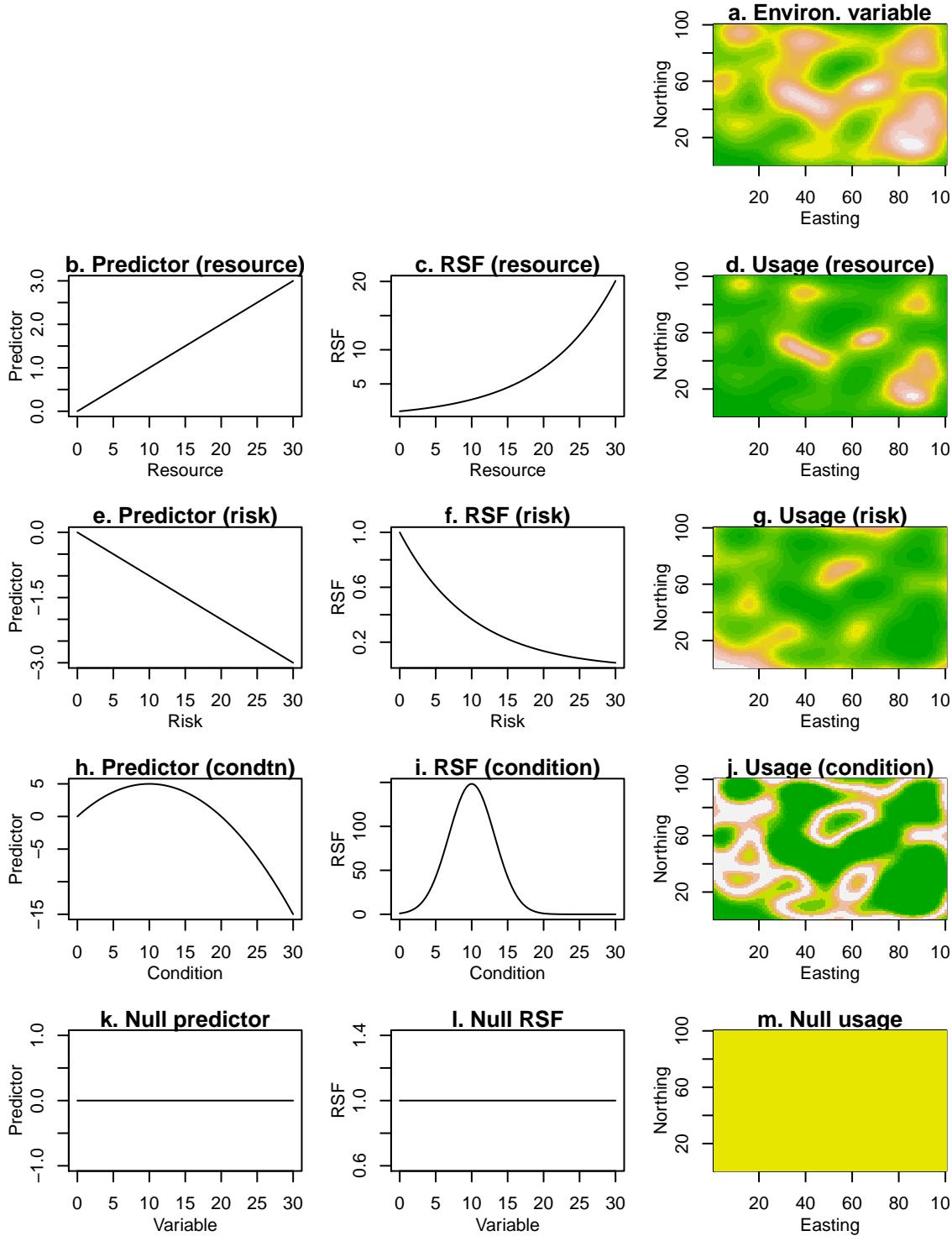
Figure 2.3: Usage in response to different types of variables. Consider the generic layer in (a) which representes the spatial distribution of an environmental variable. In the first and second columns we show the predictor and selection functions assuming that the organism responds to the variable as if it is a resource (b, c), a risk (e, f), a condition (h, i). In the third column (d, g, j) we show the expected distribution of the organism in each of these three scenarios. For completeness, in the bottom row, we show the predictor, selection functions and homogeneous spatial expectation for an organism that is unresponsive to the environmental variable.

will assume that the form of the preference and selection functions are fixed (as defined in sections 2.7 and 2.8), and we will instead focus on the predictor function. It turns out that, by incorporating simple model features, such as multiple covariates, quadratic terms and interactions, the predictor function can capture a lot of interesting biology.

2.10.1   Single-variable models

Let us start by examining one-dimensional environmental spaces, in the variable x (Fig. 2.3a shows an example of such a variable in G-space). Our basic classification of habitat characteristics into resources, risks and conditions has an easy mathematical representation (Matthiopoulos et al., 2015). Resources are consistently desirable, so high values of resources should be preferred over low ones. We should therefore expect a positive response to these variables. Risks (assuming they are real and perceived, or just perceived) should have a negative response, indicating avoidance. Therefore, at the scale of the predictor function, we can write

$$g(x) = \alpha_1 x \tag{2.24}$$

where $\alpha_1 > 0$ if x is a resource and $\alpha_1 < 0$ if it is a risk. This assumes the simplest possible form of monotonic relationship between the variable x and the predictor. Other, more complicated monotonic functions are possible (see e.g. section 2.10.6 on saturating responses), but the biology will need to warrant the use of such additional complexity. However, the proportionality of eq. (2.24) does not mean that such a model's spatial predictions will be linear. Figs 2.3b & e, show (linear) plots of the predictor for a resource and a risk respectively. The panels right next to them (Figs 2.3c & f) show the exponential version of the same functions (i.e. $w(x)$, defined, above as the selection function). Finally, the rightmost panels (Figs 2.3d & g) show expected usage in G-space. These maps inherit the nonlinearity and complexity of the underlying environmental variable, but as you might expect, peaks in the resource distribution match the hotspots in expected usage, whereas peaks in risk are coldspots in expected usage.

Response to environmental conditions can be thought of as a combination of the above two scenarios. Consider the example of water availability to an organism. Complete drought is a problem for any plant or animal, so preference should increase together with water availability, initially. After a certain point however, inundation becomes a risk, and preference should decrease with water availability. We therefore need a function that describes such a modal response. The easiest one is a quadratic function

$$g(x) = \alpha_1 x + \alpha_2 x^2 \qquad \alpha_2 < 0 \tag{2.25}$$

The negative value of α_2 ensures that the graph of $g(x)$, is a downward-pointing parabola (see Fig. 2.3h). Viewed as an RSF (Fig. 2.3i), this has a rather appealing uni-modal form that anticipates little usage at extreme values of the environmental variable x. When applied to spatial prediction, this model gives us bands of high expected usage (the yellow-white areas in Fig. 2.3j) at intermediate (optimal) values of the condition x. Again, this is the simplest function that can create such a "peaked" response. More complicated functions could be used to ensure, for example, that the curve in Fig. 2.3i is not symmetric, or that it has thicker tails, but an enthusiastic biologist will really need to present a critical modeler with overwhelming evidence to justify such increases in complexity. Even more demanding would be the biological case for a multimodal response to a condition. It is not inconceivable, but seems somewhat less likely (Austin, 1999), that an organism has two optima at different regions of the same variable. It may be argued that such apparent multimodality is the result of responses to more than one variables that have an interactive effect on the organism, so that such phenomena are best modeled in the setting of more than one variables (see section 2.10.3 below).

There is a special case of univariate relationships, indicating absence of preference. An unresponsive relationship $g(x) = 0$ is obtained by setting the coefficients α_1 and α_2 to zero. According to eq. (2.12), this yields a resource-selection function $w(x) = 1$. Placing this into eq. (2.13), we can show that $h(x) = 1$. What does this imply? In environmental space, it means that values of x are used in proportion to their availability (from eq.(2.7)). In geographical space (see, eq.(2.22)), it means that all points are expected to be used uniformly. These facts are depicted in the bottom three panels of Fig. 2.3.

The above framework captures a wealth of biology with considerable mathematical economy. Using a 2nd-order polynomial $g(x) = \alpha_1 x + \alpha_2 x^2$, one of the best-behaved functions in high-school mathematics, we can represent qualitatively different types of variables by selectively switching different α parameters on and off.
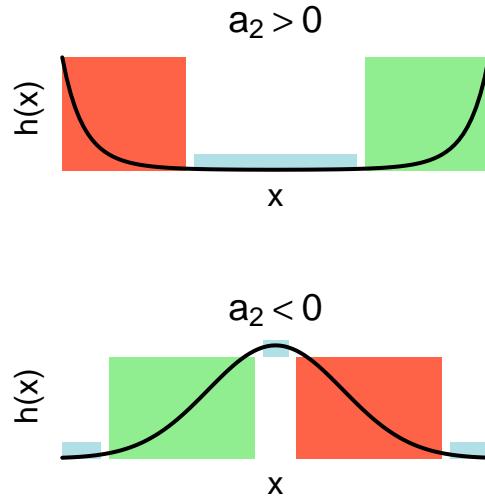

Figure 2.4: Upward and downward pointing parabolas (in the exponential scale corresponding to an RSF) and their biological interpretation over different segments of the environmental variable. The curves behave like resources in the green windows and like risks in the red windows. They appear approximately unresponsive to the environmental variable in the blue regions.

At the same time, this framework also illustrates why it is better to motivate model development from biological first principles rather than mathematical completeness. For example, it is hard to think of a biological reason why an upward-pointing parabola (obtained when $\alpha_2 > 0$) would make sense across the entire range of values of x. However, such a curve might make sense for smaller windows of the full range. In Fig.2.4 we illustrate this point for both upward and downward-pointing parabolas. Different parts of the curve can be used to model different types of responses, if we are prepared to restrict the values of x (i.e. if we zoom in at selected windows of these plots).

This observation is particularly relevant when data and uncertainty are involved (see Chapter 3). If the window of observed values for x is sufficiently narrow, and if there is enough fuzziness in the data, then a condition can be misconstrued as a risk and a resource or an irrelevant variable (indicating no response). Indeed, a parabola with a positive value of α_2 may be supported by the data in describing a resource or a condition.

Therefore, it is necessary to use biology to supervise the interpretation of these models. This means that only the first of the following two statements is advisable:

- *"I know that my species cannot thrive in the absence or in the superabundance of water. I will therefore use a quadratic model to describe its relationship to this variable"*

- *"Within my limited window of data, the species shows a negative/positive response to water. I will therefore conclude that water is a risk/resource for this species"*

2.10.2 Two-variable models: Additive predictors

We can now move on to models with more than one variable. We will generalize on the idea of eq.(2.25), allowing us to look at how combinations of the three basic categories of environmental variables (resources, risks and conditions) behave when they act in tandem. Notionally, to keep track of the extra complexity,
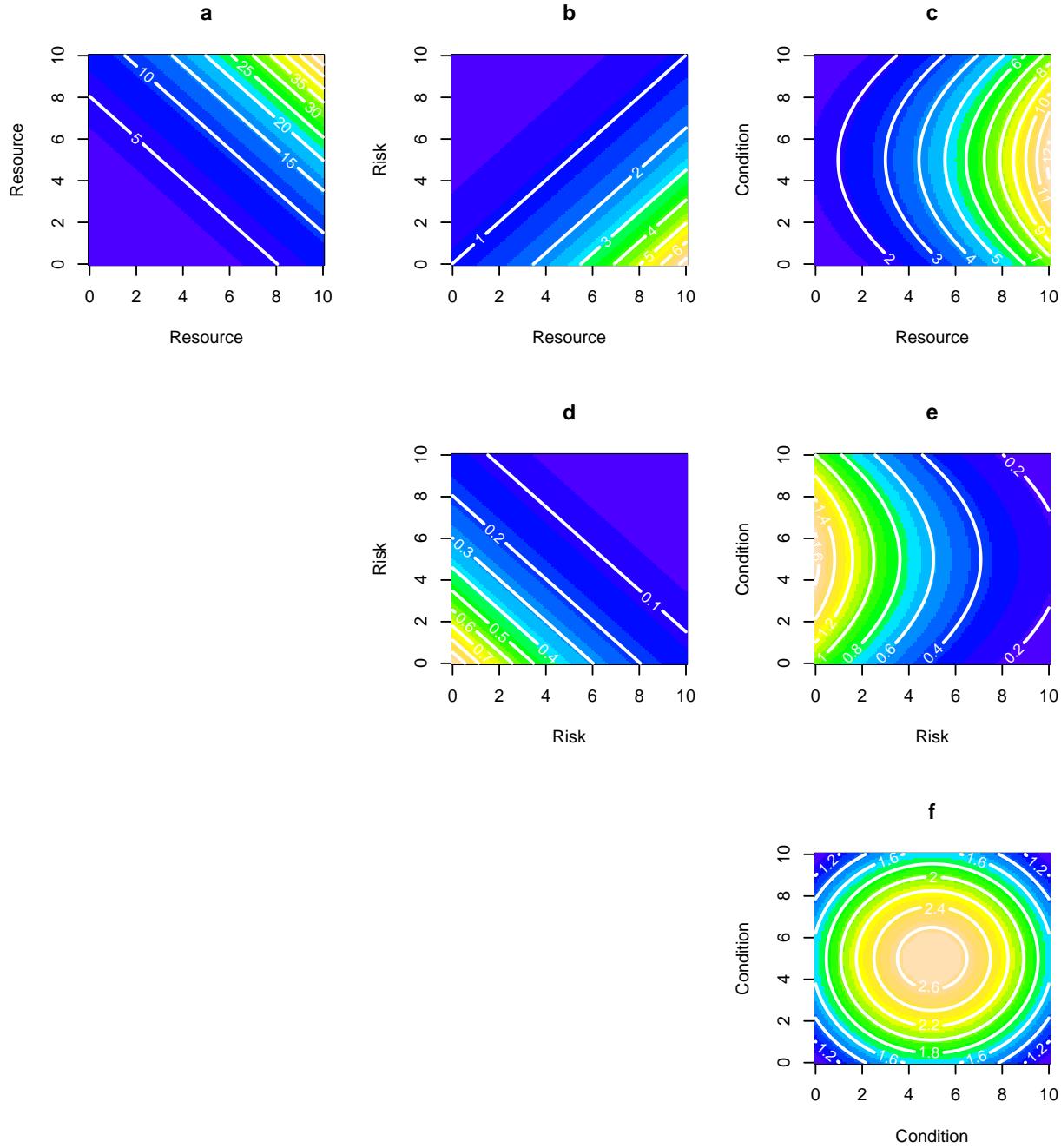
Figure 2.5: Two-variable combinations, generated by pairwise models of resources, risks and conditions. The ramp from cold to warmer colours indicates increasing preference.

we can use a double-subscript $\alpha_{i,j}$ to indicate that the coefficient belongs to the jth-order term of the ith environmental variable.

$$g(x_1, x_2) = \alpha_{1,1}x_1 + \alpha_{1,2}x_1^2 + \alpha_{2,1}x_2 + \alpha_{2,2}x_2^2 \qquad \alpha_{1,2}, \ \alpha_{2,2} < 0 \tag{2.26}$$

By setting the α's to positive, negative or zero values, we can create combinations between resources, risks and conditions. These are shown in Fig.(2.5). There are two qualitative points to make here that connect these models with ecological principles.

The shape of fundamental niches: Because SHA models are specified in E-space, there is an understandable tendency to try and interpret them in terms of the ecological niche of a species (Pulliam, 2000; Bahn & Mcgill, 2007; Hirzel & Le Lay, 2008; Holt, 2009; Warren, 2012, 2013; McInerny & Etienne, 2013). We will examine niche theory in more detail in later chapters, but it is worth broaching the issue early on. The fundamental niche is often described by its Hutchinsonian definition of an "*n-dimensional hypervolume*". The image that most textbooks use to visualize it, is akin to our Fig.(2.5f) - areas (or volumes) contained within closed contours in E-space (Blonder, 2018). However, as all the other plates in Fig.(2.5) clearly indicate, functions in E-space need not give rise to closed contours. For instance, there is no reason why the fundamental niche of an organism comprising just resources should be a closed hypervolume, because that would exclude perfectly desirable habitats that just happen to be superabundant in resources (Chase & Leibold, 2003).

There may be a couple of reasons for this misleading imagining of niches in E-space. The botanical origins of these early studies had concentrated on envelopes of temperature, or soil pH that allowed plants to survive. This immediately funnels our thinking into conditions, placing less emphasis on (usually perishable, or depletable) resources and (even less easy to measure) risks. Alternatively, the picture of a closed niche may stem from a confusion between the fundamental niche (what the organism could ideally use) and the realized niche (what it can use within the constraints of its environment). If you refer back to Fig.2.2d you see what appears to be a closed region in E-space (echoing the textbook depictions of Hutchinsonian niches). However, Fig.2.2f more clearly indicates that, once we account for habitat availability, we have a monotonically and diagonally increasing preference, identical in form to our Fig. 2.5a.

Null models for resources and risks: The contours obtained by the combination of linear terms (i.e. resources and risks) are linear (see Figs 2.5a,b and d). This is a useful trait to remember because it gives rise to two ecological null models. The first, is the model of a perfect *trade-off* between a resource and a risk (Fig.2.5b), which presents increasing contours of preference with a common, constant, slope. Biologically, the slope tells us how many additional units of the resource are needed to balance one additional unit of risk. The second, is the null model of *perfectly substitutable resources* (Fig.2.5a), a term introduced by Tilman (1982) in his classification of relationships between resources. The linear contours here have a negative slope which represents the "exchange rate" between the two resources (i.e. how much of one resource is needed to replace the loss of a unit of the other resource). We will follow up on this null model in the next section, to relax the assumption of perfect substitutability.

2.10.3 Two-variable models: Multiplicative predictors

In this section we stay with models that contain solely two resources. Tilman (1982) had discussed two further scenarios, for the joint action of resources on organisms, antagonism and complementarity (Fig. 2.6). Two or more resources that can substitute for one another, are called *antagonistic* if, when taken together, they may partially off-set the effects of each-other. The consumer requires more of the resources when they are taken together than when they are taken separately (Fig. 2.6b). In contrast, *complementary (or synergistic) resources* augment one another, so that the consumer requires less of them when taken together than when taken separately (Fig. 2.6c). These features can be captured by the general model of two resources with the addition of an interaction term.

$$g(x_1, x_2) = \alpha_1 x_1 + \alpha_2 x_2 + \beta_{1,2} x_1 x_2 \tag{2.27}$$

Where $\beta_{1,2}$ is negative for antagonism and positive for complementarity. A rather appealing property of this general model is that the null model of perfect substitutability is nested within it (just set $\beta_{1,2} = 0$). However, as with earlier models in this chapter, only some parameterizations of eq.(2.27) make biological sense. In particular, the model for antagonism only makes sense for certain values of β which, rather concerningly, depend on the upper limits of values considered for the resource axes[4].
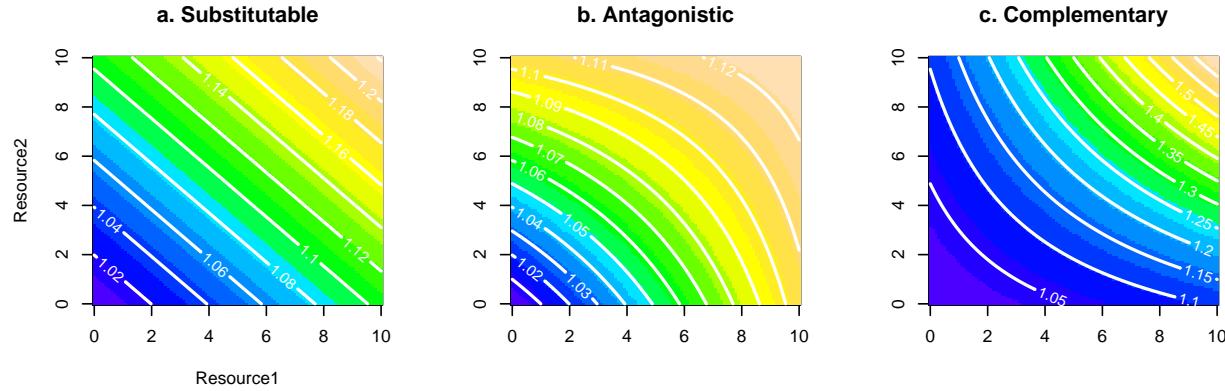

Figure 2.6: Reaxing the assumption of perfect substitutability (shown in a) between two resources, to create antagonism (part b), or complementarity (part c).

2.10.4   Multivariable models

The above principles, readily generalize to more than two habitat variables. For example, the following model

$$g(x_1, x_2, x_3, x_4) = 0.1x_1 - 0.001x_1^2 - 0.2x_2 + 0.4x_3 + 0.002x_3x_4 \qquad (2.28)$$

is rich in biological information and could be decoded as follows: 1. x_1 is a condition; 2. x_2 is a risk, 3. x_3 is a resource; 4. There is a perfect trade-off between x_2 and x_3; 5. For low values of x_1, this condition behaves as a substitutable resource to x_3; 6. High values of x_1 aggravate the effect of the risk x_2 and trade-off against the resource x_3; 7. The organism has no direct response to the variable x_4 (since there is no main term for x_4 the slope of g in response to x_4 is zero); 8. However, the variable x_4 is able to moderate the response of the organism to variable x_3.

---

[4]To see why that is, consider eq.(2.27). If this ever takes negative values, then we end up with the biologically unrealistic scenario of two resources, acting together as a risk. To avoid that, we need to stipulate that, when $\beta_{1,2} < 0$ (as it must be, to capture antagonistic resources), the net result is always positive. If $x_{1,max}$ and $x_{2,max}$ are the upper limits that we are prepared to examine for the two resources, then we require that

$$\alpha_1 x_1 + \alpha_2 x_2 + \beta_{1,2} x_1 x_2 \geq 0.$$

This means that

$$\beta_{1,2} > -\left(\frac{\alpha_1}{x_{2,max}} + \frac{\alpha_2}{x_{1,max}} \right).$$

In addition, we would like to make sure that our general model is strictly increasing (because it is a response to two resources), so that

$$\frac{\partial g}{\partial x_1}, \frac{\partial g}{\partial x_2} > 0.$$

This implies the additional (and more stringent) constraints

$$\beta_{1,2} > max\left(-\frac{\alpha_1}{x_{1,max}}, -\frac{\alpha_2}{x_{2,max}} \right).$$

Interpretation of these more complex models becomes much harder as additional terms are introduced, and it becomes increasingly difficult to specify constraints for the parameter values that can keep the behavior of the model within desired biological requirements[5].

2.10.5 Functional and multiscale responses in SHA models

All of the extensions of the basic linear model that we have examined until now refer back to a special case of the habitat preference function. In eq.(2.7) we made the explicit assumption that habitat preference is $h(\mathbf{x})$, i.e. purely a function of local habitat $\mathbf{x}$. However, the original expression of habitat preference in (2.5) was more flexible than that, allowing for other, e.g. non-local arguments to be used. Two particular topics that have often made their appearance in discussions about SHA models are *habitat context* and *spatial scale*. We look at those in turn.

The assumption made by most current studies of habitat selection is that the choices made by organisms (particularly animals) are constant, irrespective of the availability of habitats in their broader neighborhood. However, a mounting body of literature, reviewed in Matthiopoulos et al. (2011) and Holbrook et al. (2019), is arguing from the basis of both biological data and theoretical principles that animals modify their choices depending on habitat context. This phenomenon is known as a *functional response* in habitat selection (Mysterud & Ims, 1998). In the vaguest, most unhelpful terms, we can formalize this as a dependence of habitat preference on the makeup of E-space as a whole

$$h(\ ) = h(\mathbf{x}, f_a)$$

However, this does not readily imply a way forward for mathematical modeling. Without throwing the models of the preceding sections to the waste bin, how can we extend our existing mathematical framework of habitat preference functions to account for the effect on space-use of the availability of **all** habitats in the neighborhood of an animal? Four difficult questions present themselves.

How should we summarize habitat availability? This question cannot be answered easily, but not for lack of answers (far too many possibilities present themselves). The Rolls-Royce approach would use a detailed parametric description of availability $f_a(\mathbf{x})$ across E-space, such that every little nuance of the variations in habitat availability is captured. A less luxurious approach would focus on each individual habitat variable and describe its frequency distribution $f_a(x)$ parametrically. A less immediate, but potentially quite detailed method would use a sequence of statistical moments from the distribution of each variable $E(X), E(X^2), E(X^3), ...$ The least powerful approach is to simply use the first of those moments, i.e. the average value of a variable $\bar{x} = E(X)$. For the rest of this section, let us represent availability by this simplest of summaries. The average value of a variable in the neighborhood of an animal.

How should the availability of a single habitat affect usage? In their review (Holbrook et al., 2019), present four apparently distinct ways in which habitat availability can affect usage. One of those, originally proposed by Boyce & McDonald (1999), suggested that the coefficients of a selection function could be modeled as functions of habitat availability. In statistics, this is known as a varying-coefficient model (Hastie & Tibshirani, 1993). In fact, as we will see in a future chapter, all four of the approaches outlined by Holbrook et al. (2019) can be reduced to a model of varying coefficients. For example, in two of their possible approaches, they debate implicitly the merits of having a direct or interactive effect of availability on the linear predictor of a model. Consider a simple univariate model with predictor

$$g(x) = \alpha_0 + \alpha_1 x \tag{2.29}$$

The availability of habitat in the neighborhood of the animal can be introduced by making the coefficients dependent on availability (i.e. $\alpha_0(\bar{x})$ and $\alpha_1(\bar{x})$). We can do this most easily by recasting the α's as linear functions of $\bar{x}$

---

[5]The role of constrained optimization for fitting SHA models and SDMs is underexplored. As we begin to think about these models more explicitly from the point of view of fitness, it will be sensible to constrain their parameters to plausible ranges (in likelihood approaches) or provide them with more informative priors (in Bayesian approaches).

$$\alpha_0(\bar{x}) = \beta_{0,0} + \beta_{0,1}\bar{x} \qquad \alpha_1(\bar{x}) = \beta_{1,0} + \beta_{1,1}\bar{x}$$

Placing these expressions back into eq. (2.29) gives us both main and multiplicative effects of availability on the linear predictor, in the new β coefficients:

$$g(x) = \beta_{0,0} + \beta_{0,1}\bar{x} + \beta_{1,0}x + \beta_{1,1}\bar{x}x,$$

indicating that we don't really need to step outside the simple varying-coefficient paradigm to be able to have direct and multiplicative effects of availability on usage. However, from the point of view of interpretation, the existence of a direct term (i.e. $\beta_{0,1}\bar{x}$) implies an effect on the intercept (α_0) of the response, whereas an interactive term ($\beta_{1,1}\bar{x}x$) comes from a change in the slope (α_1) of the predictor.

How should the availability of all habitats on all other habitats be considered simultaneously? If we generalize eq.(2.27) to two habitat variables,

$$g(x_1, x_2) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$$

we can also generalize the dependence of model coefficients to more than one expectations, so that

$$\alpha_0(\bar{x}_1, \bar{x}_2) = \beta_{0,0} + \beta_{0,1}\bar{x}_1 + \beta_{0,2}\bar{x}_2$$

$$\alpha_1(\bar{x}_1, \bar{x}_2) = \beta_{1,0} + \beta_{1,1}\bar{x}_1 + \beta_{1,2}\bar{x}_2$$

$$\alpha_2(\bar{x}_1, \bar{x}_2) = \beta_{2,0} + \beta_{2,1}\bar{x}_1 + \beta_{2,2}\bar{x}_2$$

Or, more generally, if we introduce

$$\mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_K \end{bmatrix}, \quad \bar{\mathbf{x}} = \begin{bmatrix} 1 \\ \bar{x}_1 \\ \vdots \\ \bar{x}_K \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_{0,0} & \cdots & \beta_{0,1} \\ \vdots & \ddots & \vdots \\ \beta_{K,0} & \cdots & \beta_{K,1} \end{bmatrix}, \quad \alpha = \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_K \end{bmatrix}$$

Then the predictor model can be written as

$$g(\mathbf{x}, \bar{\mathbf{x}}) = \alpha^T \mathbf{x} = (\beta\bar{\mathbf{x}})^T \mathbf{x} \tag{2.30}$$

This is a good starting point for the n-variable, varying-coefficient model, assuming that both the predictor function and the variations in the coefficients are linear functions.

What is a relevant neighborhood over which to consider availability? As proffered by (Allen & Hoekstra, 1992), "*In ecology, looking for the right thing is easier than looking for the right size*". While focusing on such exotica as functional responses, many papers (including several of ours) sweep the question of the scale of availability under the carpet. However, it is important to try and consider, from first principles, what is a relevant scale for calculating habitat availability. For a given temporal scale, presumably the easiest thing is to define a spatial buffer (e.g. a disc), determined by the speed of movement of an animal. However, driving the definition purely by movement capabilities may be misleading. For one, the environment may present obstacles to movement, so that the buffer is not radially symmetric or constant in size. But higher cognitive processes can also interfere. For a large temporal scale, yielding a buffer much larger than an animal's range of perception, several far-away habitats may be accessible but not perceived. Does this mean they are still available, if they are not explicitly included in the menu from which the individual is making a choice? Alternatively, it can be argued that information is not purely dependent on immediate perception. Memory and communication play a role, effectively extending the range of perception of an animal in non-trivial ways. We will return to these issues of availability in later chapters where we will broaden our examination of functional responses.

Nevertheless, this last question (on scale) prompts the issue of multi-scale selection. Habitat selection happens at multiple spatiotemporal scales (Johnson, 1980; Mayor, Schneider, Schaefer, & Mahoney, 2009) and statistically, the assumptions of scale, regarding availability can seriously affect our biological inferences (Beyer et al., 2010; Paton & Matthiopoulos, 2018). So, why have to choose? Why not develop a method that captures multiple scales at the same time? The main challenge here is that every choice made today, necessarily limits the options available for tomorrow. More specifically (Decesare et al., 2012), there must be a hierarchy of scales so that choices made at finer scales (e.g. minute-to-minute and step-by-step movements between fine-grain habitats) are *conditional* on choices made at larger scales (e.g. annual or interannual decisions on which broad area to live in).

2.10.6 Saturating responses

One of the main features missing from current formulations of SHA models (but which is quite frequently encountered in nature) is the principle of diminishing returns, the idea that preference does not increase at the same rate with additional increases in resources. The polynomial gold vein we have been mining until now (and that has taken us quite far biologically), becomes exhausted at this point. It is possible to caricature the behavior of a saturating function, using high-order polynomials, within a strictly defined window of habitat values, but that behavior inevitably breaks down under extrapolation (Fig. 2.7).

To capture saturating responses economically and robustly, we need to extend the predictor function to take as arguments more general functions for some of the habitat variables:

$$g(\mathbf{x}) = \alpha_0 + \sum_{k=1}^{K_1} \alpha_k x_k + \sum_{k=K_1+1}^{K_2} f_k(x_k)$$

In theoretical ecology, saturating responses have been captured with a wide range of specific functions, but most often we use fractional and exponential formulae, e.g.

$$f(x) = \frac{\alpha x^c}{\gamma^c + x^c}, \qquad f(x) = \alpha(1 - exp(-\gamma x^c)), \quad (\alpha, \gamma, c > 0)$$

If you are familiar with consumer-resource models (Murdoch, Briggs, & Nisbet (2003)), you may recognize the first of those functions as the most general (i.e.type III) Holling functional response (Holling (1959)). For example, the curve in Fig 2.7) is derived from the function $f(x) = 10x/(4 + x)$.

There is considerable potential for exploring how such curves interplay in predictor functions of multiple variables, particularly given the normalization of the selection function into a preference function.

2.11 Concluding remarks

SHA models are formulated in *E*-space and generate predictions about the expected distribution of a species in *G*-space. The transition from *G*-spaces to *E*-spaces and back again is a central theme to SHA approaches, although it is often done with minimal acknowledgment. Knowing which type of space we are operating in is important, because the transition is not always loss-free. For example, several difficult decisions on how to define habitat availability (an *E*-space concept) are ultimately questions of geographic accessibility (a *G*-space concept), but the two are not interchangeable. A model that talks about habitat availability without recognizing the geographic proximity between habitats is, by the same token, a simpler and information-poorer model. The mathematical phrasing of SHA models in *E*-spaces can be kept quite simple but yet capture a remarkable number of biological phenomena. There are, however, several places where these simple mathematical formulations break down biologically and need to be constrained or made more complex. Traditionally, neither constraints nor complexity have been willingly pursued with SHA models because analysts have always kept foremost in their mind that the of SHA models would need to be derived from field data (and not out of the blue, as we mostly did in this chapter). When fitting models to data, we are bound to look at the resulting responses and interpret them biologically. However, model fitting (the subject of the next chapter) is liable to return parameter values that challenge the interpretive abilities of even the most
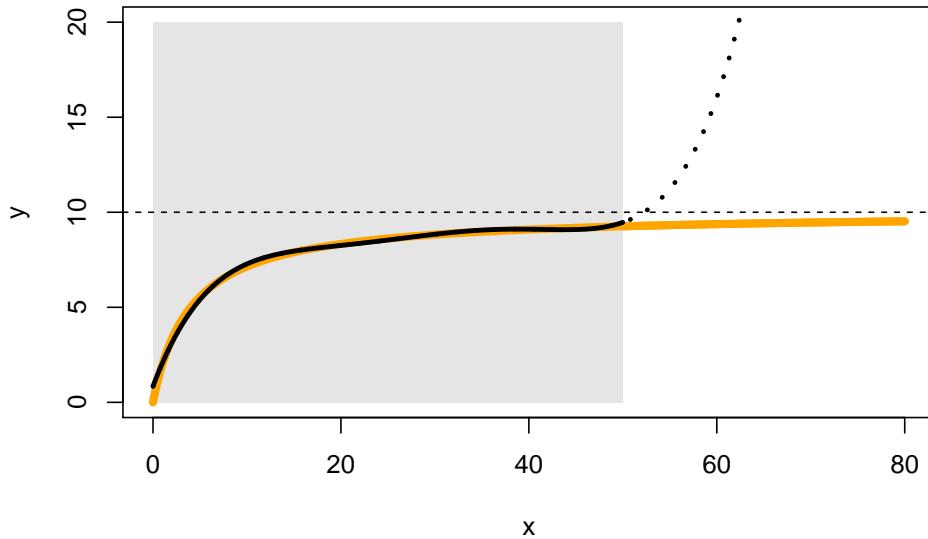
Figure 2.7: Our ability to describe saturating responses with polynomials is limited to the window of approximation (shown in grey) and the approximation is parameter-greedy. The saturating curve shown in orange, approaches an asymptote at the value 10. The approximating polynomial is based on the values of the orange curve within the grey window. It takes a 5th-order polynomial to achieve a convincing approximation (i.e. 6 parameters in total) and the approximation goes awry, immediately after we exit the grey window.

imaginative biologist. Rather than taking the results of model fitting on-faith (and even worse, generating uncritical predictions from such models), this chapter has hopefully provided the mathematical insights to specify meaningful models, cast meaning to models derived from data, and pursue the development of mathematical constraints and functional complexity to tread the fine line between empiricism and mechanism.

Chapter 3

Observation models for different types of usage data

3.1 Objectives

The objectives of this chapter are to:

1. Demonstrate how different types of habitat-use data (e.g. point locations in space, gridded counts, presence/absence data, telemetry data) can be used to inform *species-habitat-association* models.

2. Unveil commonalities and connections among different statistical approaches for quantifying species-habitat associations. In particular, *weighted distribution theory* and the *Inhomogeneous Poisson Process (IPP)* model provide suitable, overarching frameworks, encompassing the most widely used methods for modeling species distributions and their relationship with the environment.

3. Given the importance of weighted distributions and the IPP model as unifiers of various seemingly disparate methods, we will describe the main features of these frameworks and their biological and statistical assumptions.

4. By highlighting some of the challenges encountered in natural systems, we will also hint at some of the solutions for handling messy data that will be covered in later chapters of the book (e.g. methods for dealing with various sampling biases, imperfect detection, autocorrelated data, and unequal accessibility of different habitats).

3.2 You are here

Chapter 2 focused on process models for linking locations of individuals to their environment. This chapter will focus on linking process models to data. We will accomplish this by fitting statistical models that link *observations* of habitat use to resources, risks, and conditions. As an example, consider a situation where we have counts of individuals within 2 km × 2 km plots, but we fail to detect some individuals that are present. To estimate parameters in our process model, we will need a statistical model that links our imperfect count data in discrete space to our process model describing use of habitat, possibly in continuous space.

There are many different ways to collect information on habitat use (Table 3.1). These different data types have fundamentally different characteristics that need to be considered when linking habitat use to environmental variables. First and foremost, it is important to consider the sample units and how they are selected − either randomly from a larger population or opportunistically; random sampling is key to being able to generalize to a larger population.

As the name suggests, observation models also need to capture important features related to *how data were*

Table 3.1: Different types of habitat-use data used to parameterize species-habitat-association (SHA) models.

Data Collection Method	Sample Unit/Sampling Effort in Space and Time	Typical Sample Selection
Presence-only surveys	Spatial and temporal dimensions are often not well defined	Opportunistic
Presence-absence or count-based surveys	Spatial areas at fixed points in time	Random
Camera trap	Areas within the camera's field of view, followed over time	Systematic, Random
Telemetry	Individuals followed over time	Opportunistic

collected, including processes that may lead to a biased sample of habitat-use data (e.g. variable sampling effort, imperfect detection). Frequently, the data used to parameterize SHA models come from haphazardly collected sampling efforts (e.g. museum specimens or citizen science data without any quantitative measure of survey effort). These data will rarely be representative of a population's habitat use, and care must be taken when constructing and interpreting SHA models fit to these data. If no information is available to understand or adequately model spatial variability in sampling effort and detection probabilities (e.g. presence-only surveys in Table 3.1), then we may have to reckon with the words of R.A. Fisher, "To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of."

As Fisher notes, challenges associated with collecting a representative sample of habitat-use data should ideally be considered *before collecting data*. With some forethought, we can use specialized data collection protocols and associated models to estimate and correct for observation biases (e.g. imperfect detection). Alternatively, it may be possible to combine haphazardly collected data with data collected in a structured fashion to more fully understand and model sampling biases (Dorazio, 2014; Fithian, Elith, Hastie, & Keith, 2015; Koshkina et al., 2017; Isaac et al., 2020).

We begin this chapter by considering an idealized scenario where we record the locations of all individuals within a geographic domain, G, at a specific point in time. We will further assume these locations are measured without error and that any clustering in space is due solely to the species association with spatially autocorrelated covariates. Although this simplest of cases may seem unrealistic, it encapsulates the main features of the Inhomogeneous Poisson process (IPP) model, which serves to unite most popular analyses of habitat-use data under a common umbrella (Aarts et al., 2012; Wang & Stone, 2019), including resource-selection functions estimated using logistic regression (Boyce & McDonald, 1999) and Maximum Entropy (or MaxEnt) (Phillips & Dudik, 2008). The IPP model has also served as a useful framework for synthesizing information from multiple data sources using joint likelihoods (Miller et al., 2019). This idealized scenario may be reasonable for some plant and tree surveys, but most applications of SHA models will require more complicated observation models to account for violations of the assumptions of perfect detection, error-free measurement, independent locations, and constant sampling effort.

We note that wildlife telemetry data have often been modeled using logistic regression, which can be shown to be equivalent to fitting an IPP model (Warton et al., 2010) and discuss these connections here. It is reasonable to question whether telemetry data should be treated in this way because: 1) this approach ignores the temporal information in wildlife tracking data; and 2) locations are treated as though they are independent or at least exchangeable (equally correlated) within an individual. This assumption is problematic for most modern telemetry data sets since observations close in time will tend to also be close in space (Fleming et al., 2014). We briefly discuss these issues and potential solutions in the context of telemetry studies, and SHA models more generally; nonetheless, we leave full treatment of these topics to later chapters.

Lastly, we note that our examples in this chapter are far from exhaustive. Indeed, whole books have been dedicated to specialized methods for collecting and analyzing data so as to allow for the estimation of detection probabilities. Examples include Buckland, Anderson, Burnham, & Laake (2005) (distance sampling), J. A.

Royle et al. (2013a) (spatial mark-recapture), MacKenzie et al. (2017) (occupancy models). Interested readers may want to follow up with these resources as well as more general books discussing hierarchical models for estimating trends in population abundance (e.g. Kéry & Royle, 2015).

3.3   Homogeneous Point-Process Model

At any given moment in time, individuals in a population occupy a set of locations in space. Our goal is to understand why they are at those particular places. The field of spatial statistics, and in particular *spatial point-process models*, offer a formal approach for modeling the number and distribution of locations in space. Consider sampling an area G, and observing n individuals at locations $s_1, s_2, \ldots, s_n$ (more generally, we may refer to these locations as *points* or *events*, or in the context of telemetry studies, *used locations*). The simplest model for the data that we might consider is a Homogeneous Poisson-Process, which has the following features:

1. Let N be a random variable describing the number of individuals or points in a spatial domain, G. N is a Poisson random variable with mean $= \mu = \lambda|G|$, where λ is the density of the points in G (i.e. points per unit area) and $|G|$ is the area of G. Thus, the probability of observing n individuals in G can be described using the Poisson probability mass function:

$$P(N = n) = \frac{\exp^{-\mu}(\mu)^n}{n!} = \frac{\exp^{-\lambda|G|}(\lambda|G|)^n}{n!}, n = 0, 1, 2, \ldots \tag{3.1}$$

2. Conditional on having observed n individuals, the spatial locations are independent and follow a uniform distribution on G, $f(s_i) \sim \frac{1}{|G|}$. Because the locations are assumed to be exchangeable (unordered), there are $n!$ different ways to arrange them, all resulting in the same joint likelihood. Thus, the conditional likelihood for the spatial locations is given by:

$$f(s_1, s_2, \ldots, s_n | N = n) = n! \frac{1}{|G|^n} \tag{3.2}$$

Combining eq. (3.1) and eq. (3.2) gives us the unconditional likelihood for the data:

$$
\begin{aligned}
L(s_1, s_2, \ldots, s_n, n | \lambda) &= \frac{\exp^{-\lambda|G|}(\lambda|G|)^n}{n!} n! \frac{1}{|G|^n} \tag{3.3} \\
&= \lambda^n \exp^{-\lambda|G|} \tag{3.4}
\end{aligned}
$$

To estimate λ using Maximum Likelihood, we rewrite this expression so that it is stated as a function of the unknown parameter, λ, conditional on the observed data, $L(\lambda|s_1, s_2, \ldots, s_n, n)$. We then choose the value of λ that makes the likelihood of the observed data as large as possible. This step requires taking derivatives of the likelihood function with respect to λ, setting the expression equal to 0, and solving for λ. It is customary to maximize the log-likelihood rather than the likelihood. The log-likelihood is usually easier to work with and maximize numerically, and the value of λ that maximizes the log-likelihood will also maximize the likelihood. The log-likelihood in this case is given by:

$$\log L(\lambda|s_1, s_2, \ldots, s_n, n) = n \log(\lambda) - \lambda|G|$$

Taking the derivative of the above expression with respect to λ, setting the expression equal to 0, and solving gives us the intuitive maximum likelihood estimator for λ equal to the number of observed individuals divided by the area sampled:

$$\frac{\mathrm{d}\log L}{\mathrm{d}\lambda} = \frac{n}{\lambda} - |G| = 0 \tag{3.5}$$

$$\implies \hat{\lambda} = \frac{n}{|G|} \tag{3.6}$$

3.4 The Inhomogeneous Poisson Point-Process Model

To move from the Homogeneous Poisson Point-Process model to more realistic models, we need to allow λ to vary spatially. Such variations are, in most cases, generated by underlying biological processes, that may be adequately captured by co-variations with environmental variables. The Inhomogeneous Poisson Point-Process (IPP) model provides a simple framework for modeling the density of points in space as a log-linear function of environmental predictors through a spatially-varying intensity function:

$$\log(\lambda(s)) = \beta_0 + \beta_1 x_1(s) + \ldots \beta_p x_p(s) \tag{3.7}$$

where $\lambda(s)$ is the intensity function at location s, $x_1(s), \ldots, x_p(s)$ are spatial predictors associated with location s, and $\beta_1, \ldots, \beta_p$ describe the effect of spatial covariates on the relative density of locations in space; $\lambda(s)ds$ gives the expected number of points occurring within an infinitesimally small area of geographic space of size ds centered on location s.[1] The IPP model has the following features:

1. The number of points, n, in an area G is given by a Poisson random variable with mean $\mu = \int_G \lambda(s)ds$ (i.e. the expected number of points is determined by the average of the intensity function over the spatial domain, G).

$$P(N = n) = \frac{\exp^{-\int_G \lambda(s)ds}(\int_G \lambda(s)ds)^n}{n!}, n = 0, 1, 2, \ldots$$

2. Conditional on having observed n points, their spatial locations are assumed to be independent and have the following joint likelihood:

$$f(s_1, s_2, \ldots, s_n | N = n) = n! \frac{\prod_{i=1}^n \lambda(s_i)}{(\int_G \lambda(s)ds)^n} \tag{3.8}$$

Combining these two expressions gives the unconditional likelihood of the data:

$$L(s_1, s_2, \ldots, s_n, n | \beta_0, \beta_1, \ldots, \beta_p) = \exp^{-\int_G \lambda(s)ds} \prod_{i=1}^n \lambda_i \tag{3.9}$$

Thus, the log-likelihood is given by:

$$\log(L(\beta_0, \beta_1, \ldots, \beta_p | s_1, s_2, \ldots, s_n, n)) = \sum_{i=1}^n \log(\lambda(s_i)) - \int_G \lambda(s)ds, \tag{3.10}$$

where again, we have rewritten the likelihood as a function of parameters, $\beta_0, \beta_1, \ldots, \beta_p$, conditional on the observed data.

3.4.1 Estimating Parameters in the IPP Model

The IPP model allows us to relate the density of points to spatial covariates. To estimate parameters in the IPP model, we must be able to solve the integral in equation (3.10). Although, in principle, it is possible to formulate parametric models for the distribution of spatial covariates in G, which may provide an opportunity to solve the integral analytically, this is almost never done (but see Matthiopoulos et al., 2020). Thus, in practice, we will almost always need to approximate the integral in eq. (3.10) using numerical methods. We can accomplish this goal using:

---

[1]We can also write this equation using a summation, $\log(\lambda(s)) = \beta_0 + \sum_{i=1}^p \beta_i x_i(s)$, or matrix notation, $\log(\lambda(s)) = X(s)\beta$ where $X(s)$ is an $n \times p$ matrix and β is a $1 \times p$ vector of parameters.

1. **Monte-Carlo integration**: we can use a large number of randomly generated locations from within G to evaluate the integral (e.g. Lele & Keim, 2006; Lele, 2009). This approach estimates $\mu = \int_G \lambda(s)ds$ with the sample mean:

$$\int_G \lambda(s)ds \approx \sum_{j=1}^{m} \frac{\lambda_j}{m} \tag{3.11}$$

Although this approach is commonly used, it can be inefficient compared to numerical quadrature methods that evaluate λ more frequently in areas where it is changing most rapidly.

2. **Numerical integration**: a set of quadrature points can be used to evaluate the integral in the log-likelihood function (if this sounds foreign, perhaps you will recall calculating areas under the curve using a series of rectangles from a first course in calculus). This approach replaces:

$$\int_G \lambda(s)ds \text{ with } \sum_{j=1}^{m} w_j \lambda_j \tag{3.12}$$

where w_j is a set of quadrature weights. If the m quadrature points are placed on a regular grid with spacing (Δ_x, Δ_y), then we can set $w_j = \Delta_x \Delta_y$. This is the simplest approach, but more efficient integration schemes can be developed by placing points non-uniformly in space. For example, it may be advantageous to include more points in areas where $\lambda(s)$ is changing quickly and also to include the locations of observed individuals (i.e. $s_1, s_2, \ldots, s_n$) as quadrature points. Both are possible with variable quadrature weights (e.g. see Warton et al., 2010). One option is to create a Dirichlet tessellation of the spatial domain (G) using the set of gridded and observed locations, and then let w_j equal the area associated with each tile in the tessellation; this approach is available in the `spatstat` library (Baddeley & Turner, 2005). Below, we demonstrate how to calculate a Dirichlet tessellation in R from a random set of points in $\mathbb{R}^2$.

```
library(spatstat)
X <- runifpoint(42)
plot(dirichlet(X), main="Dirichlet Tessellation")
plot(X, add=TRUE)
```

**Dirichlet Tessellation**
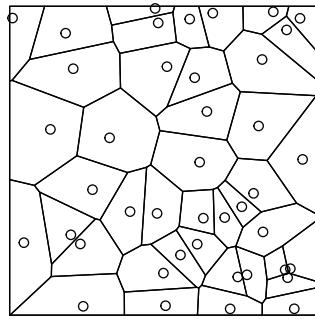


Figure 3.1: Dirichlet tesselation of $\mathbb{R}^2$ using a set of randomly generated points.

Note that with either of these two approaches, the value of $\lambda(s)$ needs to be evaluated at a sample of points distributed throughout the spatial domain, $G$. Readers familiar with fitting species distribution models using MaxEnt or with estimating resource-selection functions using logistic regression hopefully see a connection here (i.e. the need for a representative sample from a well-defined spatial domain). Indeed, both MaxEnt and resource-selection functions have been shown to be equivalent to fitting an IPP model (Warton et al., 2010; Aarts et al., 2012; Fithian & Hastie, 2013; Renner & Warton, 2013; Renner et al., 2015), and it is useful to consider the *background* or *available* locations required by these methods as also approximating the integral in the IPP likelihood. This understanding, for example, suggests that the number of available points should be increased until the likelihood converges to a constant value (Warton et al., 2010; Renner et al., 2015).

## 3.5 Weighted Distributions: Connecting SHA Process Models to the IPP

Weighted distribution theory (Lele & Keim, 2006) provides a nice way to connect the SHA models discussed in Chapter 2 to the IPP model introduced here (Aarts et al., 2012). As in Chapter 2, let:

- $f_u(x)$ = the frequency distribution of habitat covariates at locations where animals are observed.

- $f_a(x)$ = the frequency distribution of habitat covariates at locations assumed to be available to our study animals.

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

$$f_u(x) = \frac{w(x, \beta)f_a(x)}{\int_{z \in E} w(z, \beta)f_a(z)dz} \tag{3.13}$$

The denominator of (3.13) ensures that the left hand side integrates to 1 and thus, $f_u(x)$ is a proper probability distribution.

We can also write the model in $G$-space:

$$f_u(s) = \frac{w(x(s), \beta)f_a(s)}{\int_{g \in G} w(x(g), \beta)f_a(g)dg} \tag{3.14}$$

In this case, $f_u(s)$ tells us how likely we are to find an individual at location $s$ in $G$-space, which depends on the environmental covariates associated with location $s$, $x(s)$. Typically, $f_a(s)$ is assumed to be uniform (i.e. all areas within $G$ are assumed to be equally available to the species). Then, if we let $w(x(s), \beta) = \exp(x\beta)$, we end up with the conditional likelihood of the IPP model (eq. (3.8)) (Aarts et al., 2012).

Thinking in terms of weighted distributions will sometimes be advantageous. In particular:

1. it makes it immediately clear that we are modeling the spatial distribution of observed (or, for telemetry studies, 'used') locations, $f_u(s)$, as a function of covariates (through $w(x(s), \beta)$), while assuming all habitat in $G$ is equally available/accessible.

2. it highlights a potential issue with applying this modeling framework in the context of species distribution models in that not all habitat may be equally accessible. In later chapters, we will relax this assumption by modeling $f_a(s)$ using non-uniform distributions, recognizing that nearby locations are more accessible than far away locations.

## 3.6   Simulation Example: Fitting the IPP model

To elucidate the above theoretical sections, we demonstrate how to simulate data from an IPP model. We then use the two approximations to the likelihood discussed above to see if we can recover the parameters used to simulate the data.

First, we create a landscape consisting of a risk and a resource. There are several ways to generate spatially correlated random variables. Here we simulate these environmental variables as independent unconditional Gaussian Random Fields using the `grf` function in the `geoR` package (Ribeiro Jr & Diggle, 2018)[2]. The resulting output will be formatted as a `raster` (i.e. a grid of values with an associated spatial extent and resolution or cell size).

```
library(geoR) # for gfr function
library(raster) # to work with rasters
library(dplyr) # for manipulating data
```

```
#' Generate 2 covariates
risk <- raster(grf(ncell(r), grid="reg", nx=2000, ny=2000,
                   xlims=c(0,2000), ylims=c(0,2000),
                   cov.pars=c(0.1, 100), messages=FALSE))

resource <- raster(grf(ncell(r), grid="reg", nx=2000,
                       ny=2000, xlims=c(0,2000),
                       ylims=c(0,2000),
                       cov.pars=c(0.1, 100), messages=FALSE))
env <- stack(risk, resource) # environmental layers
names(env)<-c("risk", "resource")
```

We set the true intensity function to:

$$\log(\lambda(s)) = 6 - x_1(s) + 2x_2(s)$$

where $x_1(s)$ and $x_2(s)$ contain the values of the risk and resource, respectively, at each location, s, within G (Fig. 3.2).

```
# Coefficients for the IPP model:
betas<-matrix(c(6,-1,2), nrow=3, ncol=1)
# Intensity function
lambdas<- exp(betas[1] + betas[2]*risk + betas[3]*resource)
# Point version
lambdas_m<-rasterToPoints(lambdas)
```

We then simulate random points within G using two steps:

1. We determine a random number of points, n_u, to generate using the fact that n_u is a Poisson random variable with mean $\mu = \int_G \lambda(s)ds$.

2. Conditional on having n_u points within G, we determine their spatial locations by sampling from the original landscape (approximated here using a really fine grid) with cell probabilities proportional to $\lambda(s)$ (eq. (3.8)).

Equivalently, we could have generated independent Poisson random variables for each cell in our fine grid (see Chapter 3.7.1 for more on the connection between gridded counts and the IPP model).

---

[2]the spatial correlation of the Gaussian Random field is determined by its parameters: the range, $\phi = 50$, determines the distance at which locations are no longer autocorrelated, and the sill, $\sigma^2 = 0.1$, determines the spatial or temporal variance for locations that far enough away to be statistically independent.
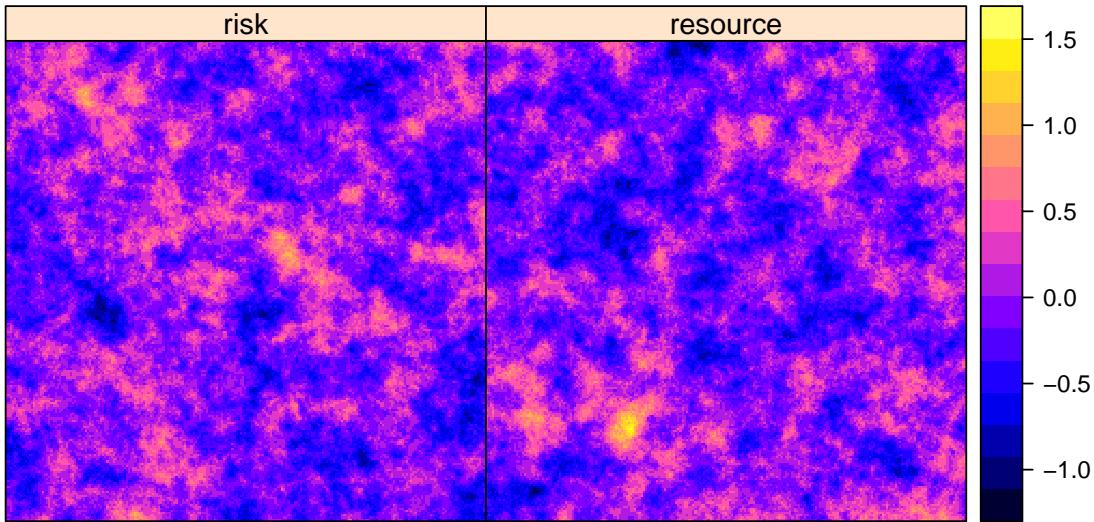
Figure 3.2: Geographical distribution of a risk (x_1) and resource (x_2) on a landscape.

```
# Step 1: determine number of random points to generate
# mu = mean(lambda(s)) over the spatial domain
mu <- cellStats(lambdas, stat='mean')
n_u <- rpois(1, lambda=mu) # number of points to generate
cat("mu = ", mu, "n_u = ", n_u)
```

```
## mu =  476.999 n_u =  486
```

```
# Step 2: determine their locations by sampling with
# probability proportional to labmda(s)
# x = design matrix holding the risk and resource values in the spatial domain
# n = number of grid cells in the spatial domain
# obs = index for each chosen observation
x <- cbind(rasterToPoints(env))
n <- nrow(x)
obs<-sample(1:n, size=n_u, prob=lambdas_m[,3], replace=TRUE)
# Step 3: create data frame with observed point-pattern
ppdat<-as.data.frame(x[obs,])
```

We can then look at the distribution of the locations in both G- and E-space (Figure 3.3).

```
gspace <- ggplot(data = ppdat, aes(x = x, y = y))+geom_point()+
    stat_density2d() +xlab("Easting")+ylab("Northing")+
    ggtitle("a. Geographical Space")
espace<- ggplot(data = ppdat, aes(x = risk, y = resource))+geom_point()+
    stat_density2d() +
    ggtitle("b. Environmental Space")
```

```
grid.arrange(gspace, espace, ncol=2)
```
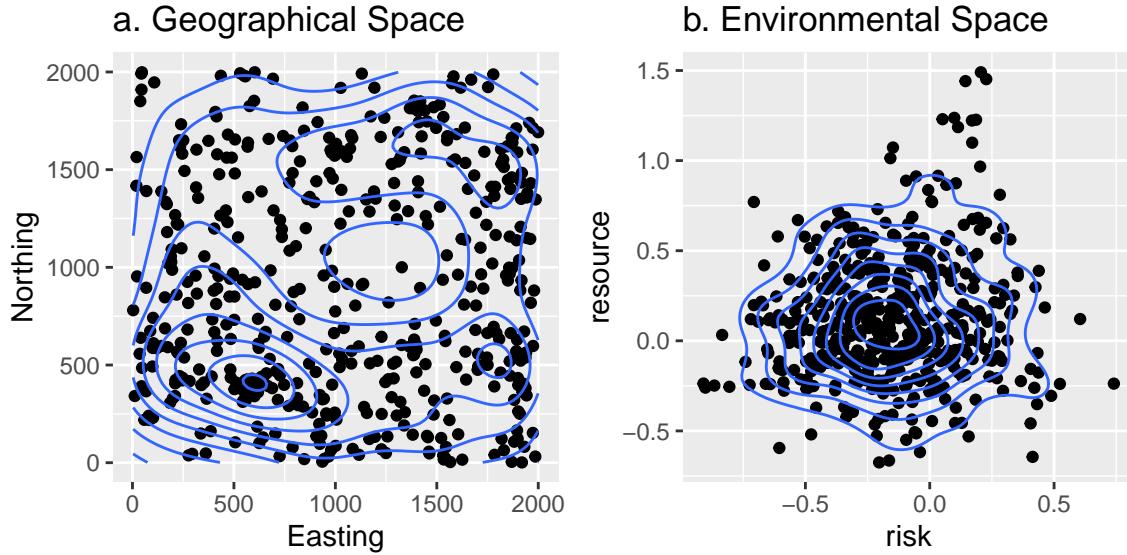

Figure 3.3: Plot of simulated locations in (a) geographical and (b) environmental space.

3.6.1   Maximum Likelihood Estimation (MLE) using Monte-Carlo Integration

To calculate the log-likelihood, we need to evaluate the intensity function at the observed points (eq. (3.10)). In addition, when estimating parameters using the approximation to the likelihood given by eq (3.11), we will need to evaluate the intensity function at a set of random (available or background) points from within G. We will use matrices and matrix multiplication to perform these calculations.[3] We begin by creating a design matrix (`xmat_u`) for the observed points. Each row of the design matrix captures the values of the explanatory variables for one of the points. A column of `1`'s is also included for the intercept (eq. (3.7)).

```
# Create design matrix containing covariates at observed locations
xmat_u <- cbind(1,ppdat[, c("risk", "resource")])
head(xmat_u, n=2) # Look at the first two observations
```

```
##   1       risk  resource
## 1 1  0.13298777 0.2028965
## 2 1 -0.06392881 0.2424878
```

We next generate a random sample of available points and create a design matrix for these points, `xmat_a`.

```
#' First, generate n_a random available points
n_a<-10000
randobs<-sample(1:n, size=n_a, replace=TRUE)
xmat_a <- cbind(1,x[randobs, c("risk", "resource")])
head(xmat_a, n=2)
```

```
##              risk   resource
## [1,] 1 -0.2958577 -0.3178855
## [2,] 1  0.3858838 -0.1474475
```

We then write a function to calculate the negative log-likelihood using Monte Carlo integration (eq. (3.11)).

---

[3]in R, we can use `%*%` to perform matrix multiplication. If X is an $n \times p$ matrix and β is a $p \times 1$ matrix, then $X\beta$ will be an $n \times 1$ matrix (a column vector). The ith element of this matrix will equal $X_{i1}\beta_1 + X_{i2}\beta_2 + \ldots + X_{ip}\beta_p$.

```r
logL_MC<-function(par, xmat_u, xmat_a){
  betas<-par
# log(lambda) at observed locations for first term of logL
# note %*% is used for matrix multiplication in R
  loglam.u<-as.matrix(xmat_u) %*% betas
# lambda(s) at background locations for second term of LogL
  lambda.a<-exp(as.matrix(xmat_a) %*% betas)
# Now, sum terms and add a negative sign
# to translate the problem from "maximization" to a "minimization" problem
  -sum(loglam.u)+mean(lambda.a) #negative log L
}
```

To find the values of β that minimize this function (equivalent to maximizing the likelihood), we will use the `optim` function in R, which offers a number of built-in optimization routines. `optim` requires a list of starting values for the numerical optimization routine (`par = c(4, 0, 0)`), the function to be minimized (`logL_MC`; note the first argument of this function needs to be the parameter vector that we are varying in order to find values that minimize our function), and an optimization method used to find the minimum (we choose to use `BFGS` but other options are also available). We can pass other arguments (e.g. `xmat_u`, `xmat_a`), which are needed in this case to calculate the negative log-likelihood at a given set of β's.

```r
#' Now estimate parameters:
mle_MC<-optim(par=c(4, 0, 0), fn=logL_MC, method="BFGS",
              xmat_u = xmat_u, xmat_a=xmat_a,
              hessian=TRUE)
```

The argument `hessian=TRUE` allows us to calculate asymptotic standard errors using standard theory for maximum likelihood estimators (Casella & Berger, 2002).[4] The `solve` function in R will find the inverse of the Hessian, which provides an estimate of the variance-covariance matrix for our parameters. The square root of the diagonal elements of this matrix give us our estimated standard errors.

```r
varbeta_MC<-solve(mle_MC$hessian)
ests_MC<-data.frame(beta.hat=mle_MC$par,
                    se.beta.hat=sqrt(diag(varbeta_MC)) )
#' Add the confidence limits to the data set using the mutate function
ests_MC <- ests_MC %>% mutate(lowerCI=beta.hat - 1.96*se.beta.hat,
                              upperCI=beta.hat + 1.96*se.beta.hat)
round(ests_MC, 2)
```

```
##   beta.hat se.beta.hat lowerCI upperCI
## 1     6.01        0.05    5.91    6.12
## 2    -1.14        0.16   -1.46   -0.83
## 3     1.89        0.14    1.61    2.17
```

We see that the parameters used to simulate the data, $(\beta_0, \beta_1, \beta_2) = (6, -1, 2)$ all fall within their respective 95% confidence intervals.

3.6.2 MLE using numerical integration

Alternatively, we could estimate parameters using the approximation to the likelihood given by eq (3.12). To do so, we need to generate available locations on a regular grid and then determine appropriate quadrature weights. Quadrature weights reflect the area associated with the tile surrounding each available location and should sum to the total area of G which we have arbitrarily set to 1. We use the resolution of the sampled grid to determine the quadrature weights.

---

[4]As our sample size gets large, the sampling distribution of maximum likelihood estimators, $\hat{\theta}$, converges to $N(\theta, I^{-1}(\theta))$, where $I^{-1}(\theta)$ is the inverse of the Hessian matrix.

```r
# Generate available data using a gridded background with
# n_a available points.
x_samp<-sampleRegular(env, size=n_a, asRaster=TRUE)
# Calculate quadrature weights:
# weights = area of grid cell/Total area
weights <-xres(x_samp)*yres(x_samp)/prod(dim(resource))
cat("weights = ", weights, ";", "sum of weights = ",
    sum(rep(weights, n_a)),  ";",
    "area of omega =" , sum(xres(resource)*yres(resource)))
```

```r
## weights =  1e-04 ; sum of weights =  1 ; area of omega = 1
```

We then create a design matrix for calculating the intensity function at the quadrature points.

```r
#' Create design matrix containing covariates at background locations
xmat_a2<-cbind(1, rasterToPoints(x_samp)[, c("risk", "resource")])
head(xmat_a2, n=2) # look at the first two observations
```

```r
##                 risk    resource
## [1,] 1 -0.13756791 -0.1481310
## [2,] 1 -0.05136795  0.1363807
```

We then write a function to calculate the negative log-likelihood for a given set of β's, using the set of gridded available locations and their associated quadrature weights via eq. (3.12).

```r
# Likelihood with numerical integration
logL_num<-function(par, xmat_u, xmat_a2,  weights){
  betas<-par
# log(lambda) at observed locations for first term of logL
  loglam.u<-as.matrix(xmat_u) %*% betas
# lambda(s) at background locations for second term of LogL
  lambda.a<-exp(as.matrix(xmat_a2) %*% betas)
# Now, calculate the negative log L
  -sum(loglam.u)+sum(lambda.a*weights)
}
```

And, again we use `optim` to minimize this function.

```r
#Maximize likelihood
mle_num<-optim(par=c(4,0,0), fn=logL_num, method="BFGS",
               xmat_u = xmat_u, xmat_a2=xmat_a2,
               weights=weights, hessian=TRUE)
#' Estimates and SE's
varbeta_num<-solve(mle_num$hessian) # var/cov matrix of beta^
ests_num <- data.frame(beta.hat=mle_num$par,
                       se.beta.hat=sqrt(diag(varbeta_num)))
#' Add the confidence limits to the data set using the mutate function
ests_num <- ests_num %>% mutate(lowerCI=beta.hat - 1.96*se.beta.hat,
                                upperCI=beta.hat + 1.96*se.beta.hat)
round(ests_num,2)
```

```r
##   beta.hat se.beta.hat lowerCI upperCI
## 1     6.01        0.05    5.90    6.12
## 2    -1.15        0.16   -1.47   -0.84
## 3     1.91        0.14    1.64    2.19
```

We get nearly identical results as when using Monte Carlo integration, which is reassuring.

3.7 Fitting IPP models: Special Cases

3.7.1 Special Case I: Gridded Data

In many applications of SHA models, covariate data come from remote sensing data products that exist as a raster. In this case, we can estimate parameters of the IPP model by fitting a Poisson regression model to the count of individuals in each grid cell using the `glm` function in R (Aarts et al., 2012). A Poisson glm is appropriate because the IPP model implies that spatial locations of the individuals are independent and that the number of individuals in grid cell i is a Poisson random variable with mean, $\mu_i = \int_{G_i} \lambda(s)ds$. When all predictors are grid-based, then $\mu_i = \lambda_i|G_i|$, where λ_i is the value of $\lambda(s)$ in grid cell i and $|G_i|$ is the area of grid cell i. This model could be specified using: `glm(y ~ x + offset(log(area)), family = poisson())`, where `area` measures the area of each grid cell. For readers that may not be familiar with an offset, this term is similar to including `log(area)` as a predictor, but forcing its coefficient to be 1. There are two reasons to include an offset: 1) it allows us to directly model densities (number of individuals per unit area); and 2) it provides a way to account for varying cell sizes in the case of irregular grids or meshes.

Let N_i and D_i represent the number and density of individuals in grid cell i. We can formulate our model in terms of density:

$$E[D_i] = \frac{E[N_i]}{G_i} = \lambda_i \tag{3.15}$$

$$\implies \log\left(\frac{E[N_i]}{|G_i|}\right) = \log(\lambda_i) \tag{3.16}$$

$$\implies \log(E[N_i]) - \log(|G_i|) = \log(\lambda_i) \tag{3.17}$$

$$\implies \log(E[N_i]) = \log(|G_i|) + \beta_0 + \beta_1 x_1(s_i) + \dots \beta_p x_p(s_i) \tag{3.18}$$

If all cells are equally sized, then including an offset will only affect the intercept β_0 but not the other regression parameters. In the case of unequal cell sizes, however, omitting the offset can lead to biased parameter estimates.

3.7.2 Special Case II: Presence-Absence Data

What if our usage data consists of binary (0 or 1) presence-absence data within the set of grid cells? Let's assume the IPP model is again appropriate. Consider grid cell i that is of area $|G_i|$, and let $y_i = 1$ if there is at least 1 observed individual in cell i and 0 otherwise. $P(y_i = 1)$ is given by:

$$P(y_i = 1) = 1 - P(y_i = 0). \tag{3.19}$$

If the IPP model is appropriate, then the number of individuals in cell i is a Poisson random variable with mean $\lambda_i|G_i| = \exp(\beta_0 + \beta_1 x_1(s_i) + \dots \beta_p x_p(s_i))|G_i|$. Thus,

$$P(y_i = 0) = \frac{y_i^0 \exp(-\lambda_i|G_i|)}{0!} \tag{3.20}$$

$$= \exp(-\lambda_i|G_i|) \tag{3.21}$$

and, we have:

$$P(y_i = 1) = p_i = 1 - \exp(-\lambda_i|G_i|) \tag{3.22}$$

$$\implies p_i = 1 - \exp(-\exp(\beta_0 + \beta_1 x_1(s_i) + \dots \beta_p x_p(s_i))|G_i|) \tag{3.23}$$

Thus, it is natural to model binary presence-absence data using the so-called "complementary log-log link" (Baddeley et al., 2010):

$$y_i \sim \text{Bernoulli}(p_i), \tag{3.24}$$

$$\log(-\log(1 - p_i)) = \beta_0 + \beta_1 x_1(s) + \ldots \beta_p x_p(s) + log(|G_i|) \tag{3.25}$$

We can fit this model in R using:

```
glm(y~x1+x2+...xp + offset(log(area)), family=Binomial(link=cloglog))
```

Although it is more common to fit binary regression models with a logit link (i.e. $\log(\frac{p_i}{1-p_i}) = x\beta$), the complementary log-log link is appealing because of its connection with the IPP model. Further, it provides an opportunity to connect different types of usage data (presence-absence and point or count data) under a common process model (e.g.@Fithian2013; Koshkina et al., 2017; Scotson, Fredriksson, Ngoprasert, Wong, & Fieberg, 2017). Like the logit link, the complementary log-log link ensures p is constrained to lie between 0 and 1 as our predictor function goes from $-\infty$ to ∞ (Figure 3.4).
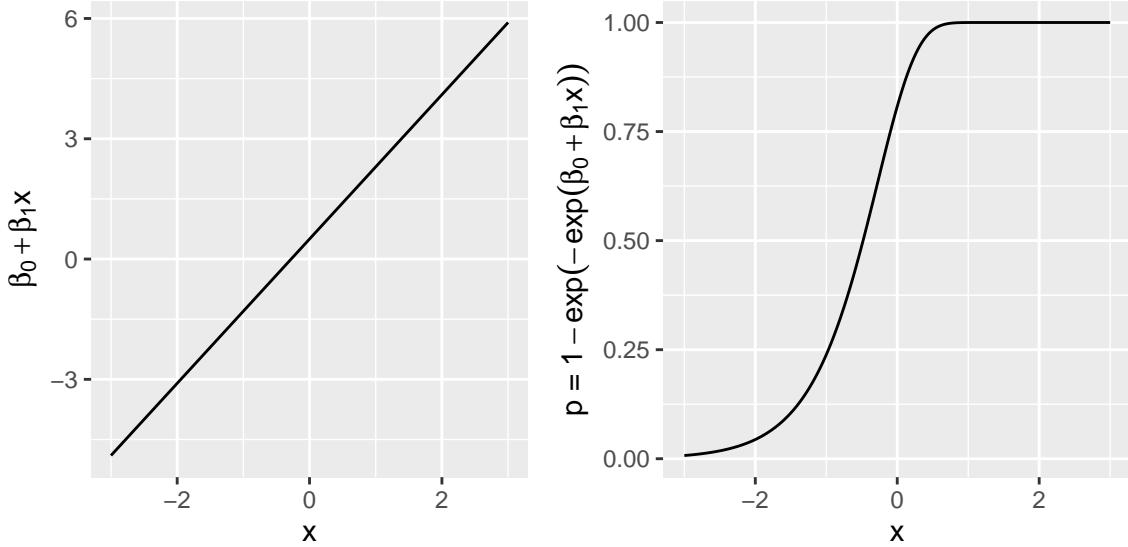

Figure 3.4: Plot of the predictor function and $P(y = 1) = p$ when using a complementary log-log link function with a single predictor x.

Note, although presence-absence data are frequently used to model species distributions, there is a potential loss of information when summarizing counts using binary observations. The loss of information by summarizing counts into binary observations becomes more severe when grid cells are large and species density is high. For example, consider the extreme scenario where all grid cells include at least one observed individual. In this case, there will be no spatial variation in species occurrence, and we will have lost the ability to estimate species-habitat associations. It is also important, though not always recognized, that models of the probability of occurrence are scale dependent unless formulated under a continuous point-process framework (i.e. with complementary log-log link and offset equal to the log cell size; Baddeley et al., 2010). This scale dependence, and potential loss of information, can easily be seen by binning our simulated count data using a series of coarsening cell sizes (see Figure 3.5). The expected counts and the probability of presence both depend on the size of the grid cell. In addition, the probability of presence approaches 1 as the size of the grid cells increases.
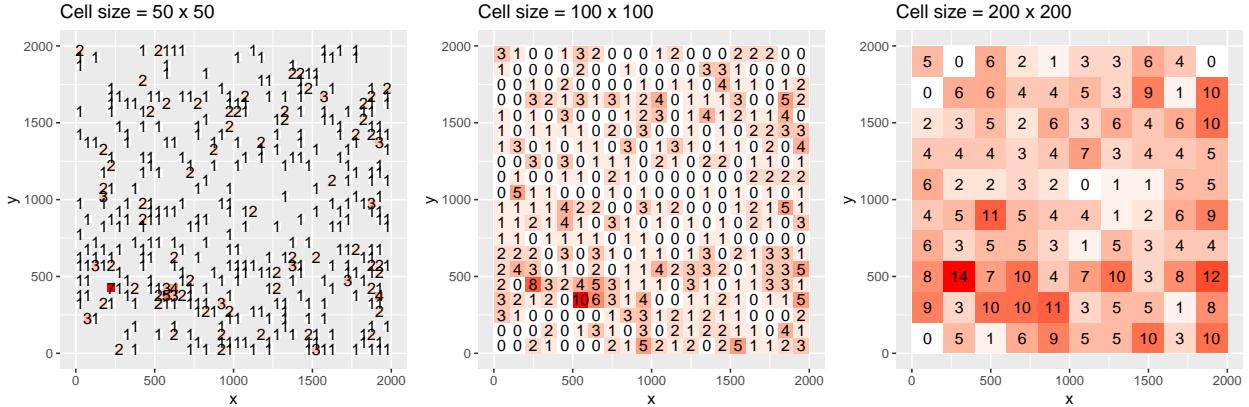
Figure 3.5: Gridded counts of the simulated locations shown in Figure 3.3a using grid cells of size a) 50 x 50, b) 100 x 100, and c) 200 x 200. Gray areas in panel (a) represent cells with 0 locations, which are too numerous to annotate. Note that the expected counts and the probability of presence both depend on the size of the grid cell, and the probability of presence approaches 1 as the size of the grid cells increases.

3.8 Imperfect Detection and Sampling Biases

The absolute density of observed individuals will often be driven by various observation-level processes (e.g. sampling effort, duration of data collection). In this case, our models effectively capture the *density of sightings* rather than the *density of individuals on the landscape*. As a result, the intercept in the IPP model is often difficult to interpret, and estimates of slope coefficients may be impacted by spatially varying observation biases. Given these issues, many applications of SHA models include covariates to adjust for observation biases and (rightly) abandon the idea of estimating absolute densities. In such cases, it can be convenient to work with the conditional likelihood of the IPP model (eq. (3.8); Aarts et al., 2012) – i.e. the likelihood for the distribution of the locations as a function of covariates, conditional on having observed n_u locations. The conditional likelihood will also prove useful for extending our SHA model to address problems of unequal availability or accessibility of various habitats (see Section 3.5).

3.8.1 Conditional likelihood for IPP model

Here, we build on the simulation example from Section 3.6, demonstrating that we can recover the slope parameters by maximizing the conditional likelihood (eq. (3.8)), which we define below:

```r
#' Conditional log-likelihood
logLCond<-function(par, xmat_u, xmat_a, MC=FALSE, weights=NULL){
  # MC = TRUE for Monte Carlo Integration
  # MC = FALSE for numerical integration
  betas<-par
  n<-nrow(xmat_u) # number of observed individuals
# log(lambda) at observed locations for first term of logL
# (use -1 to remove the term for the intercept)
  loglam.u<-as.matrix(xmat_u)[,-1] %*% betas
# lambda(s) at background locations for second term of LogL
  lambda.a<-exp(as.matrix(xmat_a)[,-1] %*% betas)
  if(MC==FALSE){
    -sum(loglam.u)+n*log(sum(lambda.a*weights)) #negative log L
  }else{
    -sum(loglam.u)+n*log(mean(lambda.a)) #negative log L
  }
}
```

Table 3.2: Estimates of Slope Parameters

Method	$\hat{\beta}_1$	SE	$\hat{\beta}_2$	SE
Full Monte Carlo	-1.14	0.16	1.89	0.14
Full Quadrature	-1.15	0.16	1.91	0.14
Conditional Monte Carlo	-1.14	0.16	1.89	0.14
Conditional Quadrature	-1.15	0.16	1.91	0.14

```r
# Using Monte Carlo Integration
mle_Cond_MC<-optim(par=c(0,0), fn=logLCond, method="BFGS",
                   xmat_u=xmat_u, xmat_a=xmat_a, MC=TRUE,
                   hessian=TRUE)
varbeta_Cond_MC<-solve(mle_Cond_MC$hessian)
ests_Cond_MC<-data.frame(beta.hat=mle_Cond_MC$par,
                   se.beta.hat=sqrt(diag(varbeta_Cond_MC)) )

# Now using numerical integration
mle_Cond_NI<-optim(par=c(0,0), fn=logLCond, method="BFGS",
                   xmat_u = xmat_u, xmat_a=xmat_a2,
                   weights=weights, MC=FALSE, hessian=TRUE)
varbeta_Cond_NI<-solve(mle_Cond_MC$hessian)
ests_Cond_NI<-data.frame(beta.hat=mle_Cond_NI$par,
                   se.beta.hat=sqrt(diag(varbeta_Cond_NI)) )
```

The results are nearly identical to those obtained from maximizing the full unconditional likelihood (Table 3.2).

3.8.2   Thinned Point-Process Models

Our observations of habitat use will often be influenced by various sampling biases and uneven sampling effort (e.g. we may oversample areas near roads that are easily accessible by human observers; Sicacha-Parada, Steinsland, Cretois, & Borgelt, 2020). These issues are particularly relevant to data collected haphazardly or opportunistically as part of museum collections or citizen science programs (Fithian et al., 2015). In such cases, it will often be fruitful to model the data as a *thinned point-process model* (Warton, Renner, & Ramp, 2013; Dorazio, 2014; Fithian et al., 2015; Sicacha-Parada et al., 2020). Specifically, we may think of the observed data as being generated by 2 processes: 1) the biological process model (as discussed in Chapter 2) written in terms of an Inhomogeneous Poisson Point-Process, and 2) an independent observation model that "thins" the data, specified using a separate function $p(s)$ that describes the likelihood of observing an individual, conditional on that individual being present at location s. Typically, $p(s)$ will also be written as a log-linear function of spatial (and potentially non-spatial) covariates $(z_1, z_2, \ldots, z_m)$ and regression parameters $(\gamma_1, \gamma_2, \ldots, \gamma_m)$[5]:

$$log(p(s)) = \gamma_0 + \gamma_1 z_1(s) + \ldots \gamma_m z_m(s)$$

This leads to the following likelihood for the *observed* sightings:

$$\log(L(\lambda, \beta\gamma|s_1, s_2, \ldots, s_n, n)) = \sum_{i=1}^{n} \log(\lambda(s_i)p(s_i)) - \int_G \lambda(s)p(s)ds \qquad (3.26)$$

---

[5]As a thinning process, we expect $0 \leq p(s) \leq 1$, which is not strictly enforced for the log link. Other links could be used, but the log-link is convenient as it leads to an additive model.

If the covariates affecting $\lambda(s)$ and $p(s)$ are distinct, then we can think of the thinned point-process model as one where the intensity function includes both environmental drivers (x) and "observer bias variables" (z) (Warton et al., 2013). A benefit of the log-linear formulation of both $\lambda(s)$ and $p(s)$ is that it results in an additive model:

$$
\begin{aligned}
\log(\lambda(s)p(s)) &= \log(\lambda(s)) + \log(p(s)) \\
&= (\beta_0 + \gamma_0) + \beta_1 x_1(s) + \ldots \beta_p x_p(s) + \gamma_1 z_1(s) + \ldots \gamma_m z_m(s)
\end{aligned}
\tag{3.27}
$$

In this case, the intercepts of the two processes are not separately identifiable (i.e. we can only estimate their sum). Thus, we are unable to the model absolute intensity, $\lambda(s)$, or absolute density of individuals. Similar issues arise if covariates influence both density and detectability, i.e. we will only be able to estimated the sum of these effects (but see Section 3.9 for possible solutions using data-integration techniques). If, on the other hand, the covariates that influence density and detectability are distinct, then we can obtain unbiased predictions for the distribution of locations, conditional on having observed a set of n_u locations, by setting the z's to common values (Warton et al., 2013).

Quantitative ecologists have also developed a variety of methods to estimate absolute densities in situations where we cannot detect all individuals, but these methods will typically require specialized data collection methods (recall the Fisher quote from the start of this chapter) and additional modeling assumptions. We will demonstrate one of these approaches, *distance sampling*, in the next Section (3.8.3).

3.8.3 Simulation Example: Distance Sampling as a Thinned Point Process

To demonstrate fitting of a thinned point-process model, we will consider surveying our simulated individuals from Section 3.6 using a set of line transects. Specifically, we will place 40 equally-spaced transects throughout the survey region and assume:

1. all individuals that fall on the transect line are detected.
2. the probability of detecting individuals decreases with their distance from the transect line. A variety of models can be used to describe how detection drops with distance (Buckland et al., 2005). Here, we assume that detection probabilities follow a half-normal curve.

These assumptions result in a simple model for $p(s)$, written in terms of a single covariate that depends on the perpendicular distance, z, between each individual and the nearest transect line (Borchers & Marques, 2017)[6]:

$$
p(s) = \exp\left(\frac{-z(s)^2}{2\sigma^2}\right)
\tag{3.28}
$$

This formulation of the half normal model has one parameter, σ, which determines how quickly detection probabilities decline with distance. It is also frequently used to model detection probabilities in spatial mark-recapture studies, with $z(s)$ measuring the distance between an individual's latent (i.e. unobserved) activity center and a set of traps (J. A. Royle et al., 2013b).

We can write our model for $p(s)$ as a log-linear model:

$$
log(p(s)) = \gamma_0 + \gamma_1 z_1(s)
\tag{3.29}
$$

with $\gamma_0 = 0$ (to ensure $p(s) = 1$ for observations at 0 distance), $z_1(s) = z(s)^2$, and $\gamma_1 = 1/(2\sigma^2)$.

Here, we simulate data and show that we can again recover the parameters describing the intensity function. In the code below, we:

---

[6]in truth, the individual could potentially be detected from multiple transect lines if they are close together, but the probability is ≈ 0 for all but the nearest transect line

1. Place vertical transects along G using a systematic sampling design with a random starting location for the first transect.
2. Calculate perpendicular distances between each individual $(s_1, s_2, \ldots, s_{511})$ and the nearest transect to determine the probability of observing each individual.
3. Use a binomial random number generator to determine which of the $n_u = 511$ individuals are observed.

```r
# place 40 transects systematically with a random start
transects.x<-seq(from=runif(1,0,50),to=2000, length=40)
# determine distance between each observation and nearest
# transect
obs_x<-ppdat[, "x"] # x-coordinates of the n_u individuals
rownames(xmat_u)<-NULL # get rid of rownames and treat as data frame
xmat_u<-as.data.frame(xmat_u)
# Calculate perpendicular distance to nearest transect and square it
xmat_u$z<-sapply(1:n_u, FUN=function(x){min((obs_x[x]-transects.x)^2)})
# Determine detection probabilities using half-normal
sig2 <- 100 # sigma^2
ps<-exp(-xmat_u$z/(2*sig2)) # p(s)
ppdat$detect<-rbinom(n_u, 1, ps)
#observed data
xmat_uobs<-xmat_u[ppdat$detect==1,]
# Available locations
gridobs_x<-rasterToPoints(x_samp)[,1]
rownames(xmat_a)<-NULL
xmat_a_th<-as.data.frame(xmat_a)
# Add (distances to transects)^2 to the dataframe containing available data
xmat_a_th$z<-sapply(1:n_a, FUN=function(x){min((gridobs_x[x]-transects.x)^2)})
```
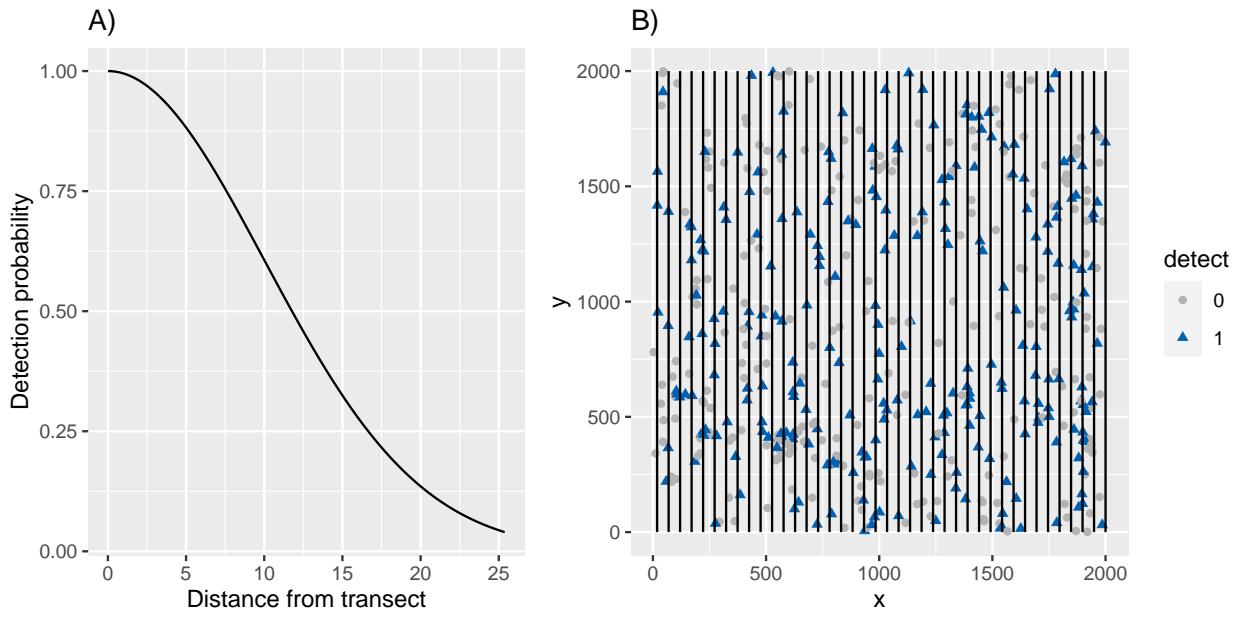

Figure 3.6: Distance sampling applied to the simulated data in Chapter 3. A) Detection probability as a function of distance from the transect line. B) Line transects (black lines) along with the underlying true and observed spatial point pattern.

Now, we write the likelihood of the thinned data and again use `optim` to estimate parameters of the process and observation model.

Table 3.3: Estimates of IPP and Distance Sampling Parameters

Parameter	True value	$\hat{\beta}_1$	SE
β_0	6	6.01	0.09
β_1	-1	-1.27	0.23
β_2	2	1.78	0.20
σ	10	10.29	0.58

```r
LogL_thin<-function(pars, xmat_u, xmat_a, MC=FALSE, weights=NULL){
  # MC = TRUE for Monte Carlo Integration
  # MC = FALSE for numerical integration
  betas[1:3]<-pars[1:3] # parameters in lambda(s)
  betas[4]<- -1/(2*pars[4]^2) # par[4] = sigma
  n<-nrow(xmat_u)
# log(lambda(s)*p(s)) at observed locations for first term of logL
  loglam.u<-as.matrix(xmat_u) %*% betas
# lambda(s)*p(s) at background locations for second term of LogL
  lambda.a<-exp(as.matrix(xmat_a) %*% betas) #second term
  if(MC==FALSE){ #quadrature
     -sum(loglam.u)+sum(lambda.a*weights) #negative log L
  }else{ # Monte Carlo integration
     -sum(loglam.u)+mean(lambda.a) #negative log L
  }
}
mle_thin<-optim(par=c(4, 0,0,20), fn=LogL_thin, method="BFGS",
             xmat_u = xmat_uobs, xmat_a=xmat_a_th, weights=weights, MC=FALSE,
             hessian=TRUE)
varbeta_thin<-solve(mle_thin$hessian)
```

We see (Table 3.3) that we are able to obtain accurate estimates of the parameters of the intensity function and the detection function. Note that it is the strong assumption of perfect detection on the transect line that allows us to estimate absolute density in this case. If this assumption fails, $\gamma_0 \neq 0$ and we will underestimate density.

3.9 Looking Forward: Data Integration for Addressing Sampling Biases and Imperfect Detection

Of course, it is possible (and maybe quite likely) that $\lambda(s)$ and $p(s)$ will be influenced by one or more shared covariates (e.g. species that prefer forest cover may also be more difficult to detect when present in the forest compared to when they are in clearings). Unfortunately, in this case, we will not generally be able to separate out the effect of shared covariates on detection from their effect on species distribution patterns; we will only be able to estimate a single parameter for each shared covariate equal to $\beta_i + \gamma_i$ (Dorazio, 2012; Fithian et al., 2015). There are some situations, however, where we can take advantage of multiple data sources and standardized survey protocols to estimate covariate effects on species distribution patterns in the presence of sampling biases. We mention a few possibilities here, but leave detailed discussions of these approaches to later chapters.

Quantitative ecologists have developed numerous methods for estimating detection probabilities in wildlife surveys (Lancia, Kendall, Pollock, & Nichols, 2005) − we saw one such approach in Section 3.8.3 (i.e. distance sampling). Other popular approaches typically require marked individuals or repeated sampling efforts either by multiple observers or the same observer at multiple points in time. Examples of methods that use marked

individuals include spatial or non-spatial mark-recapture estimators (Royle et al., 2013a; McCrea & Morgan, 2014) and sightability models (Steinhorst & Samuel, 1989; Fieberg, 2012; Fieberg, Alexander, Tse, & Clair, 2013). Methods that rely on repeated sampling include double-observer methods (e.g. Nichols et al., 2000), occupancy models (MacKenzie et al., 2017), and N-mixture models (Royle, 2004). These latter methods work by assuming population abundance (or presence-absence in the case of occupancy models) does not change between sampling occasions (referred to as an assumption of *population closure*). If this assumption holds, repeated sampling can provide the information necessary to model detection probabilities. In particular, if we see a species present at a site during one sampling occasion and not another, then we know that we failed to detect it during the latter occasion. N-mixture models work similarly, but rely on variability in counts over time to inform the model of detection probabilities. Yet, both approaches can break down when their assumptions, which are difficult to verify, do not hold (e.g. Welsh, Lindenmayer, & Donnelly, 2013; Barker, Schofield, Link, & Sauer, 2018; Link, Schofield, Barker, & Sauer, 2018).

Recently, there have been many efforts to integrate data from unstructured ("presence-only") surveys with data from structured surveys that allow estimation of detection probabilities (see recent reviews by Miller et al., 2019; and Fletcher et al., 2019; Isaac et al., 2020). Data integration has the potential to increase spatial coverage, precision, and accuracy of estimators of species distributions (Pacifici, Reich, Miller, & Pease, 2019). Most often, data integration is done via a joint likelihood with shared parameters (e.g. treating the presence-only data as a thinned version of the same point-process model used to model data from structured surveys; Dorazio, 2014; Fithian et al., 2015; Koshkina et al., 2017; Schank et al., 2017; Isaac et al., 2020). As an alternative to using joint likelihoods, Pacifici et al. (2017) suggested using "data fusion approaches" that share information by using one data set as a predictor when modeling another or by using spatial random effects that are shared across multiple data sets (see also Gelfand & Shirota, 2019; Gelfand, 2020). These latter approaches should be more robust to differences in underlying intensity functions since they make fewer assumptions about shared parameters, and they are also likely to be preferable in cases where one data source is clearly superior to others. Data may also be integrated across multiple species to infer spatial variability in detection probabilities, either using data from non-target species as a surrogate for survey effort (Manceur & Kühn, 2014) or through joint modeling under the assumption that observation processes are common to multiple species (e.g. Fithian et al., 2015; Giraud, Calenge, Coron, & Julliard, 2016; Coron, Calenge, Giraud, & Julliard, 2018). Other work in this area has included developing methods to deal with misaligned spatial data (Pacifici et al., 2019) and integrating data that are clustered in space (Renner, Louvrier, & Gimenez, 2019).

Simmonds, Jarvis, Henrys, Isaac, & O'Hara (2020) used simulations to evaluate the performance of different data integration frameworks when faced with small sample sizes, low detection probabilities, correlated covariates, and poor understanding of factors driving sampling biases. They found that data integration was not always effective at correcting for spatial bias, but that model performance improved when either covariates or flexible spatial terms were included to account for sampling biases. In another study, Fletcher et al. (2019) found that models fit to combined presence-only and standardized survey data were nearly identical to models fit to just the presence-only data due to their being much more presence-only than standardized data. Using cross-validation techniques, they found that weighted likelihoods led to improved predictions. Much work still needs to be done to evaluate the performance of various data integration methods under different data generating scenarios (Isaac et al., 2020; Simmonds et al., 2020).

3.10   Looking Forward: Relaxing the Independence Assumption

As discussed in Chapter 1, locations of individuals will often be clustered in space, and some of the clustering may be driven by demographic rather than environmental processes. A classic example is the clustering in space of forest communities driven by limited dispersal capabilities (Seidler & Plotkin, 2006). Autocorrelation is also an inherent property of telemetry data (as will be discussed in the next Section, 3.11). A key assumption of the Inhomogeneous Poisson Point-Process model is that any clustering of observations in space can be explained using spatial covariates. Or, in other words, locations of the individuals in the population are independent after conditioning on $x(s)$. There are several ways that this assumption can be relaxed. In particular, there are formal point-process models that include various forms of dependence (attraction or repulsion of points), and several of these can be fit using functions in the `spatstat` library (Baddeley &

Turner, 2005). For an example in the context of species distribution modeling, see Renner et al. (2015). Alternatively, we can allow for dependencies by adding a spatially correlated random effect, $\eta(s)$, to the intensity function:

$$\log(\lambda(s)) = \beta_0 + \beta_1 x_1(s) + \ldots \beta_p x_p(s) + \eta(s) \tag{3.30}$$

Usually, $\eta(s)$ is modeled as a multivariate Gaussian process with $E[\eta(s)] = 0$ and $Cov[\eta(s), \eta(s')]$ depending on the distance between s and s'; we used a similar approach to simulate data in Section 3.6.[7]

Although adding spatially correlated random effects may at first appear to be a simple extension to the IPP model, these models can be more difficult to fit to real data, particularly in a frequentist framework (Renner et al., 2015). Thus, we will leave further consideration of these approaches until later chapters. We do, however, wish to make two more important points. First, there may be times when we are willing to estimate parameters (or summarize location data from each of several tagged individuals) using models that assume independence, but use alternative methods for calculating SE's that are robust to the independence assumption (e.g. using a block or cluster-level bootstrap; Fieberg, Vitense, & Johnson, 2020; Renner et al., 2015). Second, residual spatial correlation may differ across different data sets, and this may have important implications for integrated models that combine data under the assumption of a constant intensity function as discussed in Section 3.9 (see also Renner et al., 2019).

3.11 Telemetry data

3.11.1 Logistic Regression and Resource-Selection Functions

As briefly alluded to in the introduction of this chapter (Table 3.1), telemetry data differ from most survey data sets in that individuals are repeatedly followed over time. Despite this inherent difference, data from telemetry studies are often pooled and analyzed using methods similar to those described earlier in this chapter. In particular, logistic regression is often applied to data sets consisting of observed locations (with $y_i = 1$) and randomly generated background (referred to as available) locations (with $y_i = 0$). Logistic regression models are parameterized using a logit link ($i.e. \log(\frac{u}{1-u})$) rather than complementary log-log link as described in Section 3.7.2:

$$\log(P(Y_i = 1|x)/[1 - P(Y_i = 1|x)]) = \beta_0 + \beta_1 x_1(s) + \beta_2 x_2(s) + \ldots \beta_p x_p(s) \tag{3.31}$$

Rather than use the inverse-logit transform ($\frac{e^\mu}{1+e^\mu}$) to estimate $P(Y_i = 1|x)$, which largely reflects the number of available locations used to fit the model, most analyses of telemetry data focus on the exponential of the linear predictor without the intercept:

$$\exp(\beta_1 x_1(s) + \beta_2 x_2(s) + \ldots \beta_p x_p(s)) \tag{3.32}$$

referred to as a resource-selection function (RSF) (Section 2.8; Boyce & McDonald, 1999).

Historically, RSFs were described as providing estimates of "relative probabilities of use." We find this terminology problematic, however, since the probability of use requires first defining an appropriate spatial and temporal unit (e.g. use of a 100 km^2 area over a 3-month time period); the probability of finding at least one telemetry location in a grid cell will increase as we increase the grid cell size or the length of a telemetry study (see discussion in Lele & Keim, 2006).

Ecologists debated for many years whether logistic regression was appropriate for analyzing use-availability data (e.g. Keating & Cherry, 2004; Johnson, Nielsen, Merrill, McDonald, & Boyce, 2006). Warton et al. (2010) ended this debate by showing that the slope parameters of a logistic regression model will converge to those

---

[7]The model given in eq. (3.30) is referred to as a log-Gaussian Cox process (Møller, Syversveen, & Waagepetersen, 1998). Alternatively, spatial correlation can be modeled using conditional auto-regressive random effects in a Bayesian context (Pacifici et al., 2017, 2019).

Table 3.4: Regression coefficients in infinitely-weighted logistic and down-weighted Poisson regression models fit to simulated data.

	Infinitely-weighted Logistic		Down-weighted Poisson	
	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	NA	NA	6.01	0.05
risk	-1.14	0.16	-1.14	0.16
resource	1.89	0.14	1.89	0.14

of an IPP model as the number of available points increases to infinity. The connection to the IPP model also clarified the role of available points in the model fitting process (they allow us to approximate the integral in the denominator of the IPP likelihood). With logistic regression, the used locations also contribute to this integral, but the contribution becomes negligible as the number of available locations increases. Fithian & Hastie (2013) later showed that convergence is only guaranteed if the model is correctly specified or if available points are assigned "infinite weights." Therefore, when fitting logistic regression or other binary response models (e.g. boosted regression trees), one should assign a large weight (say 5000 or more) to each available location and a weight of 1 to all observed locations (larger values can be used to verify that results are robust to this choice). Alternatively, instead of increasing the contribution of the availability points to the integral, we can decrease the contribution of the used locations and switch to a Poisson likelihood function as discussed in Section 5, Supplement S1 of Renner et al. (2015). The potential advantage of this approach is that it then becomes possible to estimate the intercept, and thus, the absolute density of observations.

To demonstrate that infinitely-weighted logistic regression and down-weighted Poisson regression can be used to estimate parameters in an IPP model, we apply these approaches to our simulated data from Section 3.6. We begin with infinitely-weighted logistic regression:

```r
# Create a data frame with used and available locations
rsfdat<-rbind(xmat_u[,c("risk", "resource")], xmat_a[, c("risk", "resource")])
# Create y (1 if used, 0 otherwise)
rsfdat$y<-c(rep(1, nrow(xmat_u)),rep(0, nrow(xmat_a)))
# Assign large weights to available locations
rsfdat$w<-ifelse(rsfdat$y==0,5000,1)
# Fit weighted logistic model and look at summary
rsf<-glm(y~risk+resource, data=rsfdat, weight=w, family=binomial())
```

We see that we get identical estimates of the coefficients and their SEs as obtained when fitting the IPP model using the conditional likelihood (compare Tables 3.2, 3.4).

To implement the down-weighted Poisson approach, we create small weights ($w = 1^{-6}$) for the observed locations and weights equal to the area sampled divided by the number of random points ($w = \frac{|G|}{n_a}$) for the available locations. We then model $Y_i = 0$ for the available locations or $Y_i = 1/w$ for the observed locations.

```r
# A = area sampled
A=1
# Create weights
rsfdat$w2<-ifelse(rsfdat$y==0, A/sum(rsfdat$y==0), 1.e-6)
# Fit down-weighted Poisson regression model
rsfDWP<-glm(y/w2~risk+resource, data=rsfdat, weight=w2, family=poisson())
```

We see that we again recover the parameter values used to simulate the data, including the intercept (Table 3.4).

3.11.2 Connection to Utilization Distributions and Animal Home Ranges

In the context of home-range modeling, the distribution of habitat use in geographic space, $f_u(s)$, is often refereed to as a *utilization distribution* (Jennrich & Turner, 1969; Van Winkle, 1975). We can estimate this distribution directly using various non-parametric methods, including 2-dimensional kernel density estimators (Figure 3.7).

```
# Create data set with locations (x,y)
dataHR = as.data.frame(x[obs,c("x","y")])
# Calculate and plot 2-D density estimate f_u(s) using kde
kd <- ks::kde(dataHR, compute.cont=TRUE)
```
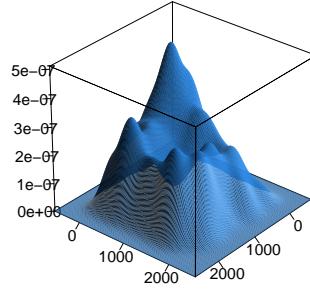

Figure 3.7: Kernel density estimate of $f_u(s)$ (z-axis) for the simulated point pattern data in Figure 3.3; $f_u(s)$ depicts the distribution of habitat use in geographic space, also known as the utilization distribution.

In Section 2.9, we demonstrated that we can also estimate $f_u(s)$ using a fitted RSF. What we omitted at the time, however, was that we first need to determine an appropriate spatial domain of availability, G, in geographic space. It is common to estimate G using a minimum convex polygon or the 95% isopleth of a kernel density estimator (KDE) applied to an individual's locations in geographic space (Figure 3.8); these same methods are frequently used to determine an outer contour when estimating the size of an individual's home range (see e.g. Worton, 1989; Kie et al., 2010; Fieberg & Börger, 2012; Horne et al., 2020).

Once we have determined an appropriate domain of habitat availability, G, we can fit an IPP model using the methods described in this chapter. We can then estimate $f_u(s)$ using eq. (3.14) as demonstrated in Section 2.9.

Although the RSF approach allows us to model $f_u(s)$ as a function of spatial predictors, it is somewhat circular since we have to first estimate G (often with an initial estimate of $f_u(s)$ obtained from a KDE) before estimating β, $w(x(s), \beta)$, and then $f_u(s)$ again via equation (3.14). Others have suggested estimating $f_u(s)$ using a KDE, which then may be regressed against spatial predictors to infer the importance of environmental characteristics on habitat use (Marzluff, Millspaugh, Hurvitz, & Handcock, 2004; Millspaugh et al., 2006). This approach, when used with a log-transformation, can again be shown to approximate an IPP model (Hooten, Hanks, Johnson, & Alldredge, 2013). Alternatively, Horne, Garton, & Rachlow (2008) suggested we should simultaneously estimate parameters in $f_a(s)$ and $w(x(s), \beta)$ by directly maximizing the weighted distribution likelihood formed using eq. (3.14). Horne et al. (2008) referred to this approach as fitting a *synoptic model* since it models space use by incorporating both home-range and resource-selection processes. If we assume $f_a(s)$ is described by bi-variate normal distribution, $\Phi(s, \mu, \sigma^2)$, then we can simultaneously estimate the mean (μ) and variance parameters (σ^2) in $f_a(s)$ and the regression parameters in $w(x(s), \beta)$ by maximizing the likelihood given by (3.33):
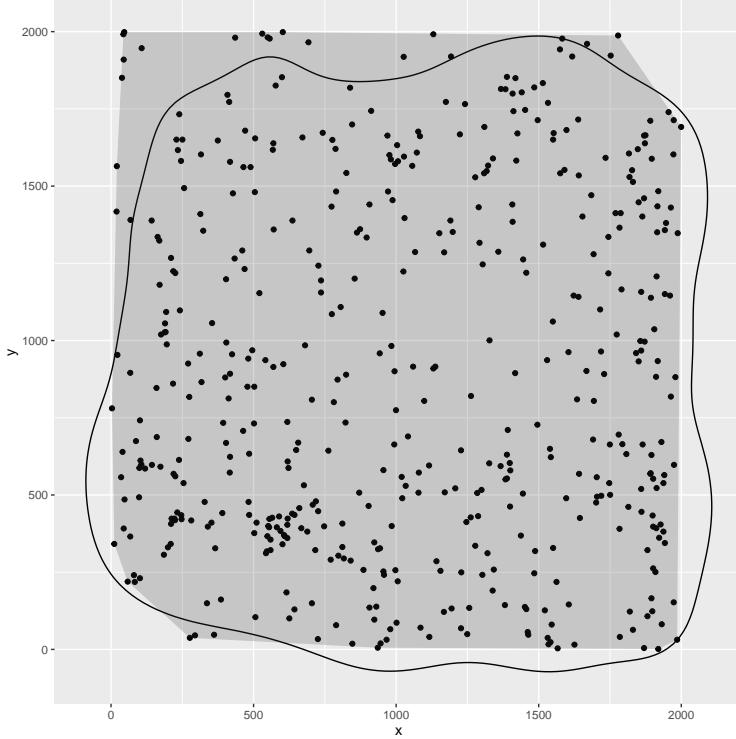
Figure 3.8: Minimum convex polygon (gray area) and 95th percentile isopleth of a kernel density estimate of $f_u(s)$ (black line) for the simulated point pattern data in Chapter 3.

$$L(\mu, \sigma, \beta | s_1, \ldots, s_n) = \prod_{i=1}^{n} \frac{w(x(s_i), \beta)\Phi(s_i, \mu, \sigma^2)}{\int_{g \in G} w(x(g), \beta)\Phi(g, \mu, \sigma^2)dg} \tag{3.33}$$

As with IPP models, the integral in the likelihood must be approximated using Monte Carlo or numerical integration techniques. The same approach has been used to motivate a joint likelihood model for integrating telemetry data and spatial mark-recapture data (Royle et al., 2013b).

In the home-range literature, an important distinction has been made between the distribution of locations an animal uses during a specific observation window, referred to as an *occurrence distribution*, and the asymptotic (or equilibrium) distribution that results from assuming animals move consistently in a range-restricted manner, referred to as the *range distribution* (Fleming et al., 2015, 2016; Horne et al., 2020). These two distributions represent different statistical estimation targets, which has important implications when choosing a home-range estimator and when considering the effect of autocorrelation on home-range estimators (Fleming et al., 2014; Noonan et al., 2019). Autocorrelation is an asset when estimating the occurrence distribution using time-series kriging, a method that uses autocorrelation to better predict "nearby" locations in time (Fleming et al., 2016). On the other hand, autocorrelation may impact effective sample sizes and thus optimal levels of smoothing when using kernel density estimators to estimate the range distribution (Fleming et al., 2015).

This distinction between describing, retrospectively, the distribution of habitat used by an animal and, prospectively, the distribution of habitat that an animal will likely use in the future has received less attention in the SHA literature. Our view is that RSFs capture historic habitat use (analogous to an *occurrence distribution*), which when combined across various sampling instances, can provide insights that may allow us to predict future habitat use in novel environments. In later chapters we will encounter movement-based habitat selection models that can be used to generate predictive distributions of animal space-use in novel

environments. In both cases, we will need to consider the role autocorrelation plays in determining appropriate estimators, including estimators of parameter uncertainty.

3.11.3   Looking Forward: Non-Independence of Telemetry Locations

As alluded to in the previous section, it is important to consider whether the assumption that observed locations are independent is reasonable when modeling home ranges or habitat selection from telemetry data. Strictly speaking, this assumption will almost never be met, particularly with modern-day telemetry studies that allow several locations to be collected on the same day. In particular, observations close in time are likely to be close in space (Aarts et al., 2008; Fleming et al., 2014). Elsewhere Fieberg (2007) argued that a representative sample of locations may be sufficient for consistent (asymptotically unbiased) estimation of utilization distributions that describe habitat use during a fixed observation window (e.g. a breeding season or annual cycle), analogous to estimating the *occurrence distribution*. Similarly, if telemetry observations are collected regularly (or randomly) in time, then the locations may be argued to provide a representative sample of habitat use from a specific observation window. In these cases, it may be helpful to view our estimates of RSF parameters, $\hat{\beta}$, as useful summaries of habitat use for tagged individuals during these fixed time periods. The question then becomes, how should we use these summaries to make inferences about a population or species? The assumption of independence of our locations is clearly problematic, and ignoring statistical dependence will bias our estimates of uncertainty. What should we do about this? We have a few options:

1. Calculate a time to independence between observations and either down-weight the contribution of each observation or subsample from the autocorrelated observations to obtain an independent sample (Swihart & Slade, 1985). Yet, subsampling involves the removal of valuable data, and it may not always be possible to achieve independence through subsampling (Noonan et al., 2020).
2. If we are interested in population-level summaries of habitat use, then we may choose to ignore within-individual autocorrelation when estimating individual-specific coefficients, but use a robust form of SE that treats individuals as independent when describing uncertainty in population-level parameters (e.g. Craiu, Duchesne, & Fortin, 2008; Fieberg et al., 2010).
3. Alternatively, we may be able to address issues related to autocorrelation by choosing a framework that allows us to model habitat selection and animal movement together (Hanks, Hooten, Alldredge, & others, 2015; Signer, Fieberg, & Avgar, 2017, 2019). These approaches can then be combined with hierarchical modeling frameworks that allow for individual-specific coefficients (Muff, Signer, & Fieberg, 2020). We will revisit these ideas in later chapters of the book.

An advantage of the last approach is that it provides a mechanistic link between spatial covariates and animal movements, which can be used to explore space-use patterns in novel environments (Signer et al., 2017; Michelot, Blackwell, & Matthiopoulos, 2019).

3.12   Camera Traps

Camera traps have become a popular and relatively inexpensive tool for monitoring the use of habitats by multiple species (O'Connell, Nichols, & Karanth, 2010; Wearn & Glover-Kapfer, 2019). Data from camera traps have been analyzed in many different ways (for an overview, see Sollmann, 2018), but it is not uncommon to fit models to counts of observations per unit time or to binary (detect/non-detect) summaries created for pre-specified time periods (e.g. days). Both approaches can be formulated in terms of an underlying spatial point-process or spatial-temporal point-process model. This feature has facilitated joint modeling of camera trap and other survey data (e.g. Scotson et al., 2017; Bowler et al., 2019). If individuals can be uniquely identified, then it is possible to fit spatial-mark-recapture models to camera trap data as thinned point processes (Borchers & Marques, 2017); it is also possible to integrate camera trap and telemetry data using a joint likelihood approach (Royle et al., 2013b; Proffitt et al., 2015; Sollmann, Gardner, Belant, Wilton, & Beringer, 2016; Linden, Sirén, & Pekins, 2018). Given the ubiquity of camera trap data, we expect to see further development of approaches to integrating camera trap data with other data types in the future.

One important consideration when fitting models to camera trap data is that cameras sample very small areas defined by the camera's field of view, and therefore, effectively sample points in space. Thus, although camera

trap data are often analyzed using occupancy models, there is no clear connection to a "site" with well-defined spatial extent (Sollmann, 2018). Further, interpretation of estimates of occupancy is challenging since this metric will be influenced by both density of individuals and their average home-range size (Efford & Dawson, 2012; but see Steenweg, Hebblewhite, Whittington, Lukacs, & McKelvey, 2018 for an alternative point of view). Another important feature to consider when analyzing camera trap data (or other mark-recapture data) is whether or not lures or attractants have been used to increase detection rates (see e.g. Garrote et al., 2012; Braczkowski et al., 2016; Iannarilli, Erb, Arnold, & Fieberg, 2017; Maxwell, 2018; Mills, Fattebert, Hunter, & Slotow, 2019). These effects should also be considered when modeling habitat use-data as they may induce clustering rather than a thinning of observed data.

3.13   Software for fitting IPP models

In this chapter, we have demonstrated how the IPP model can be fit to simulated data using the full likelihood, conditional (weighted distribution theory) likelihood, and using infinitely-weighted logistic regression or down-weighted Poisson regression. Here, we briefly mention open-source software available for fitting the IPP model to spatial usage data. Advantages and disadvantages of different options for fitting IPP models are discussed in more detail in Renner et al. (2015), and we encourage interested readers to follow up by reading their paper (along with its associated supplementary material).

- the `spatstat` package in R uses numerical integration to fit a wide range of point-process models, including models that allow for clustering or competition in space (Baddeley & Turner, 2005). An advantage of using the `spatstat` library is that it also has built-in functions for evaluating model assumptions (e.g. point independence) and fit of the model (e.g. using residual plots). These functionalities will be illustrated in a later chapter.

- the `ppmlasso` package in R provides methods for fitting a range of spatial point-process models while also using regularization techniques to constrain model complexity and avoid overfitting.

- `MaxEnt` is a popular software program for fitting species distribution models using a maximum entropy approach (Phillips & Dudik, 2008; Merow et al., 2013). This approach is equivalent to fitting a Poisson process model but using highly flexible predictors to allow for complex species-habitat associations (non-linear responses and interactions). As with `ppmlasso`, a penalization term is used to constrain model complexity. `MaxEnt` also includes nice graphical features that allow one to explore areas in geographical space that require extrapolating outside the range of the data used to fit the model, and to evaluate which predictors are driving the predictions at any location in space. It is possible to fit models using `MaxEnt` using the `dismo` package in R (Hijmans, Phillips, Leathwick, & Elith, 2017).

- `INLA` is a general purpose Bayesian software platform that approximates posterior marginal distributions using integrated nested Laplace approximations rather than using Markov chain Monte Carlo (MCMC) sampling. It has become a popular method for fitting spatial models in a Bayesian framework (Illian, Sørbye, & Rue, 2012; Lindgren & Rue, 2015). The `INLA` R package makes it is possible to fit models using `INLA` via an R interface (for more information about `INLA`, see: http://www.r-inla.org/).

- The `glmmTMB` package in R (Brooks et al., 2017) can be used to fit linear and generalized linear mixed models using maximum likelihood via Template Model Builder (TMB) (Kristensen, Nielsen, Berg, Skaug, & Bell, 2015). TMB uses automatic differentiation to calculate likelihood gradients, which makes it easier to fit complex models. `glmmTMB` can fit IPP models to use-availability data via infinitely-weighted logistic regression or down-weighted Poisson Regression (Renner et al., 2015). Similar to `INLA`, `glmmTMB` can include spatial random effects using a Laplace approximation. The appeal is that it is relatively stable and fast for larger data sets.

3.14   Concluding remarks

In this chapter, we introduced the Inhomogeneous Poisson Point-Process model as a general framework for modeling species distributions. The IPP model allows us to work with and integrate several different data

types (point locations, presence-absence data, gridded counts), while also accounting for observation and sampling biases via a thinning process. Thus, it provides a solid statistical foundation that allows us to link organisms to their habitat. In fact, it can be shown that most methods for modeling species distributions, including MaxEnt and resource-selection functions, are equivalent to fitting an IPP model. Nonetheless, it is important to recognize limitations and inherent challenges with interpreting the output of these models biologically. As discussed in Chapter 1, there are many reasons why the density of a species may not accurately reflect habitat suitability or fitness. Furthermore, there is a tendency among practitioners to approach model building with little thought − i.e. pass whatever remotely-sensed variables are available to MaxEnt, and let it construct an extremely flexible model that allows the "data to speak". There are many issues with this approach. In addition to ignoring sampling biases (Yackulic et al., 2013), machine learning approaches are likely to lead to overfit models that do not transfer well to other locations (Fourcade et al., 2018). As we argued in Chapter 2, it is important to carefully consider biological mechanisms when building process models; classifying predictors as resources, risks, or conditions allows us to anticipate the functional relationship between predictors and fitness and is an important first step to building more realistic process models. This simple step may take us quite far when density is proportional to fitness (i.e. when our null model in Chapter 1 is appropriate). Unfortunately, the null model will often not be appropriate, and thus, future chapters will be devoted to relaxing its assumptions.

References

Aarts, G., Fieberg, J., & Matthiopoulos, J. (2012). Comparative interpretation of count, presence-absence and point methods for species distribution models. *Methods in Ecology and Evolution*, *3*(1), 177–187. doi:10.1111/j.2041-210X.2011.00141.x

Aarts, G., MacKenzie, M., McConnell, B., Fedak, M., & Matthiopoulos, J. (2008). Estimating space-use and habitat preference from wildlife telemetry data. *Ecography*, *31*(1), 140–160. doi:10.1111/j.2007.0906-7590.05236.x

Allee, W. (1951). *Cooperation among Animals, with Human Implications.* (p. 233). New York: Henry Schuman.

Allen, T. F. H., & Hoekstra, T. W. (1992). *Toward a unified ecology* (p. 384). New York: Columbia University Press.

Augustin, N. H., Trenkel, V. M., Wood, S. N., & Lorance, P. (2013). Space-time modelling of blue ling for fisheries stock management. *Environmetrics*, *24*(2), 109–119. doi:10.1002/env.2196

Austin, M. P. (1999). A silent clash of paradigms: Some inconsistencies in community ecology. *Oikos*, *86*(1), 170–178.

Baddeley, A., Berman, M., Fisher, N. I., Hardegen, A., Milne, R. K., Schuhmacher, D., . . . Turner, R. (2010). Spatial logistic regression and change-of-support in Poisson point processes. *Electronic Journal of Statistics*, *4*, 1151–1201. doi:doi:10.1214/10-EJS581

Baddeley, A., & Turner, R. (2005). spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, *12*(6), 1–42. doi:10.18637/jss.v012.i06

Bahn, V., & Mcgill, B. J. (2007). Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography*, *16*(6), 733–742. doi:10.1111/j.1466-8238.2007.00331.x

Bailleul, F., Charrassin, J.-B., Monestiez, P., Roquet, F., Biuw, M., & Guinet, C. (2007). Successful foraging zones of southern elephant seals from the Kerguelen Islands in relation to oceanographic conditions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1487), 2169–2181. doi:10.1098/rstb.2007.2109

Barker, R. J., Schofield, M. R., Link, W. A., & Sauer, J. R. (2018). On the reliability of n-mixture models for count data. *Biometrics*, *74*(1), 369–377.

Bassar, R. D., Marshall, M. C., López-Sepulcre, A., Zandonà, E., Auer, S. K., Travis, J., . . . Reznick, D. N. (2010). Local adaptation in Trinidadian guppies alters ecosystem processes. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(8), 3616–3621. doi:10.1073/pnas.0908023107

Begon, M., Harper, J. L., & Townsend, C. R. (1996). *Ecology: Individuals, Populations and Communities (3rd Edition)* (p. 1068). Cambridge, Massachusetts, USA: Blackwell Science.

Bell, D. M., & Schlaepfer, D. R. (2016). On the dangers of model complexity without ecological justification in species distribution modeling. *Ecological Modelling*, *330*, 50–59. doi:10.1016/j.ecolmodel.2016.03.012

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 of London B: Biological Sciences, *365*(1550), 2245–2254.

Beyer, H. L. H. L., Gurarie, E., Börger, L., Panzacchi, M., Basille, M., Herfindal, I., . . . Matthiopoulos, J. (2016). 'You shall not pass!': Quantifying barrier permeability and proximity avoidance by animals. *Journal of Animal Ecology*, *85*(1), 43–53. doi:10.1111/1365-2656.12275

Biernaskie, J. M., Walker, S. C., & Gegear, R. J. (2009). Bumblebees learn to forage like bayesians. *The American Naturalist*, *174*(3), 413–423. doi:10.1086/603629

Blonder, B. (2018). Hypervolume concepts in niche- and trait-based ecology. *Ecography*, *41*(9), 1441–1455. doi:10.1111/ecog.03187

Borchers, D. L., & Marques, T. A. (2017). From distance sampling to spatial capture–recapture. *AStA Advances in Statistical Analysis*, *101*(4), 475–494.

Bowler, D. E., Nilsen, E. B., Bischof, R., O'Hara, R. B., Yu, T. T., Oo, T., . . . Linnell, J. D. (2019). Integrating data from different survey types for population monitoring of an endangered species: The case of the eld's deer. *Scientific Reports*, *9*(1), 7766.

Boyce, M. S., & McDonald, L. L. (1999). Relating populations to habitats using resource selection functions. *Trends in Ecology & Evolution*, *14*(7), 268–272.

Boyce, M. S., Vernier, P. R., Nielsen, S. E., & Schmiegelow, F. K. A. (2002). Evaluating resource selection functions. *Ecological Modelling*, *157*(2), 281–300. doi:10.1016/S0304-3800(02)00200-4

Börger, L., Matthiopoulos, J., Holdo, R. M., Morales, J. M., Couzin, I., & McCauley, E. (2013). Migration quantified: constructing models and linking them with data. *Animal Migration*, 110–128. doi:10.1093/acprof:oso/9780199568994.003.0008

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*(1855), 20170449. doi:10.1098/rspb.2017.0449

Braczkowski, A. R., Balme, G. A., Dickman, A., Fattebert, J., Johnson, P., Dickerson, T., . . . Hunter, L. (2016). Scent lure effect on camera-trap based leopard density estimates. *PLoS One*, *11*(4), e0151033.

Brasseur, S. M. J. M., van Polanen Petel, T. D., Gerrodette, T., Meesters, E. H. W. G., Reijnders, P. J. H., & Aarts, G. (2015). Rapid recovery of Dutch gray seal colonies fueled by immigration. *Marine Mammal Science*, *31*(2), 405–426. doi:10.1111/mms.12160

Broennimann, O., Fitzpatrick, M. C., Pearman, P. B., Petitpierre, B., Pellissier, L., Yoccoz, N. G., . . . Guisan, A. (2012). Measuring ecological niche overlap from occurrence and spatial environmental data. *Global Ecology and Biogeography*, *21*(4), 481–497. doi:10.1111/j.1466-8238.2011.00698.x

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., . . . Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400.

Brown, J. S. (1988). Patch use as an indicator of habitat preference, predation risk, and competition. *Behavioral Ecology and Sociobiology*, *22*(1), 37–47. doi:10.1007/BF00395696

Brown, J. S., & Kotler, B. P. (2004). Hazardous duty pay and the foraging cost of predation. *Ecology Letters*, *7*(10), 999–1014. doi:10.1111/j.1461-0248.2004.00661.x

Brown, J. S., Laundre, J. W., & Gurung, M. (1999). The Ecology of Fear: Optimal Foraging, Game Theory, and Trophic Interactions. *Journal of Mammalogy*, *80*(2), 385–399. doi:10.2307/1383287

Buckland, S. T., Anderson, D. R., Burnham, K. P., & Laake, J. L. (2005). Distance sampling. *Encyclopedia of Biostatistics*, *2*.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2). Duxbury Pacific Grove, CA.

Caughley, G., & Sinclair, A. R. E. (1994). *Wildlife Ecology and Management* (p. 334). Wiley.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, *9*(2), 129–136.

Chase, J. M., & Leibold, M. A. (2003). *Ecological Niches: Linking Classical and Contemporary Approaches* (Vol. 13, p. 212). Springer Science + Business Media BV, Formerly Kluwer Academic Publishers BV.

Ciuti, S., Northrup, J. M., Muhly, T. B., Simi, S., Musiani, M., Pitt, J. A., & Boyce, M. S. (2012). Effects of humans on behaviour of wildlife exceed those of natural predators in a landscape of fear. *PLoS ONE*, *7*(11), e50611. doi:10.1371/journal.pone.0050611

Coron, C., Calenge, C., Giraud, C., & Julliard, R. (2018). Bayesian estimation of species relative abundances and habitat preferences using opportunistic data. *Environmental and Ecological Statistics*, *25*(1), 71–93.

Craiu, R. V., Duchesne, T., & Fortin, D. (2008). Inference methods for the conditional logistic regression model with longitudinal data. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *50*(1), 97–109.

Decesare, N. J., Hebblewhite, M., Schmiegelow, F., Hervieux, D., Mcdermid, G. J., Neufeld, L., . . . Musiani, M. (2012). Transcending scale dependence in identifying habitat with resource selection functions. *Ecological Applications*, *22*(4), 1068–1083. Retrieved from http://www.jstor.org/stable/23213945

De Goeij, P., & Honkoop, P. J. C. (2002). The effect of immersion time on burying depth of the bivalve Macoma balthica (Tellinidae). *Journal of Sea Research*, *47*(2), 109–119. doi:10.1016/S1385-1101(02)00095-3

Dennis, R. L. H. (2012). What is a Habitat? An Awkward Question. In *A resource-based habitat view for conservation* (pp. 1–8). Chichester, West Sussex, UK: John Wiley & Sons, Ltd. doi:10.1002/9781444315257.ch1

Dorazio, R. M. (2012). Predicting the geographic distribution of a species from presence-only data subject to detection errors. *Biometrics*, *68*(4), 1303–1312.

Dorazio, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, *23*(12), 1472–1484. doi:10.1111/geb.12216

Dormann, C. F., McPherson, J. M., Araújo, Miguel, B., Bivand, R., Bolliger, J., Carl, G., . . . Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, *30*(5), 609–628. doi:10.1111/j.2007.0906-7590.05171.x

Efford, M. G., & Dawson, D. K. (2012). Occupancy in continuous habitat. *Ecosphere*, *3*(4), 1–15.

Elith, J., & Graham, C. H. (2009). Do they? How do they? WHY do they differ? on finding reasons for differing performances of species distribution models. *Ecography*, *32*(1), 66–77. doi:10.1111/j.1600-0587.2008.05505.x

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*(1), 43–57. doi:10.1111/j.1472-4642.2010.00725.x

Fagan, W. F., Lewis, M. A., Auger-Méthé, M., Avgar, T., Benhamou, S., Breed, G., . . . Mueller, T. (2013, October). Spatial memory and animal movement. John Wiley & Sons, Ltd. doi:10.1111/ele.12165

Fieberg, J. (2007). Kernel density estimators of home range: smoothing and the autocorrelation red herring. *Ecology*, *88*(4), 1059–1066.

Fieberg, J. (2012). Estimating population abundance using sightability models: R SightabilityModel package. *Journal of Statistical Software*, *51*(9), 1–20.

Fieberg, J., Alexander, M., Tse, S., & Clair, K. S. (2013). Abundance estimation with sightability data: A Bayesian data augmentation approach. *Methods in Ecology and Evolution*, *4*(9), 854–864.

Fieberg, J., & Börger, L. (2012). Could you please phrase 'home range' as a question? *Journal of Mammalogy*, *93*(4), 890–902.

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 of London - Series B: Biological Sciences*, *365*(1550), 2233–2244. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2894958%7B/&%7Dtool=pmcentrez%7B/&%7Drendertype=abstract

Fieberg, J. R., Shertzer, K. W., Conn, P. B., Noyce, K. V., & Garshelis, D. L. (2010). Integrated population modeling of black bears in Minnesota: implications for monitoring and management. *Plos One*, *5*(8), e12114.

Fieberg, J. R., Vitense, K., & Johnson, D. H. (2020). Resampling-based methods for biologists. *PeerJ*, *8*, e9089. doi:10.7717/peerj.9089

Fithian, W., Elith, J., Hastie, T., & Keith, D. a. (2015). Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, *6*(4), 424–438. doi:10.1111/2041-210X.12242

Fithian, W., & Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *The Annals of Applied Statistics*, *7*(4), 1917. doi:10.1214/13-AOAS667

Fleming, C. H., Calabrese, J. M., Mueller, T., Olson, K. A., Leimgruber, P., & Fagan, W. F. (2014). From fine-scale foraging to home ranges: A semivariance approach to identifying movement modes across spatiotemporal scales. *The American Naturalist*, *183*(5), E154––E167.

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*(5), 1182–1188.

Fleming, C. H., Fagan, W. F., Mueller, T., Olson, K. A., Leimgruber, P., & Calabrese, J. M. (2016). Estimating where and how animals travel: An optimal framework for path reconstruction from autocorrelated tracking data. *Ecology*, *97*(3), 576–582.

Fletcher, R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., & Dorazio, R. M. (2019). A practical guide for combining data to model species distributions. *Ecology*, *100*(6), e02710. doi:10.1002/ecy.2710

Fourcade, Y., Besnard, A. G., & Secondi, J. (2018). Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography*, *27*(2), 245–256. doi:10.1111/geb.12684

Fretwell, S. D., & Lucas, H. L. (1969). On territorial behavior and other factors influencing habitat distribution in birds - I. Theoretical development. *Acta Biotheoretica*, *19*(1), 16–36. doi:10.1007/BF01601953

Gallien, L., Douzet, R., Pratte, S., Zimmermann, N. E., & Thuiller, W. (2012). Invasive species distribution models - how violating the equilibrium assumption can create new insights. *Global Ecology and Biogeography*, *21*(11), 1126–1136. doi:10.1111/j.1466-8238.2012.00768.x

Garrote, G., Gil-Sánchez, J. M., McCain, E. B., Lillo, S. de, Tellería, J. L., & Simón, M. Á. (2012). The effect of attractant lures in camera trapping: A case study of population estimates for the iberian lynx (lynx pardinus). *European Journal of Wildlife Research*, *58*(5), 881–884.

Gelfand, A. E. (2020). Statistical challenges in spatial analysis of plant ecology data. *Spatial Statistics*, *37*, 100418. doi:10.1016/j.spasta.2020.100418

Gelfand, A. E., & Shirota, S. (2019). Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. *Ecological Monographs*, *89*(3), 1–17. doi:10.1002/ecm.1372

Giannini, T. C., Chapman, D. S., Saraiva, A. M., Alves-dos-Santos, I., & Biesmeijer, J. C. (2013). Improving species distribution models using biotic interactions: a case study of parasites, pollinators and plants. *Ecography*, *36*(6), 649–656. doi:10.1111/j.1600-0587.2012.07191.x

Giraud, C., Calenge, C., Coron, C., & Julliard, R. (2016). Capitalizing on opportunistic data for monitoring relative abundances of species. *Biometrics*, *72*(2), 649–658.

Grady, J. M., Maitner, B. S., Winter, A. S., Kaschner, K., Tittensor, D. P., Record, S., . . . Brown, J. H. (2019). Biodiversity patterns: Metabolic asymmetry and the global diversity of marine predators. *Science*, *363*(6425). doi:10.1126/science.aat4220

Grecian, W. J. J., Witt, M. J. J., Attrill, M. J. J., Bearhop, S., Becker, P. H. H., Egevang, C., . . . Votier, S. C. C. (2016). Seabird diversity hotspot linked to ocean productivity in the Canary Current Large Marine Ecosystem. *Biology Letters*, *12*(8). doi:10.1098/rsbl.2016.0024

Grohmann, C., Hartmann, J. N., Kovalev, A., & Gorb, S. N. (2019). Dandelion diaspore dispersal: frictional anisotropy of cypselae of Taraxacum officinale enhances their interlocking with the soil. *Plant and Soil*, *440*(1-2), 399–408. doi:10.1007/s11104-019-04086-x

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, *8*(9), 993–1009. doi:10.1111/j.1461-0248.2005.00792.x

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat suitability and distribution models: with applications in R*. Cambridge University Press.

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*(2-3), 147–186. doi:10.1016/S0304-3800(00)00354-9

Hall, L. S., Krausman, P. R., Morrison, M. L., Hall, L. S., Krausman, P. R., & Morrison, M. L. (1997). The habitat concept and a plea for standard terminology. *Wildlife Society Bulletin*, *25*(1), 173–182.

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

Hastie, T., & Tibshirani, R. (1993). Varying-Coefficient Models. *Journal of the Royal Statistical Society*, *55*(4), 757–796.

Hedley, S. L., & Buckland, S. T. (2004). Spatial models for line transect sampling. *Journal of Agricultural, Biological, and Environmental Statistics*, *9*(2), 181–199. doi:10.1198/1085711043578

Hefley, T. J., & Hooten, M. B. (2016). Hierarchical Species Distribution Models. *Current Landscape Ecology Reports*, *1*(2), 87–97. doi:10.1007/s40823-016-0008-7

Hijmans, R. J., Phillips, S., Leathwick, J., & Elith, J. (2017). *Dismo: Species distribution modeling*. Retrieved from https://CRAN.R-project.org/package=dismo

Hirzel, A. H., Hausser, J., Chessel, D., Perrin, N., & Jul, N. (2002). Ecological-Niche Factor Analysis : How to compute habitat-suitability maps without absence data? *Society*, *83*(7), 2027–2036.

Hirzel, A. H., & Le Lay, G. (2008). Habitat suitability modelling and niche theory. *Journal of Applied Ecology*, *45*(5), 1372–1381. doi:10.1111/j.1365-2664.2008.01524.x

Hodges, J. S., & Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *American Statistician*, *64*(4), 325–334. doi:10.1198/tast.2010.10052

Holbrook, J. D., Olson, L. E., DeCesare, N. J., Hebblewhite, M., Squires, J. R., & Steenweg, R. (2019). Functional responses in habitat selection: clarifying hypotheses and interpretations. *Ecological Applications*, e01852. doi:10.1002/eap.1852

Holling, C. S. (1959). Some characteristics of simple types of predation and parasitism. *The Canadian Entomologist*, *91*(7), 385–398.

Holt, R. D. (2009). Bringing the Hutchinsonian niche into the 21st century: Ecological and evolutionary perspectives. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(Supplement 2), 19659–19665. doi:10.1073/pnas.0905137106

Hooten, M. B., Hanks, E. M., Johnson, D. S., & Alldredge, M. W. (2013). Reconciling resource utilization and resource selection functions. *Journal of Animal Ecology*, *82*(6), 1146–1154.

Hooten, M. B., Hanks, E. M., Johnson, D. S., & Alldredge, M. W. (2014). Temporal variation and scale in movement-based resource selection functions. *Statistical Methodology*, *17*, 82–98. doi:10.1016/j.stamet.2012.12.001

Hooten, M. B., Johnson, D. S., Mcclintock, B. T., & Morales, J. M. (2017). *Animal Movement: Statistical Models for Telemetry Data* (p. 320). CRC Press.

Horne, J. S., Fieberg, J., Börger, L., Rachlow, J. L., Calabrese, J. M., & Fleming, C. H. (2020). Animal home ranges: Concepts, uses, and estimation. In *Population Ecology in Practice* (pp. 315–332). New York, NY: Wiley.

Horne, J. S., Garton, E. O., & Rachlow, J. L. (2008). A synoptic model of animal space use: Simultaneous estimation of home range, habitat selection, and inter/intra-specific relationships. *Ecological Modelling*, *214*(2-4), 338–348.

Hubbell, S. P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)*. Princeton University Press.

Hutchinson, G. (1957). Concluding remarks. In *Cold spring harbor symposia on quantitative biology* (pp. 75–96). doi:10.1201/9781315366746

Iannarilli, F., Erb, J., Arnold, T., & Fieberg, J. (2017). Evaluation of design and analysis of a camera-based multi-species occupancy survey of carnivores in Minnesota. In *Summaries of wildlife research findings 2017* (pp. 176–195). Minnesota Department of Natural Resources (MNDNR).

Illian, J. B., Sørbye, S. H., & Rue, H. (2012). A toolbox for fitting complex spatial point process models using integrated nested laplace approximation (INLA). *The Annals of Applied Statistics*, 1499–1530.

Ims, R. A. (1990). On the Adaptive Value of Reproductive Synchrony as a Predator-Swamping Strategy. *The American Naturalist*, *136*(4), 485–498. doi:10.2307/2462190

Isaac, N. J. B., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., . . . O'Hara, R. B. (2020). Data Integration for Large-Scale Models of Species Distributions. *Trends in Ecology and Evolution*, *35*(1), 56–67. doi:10.1016/j.tree.2019.08.006

Jennrich, R., & Turner, F. (1969). Measurement of non-circular home range. *Journal of Theoretical Biology*, *22*(2), 227–237.

Jesmer, B. R., Merkle, J. A., Goheen, J. R., Aikens, E. O., Beck, J. L., Courtemanch, A. B., . . . Kauffman, M. J. (2018). Is ungulate migration culturally transmitted? Evidence of social learning from translocated animals. *Science*, *361*(6406), 1023–1025. doi:10.1126/science.aat0985

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

Johnson, D. H. (1980). The comparison of usage and availability measurements for evaluating resource preference. *Ecology*, *61*(1), 65–71.

Kamil, A. C., Misthal, R. L., & Stephens, D. W. (1993). Failure of simple optimal foraging models to predict residence time when patch quality is uncertain. *Behavioral Ecology*, *4*(4), 350–363. doi:10.1093/beheco/4.4.350

Kearney, M., & Porter, W. (2009). Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters*, *12*(4), 334–350. doi:10.1111/j.1461-0248.2008.01277.x

Keating, K. A., & Cherry, S. (2004). Use and interpretation of logistic regression in habitat-selection studies. *Journal of Wildlife Management*, *68*(4), 774–789.

Kery, M., & Royle, J. A. (2020). *Applied Hierarchical Modeling in Ecology: Analysis of Distribution, Abundance and Species Richness in R and BUGS: Volume 2: Dynamic and Advanced Models* (p. 890). Academic Press.

Kéry, M., & Royle, J. A. (2015). *Applied hierarchical modeling in ecology: Analysis of distribution, abundance and species richness in r and bugs: Volume 1: Prelude and static models*. Academic Press.

Kie, J. G., Matthiopoulos, J., Fieberg, J., Powell, R. A., Cagnacci, F., Mitchell, M. S., . . . Moorcroft, P. R. (2010). The home-range concept: Are traditional estimators still relevant with modern telemetry technology? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1550), 2221–2231.

Kirk, D. A., Park, A. C., Smith, A. C., Howes, B. J., Prouse, B. K., Kyssa, N. G., . . . Prior, K. A. (2018). Our use, misuse, and abandonment of a concept: Whither habitat? *Ecology and Evolution*, *8*(8), 4197–4208. doi:10.1002/ece3.3812

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R. M., White, M., & Stone, L. (2017). Integrated species distribution models: combining presence-background data and site-occupany data with imperfect detection. *Methods in Ecology and Evolution*, *8*(4), 420–430. doi:10.1111/2041-210X.12738

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. (2015). TMB: Automatic differentiation and laplace approximation. *arXiv Preprint arXiv:1509.00660*.

Křivan, V., Cressman, R., & Schneider, C. (2008). The ideal free distribution: A review and synthesis of the game-theoretic perspective. *Theoretical Population Biology*, *73*, 403–425. doi:10.1016/j.tpb.2007.12.009

Lancia, R. A., Kendall, W. L., Pollock, K. H., & Nichols, J. D. (2005). Estimating the number of animals in wildlife populations. In *Techniques for wildlife investigations and management* (pp. 106–153). Bethesda, MD: Wildlife Society.

Laundré, J. W., Hernández, L., & Ripple, W. J. (2010). *The Landscape of Fear: Ecological Implications of Being Afraid* (Vol. 3, pp. 1–7).

Law, G. R., Feltbower, R. G., Taylor, J. C., Parslow, R. C., Gilthorpe, M. S., Boyle, P., & McKinney, P. A. (2008, August). What do epidemiologists mean by 'population mixing'? Pediatr Blood Cancer. doi:10.1002/pbc.21570

Lele, S. R. (2009). A new method for estimation of resource selection probability function. *Journal of Wildlife Management*, *73*(1), 122–127. doi:10.2193/2007-535

Lele, S. R., & Keim, J. L. (2006). Weighted distributions and estimation of resource selection probability functions. *Ecology*, *87*(12), 3021–3028.

Levin, S. A. (1992). The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. *Ecology*, *73*(6), 1943–1967. doi:10.2307/1941447

Linden, D. W., Sirén, A. P., & Pekins, P. J. (2018). Integrating telemetry data into spatial capture–recapture modifies inferences on multi-scale resource selection. *Ecosphere*, *9*(4), e02203.

Lindgren, F., & Rue, H. (2015). Bayesian Spatial Modelling with R-INLA. *Journal of Statistical Software*, *63*(19), 1–25. doi:10.18637/jss.v063.i19

Link, W. A., Schofield, M. R., Barker, R. J., & Sauer, J. R. (2018). On the robustness of N-mixture models. *Ecology*, *99*(7), 1547–1551.

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L., & Hines, J. E. (2017). *Occupancy estimation and modeling: Inferring patterns and dynamics of species occurrence*. Elsevier.

Manceur, A. M., & Kühn, I. (2014). Inferring model-based probability of occurrence from preferentially sampled data with uncertain absences using expert knowledge. *Methods in Ecology and Evolution*, *5*(8), 739–750.

Manly, B. F. J., Thomas, D., McDonald, T. L., & Erickson, W. P. (2002). *Resource Selection by Animals* (p. 222). Springer Netherlands. doi:10.1007/0-306-48151-0

Marzluff, J. M., Millspaugh, J. J., Hurvitz, P., & Handcock, M. S. (2004). Relating resources to a probabilistic measure of space use: Forest fragments and steller's jays. *Ecology*, *85*(5), 1411–1427.

Matthiopoulos, J. (2003a). Model-supervised kernel smoothing for the estimation of spatial usage. *Oikos*, *102*(2), 367–377. doi:10.1034/j.1600-0706.2003.12528.x

Matthiopoulos, J. (2003b). The use of space by animals as a function of accessibility and preference. *Ecological Modelling*, *159*(2-3), 239–268. doi:10.1016/S0304-3800(02)00293-4

Matthiopoulos, J. (2011). *How to be a Quantitative Ecologist: The 'A to R' of Green Mathematics and Statistics*. doi:10.1002/9781119991595

Matthiopoulos, J., & Aarts, G. (2007). The Spatial analysis of marine mammal abundance. In I. L. Boyd, W. D. Bowen, & S. J. Iverson (Eds.), *Marine mammal ecology and conservation: A handbook of*

techniques (oxford biology) (techniques in ecology & conservation) (pp. 27–33). Oxford University Press. doi:10.1007/978-94-009-5211-9_4

Matthiopoulos, J., Cordes, L., Mackey, B., Thompson, D., Duck, C., Smout, S., . . . Thompson, P. (2014). State-space modelling reveals proximate causes of harbour seal population declines. *Oecologia*, *174*(1), 151–162. doi:10.1007/s00442-013-2764-y

Matthiopoulos, J., Fieberg, J., Aarts, G., Barraquand, F., & Kendall, B. E. (2020). Within reach? Habitat availability as a function of individual mobility and spatial structuring. *The American Naturalist*, *195*(6), 1009–1026. doi:10.1086/708519

Matthiopoulos, J., Fieberg, J., Aarts, G., Beyer, H. L., Morales, J. M., & Haydon, D. T. (2015). Establishing the link between habitat selection and animal population dynamics. *Ecological Monographs*, *85*(3), 413–436. doi:10.1890/14-2244.1

Matthiopoulos, J., Field, C., & MacLeod, R. (2019). Predicting population change from models based on habitat availability and utilization. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1901). doi:10.1098/rspb.2018.2911

Matthiopoulos, J., Harwood, J., & Thomas, L. (2005). Metapopulation consequences of site fidelity for colonially breeding mammals and birds. *Journal of Animal Ecology*, *74*(4), 716–727. doi:10.1111/j.1365-2656.2005.00970.x

Matthiopoulos, J., Hebblewhite, M., Aarts, G., & Fieberg, J. (2011). Generalized functional responses for species distributions. *Ecology*, *92*(3), 583–589. doi:10.1890/10-0751.1

Maxwell, J. (2018). *The allure of lure and its impact on perceived community composition when monitoring tropical mammalian biodiversity* (PhD thesis). University of Delaware.

Mayor, S. J., Schneider, D. C., Schaefer, J. A., & Mahoney, S. P. (2009). Habitat selection at multiple scales. *Écoscience*, *16*(2), 238–247. doi:10.2980/16-2-3238

McClintock, B. T., King, R., Thomas, L., Matthiopoulos, J., McConnell, B. J., & Morales, J. M. (2012). A general modelling framework for animal movement and migration using multistate random walks. *Ecological Monographs*, *82*(3), 1–52. doi:10.1890/11-0326.1

McCrea, R. S., & Morgan, B. J. (2014). *Analysis of capture-recapture data*. Chapman; Hall/CRC.

McInerny, G. J., & Etienne, R. S. (2013). 'Niche' or 'distribution' modelling? A response to Warren. *Trends in Ecology and Evolution*, *28*(4), 191–192. doi:10.1016/j.tree.2013.01.007

McNamara, J. M., & Houston, A. I. (1987). Starvation and predation as factors limiting population size. *Ecology*, *68*(5), 1515–1519. doi:10.2307/1939235

Merkle, J. A., Sawyer, H., Monteith, K. L., Dwinnell, S. P., Fralick, G. L., & Kauffman, M. J. (2019). Spatial memory shapes migration and its benefits: Evidence from a large herbivore. *Ecology Letters*, *22*(11), 1797–1805.

Merow, C., Smith, M. J., Jr, T. C. E., Guisan, A., Mcmahon, S. M., Normand, S., . . . Elith, J. (2014). What do we gain from simplicity versus complexity in species distribution models ?, (July), 1267–1281. doi:10.1111/ecog.00845

Merow, C., Smith, M. J., & Silander, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, *36*(10), 1058–1069.

Michelot, T., Blackwell, P. G. G., & Matthiopoulos, J. (2019). Linking resource selection and step selection models for habitat preferences in animals. *Ecology*, *100*(1), 1–12. doi:10.1002/ecy.2452

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, *10*(1), 22–37. doi:10.1111/2041-210X.13110

Mills, D., Fattebert, J., Hunter, L., & Slotow, R. (2019). Maximising camera trap data: Using attractants to improve detection of elusive species in multi-species surveys. *PloS One*, *14*(5), e0216447.

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

Moilanen, A., Wilson, Kerrie, A., & Possingham, H. P. (2008). *Spatial Conservation Prioritization: Quantitative Methods and Computational Tools* (p. 328). Oxford UP.

Monsarrat, S., Pennino, M. G., Smith, T. D., Reeves, R. R., Meynard, C. N., Kaplan, D. M., & Rodrigues, A. S. L. (2015). Historical summer distribution of the endangered North Atlantic right whale (Eubalaena glacialis): a hypothesis based on environmental preferences of a congeneric species. *Diversity and Distributions*, *21*(8), 925–937. doi:10.1111/ddi.12314

Moorcroft, P. R. (2012). Mechanistic approaches to understanding and predicting mammalian space use: recent advances, future directions. *Journal of Mammalogy*, *93*(4), 903–916. doi:10.1644/1

Moorcroft, P. R., & Barnett, A. (2008). Mechanistic home range models and resource selection analysis: a reconciliation and unification. *Ecology*, *89*(4), 1112–1119.

Moorcroft, P. R., Lewis, M. A., & Crabtree, R. L. (1999). Home Range Analysis Using a Mechanistic Home Range Model, *80*(5), 1656–1665.

Moorcroft, P. R., Lewis, M. A., & Crabtree, R. L. (2006). Mechanistic home range models capture spatial patterns and dynamics of coyote territories in Yellowstone. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1594), 1651–1659. doi:10.1098/rspb.2005.3439

Morris, L. R., Proffitt, K. M., & Blackburn, J. K. (2016). Mapping resource selection functions in wildlife studies: Concerns and recommendations. *Applied Geography*, *76*, 173–183. doi:10.1016/j.apgeog.2016.09.025

Muff, S., Signer, J., & Fieberg, J. (2020). Accounting for individual-specific variation in habitat-selection studies: Efficient estimation of mixed-effects models using bayesian or frequentist computation. *Journal of Animal Ecology*, *89*(1), 80–92.

Murdoch, W. W., Briggs, C. J., & Nisbet, R. M. (2003). *Consumer-resource dynamics* (p. 464). Princeton University Press.

Murray, J. D. (2013). *Mathematical Biology* (p. 770). Springer.

Mysterud, A., & Ims, R. A. (1998). Functional Responses in Habitat Use : Availability Influences Relative Use in Trade-Off Situations. *Ecology*, *79*(4), 1435–1441.

Møller, J., Syversveen, A. R., & Waagepetersen, R. P. (1998). Log gaussian cox processes. *Scandinavian Journal of Statistics*, *25*(3), 451–482.

Newman, K. B., Buckland, S. T., Morgan, B. J., King, R., Borchers, D. L., Cole, D. J., . . . Thomas, L. (2014). *Modelling Population Dynamics: Model Formulation, Fitting and Assessment Using State-Space Methods* (p. 228). Springer.

Nichols, J. D., Hines, J. E., Sauer, J. R., Fallon, F. W., Fallon, J. E., & Heglund, P. J. (2000). A double-observer approach for estimating detection probability and abundance from point counts. *The Auk*, *117*(2), 393–408.

Nisbet, R. M., & Gurney, W. S. (2004). *Modelling fluctuating populations* (p. 396). Blackburn.

Noonan, M. J., Fleming, C. H., Tucker, M. A., Kays, R., Harrison, A.-L., Crofoot, M. C., . . . others. (2020). Effects of body size on estimation of mammalian area requirements. *Conservation Biology*, *34*(4), 1017–1028.

Noonan, M. J., Tucker, M. A., Fleming, C. H., Akre, T. S., Alberts, S. C., Ali, A. H., . . . others. (2019). A comprehensive analysis of autocorrelation and bias in home range estimation. *Ecological Monographs*, *89*(2), e01344.

Nur, N. (1987). Population Growth Rate and the Measurement of Fitness: A Critical Reflection. *Oikos*, *48*(3), 338. doi:10.2307/3565523

O'Connell, A. F., Nichols, J. D., & Karanth, K. U. (2010). *Camera traps in animal ecology: Methods and analyses*. Springer Science & Business Media.

Odling-Smee, F. J., Laland, K. N., & Feldman, M. W. (2013). *Niche construction: the neglected process in evolution*. Princeton: Princeton University Press.

Oliveira-Santos, L. G. R., Forester, J. D., Piovezan, U., Tomas, W. M., & Fernandez, F. A. (2016). Incorporating animal spatial memory in step selection functions. *Journal of Animal Ecology*, *85*(2), 516–524.

Ollason, J. G. (1980). Learning to Forage-Optimally? *Theoretical Population Biology*, *18*, 44–56.

Oppel, S., Meirinho, A., Ramírez, I., Gardner, B., O'Connell, A. F., Miller, P. I., & Louzao, M. (2012). Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. *Biological Conservation*, *156*, 94–104. doi:10.1016/j.biocon.2011.11.013

Otto, S. P., & Day, T. (2011). *A Biologist's Guide to Mathematical Modeling in Ecology and Evolution* (p. 744). Princeton UP.

Ovaskainen, O., & Abrego, N. (2020). *Joint Species Distribution Modelling: With Applications in R (Ecology, Biodiversity and Conservation)* (p. 388). Cambridge University Press.

Ovaskainen, O., Hottola, J., & Shtonen, J. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, *91*(9), 2514–2521. doi:10.1890/10-0173.1

Pacifici, K., Reich, B. J., Miller, D. A., Gardner, B., Stauffer, G., Singh, S., . . . Collazo, J. A. (2017). Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology*, *98*(3), 840–850.

Pacifici, K., Reich, B. J., Miller, D. A., & Pease, B. S. (2019). Resolving misaligned spatial data with integrated species distribution models. *Ecology*, e02709.

Palmer, M. S., Fieberg, J., Swanson, A., Kosmala, M., & Packer, C. (2017, November). A 'dynamic' landscape of fear: prey responses to spatiotemporal variations in predation risk across the lunar cycle. Blackwell Publishing Ltd. doi:10.1111/ele.12832

Paton, S. R., & Matthiopoulos, J. (2018). Defining the scale of habitat availability for models of habitat selection. *Ecology*, *97*(July), 1113–1122. doi:10.1002/ecy.2446

Pearce, J. L., & Boyce, M. S. (2006). Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, *43*(3), 405–412. doi:10.1111/j.1365-2664.2005.01112.x

Pearl, R., & Reed, L. J. (1920). On the Rate of Growth of the Population of the United States Since 1790 and its Mathematical Representation. *Proceedings of the National Academy of Sciencesatural Academy of Sciences*, *6*, 275–288. doi:10.1073/pnas.6.6.275

Pearman, P. B., Guisan, A., Broennimann, O., & Randin, C. F. (2008). Niche dynamics in space and time. *Trends in Ecology and Evolution*, *23*(3), 149–158. doi:10.1016/j.tree.2007.11.005

Pearson, R. G., & Dawson, T. P. (2003). Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography*, *12*(5), 361–371. Retrieved from http://dx.doi.org/10.1046/j.1466-822X.2003.00042.x

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martinez-Meyer, E., Nakamura, M., . . . Araújo, M. B. (2011). *Ecological niches and geographic distributions* (Vol. 56, p. 314). Princeton University Press.

Phillips, S. J., & Dudik, M. (2008). Modeling of species distributions with {Maxent}: new extensions and a comprehensive evaluation. *Ecography*, *31*(2), 161–175.

Proffitt, K. M., Goldberg, J. F., Hebblewhite, M., Russell, R., Jimenez, B., Robinson, H. S., . . . Schwartz, M. K. (2015). Integrating resource selection into spatial capture-recapture models for large carnivores. *Ecosphere*, *6*(11), 1–15.

Pulliam, H. R. (2000). On the relationship between niche and distribution. *Ecology Letters*, *3*(4), 349–361. doi:10.1046/j.1461-0248.2000.00143.x

Qiao, H., Feng, X., Escobar, L. E., Peterson, A. T., Soberón, J., Zhu, G., & Papeş, M. (2019). An evaluation of transferability of ecological niche models. *Ecography*, *42*(3), 521–534. doi:10.1111/ecog.03986

Randin, C. F., Dirnböck, T., Dullinger, S., Zimmermann, N. E., Zappa, M., & Guisan, A. (2006). Are niche-based species distribution models transferable in space? *Journal of Biogeography*, *33*(10), 1689–1703. doi:10.1111/j.1365-2699.2006.01466.x

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., . . . Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, *6*(4), 366–379.

Renner, I. W., Louvrier, J., & Gimenez, O. (2019). Combining multiple data sources in species distribution models while accounting for spatial dependence and overfitting with combined penalised likelihood maximisation. *bioRxiv*, 615583.

Renner, I. W., & Warton, D. I. (2013). Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics*, *69*(1), 274–281.

Ribeiro Jr, P. J., & Diggle, P. J. (2018). *GeoR: Analysis of geostatistical data.* Retrieved from https://CRAN.R-project.org/package=geoR

Riotte-Lambert, L., & Matthiopoulos, J. (2019). Communal and efficient movement routines can develop spontaneously through public information use. *Behavioral Ecology*, *30*(2), 408–416. doi:10.1093/beheco/ary180

Robertson; M. P., Caithness, N., & Villet, M. H. (2012). A PCA-based modelling for predicting technique environmental suitability for organisms from presence records. *Biodiversity Research*, *7*(1), 15–27.

Robinson, A. L. M., Elith, J., Hobday, A. J., Pearson, R. G., Kendall, B. E., & Richardson, A. J. (2017). Pushing the limits in marine species distribution modelling: lessons from the land present challenges and opportunities. *Global Ecology and Biogeography Biogeography*, *20*(6), 789–802.

Royle, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, *60*(1), 108–115.

Royle, J. A., Chandler, R. B., Sollmann, R., & Gardner, B. (2013a). *Spatial capture-recapture.* Academic Press.

Royle, J. A., Chandler, R. B., Sun, C. C., & Fuller, A. K. (2013b). Integrating resource selection information with spatial capture–recapture. *Methods in Ecology and Evolution*, *4*(6), 520–530.

Russell, D. J. F., Hastie, G. D., Thompson, D., Janik, V. M., Hammond, P. S., Scott-Hayward, L. A. S., . . . McConnell, B. J. (2016). Avoidance of wind farms by harbour seals is limited to pile driving activities. *Journal of Applied Ecology*, *53*(6). doi:10.1111/1365-2664.12678

Sarker, S. K., Reeve, R., Thompson, J., Paul, N. K., & Matthiopoulos, J. (2016). Are we failing to protect threatened mangroves in the Sundarbans world heritage ecosystem? *Scientific Reports*, *6*. doi:10.1038/srep21234

Schank, C. J., Cove, M. V., Kelly, M. J., Mendoza, E., O'Farrill, G., Reyna-Hurtado, R., . . . others. (2017). Using a novel model approach to assess the distribution and conservation status of the endangered baird's tapir. *Diversity and Distributions*, *23*(12), 1459–1471.

Scheele, B. C., Foster, C. N., Banks, S. C., & Lindenmayer, D. B. (2017). Niche Contractions in Declining Species: Mechanisms and Consequences. *Trends in Ecology and Evolution*, *32*(5), 346–355. doi:10.1016/j.tree.2017.02.013

Scotson, L., Fredriksson, G., Ngoprasert, D., Wong, W.-M., & Fieberg, J. (2017). Projecting range-wide sun bear population trends using tree cover and camera-trap bycatch data. *PloS One*, *12*(9), e0185336.

Seale, M., Zhdanov, O., Cummins, C., Kroll, E., Blatt, M., Busse, A., . . . Nakayama, N. (2019). Moisture-Dependent Morphing Tunes the Dispersal of Dandelion Diaspores. *SSRN Electronic Journal.* doi:10.2139/ssrn.3334428

Seidler, T. G., & Plotkin, J. B. (2006). Seed dispersal and spatial pattern in tropical trees. *PLoS Biol*, *4*(11), e344.

Sequeira, A. M. M., Bouchet, P. J., Yates, K. L., Mengersen, K., & Caley, M. J. (2018). Transferring biodiversity models for conservation: Opportunities and challenges. *Methods in Ecology and Evolution*, *9*(5), 1250–1264. doi:10.1111/2041-210X.12998

Sicacha-Parada, J., Steinsland, I., Cretois, B., & Borgelt, J. (2020). Accounting for spatial varying sampling effort due to accessibility in Citizen Science data: A case study of moose in Norway. *Spatial Statistics*, 100446. doi:10.1016/j.spasta.2020.100446

Signer, J., Fieberg, J., & Avgar, T. (2017). Estimating utilization distributions from fitted step-selection functions. *Ecosphere*, *8*(4). doi:10.1002/ecs2.1771

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*, (July 2018), 880–890. doi:10.1002/ece3.4823

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis* (p. 175). CRC Press.

Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J., & O'Hara, R. B. (2020). Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography*, *43*, 1413–1422.

Soberón, J. (2007). Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters*, *10*(12), 1115–1123. doi:10.1111/j.1461-0248.2007.01107.x

Sollmann, R. (2018). A gentle introduction to camera-trap data analysis. *African Journal of Ecology*, *56*(4), 740–749.

Sollmann, R., Gardner, B., Belant, J. L., Wilton, C. M., & Beringer, J. (2016). Habitat associations in a recolonizing, low-density black bear population. *Ecosphere*, *7*(8).

Steenweg, R., Hebblewhite, M., Whittington, J., Lukacs, P., & McKelvey, K. (2018). Sampling scales define occupancy and underlying occupancy–abundance relationships in animals. *Ecology*, *99*(1), 172–183.

Steinhorst, R. K., & Samuel, M. D. (1989). Sightability adjustment methods for aerial surveys of wildlife populations. *Biometrics*, 415–425.

Stephens, P. A., & Sutherland, W. J. (1999). What is the Allee effect? *Oikos*, *87*(1), 185–190. doi:10.2307/3547011

Street, G. M., Vander Vennen, L. M., Avgar, T., Mosser, A., Anderson, M. L., Rodgers, A. R., & Fryxell, J. M. (2015). Habitat selection following recent disturbance: model transferability with implications for management and conservation of moose (\emph{Alces alces}). *Canadian Journal of Zoology*, *93*(999), 813–821.

Swihart, R. K., & Slade, N. A. (1985). Testing for independence of observations in animal movements. *Ecology*, *66*(4), 1176–1184.

Šmejkal, M., Souza, A. T., Blabolil, P., Bartoň, D., Sajdlová, Z., Vejřík, L., & Kubečka, J. (2018). Nocturnal spawning as a way to avoid egg exposure to diurnal predators. *Scientific Reports*, *8*(1), 1–7. doi:10.1038/s41598-018-33615-4

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

Tilman, D. (1982). Resource competition and community structure. *Monographs in Population Biology*, *17*, 1–296.

Tuanmu, M.-N., Viña, A., Roloff, G. J., Liu, W., Ouyang, Z., Zhang, H., & Liu, J. (2011). Temporal transferability of wildlife habitat models: implications for habitat monitoring. *Journal of Biogeography*, *38*(8), 1510–1523. doi:10.1111/j

Van Gils, J. A., Edelaar, P., Escudero, G., & Piersma, T. (2004, January). Carrying capacity models should not use fixed prey density thresholds: A plea for using more tools of behavioural ecology. doi:10.1111/j.0030-1299.2003.12214.x

Van Leeuwen, A., De Roos, A. M., & Persson, L. (2008). How cod shapes its world. *Journal of Sea Research*, *60*(1-2), 89–104. doi:10.1016/j.seares.2008.02.008

Van Winkle, W. (1975). Comparison of several probabilistic home-range models. *The Journal of Wildlife Management*, 118–123.

Verhulst, P. (1845). Resherches mathematiques sur la loi d'accroissement de la population. *Nouveaux Memoires de L'Academie Royale Des Sciences*, *18*, 1–41.

Wang, Y., & Stone, L. (2019). Understanding the connections between species distribution models for presence-background data. *Theoretical Ecology*, *12*(1), 73–88.

Warren, D. L. (2012). In defense of 'niche modeling'. *Trends in Ecology and Evolution*, *27*(9), 497–500. doi:10.1016/j.tree.2012.03.010

Warren, D. L. (2013). 'Niche modeling': That uncomfortable sensation means it's working. A reply to McInerny and Etienne. *Trends in Ecology and Evolution*, *28*(4), 193–194. doi:10.1016/j.tree.2013.02.003

Warton, D. I., Renner, I. W., & Ramp, D. (2013). Model-based control of observer bias for the analysis of presence-only data in ecology. *PloS One*, *8*(11), e79168.

Warton, D. I., Shepherd, L. C., & Others. (2010). Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *The Annals of Applied Statistics*, *4*(3), 1383–1402.

Watts, K., Whytock, R. C., Park, K. J., Fuentes-Montemayor, E., Macgregor, N. A., Duffield, S., & McGowan, P. J. K. (2020). Ecological time lags and the journey towards conservation success. *Nature Ecology and Evolution*, *4*(3), 304–311. doi:10.1038/s41559-019-1087-8

Wearn, O. R., & Glover-Kapfer, P. (2019). Snap happy: Camera traps are an effective sampling tool when compared with alternative methods. *Royal Society Open Science*, *6*(3), 181748.

Weddell, B. J. (2002). *Conserving Living Natural Resources.* Cambridge University Press. doi:10.1017/cbo9780511804298

Welsh, A. H., Lindenmayer, D. B., & Donnelly, C. F. (2013). Fitting and interpreting occupancy models. *PloS One*, *8*(1), e52015.

Wenger, S. J., & Olden, J. D. (2012). Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, *3*(2), 260–267. doi:10.1111/j.2041-210x.2011.00170.x

Wilson, R. P., Börger, L., Holton, M. D., Scantlebury, D. M., Gómez-Laich, A., Quintana, F., . . . Shepard, E. L. C. (2020). Estimates for energy expenditure in free-living animals using acceleration proxies: A reappraisal. *Journal of Animal Ecology*, *89*(1), 161–172. doi:10.1111/1365-2656.13040

Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Damgaard, C. F., Dormann, C. F., . . . Svenning, J.-c. (2013). The role of biotic interactions in shaping distributions and realised assemblages of species : implications for species distribution modelling, *88*, 15–30. doi:10.1111/j.1469-185X.2012.00235.x

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R.* Chapman & Hall/CRC.

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

Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Campbell Grant, E. H., & Veran, S. (2013). Presence-only modelling using maxent: When can we trust the inferences? *Methods in Ecology and Evolution*, *4*(3), 236–243.

Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., . . . Sequeira, A. M. M. (2018). Outstanding Challenges in the Transferability of Ecological Models. *Trends in Ecology and Evolution*, *33*(10), 790–802. doi:10.1016/j.tree.2018.08.001

Yen, J. D. L., Tonkin, Z., Lyon, J., Koster, W., Kitchingman, A., Stamation, K., & Vesk, P. A. (2019). Integrating multiple data types to connect ecological theory and data among levels. *Frontiers in Ecology and Evolution*, *7*, 95. doi:10.3389/fevo.2019.00095

Zipkin, E. F., & Saunders, S. P. (2018). Synthesizing multiple data types for biological conservation using integrated population models. *Biological Conservation*, *217*, 240–250. doi:10.1016/j.biocon.2017.10.017