Applying a Multiverse to Habitat Analyses

Benjamin Michael Marshall*[1] and Alexander Bradley Duthie[1]**

[1]**Biological and Environmental Sciences, Faculty of Natural Sciences, University of Stirling, Stirling, FK9 4LA, Scotland, UK**

*benjaminmichaelmarshall@gmail.com
**alexander.duthie@stir.ac.uk

Abstract

—

Keywords

Movement ecology, simulation

1 Multiverse Introduction

1.1 Research flexibilty and the justification for a multiverse approach

Researchers are intrinsically part of the research process (Levins & Lewontin, 1985; Tang-Martínez, 2020), and our expectations can shape the answers we find (Holman et al., 2015). While we can strive to conduct research objectively, there are frequent moments during research that require judgement calls (Steegen et al., 2016). Such calls, choices, or decisions occur throughout the research process, encompassing everything from study design (e.g., sample size, sampling intensity, sample stratification) to analysis (e.g., Bayesian or frequentist, ex/inclusion of outliers). We draw on our own experience and the input of peers to try and ensure the best choices are made to produce robust and reliable results, but we also contend with data, expertise, and interpretability constraints (Liu, Althoff & Heer, 2020).

Researchers are also influenced by the incentive system around them (Anderson et al., 2007). We cannot undertake research in a vacuum; we often require institutions, funders, and scientific journals to produce and share research. These bodies can influence what research is conducted; they are in a position to incentivise or disincentivise research of certain topics or methodologies (Fanelli, 2010a; Ware & Munafò, 2015; Smaldino & McElreath, 2016). The use of impact factor (and other similar metrics) is an example of citations being used as a measure of quality or research worth. But when examined closer, the impact factor appears detached from the robustness or reliability of the research (Brembs, 2018). This decrease in robustness can be seen in the increases in effect sizes inflation, p-value misreporting, among a number of other measures of quality (Brembs, 2018). Similarly, novelty has been fetishised by journals to the detriment of replications studies (Vinkers, Tijdink & Otte, 2015; Forstmeier, Wagenmakers & Parker, 2017; Brembs, 2019), despite widespread agreement on the importance of replications (Fraser et al., 2020). There is a bias towards positive, statistically significant results (Jennions & Møller, 2002; Cassey et al., 2004). Due to the nature of frequentist statistical significance, a prioritisation of significant results can elevate underpowered studies and boost false positive rates (Forstmeier & Schielzeth, 2011; Albers, 2019).

Unfortunately, there is evidence that the system of incentives trickle down to impact the judgement calls and decisions of researchers while undertaking research and when publishing those results. The more detrimental of these decisions have been termed questionable research practices (Fraser et al., 2018; Bishop, 2019). They chiefly come in three forms: HARKing, cherry-picking, and p-hacking. Hypothesising After Results Known (HARKing) is where the research can present the results as a confirmatory result, despite originally there being no or contrary hypothesis. HARKing can sometimes be further enabled and rationalised by hindsight bias, where unexpected results are perceived as more likely once they have been observed (Gelman & Loken, 2013; Forstmeier, Wagenmakers & Parker, 2017). Cherry-picking is the removal or non-reporting of data points or (co-)variables, that did not yield significant results. P-hacking is the repeated use of statistical tests, with different settings, to achieve statistical significance. Arguably, the existence of p-hacking is enabled by an over-reliance on p-value thresholds, rather than flexible p-value thresholds that are predefined based on effect size of interest, sample size, and desired accuracy of estimation (Lakens et al., 2018). Questionable research practices can be viewed as methods to achieve a neat, statistically significant, and publishable narrative (O'Boyle, Banks & Gonzalez-Mulé, 2014); in the worse cases, narratives could be prioritised over transparently reported results.

There is a fear that questionable research practices, and the broader incentives they are connected to, are responsible for the replicability crisis (often referred to as the "reproducibility crisis"). Across many disciplines, there are examples of replication studies being unable to replicate prior research (Freedman, Cockburn & Simcoe, 2015; Open Science Collaboration, 2015; Kelly, 2019). Often these replication efforts are conducted with larger sample sizes, or rely on the consolidation of many independent studies (often in the form of meta-analyses). The implication is not that the original studies were necessarily flawed; but – in the absence of questionable research practices – sufficient variation exists in the study subjects to obscure a consistent effect [i.e., variation beyond variation stemming from sampling; Simonsohn (2015)].

However, variation can also stem from analysis flexibility (i.e., the presence of many ways to analyse the same data to answer the same question). This flexibility helps enable questionable research practices (Fraser et al., 2018), and is potentially steered by publication bias (Jennions & Møller, 2002; Cassey et al., 2004) if results that produce more publishable results are prioritised/rewarded over less exciting but robust results. Given the prominence of questionable research practices and publication bias, the inconsistencies between initial and replication studies warrant investigation (especially when analysis flexibility is also implicated in potentially flawed replications Bryan, Yeager & O'Brien, 2019). It is key to note that analysis flexibility can still lead to variable results in the absence of any undesirable incentives simply as the result of researchers considering different approaches of differing validities for a given dataset (Gelman & Loken, 2013).

Scientific progress requires building upon past results, and therefore requires confidence in past results. Issues arise when subsequent research is based upon weak foundations –i.e., studies with a limited capacity to be replicated because of questionable practices or over-generalisation. Early significant results can dictate the direction of research and grow resistant to later contradictory results (Barto & Rillig, 2012); therefore, early diagnoses of overly confident results or previously unknown sources of variation becomes a priority.

In medical fields, a lack of replicability comes with direct monetary and well-being costs (Freedman, Cockburn & Simcoe, 2015). Like the medical field, ecological studies can come with well-being costs to the study subjects [e.g., direct surgery/marking of the animal (Reinert & Cundall, 1982; Winne et al., 2006)], as well as impacts on stakeholders stemming from management decisions. There are fears that the lack of replicability will feed distrust of science more generally (Anvari & Lakens, 2018). Therefore, maximising replicability in ecology is key to minimising research waste (Grainger et al., 2019) and the negative impacts on systems and subjects studied.

The impacts on the study subjects, paired with the often high monetary costs of ecological studies (particularly bio-logging where animals may undergo surgery, Weaver, Westphal & Taylor, 2021) means that replications can be more difficult to justify. When paired with fact that ecological systems are complex and in constant flux –often frustrating perfect replications due to changes in space and time (Nakagawa & Parker, 2015; Schnitzer & Carson, 2016) –we are left with a distinct lack of direct replications in ecology (Kelly, 2019).

The low prevalence of replications in ecology make it difficult to assess the overall irreplicabilty situation in ecology (Kelly, 2019); but there are several examples that suggest irreplicabilty is something ecologists should be wary of (Wang et al., 2018; Sánchez-Tójar et al., 2018; Roche et al., 2020; Clark et al., 2020). The potential for irreplicabilty is further supported by evidence of positive publication bias (Fanelli, 2010b, 2012), and links between smaller sample sizes and inflated effect sizes (Lemoine et al., 2016).

In the absence of direct replications, ecology is often left to assess replicability via conceptual replications (Fraser et al., 2020) or efforts broadly referred to as quasi-replications (Palmer, 2000). Replications range in intensity. Direct (or exact) replications are attempts to replicate a tightly defined concept/hypothesis while duplicating of all characteristics of the original study. Partial replications are a step looser, where the concept/hypothesis tested is less clearly defined (e.g., applicable to a broader scale) but efforts are made to repeat the same methodology. The most general category are conceptual replications, where the subject and method of study varies from the original study, but the replication targets a the same concept/hypothesis (Nakagawa & Parker, 2015; Kelly, 2019). Both partial and conceptual can be classed as quasi-replications if the concept and scale is broadly defined (Nakagawa & Parker, 2015).

Conceptual replications are extremely valuable, but rely on our ability to compare replication efforts to previous findings. An important aspect of those comparisons is accounting for factors differing between the studies that are not salient to the effect of interest (Forstmeier, Wagenmakers & Parker, 2017); e.g., those linked to sampling differences (Simonsohn, 2015). An example of sampling differences leading to differences in final results can be seen in the case of reptile space use. Silva et al. (2020) showed how frequently a reptile was located by a researcher interacted with the space-use estimation method, leading to large differences in area estimates even when using the same estimation method. What is revealing is not only how the choices during analysis (e.g., choice of area estimation method) impacted results, but how the error introduced by those choices changed depending on the sampling. It presents a scenario where the *correct* choice was dependent on preceding decisions when designing the study; therefore, highlighting the need to explore the impacts of multiple decisions simultaneously.

As seen in the reptile space use example, the choices made by the researcher [researcher degrees of freedom; Simmons, Nelson & Simonsohn (2011)] is a key source of variation among studies. It would be advantageous to understand which choices have a significant impact and whether we can account for differences in choice during comparisons. An understanding of choice could better guide decisions during a study and potentially be used to gauge the robustness of a given dataset in answering a given question.

Research degrees of freedom [or flexibility in analysis; Forstmeier, Wagenmakers & Parker (2017)] have been elegantly demonstrated by a number of "many analysts" studies (e.g., Silberzahn et al., 2018; Huntington-Klein et al., 2021). In these studies, a number of researchers, or research groups, are tasked with answering the same question. Naturally each participant takes a slightly different approach, both in how the question is interpreted (Auspurg & Brüderl, 2021), and the analysis approach chosen (Gelman & Loken, 2013; Bastiaansen et al., 2020), resulting is different final results. The variation in final results can be considered originating from six sources of uncertainty/variation (Hoffmann et al., 2021): measurement (randomness from the act of measuring), data preprocessing (decisions on data inclusion/exclusion and transforming), parameter (decisions on which parameters used as covariates/predictors), model (decisions on model structure and specification), method (decisions on method choice and parameterisation), and sampling (randomness as a result of sampling a wider population). Several sources of variation (data preprocessing, parameter, model, and method) are likely to be particularly key to defining researcher degrees of freedom post data collection. In some cases, the cause behind the variability in results is hard to diagnose (Breznau et al., 2021), or will be less likely to be questioned because of the agreement with existing theory (Gelman & Loken, 2013). There are examples where the variation in results is sufficient to change the final conclusions (Salis, Lena & Lengagne, 2021), and others where it alters the strength of an estimated effect (Desbureaux, 2021). The importance of the effect size variation is context specific, i.e., how variation relates to the overall effect size, and can impact results pertaining to real-world scenarios (Desbureaux, 2021).

1.2 Multiverse analysis

A rising approach to address the unknown impacts of undisclosed researcher degrees of freedom is to fully explore all plausible or reasonable analysis choices open to researchers – to explore a multiverse of design choices (Steegen et al., 2016). This multiverse analysis – closely linked to vibration of effects (Patel, Burford & Ioannidis, 2015), multi-model analysis (Young & Holsteen, 2017), and specification curve analysis (Simonsohn, Simmons & Nelson, 2020)– has the potential to demonstrate and quantify the variation stemming from researcher's analyses choices (Rijnhart et al., 2021). Choices can include everything from from sample sizes and splits (e.g., Webb & Demeyere, 2021) to measurement and summary statistics (e.g., Parsons, 2020), but crucially should only include options that are reasonable (Simonsohn, Simmons & Nelson, 2020; Del Giudice & Gangestad, 2021). What counts as reasonable is not necessarily simple, and inclusion of irrelevant choices can easily mask important choices because of the multiplicative nature of a branching path network (Del Giudice & Gangestad, 2021) (Fig. 1). Construction of a multiverse requires justification of which decisions are treated as variable, and why there is not an *a priori* and defensible single solution (Del Giudice & Gangestad, 2021). A multiverse populated with well-justified decisions allows the exploration of which choices inflate variation between analysis universes, while also offering insights into how to deflate variation [e.g., refinement of initial study design, the removal of ambiguities like tightening categories definitions; Steegen et al. (2016)].

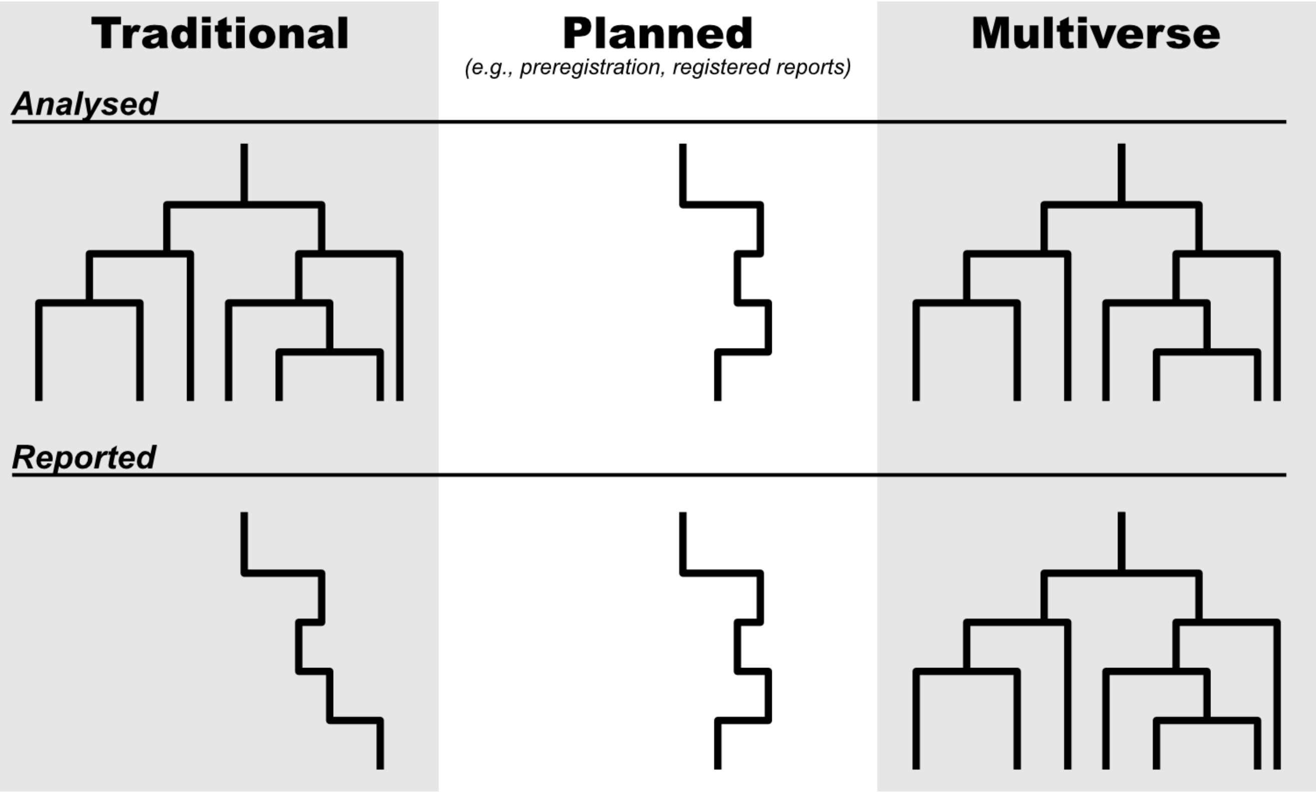

Figure 1. Diagram showing how multiverse analysis differs from other approaches. Each branch node represents a choice made during aysnalis

Ecological systems are complex to study and frustrate replication efforts (Nakagawa & Parker, 2015; Schnitzer & Carson, 2016), and in the case of movement ecology, the data analysed (i.e., derived data, such as step length, speed, and turn angle derived from timestamped coordinate data) require multiple stages of preprocessing. Therefore, multiverse analysis is an avenue to explore causes of variation between studies without the additional costs of practical studies, while also being capable of exploring data processing decisions that may not have immediately apparent impacts on final results.

If the data entropy (the process in which as data ages the chances of irreversible loss increase, Vines et al., 2014) and resistance to data sharing (Miyakawa, 2020; Tedersoo et al., 2021) can be overcome, we will be able to retroactively explore the impact of researcher degrees of freedom on ecological studies (Rijnhart et al., 2021). Such retroactive assessment is an attractive option when other methods to explore false positive rates (Hoffmann et al., 2021), such as preregistration and registered reports (Kaplan & Irvin, 2015; Scheel, Schijen & Lakens, 2021), will require more time to yield results. Ideally we can use multiverse analysis with preregistrations to boost transparency surrounding the inclusion of decisions and the rationale behind others exclusions (Dragicevic et al., 2019; Simonsohn, Simmons & Nelson, 2020). Given the success of meta-analyses to overcome short-comings in the publication record [e.g., p-hacking; Head et al. (2015)], multiverse analysis may aid the direction of future research efforts by providing a means of meeting calls to replicate results before collecting more (Nuijten et al., 2018).

However, as not all choices are equally valid, so multiverse analysis cannot simply provide a correct answer (Steegen et al., 2016) –the "average" result is not necessarily the closest to the truth. If we were to undertake a multiverse analysis in a scenario with a "known truth", i.e., using a simulated dataset (Bastiaansen et al., 2020), we may be able to detect identify the amount of variation from different sources [e.g, biological variation vs study design variation; Breznau et al. (2021)], and potentially the systematic biases stemming from specific choices (potentially via Bayesian Causal Forests Bryan, Yeager & O'Brien, 2019).

Here, using a multiverse approach, we explore how decisions concerning the design and analysis of an animal tracking study can impact the findings on habitat preference.

2 Methods

2.1 Simulating the Scenarios

We simulated three scenarios/species, comprising of different animals and landscapes. We simulated the landscapes using NLMR, and the animal movement using abmAnimalMovement. The abmAnimalMovement simulations are a discrete time, agent-based modelling approach for simulating animal movement. The full details of how the scenarios were parametrised can be found in [CITE], and also in the abmAnimalMovement github repository [LINK].

In brief: - Species 1, Badger: - Species 2, Vulture: - Species 3, King cobra:

Using these setting we simulated a population individuals of each species. Each simulation contained sufficient time steps (i.e., movement choices made by the animal) to provide data for model generation and validation.

Our landscapes comprise of three elements, which are considered by the animal differently depending on the behavioural state the animal is in. The three elements are matrices where each cell describes either foraging quality, shelter site quality, and movement ease. The three are non-independent, and based on a single initial random generation using a Gaussian field.

From the initial Gaussian field we altered the values to exaggerate the difference between high resource areas and low value areas. Broadly, we created: core resource areas (e.g., forests) with higher foraging quality but lower movement ease, edge areas that overlap with the forest with better movement ease and also housing higher shelter quality, and more barren areas with high movement ease but with minimal foraging value.

While our landscapes comprise of continuous values that determine the probability of a simulated animal using a cell or not, many habitat selection studies use categoric habitat metrics or land use types. Therefore, to simply the interpretation of the multiverse we simplify out landscape into three distinct categories for analysis.

2.2 Sampling and Analysis Options

Ultimately number of variations and choices we are able to explore will be dictated by computational costs and therefore time; however, below we describe the currently planned decision nodes.

2.2.1 Sampling

As stated we have three species simulations planned using the abmAnimalMovement package. The first decision point of a study would be the number of individuals sampled from that population (i.e., sample size). However, as many of the habitat selection analyses focus on individual selection we will not explore study sample size initially.

```
# possible way of generating a range that is skewed to more realistic number while still hitting
exp(seq(log(1), log(200), length.out = 8))
```

```
## [1]   1.000000   2.131663   4.543988   9.686251  20.647824  44.014204  93.823456
## [8] 200.000000
```

The first decision point decisions concern data quantity. We aim to vary tracking duration and tracking frequency primarily, while keep consistency fixed. While tracking consistency, or random/systematic data loss during tracking, is an element affecting data quantity the numerous ways of defining consistency means we will not explore tracking consistency at this time. Tracking timing is also an important consideration; for example, recording the locations of a diurnal animal only at night is highly unlikely to be a viable way of determining foraging choices. For this exploration we will assume that the researcher has sufficient knowledge about the animal's ecology to prioritise tracking during active hours. This only becomes a consideration when tracking frequency lowers to the point where the tracks performed during a day are all outside of the active period.

Data or location quality is another point of sampling variation. Different bio-logging equipment, terrain, animal behaviour, and weather can all impact the location error when tracking an animal. As the causes and measurement of error, as well as the solutions to, location error are numerous we will not be exploring the impact of error during this exploration in favour of more variation in other decision nodes.

2.2.2 Analysis

For the analysis decisions we focus entirely on decisions that can come out of an R analysis workflow. R presents the number one tool for analysing animal movement data [JOO CITE]. Joo et al., [CITE] provides a review of the R packages available to analysis movement data in R, and we use this review as a resource for determining the options for habitat analysis. Specifically, we will explore the decisions made during the workflows using adeHabaitatHS, amt, and ctmcmove R packages. Combined the packages offer many options for habitat analysis, we will focus on six. Three from the adeHabitatHS package: resource selection ratios (Wides), Compositional Analysis of Habitat Use (compana), Eigenanalysis of Selection Ratios (eisera). Two from the amt package: resource selection function (RSF), step selection function (SSF); and one from the ctmcmove package using a continuous time Markov chain (CTMC) framework. Each of the methods require downstream decisions resulting in a "garden of forking paths" to a final estimate of habitat selection

The methods from the adeHabitatHS package all require similar choices to be made. First is whether to approach the habitat selection analysis as a type II or III design (we are ignoring type I as we are aiming to estimate individuals' habitat selection). Type II is the habitat selection of an individual animal in relation to the available habitats on a population or landscape scale. The type II designs compares individual usage, to a universal availability. Type III differs by using individually defined available habitat. Wides, compana, and eisera analyses can be used in either type II or III scenarios.

Critical to these designs are the definition of availability. Type III or animal-specific availability is more simply defined. We will use a range of methods are build areas from the recorded locations of the animal. Minimum convex polygons (MCPs) are a simply create a polygon based on the out most locations. Kernel Density Estimations (KDEs) use kernel smoothing to build a heat map of use. Critical to their operation is a bandwidth or smoothing factor (h) that can alter the resulting area treated as used by the animal [CITE]. To account for the variation potentially introduced from smoothing factor choice, we will included two bandwidth selection processes commonly used in home range analyses: the reference bandwidth (href), and Least Squares Cross Validation (LSCV). The limitations of MCPs and KDEs have prompted the development of newer methods of area use estimation, which better account for the non-independence and autocorrelative structures within animal movement data. We use two of these methods to define availability. First is the Autocorrelated Kernel Density Estimations (AKDEs) from the ctmm package [CITE]. The ctmm package provides a workflow for creating an area estimate based on a number of movement models. The movement models (Ornstein-Uhlenbeck, Ornstein–Uhlenbeck Foraging, and Independent Identically Distributed) include different levels of autocorrelative structure. The movement model used would constitute a decision during analysis; however, the ctmm package allows for model comparisons (using AICc) to chose a single best model. Therefore, we will use the guidance from [CITE SILVA] to generate weighted AKDEs using PHREML, and select the best performing by AICc for inclusion in further analysis. The second movement-specific method we will use is dynamic Brownian Bridge Movement Models (dBBMMs), that estimates movement capacity of the animal to calibrate repeated random walks between known locations. DBBMMs require a window and margin size that defines the number of data points over which movement capacity (motion variance) is calculated. Fortunately the areas resulting from dBBMMs are largely insensitive to this choice, especially when large windows and margins are used. As window and margin are defined by data points, to keep the time they represent the same between different tracking frequencies we need to change the value for each tracking frequency. As our most infrequent tracking is 168 hours (1 week), we will set the window to the number of data points collected over 168 hours, and a margin of 48 hours. The broader window and margin sizes will help reduce computational costs. In a few cases the shorter tracking durations combined with the lower tracking frequencies will preclude calculation of available areas due to having too few locations. In these instances we will report an NA as a final result.

Type II designs require the availability to the same for all animals. To create population-level availability we will use the above described methods for each individual, then combine the areas into a single polygon. We will also include a "landscape" description of availability that will include the entire landscape raster.

Each of these area estimation methods require a choice to be made regarding the outermost boundary. MCPs areas are generated based a percentage of the location points; whereas the KDE, AKDE, and dBBMM methods require a contour to be extracted from the utilisation distribution (occurrence distribution for dBBMMs). We will explore the impact of using a 90, 95, and 99% contour for all the methods.

Once we have a defined availability area, we will run a number of variations for how many random points will be used to estimate the relative availability of different habitats.

The first method from the amt package we will use is a resource selection function (RSF). Like the adeHabitatHS methods, RSFs require a definition of what is available. We will use the same decision steps as for the adeHabitatHS method, including area definition methods (including the landscape level definition), contour decisions, and number of random points. The RSF method will have an additional decision node that will alter the weighting of the unused points compared to the used points.

The second amt method does not use an area to define availability. Instead step selection functions (SSF) use observed step lengths and turn angles to generate available locations at each time step (i.e., for each data point in the dataset; strata). As a result SSFs have a number of unique decisions. First is whether to run the model as an integrated SSF, where the step lengths and turn angles interact with the habitat. Second is where the habitat covariate data is extracted from along the step length: start, middle, or the end. Third is the number of available random steps to be generated at each time step.

The ctmc approach requires a number of decisions regarding the fitting of the quasi-continuous path model, and how that is imputed to a discrete space path. We will vary the spacing of the knots, the times at which the imputed path is sampled, and the precision matrix used during the path imputation. We will also vary the method used to convert the imputed path to a continuous-time discrete-space path. We will explore the impact of Linear Interpretation versus Shortest Path, and Rook's neighbourhood versus King's neighbourhood.

Once all decision pathways have resulted in a value for selection, we will normalise the results so they are directly comparable. This may require the results to be dichotomised into detection/non-detection of selection for a certain habitat.

2.3 Hypotheses

Although the study is primarily exploratory, with its scope determined by practical consideration of computational time, we register a few broad hypotheses. - increases in tracking duration will lead to more consistent habitat selection results - the most accurate estimates of habitat selection will come from tracking frequencies that match closest to the frequency of the animal's selection - the choice of method will have a larger impact on selection detection than the decisions within methods - ISSF will be the most success method as it matches closest to the underlying simulation mechanism

2.4 Possible Additional Explorations

Sample size impact on population level estimates of preference. Exploring the same decisions with different simulations of animal movement.

References

Albers C. 2019. The problem with unadjusted multiple and sequential statistical testing. *Nature Communications* 10:1921. DOI: 10.1038/s41467-019-09941-0.

Anderson MS, Ronning EA, De Vries R, Martinson BC. 2007. The Perverse Effects of Competition on Scientists' Work and Relationships. *Science and Engineering Ethics* 13:437–461. DOI: 10.1007/s11948-007-9042-5.

Anvari F, Lakens D. 2018. The replicability crisis and public trust in psychological science. *Comprehensive Results in Social Psychology* 3:266–286. DOI: 10.1080/23743603.2019.1684822.

Auspurg K, Brüderl J. 2021. Has the Credibility of the Social Sciences Been Credibly Destroyed? Reanalyzing the "Many Analysts, One Data Set" Project. *Socius: Sociological Research for a Dynamic World* 7:237802312110244.

DOI: 10.1177/23780231211024421.

Barto EK, Rillig MC. 2012. Dissemination biases in ecology: Effect sizes matter more than quality. *Oikos* 121:228–235. DOI: 10.1111/j.1600-0706.2011.19401.x.

Bastiaansen JA, Kunkels YK, Blaauw FJ, Boker SM, Ceulemans E, Chen M, Chow S-M, Jonge P de, Emerencia AC, Epskamp S, Fisher AJ, Hamaker EL, Kuppens P, Lutz W, Meyer MJ, Moulder R, Oravecz Z, Riese H, Rubel J, Ryan O, Servaas MN, Sjobeck G, Snippe E, Trull TJ, Tschacher W, Veen DC van der, Wichers M, Wood PK, Woods WC, Wright AGC, Albers CJ, Bringmann LF. 2020. Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *Journal of Psychosomatic Research* 137:110211. DOI: 10.1016/j.jpsychores.2020.110211.

Bishop D. 2019. Rein in the four horsemen of irreproducibility. *Nature* 568:435–435. DOI: 10.1038/d41586-019-01307-2.

Brembs B. 2018. Prestigious Science Journals Struggle to Reach Even Average Reliability. *Frontiers in Human Neuroscience* 12:1–7. DOI: 10.3389/fnhum.2018.00037.

Brembs B. 2019. Reliable novelty: New should not trump true. *PLOS Biology* 17:e3000117. DOI: 10.1371/journal.pbio.3000117.

Breznau N, Rinke EM, Wuttke A, Adem M, Adriaans J, Alvarez-Benjumea A, Andersen HK, Auer D, Azevedo F, Bahnsen O, Balzer D, Bauer G, Bauer P, Baumann M, Baute S, Benoit V, Bernauer J, Berning C, Berthold A, Bethke FS, Biegert T, Blinzler K, Blumenberg J, Bobzien L, Bohman A, Bol T, Bostic A, Brzozowska Z, Burgdorf K, Burger K, Busch K, Castillo JC, Chan N, Christmann P, Connelly R, Czymara CS, Damian E, Ecker A, Edelmann A, Eger MA, Ellerbrock S, Forke A, Forster AG, Gaasendam C, Gavras K, Gayle V, Gessler T, Gnambs T, Godefroidt A, Grömping M, Groß M, Gruber S, Gummer T, Hadjar A, Heisig JP, Hellmeier S, Heyne S, Hirsch M, Hjerm M, Hochman O, Hövermann A, Hunger S, Hunkler C, Huth N, Ignacz Z, Jacobs L, Jacobsen J, Jaeger B, Jungkunz S, Jungmann N, Kauff M, Kleinert M, Klinger J, Kolb J-P, Kołczyńska M, Kuk JS, Kunißen K, Sinatra DK, Greinert A, Lersch PM, Löbel L-M, Lutscher P, Mader M, Madia JE, Malancu N, Maldonado L, Marahrens H, Martin N, Martinez P, Mayerl J, Mayorga OJ, McManus P, Wagner K, Meeusen C, Meierrieks D, Mellon J, Merhout F, Merk S, Meyer D, Micheli L, Mijs JJB, Moya C, Neunhoeffer M, Nüst D, Nygård O, Ochsenfeld F, Otte G, Pechenkina A, Prosser C, Raes L, Ralston K, Ramos M, Roets A, Rogers J, Ropers G, Samuel R, Sand G, Schachter A, Schaeffer M, Schieferdecker D, Schlueter E, Schmidt K, Schmidt R, Schmidt-Catran A, Schmiedeberg C, Schneider J, Schoonvelde M, Schulte-Cloos J, Schumann S, Schunck R, Schupp J, Seuring J, Silber H, Sleegers WWA, Sonntag N, Staudt A, Steiber N, Steiner N, Sternberg S, Stiers D, Stojmenovska D, Storz N, Striessnig E, Stroppe A-K, Teltemann J, Tibajev A, Tung BB, Vagni G, Van Assche J, Linden M van der, Noll J van der, Van Hootegem A, Vogtenhuber S, Voicu B, Wagemans F, Wehl N, Werner H, Wiernik BM, Winter F, Wolf C, Yamada Y, Zhang N, Ziller C, Zins S, Żółtak T, Nguyen HHV. 2021. Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty. *MetaArXiv*. DOI: 10.31222/osf.io/cd5j9.

Bryan CJ, Yeager DS, O'Brien JM. 2019. Replicator degrees of freedom allow publication of misleading failures to replicate. *Proceedings of the National Academy of Sciences* 116:25535–25545. DOI: 10.1073/pnas.1910951116.

Cassey P, Ewen JG, Blackburn TM, Møller AP. 2004. A survey of publication bias within evolutionary ecology. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 271. DOI: 10.1098/rsbl.2004.0218.

Clark TD, Raby GD, Roche DG, Binning SA, Speers-Roesch B, Jutfelt F, Sundin J. 2020. Ocean acidification does not impair the behaviour of coral reef fishes. *Nature* 577:370–375. DOI: 10.1038/s41586-019-1903-y.

Del Giudice M, Gangestad SW. 2021. A Traveler's Guide to the Multiverse: Promises, Pitfalls, and a Framework for the Evaluation of Analytic Decisions. *Advances in Methods and Practices in Psychological Science* 4:251524592095492. DOI: 10.1177/2515245920954925.

Desbureaux S. 2021. Subjective modeling choices and the robustness of impact evaluations in conservation science. *Conservation Biology* 35:1615–1626. DOI: 10.1111/cobi.13728.

Dragicevic P, Jansen Y, Sarma A, Kay M, Chevalier F. 2019. Increasing the Transparency of Research Papers with Explorable Multiverse Analyses. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow Scotland Uk: ACM, 1–15. DOI: 10.1145/3290605.3300295.

Fanelli D. 2010b. "Positive" Results Increase Down the Hierarchy of the Sciences. *PLoS ONE* 5:e10068. DOI: 10.1371/journal.pone.0010068.

Fanelli D. 2010a. Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data. *PLoS ONE* 5:e10271. DOI: 10.1371/journal.pone.0010271.

Fanelli D. 2012. Negative results are disappearing from most disciplines and countries. *Scientometrics* 90:891–904. DOI: 10.1007/s11192-011-0494-7.

Forstmeier W, Schielzeth H. 2011. Cryptic multiple hypotheses testing in linear models: Overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology* 65:47–55. DOI: 10.1007/s00265-010-1038-5.

Forstmeier W, Wagenmakers E-J, Parker TH. 2017. Detecting and avoiding likely false-positive findings – a practical guide: Avoiding false-positive findings. *Biological Reviews* 92:1941–1968. DOI: 10.1111/brv.12315.

Fraser H, Barnett A, Parker TH, Fidler F. 2020. The role of replication studies in ecology. *Ecology and Evolution* 10:5197–5207. DOI: 10.1002/ece3.6330.

Fraser H, Parker T, Nakagawa S, Barnett A, Fidler F. 2018. Questionable research practices in ecology and evolution.

PLOS ONE 13:e0200303. DOI: 10.1371/journal.pone.0200303.

Freedman LP, Cockburn IM, Simcoe TS. 2015. The Economics of Reproducibility in Preclinical Research. *PLOS Biology* 13:e1002165. DOI: 10.1371/journal.pbio.1002165.

Gelman A, Loken E. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. :17.

Grainger M, Bolam FC, stewart G, Nilsen EB. 2019. Evidence synthesis for tackling research waste. *EcoEvoRxiv*. DOI: 10.32942/osf.io/42fkh.

Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. 2015. The Extent and Consequences of P-Hacking in Science. *PLOS Biology* 13:e1002106. DOI: 10.1371/journal.pbio.1002106.

Hoffmann S, Schönbrodt F, Elsas R, Wilson R, Strasser U, Boulesteix A-L. 2021. The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines. *Royal Society Open Science* 8:rsos.201925, 201925. DOI: 10.1098/rsos.201925.

Holman L, Head ML, Lanfear R, Jennions MD. 2015. Evidence of Experimental Bias in the Life Sciences: Why We Need Blind Data Recording. *PLOS Biology* 13:e1002190. DOI: 10.1371/journal.pbio.1002190.

Huntington-Klein N, Arenas A, Beam E, Bertoni M, Bloem JR, Burli P, Chen N, Grieco P, Ekpe G, Pugatch T, Saavedra M, Stopnitzky Y. 2021. The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry* 59:944–960. DOI: 10.1111/ecin.12992.

Jennions MD, Møller AP. 2002. Publication bias in ecology and evolution: An empirical assessment using the 'trim and fill' method. *Biological Reviews of the Cambridge Philosophical Society* 77:211–222. DOI: 10.1017/S1464793101005875.

Kaplan RM, Irvin VL. 2015. Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time. *PLOS ONE* 10:e0132382. DOI: 10.1371/journal.pone.0132382.

Kelly CD. 2019. Rate and success of study replication in ecology and evolution. *PeerJ* 7:e7654. DOI: 10.7717/peerj.7654.

Lakens D, Adolfi FG, Albers CJ, Anvari F, Apps MAJ, Argamon SE, Baguley T, Becker RB, Benning SD, Bradford DE, Buchanan EM, Caldwell AR, Van Calster B, Carlsson R, Chen S-C, Chung B, Colling LJ, Collins GS, Crook Z, Cross ES, Daniels S, Danielsson H, DeBruine L, Dunleavy DJ, Earp BD, Feist MI, Ferrell JD, Field JG, Fox NW, Friesen A, Gomes C, Gonzalez-Marquez M, Grange JA, Grieve AP, Guggenberger R, Grist J, Harmelen A-L van, Hasselman F, Hochard KD, Hoffarth MR, Holmes NP, Ingre M, Isager PM, Isotalus HK, Johansson C, Juszczyk K, Kenny DA, Khalil AA, Konat B, Lao J, Larsen EG, Lodder GMA, Lukavský J, Madan CR, Manheim D, Martin SR, Martin AE, Mayo DG, McCarthy RJ, McConway K, McFarland C, Nio AQX, Nilsonne G, Oliveira CL de, Xivry J-JO de, Parsons S, Pfuhl G, Quinn KA, Sakon JJ, Saribay SA, Schneider IK, Selvaraju M, Sjoerds Z, Smith SG, Smits T, Spies JR, Sreekumar V, Steltenpohl CN, Stenhouse N, Świątkowski W, Vadillo MA, Van Assen MALM, Williams MN, Williams SE, Williams DR, Yarkoni T, Ziano I, Zwaan RA. 2018. Justify your alpha. *Nature Human Behaviour* 2:168–171. DOI: 10.1038/s41562-018-0311-x.

Lemoine NP, Hoffman A, Felton AJ, Baur L, Chaves F, Gray J, Yu Q, Smith MD. 2016. Underappreciated problems of low replication in ecological field studies. *Ecology* 97:2554–2561. DOI: 10.1002/ecy.1506.

Levins R, Lewontin R. 1985. *The dialectical biologist*. Harvard University Press.

Liu Y, Althoff T, Heer J. 2020. Paths Explored, Paths Omitted, Paths Obscured: Decision Points & Selective Reporting in End-to-End Data Analysis. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu HI USA: ACM, 1–14. DOI: 10.1145/3313831.3376533.

Miyakawa T. 2020. No raw data, no science: Another possible source of the reproducibility crisis. *Molecular Brain* 13:24, s13041-020-0552-2. DOI: 10.1186/s13041-020-0552-2.

Nakagawa S, Parker TH. 2015. Replicating research in ecology and evolution: Feasibility, incentives, and the cost-benefit conundrum. *BMC Biology* 13:88. DOI: 10.1186/s12915-015-0196-3.

Nuijten MB, Bakker M, Maassen E, Wicherts JM. 2018. Verify original results through reanalysis before replicating. *Behavioral and Brain Sciences* 41:e143. DOI: 10.1017/S0140525X18000791.

O'Boyle EH, Banks GC, Gonzalez-Mulé E. 2014. The Chrysalis Effect: How Ugly Initial Results Metamorphosize Into Beautiful Articles. *Journal of Management* 43:376–399. DOI: 10.1177/0149206314527133.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349:aac4716–aac4716. DOI: 10.1126/science.aac4716.

Palmer AR. 2000. Quasi-Replication and the Contract of Error: Lessons from Sex Ratios, Heritabilities and Fluctuating Asymmetry. *Annual Review of Ecology and Systematics* 31:441–480. DOI: 10.1146/annurev.ecolsys.31.1.441.

Parsons S. 2020. Exploring reliability heterogeneity with multiverse analyses: Data processing decisions unpredictably influence measurement reliability. *PsyArXiv*. DOI: 10.31234/osf.io/y6tcz.

Patel CJ, Burford B, Ioannidis JPA. 2015. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology* 68:1046–1058. DOI: 10.1016/j.jclinepi.2015.05.029

Reinert HK, Cundall D. 1982. An Improved Surgical Implantation Method for Radio-Tracking Snakes. *Copeia* 1982:702. DOI: 10.2307/1444674.

Rijnhart JJM, Twisk JWR, Deeg DJH, Heymans MW. 2021. Assessing the Robustness of Mediation Analysis Results Using Multiverse Analysis. *Prevention Science*. DOI: 10.1007/s11121-021-01280-1.

Roche DG, Amcoff M, Morgan R, Sundin J, Finnøen MH, Lawrence MJ, Henderson E, Speers-Roesch B, Brown C, Clark

TD, Bshary R, Jutfelt F, Binning SA. 2020. Behavioural lateralisation in a detour test is not repeatable in fishes. *EcoEvoRxiv*:62. DOI: 10.32942/osf.io/6kcwa.

Salis A, Lena J-P, Lengagne T. 2021. How Subtle Protocol Choices Can Affect Biological Conclusions: Great Tits' Response to Allopatric Mobbing Calls. *Animal Behavior and Cognition* 8:152–165. DOI: 10.26451/abc.08.02.05.2021.

Sánchez-Tójar A, Nakagawa S, Sánchez-Fortún M, Martin DA, Ramani S, Girndt A, Bókony V, Kempenaers B, Liker A, Westneat DF, Burke T, Schroeder J. 2018. Meta-analysis challenges a textbook example of status signalling and demonstrates publication bias. *eLife* 7:e37385. DOI: 10.7554/eLife.37385.

Scheel AM, Schijen MRMJ, Lakens D. 2021. An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Advances in Methods and Practices in Psychological Science* 4:251524592110074. DOI: 10.1177/25152459211007467.

Schnitzer SA, Carson WP. 2016. Would Ecology Fail the Repeatability Test? *BioScience* 66:98–99. DOI: 10.1093/biosci/biv176.

Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, Awtrey E, Bahník Š, Bai F, Bannard C, Bonnier E, Carlsson R, Cheung F, Christensen G, Clay R, Craig MA, Dalla Rosa A, Dam L, Evans MH, Flores Cervantes I, Fong N, Gamez-Djokic M, Glenz A, Gordon-McKeon S, Heaton TJ, Hederos K, Heene M, Hofelich Mohr AJ, Högden F, Hui K, Johannesson M, Kalodimos J, Kaszubowski E, Kennedy DM, Lei R, Lindsay TA, Liverani S, Madan CR, Molden D, Molleman E, Morey RD, Mulder LB, Nijstad BR, Pope NG, Pope B, Prenoveau JM, Rink F, Robusto E, Roderique H, Sandberg A, Schlüter E, Schönbrodt FD, Sherman MF, Sommer SA, Sotak K, Spain S, Spörlein C, Stafford T, Stefanutti L, Tauber S, Ullrich J, Vianello M, Wagenmakers E-J, Witkowiak M, Yoon S, Nosek BA. 2018. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science* 1:337–356. DOI: 10.1177/2515245917747646.

Silva I, Crane M, Marshall BM, Strine CT. 2020. Reptiles on the wrong track? Moving beyond traditional estimators with dynamic Brownian Bridge Movement Models. *Movement Ecology* 8:43. DOI: 10.1186/s40462-020-00229-3.

Simmons JP, Nelson LD, Simonsohn U. 2011. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* 22:1359–1366. DOI: 10.1177/0956797611417632.

Simonsohn U. 2015. Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychological Science* 26:559–569. DOI: 10.1177/0956797614567341.

Simonsohn U, Simmons JP, Nelson LD. 2020. Specification curve analysis. *Nature Human Behaviour* 4:1208–1214. DOI: 10.1038/s41562-020-0912-z.

Smaldino PE, McElreath R. 2016. The natural selection of bad science. *Royal Society Open Science* 3:160384. DOI: 10.1098/rsos.160384.

Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. 2016. Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science* 11:702–712. DOI: 10.1177/1745691616658637.

Tang-Martínez Z. 2020. The history and impact of women in animal behaviour and the ABS: A North American perspective. *Animal Behaviour* 164:251–260. DOI: 10.1016/j.anbehav.2019.12.011.

Tedersoo L, Küngas R, Oras E, Köster K, Eenmaa H, Leijen Ä, Pedaste M, Raju M, Astapova A, Lukner H, Kogermann K, Sepp T. 2021. Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data* 8:192. DOI: 10.1038/s41597-021-00981-0.

Vines TH, Albert AYK, Andrew RL, Débarre F, Bock DG, Franklin MT, Gilbert KJ, Moore J-S, Renaut S, Rennison DJ. 2014. The Availability of Research Data Declines Rapidly with Article Age. *Current Biology* 24:94–97. DOI: 10.1016/j.cub.2013.11.014.

Vinkers CH, Tijdink JK, Otte WM. 2015. Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: Retrospective analysis. *BMJ*:h6467. DOI: 10.1136/bmj.h6467.

Wang D, Forstmeier W, Ihle M, Khadraoui M, Jerónimo S, Martin K, Kempenaers B. 2018. Irreproducible text-book "knowledge": The effects of color bands on zebra finch fitness: COLOR BANDS HAVE NO EFFECT ON FITNESS IN ZEBRA FINCHES. *Evolution* 72:961–976. DOI: 10.1111/evo.13459.

Ware JJ, Munafò MR. 2015. Significance chasing in research practice: Causes, consequences and possible solutions: Significance chasing. *Addiction* 110:4–8. DOI: 10.1111/add.12673.

Weaver SJ, Westphal MF, Taylor EN. 2021. Technology wish lists and the significance of temperature-sensing wildlife telemetry. *Animal Biotelemetry* 9:29. DOI: 10.1186/s40317-021-00252-0.

Webb SS, Demeyere N. 2021. Multiverse Analysis: A Method to Determine Researcher Degrees of Freedom in Test Validation. *PsyArXiv*. DOI: 10.31234/osf.io/nhrwq.

Winne CT, Willson JD, Andrews KM, Reed RN. 2006. Efficacy of marking snakes with disposable medical cautery units. *Herpetological Review* 37:52–54.

Young C, Holsteen K. 2017. Model Uncertainty and Robustness: A Computational Framework for Multimodel Analysis. *Sociological Methods & Research* 46:3–40. DOI: 10.1177/0049124115610347.