Solving the Sample Size Problem for Resource Selection Analysis

Garrett M. Street*[1,2], Jonathan R. Potts[3], Luca Börger[4,5], James C. Beasley[6], Stephen Demarais[1], John M. Fryxell[7], Philip D. McLoughlin[8], Kevin L. Monteith[9], Christina M. Prokopenko[10], Miltinho C. Ribeiro[11], Arthur R. Rodgers[12], Bronson K. Strickland[1], Floris M. van Beest[13], David A. Bernasconi[6], Larissa T. Beumer[13], Guha Dharmarajan[6], Samantha P. Dwinnell[14], David A. Keiter[6], Alexine Keuroghlian[15], Levi J. Newediuk[10], Júlia Emi F. Oshima[11], Olin Rhodes Jr.[6], Peter E. Schlichting[6], Niels M. Schmidt[13], and Eric Vander Wal[10]

[1]Department of Wildlife, Fisheries, and Aquaculture, Mississippi State University, Mississippi State, MS, 39762, USA

[2]Quantitative Ecology and Spatial Technologies Laboratory, Mississippi State University, Mississippi State, MS, 39762, USA

[3]School of Mathematics and Statistics, University of Sheffield, Sheffield, S3 7RH, United Kingdom

[4]Department of Biosciences, Swansea University, Swansea, SA2 8PP, United Kingdom

[5]Centre for Biomathematics, Swansea University, Swansea, SA2 8PP, United Kingdom

[6]Savannah River Ecology Laboratory, University of Georgia, Aiken, SC, 29802, USA

[7]Department of Integrative Biology, University of Guelph, Guelph, Ontario, N1G 2W1, Canada

[8]Department of Biology, University of Saskatchewan, Saskatoon, Saskatchewan, S7N 5E2, Canada

[9]Haub School of Environment and Natural Resources, University of Wyoming, Laramie, WY, 82072, USA

[10]Department of Biology, Memorial University of Newfoundland, St. John's, Newfoundland, A1B 3X9, Canada

[11]Instituto de Biosciências, Universidad Estadual Paulista, Rio Claro, São Paulo, Brazil

[12]Centre for Northern Forest Ecosystem Research, Ontario Ministry of Natural Resources and Forestry, Thunder Bay, Ontario, P7E 6S8, Canada

[13]Department of Bioscience, Aarhus University, Aarhus, Denmark

[14]Wyoming Cooperative Fish and Wildlife Research Unit, University of Wyoming, Laramie, WY, 82071, USA

[15]IUCN/SSC Peccary Specialist Group, Campo Grande, Brazil

Manuscript type: Standard Paper

Abstract word count: 150

Main text word count: 4904

Number of figures: 5

Number of tables: 0

*Corresponding Author (gms246@msstate.edu)

Abstract

Resource selection analysis (RSA) is a cornerstone approach for understanding animal distributions, yet there exists no rigorous quantification of sample sizes required to obtain reliable results. We provide closed-form mathematical expressions for both the number of animals and relocations per animal required for parameterising RSA to a given degree of precision. Required sample sizes depend on just two quantities: habitat selection strength and an index of landscape complexity, which we define rigorously. We validate our solutions using 5,678,623 GPS locations from 511 animals from 10 species (omnivores, carnivores, and herbivores from boreal, temperate, and tropical forests, montane woodlands, swamps, and tundra). Our results contradict conventional wisdom by showing that environmental effects on distributions can often be estimated with fewer animals and relocations than assumed, with far-reaching implications for ecologists, conservationists, and natural resource managers.

Keywords: bootstrap, habitat selection, p-value, power analysis, Resource Selection Function, sample size, Species Distribution Model, validation

Introduction

Resource selection analysis (RSA) is a broad framework linking the distribution of animals to their preferences for specific habitat conditions and is a fundamental tool in animal ecology (Boyce & McDonald 1999; Strickland & McDonald 2006). Obtaining sufficient locations to ascertain the distribution of animals across landscapes is a fundamental requirement for RSA. Indeed, to understand intra-specific variation in the distribution of animals – a critical research aim in basic and applied animal ecology – it is necessary to obtain repeated localizations on multiple individuals, now commonly collected using animal-attached GPS sensors (Hebblewhite & Haydon 2010). GPS data on animal movements are hence commonly employed for RSA and are often analyzed using Resource Selection Functions (RSFs; Boyce & McDonald 1999; Manly *et al.* 2002; Elith & Leathwick 2009; Hebblewhite & Haydon 2010). RSFs are a class of exponential models of space use that estimate the probability distribution of animal locations using different resources/conditions in the landscape, taking into account the availability of each resource, and thereby provide a measure of the 'strength' of (behavioral) selection for or against each resource (Manly *et al.* 2002). RSFs are easily fitted using standard statistical models (commonly logistic or conditional logistic regression) applied to data on animal locations and resource distributions in the landscape and have become a cornerstone of research in spatial ecology (Manly *et al.* 2002; Elith & Leathwick 2009; Renner & Warton 2013).

Given the prevalence of RSFs, it is surprising that the central question determining the validity of inferences obtained – how much data is needed to estimate a RSF for a given species? – has not been solved. This issue has been broached for occupancy analysis (Guillera-Arroita & Lahoz-Monfort 2012) and generalized linear mixed models (Johnson *et al.* 2015), and has been evaluated within individual RSF studies using simulations (Leban *et al.* 2001; Loe *et al.* 2012), yet no analytic expressions exist to determine the number of animals (M) and relocations per animal (N) required to obtain RSF outputs to a given degree of precision. While the accuracy and precision of RSFs generally increase with sample size, leading to a standard rule-of-thumb

of $M \geq 30$ needed for reliable ecological inference (Leban *et al.* 2001), this rough guideline is grounded in century-old thinking about statistics in the pre-computation world (James *et al.* 2013). Crucially, it is also oblivious to the ecological reality that a multitude of factors may affect selection strength and determine the required sample size (Manly *et al.* 2002; McLoughlin *et al.* 2010; Hebblewhite & Haydon 2010). These include density-dependence (i.e. certain habitats become less attractive when occupied by conspecifics; Fretwell & Lucas 1969; McLoughlin *et al.* 2010; van Beest *et al.* 2016), trade-offs in selection for forage and cover under predation risk (Fortin *et al.* 2005; McLoughlin *et al.* 2010), temporal variations in resource dynamics (McLoughlin *et al.* 2010; Paolini *et al.* 2018), or the degree of habitat availability or heterogeneity in a landscape (Mysterud & Ims 1998; McLoughlin *et al.* 2010; van Beest *et al.* 2016; Paolini *et al.* 2018). There is no consistency in RSF studies in the number of replicates used (Hebblewhite & Haydon 2010), as the only alternative approaches to establishing the number of replicates a priori are ecologically informed guesswork, or simply to collect as much data as possible.

The crux of the problem lies in the relationship between sample size and ecological complexity. It is suggested that more complex systems require more data to describe (Wisz *et al.* 2008), yet a robust power analysis (Johnson *et al.* 2015) allowing examination of the relationship between RSF estimation, system complexity, and data availability is crucially missing. This has obvious economic and ethical implications if more animals are tagged and monitored than needed and affects research aimed at the conservation of species, which requires reliable estimates of animal-habitat relationships but where it is often impossible to monitor large numbers of animals. Here, we provide a solution to the sample size problem in RSFs by deriving analytic expressions for the values of M and N (the number of animals and relocations per animal respectively) required to estimate RSFs to a required degree of accuracy, taking into account landscape complexity and the strength of selection for the resources. We validate these expressions using simulations and a large dataset of GPS-tagged animals (including 10 species from different continents and biomes) and show that the most biologically relevant effects of landscapes on animal distributions can often be estimated with far fewer animals and locations than are commonly stated.

Methods

We begin by describing mathematically how to determine the number of locations per animal (N) and the number of animals (M) for RSA. RSA seeks to parametrize a model of space use that has the following form (Manly *et al.* 2002):

$$u(\mathbf{x}) = \frac{A(\mathbf{x})W(\mathbf{x})}{\int_\Omega A(\mathbf{x}')W(\mathbf{x}')d\mathbf{x}'}, \tag{1}$$

where $u(\mathbf{x})$ is the *utilization distribution* of the study species (i.e. the probability density function of the study animals' locations), $A(\mathbf{x})$ is a function denoting the availability of the point $\mathbf{x}$ to the animals, Ω is the study area, and $W(\mathbf{x})$ is the RSF. (Note: throughout this manuscript, bold fonts imply that the quantity is a vector.) For the purposes of our analytic calculations, our RSF will be dependent upon a single resource layer $R(\mathbf{x})$. This could denote, for example, the vegetation quality or prey availability at point $\mathbf{x}$. However, in general, $R(\mathbf{x})$ represents a map of any environmental feature which is hypothesized to covary with space use. Although we only look at one resource layer at a time for our analytic calculations, we show in our empirical study (below) that the resulting formulae work when the RSF has multiple layers.

As is the standard method for RSA, we make 3 simplifying assumptions (Manly *et al.* 2002): (i) our weighting function is of the form $W(\mathbf{x}|\beta) = \exp[\beta R(\mathbf{x})]$, where β is a parameter to be estimated; (ii) the availability kernel $A(\mathbf{x})$ is a uniform distribution; and (iii) relocations are independent. Consequently, our model of space use from Equation (1) becomes:

$$u(\mathbf{x}|\beta) = \frac{\exp[\beta R(\mathbf{x})]}{\int_\Omega \exp[\beta R(\mathbf{x}')]d\mathbf{x}'}. \tag{2}$$

The aim of this section is to understand how many independent samples are required to give an accurate parametrization of the model in Equation (2).

Locations from a Single Individual (N)

We first need to phrase the question "How many locations?" in a concrete, mathematical way. Suppose we wish to test the null hypothesis $H_0 : \beta = 0$ against the alternative $H_1 : \beta \neq 0$ at a significance level $p \in (0, 1)$. An experiment to test this hypothesis involves measuring N samples and using (conditional) logistic regression to infer β and test the null hypothesis (as is the standard method for resource selection, e.g. Manly *et al.* 2002). We define $N_{\alpha,p}(\beta)$ to be the minimum number of samples required so that we expect to reject the null hypothesis in $100(1-\alpha)\%$ of experiments. An approximate analytical formula for $N_{\alpha,p}(\beta)$ is given as follows (derived in Supplementary Appendix A):

$$N_{\alpha,p}(\beta) \approx \frac{(z_\alpha + z_{p/2})^2}{\mathrm{Var}[R(X_\beta)]}\beta^{-2}. \tag{3}$$

Here, $z_\alpha = \Phi^{-1}(1-\alpha)$ where $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution (e.g. $z_{0.05} \approx 1.645$, $z_{0.025} \approx 1.96$), X_β is a random variable whose probability density function is given by Equation (2), and $\mathrm{Var}[R(X_\beta)]$ is the variance of $R(X_\beta)$. An explicit functional expression for $\mathrm{Var}[R(X_\beta)]$ can be written as follows:

$$\mathrm{Var}[R(X_\beta)] = \frac{\int_\Omega R^2(\mathbf{x}) \exp\left[\beta R(\mathbf{x})\right] \mathrm{d}\mathbf{x}}{\int_\Omega \exp\left[\beta R(\mathbf{x})\right] \mathrm{d}\mathbf{x}} - \left(\frac{\int_\Omega R(\mathbf{x}) \exp\left[\beta R(\mathbf{x})\right] \mathrm{d}\mathbf{x}}{\int_\Omega \exp\left[\beta R(\mathbf{x})\right] \mathrm{d}\mathbf{x}} \right)^2. \tag{4}$$

We call $\mathrm{Var}[R(X_\beta)]$ "landscape complexity". Critically, this form of landscape complexity is determined in part by multiplying the landscape layer by the expected β, so it should be understood as representing the landscape complexity *as viewed by the animal*.

The formula in Equation (3) is approximate due to two assumptions: (i) it relies on the standard error, σ, of the maximum likelihood function being approximately normally distributed, and (ii) it uses a standard result relating the standard error for the estimator of β to the second derivative of the log-likelihood function (see Supplementary Appendix A for more details). Therefore it is necessary to investigate the magnitude of these approximating assumptions using

simulated data.

To test how effective the approximate expression from Equation (3) is at capturing the actual number of samples required to infer β with a given level of accuracy, we constructed a simulated resource layer which describes an example of the function $R(\mathbf{x})$ (Fig. 1a). This test layer is a Gaussian random field, previously used in the context of resource selection by Potts *et al.* (2014). It was generated by the R function `GaussRF()` from the `RandomFields` package (Schlather *et al.*, 2016), using the exponential model with mean=0, variance=1, nugget=0, and scale=10, and consists of $L = 100$ by $L = 100$ pixels. By sampling N times from Equation (2) for various N with $R(\mathbf{x})$, we can compute empirical values for $N_{\alpha,p}(\beta)$ for different β (full method given in Supplementary Appendix B). Comparison of these empirically-derived values alongside the analytical expression from Equation (3) reveals remarkably strong agreement (Fig. 1b). This suggests that Equation (3) gives an accurate estimation of the number of independent samples required to estimate β.

Locations from multiple individuals (M)

Now we assume that there are M individuals and they each select resources with different β. To model this, let $\beta_1, \ldots, \beta_M \sim N(\beta, s^2)$ be independent draws from a normal distribution with mean β and variance s^2. Then β_i is the coefficient of resource selection for individual $i \in \{1, \ldots, M\}$. Suppose for each individual i we have gathered N_i locations. Let $\hat{\beta}_i$ be the maximum likelihood estimator for β_i. Then the standard deviation of $\hat{\beta}_i$ can be estimated as (Supplementary Appendix A, Equation 15):

$$\sigma_i = \frac{1}{\sqrt{N_i \mathrm{Var}[R(X_{\beta_i})]}}. \tag{5}$$

If $\hat{\beta}$ is the mean of $\hat{\beta}_1, \ldots, \hat{\beta}_M$, then $\hat{\beta}$ is normally distributed as follows (Supplementary Appendix C):

$$\hat{\beta} \sim N\left(\beta, \frac{1}{M^2}\sum_{i=1}^{M}\sigma_i^2 + \frac{s^2}{M}\right). \tag{6}$$

Thus $\hat{\beta}$ is an unbiased estimator of β. Notice that the variance decays as M increases. If the practitioner has some prior expectation of the possible values of β and s^2, Equation (6) can be used to calculate the number of animals, M, required to obtain an empirical estimate of β to a given degree of accuracy.

As well as calculating an estimate of β, it is also possible to estimate s^2. The following is an unbiased estimator of s^2 for $M \geq 2$ (Supplementary Appendix C):

$$\hat{s^2} = \frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{\beta}_i - \frac{1}{M}\sum_{j=1}^{M}\hat{\beta}_j\right)^2 - \frac{1}{M}\sum_{i=1}^{M}\sigma_i^2. \tag{7}$$

We were not able to derive a closed analytic formula for the uncertainty in the estimator given in Equation (7); however, we provide code for estimating this using random sampling (see Supplementary Appendix D). In general, the estimator becomes more precise for lower σ_i and higher M. This is shown in Supplementary Appendix D, where we also verify numerically Equations (6) and (7).

Equation (6) allows us to calculate the minimum number of animals, $M_{\alpha,p}(\beta)$, for which we would expect to reject the null hypothesis that $\beta = 0$, at significance level p, $100(1-\alpha)\%$ of the time (two-tailed test). $M_{\alpha,p}(\beta)$ is the minimum integer, M, that satisfies the following inequality:

$$M \geq \frac{s^2(z_{p/2}+z_\alpha)^2 + \sqrt{s^4(z_{p/2}+z_\alpha)^4 + 4\beta^2(z_{p/2}+z_\alpha)^2\sum_{i=1}^{M}\sigma_i^2}}{2\beta^2}. \tag{8}$$

Data and Resource Selection Functions

Equations (3) and (8) give predicted values for the number of relocations N and the number of animals M required for RSF estimation. To test our analytical predictions, we compiled GPS-based relocation datasets from 10 separate species with accompanying landscape data in raster format (Table S1; Fig. 2). Landscape data were either categorical (i.e. discrete landscover) or numeric (e.g. elevation, precipitation, etc.). To ensure comparability between model outputs for each species, we centered and scaled each numeric landscape raster in R using the `scale()` function with default parameters. We converted categorical landcover rasters to binary raster layers for each landcover classification of interest (e.g. deciduous forest, croplands, etc.) to acquire estimates of $\text{Var}[R(X_\beta)]$ for a given categorical raster.

We generated a 1:1 sample of availability (i.e. 1 available location per animal relocation) within each animal's 99% home range as estimated using the function `kernelUD()` in R package `adehabitatHR` with the default bandwidth estimator. We extracted centered-and-scaled (numeric) and binary (categorical) landscape data to animal relocations and available locations and fit a RSF to each animal in each dataset using logistic regression (i.e. 511 individual models; Table S2). For simplicity, we used only linear main effects for each predictor in a given RSF; however, we emphasize that more complex effects (e.g. non-linear and interaction terms) may be identically investigated using the appropriate non-linear transformation or multiplicative product on the resource layer(s) prior to calculation. Note that, although our equations operate on a single resource layer at a time, our analysis uses RSFs with multiple layers. This procedure thus tests whether multiple layers may be analyzed one-at-a-time to ascertain the number of animals and fixes required to estimate the β-value for each layer.

Empirical Validation: M

After fitting each RSF, we calculated the mean selection coefficient $\bar{\beta}$ for each landscape layer across individuals within a species. Assuming $\bar{\beta}$ was an accurate estimate of population-level

selection β, we asked: how many animals M were necessary to estimate β? We calculated $\text{Var}[R(X_\beta)]$ for each centered-and-scaled or binary raster within each animal's 99% range according to Equation (4) and the resulting values of N according to Equation (3). We generated empirical distributions of $\hat{\beta}$ and $\hat{s^2}$ as described in Supplementary Appendix D for $M \in \{2, \ldots, 30\}$ using the average N and $\text{Var}[R(X_\beta)]$ as population-level estimates of each. We computed the empirical 95% intervals at a given M (i.e. $\alpha = 0.05$). The value of M at which the empirical interval no longer contains 0 is the predicted minimum M necessary to estimate β with 95% confidence, M_{pred} (i.e. the minimum integer $M_{0.05,0.05}(\beta)$; Equation (8)).

For comparison with observation, we then resampled the estimated selection coefficients for each individual within a species. For a given $M \in \{2, \ldots, 30\}$ as above, we generated 4000 samples of M_i observed selection coefficients and calculated $\bar{\beta}$ for each (i.e. 4000 mean selection coefficients assuming M_i animals). This represents the observed distribution of possible $\bar{\beta}$ for M_i sampled animals, assuming the total pool of animals is a representative sample. Finally, for each M we calculated the grand mean $\bar{\beta}_G$ and the empirical 95% interval of $\bar{\beta}$. The value of M at which the empirical interval no longer contains 0 is the observed minimum M necessary to estimate β with 95% confidence, M_{obs}, and should correspond to M_{pred}

Empirical Validation: *N*

The M validation procedure described above assumes that, on average, sufficient relocations N were available to estimate M. Now we consider: for a given individual-level selection coefficient β, do we have sufficient N to reject the null hypothesis for a given animal? We randomly sampled 1 animal from each dataset and calculated $\text{Var}[R(X_\beta)]$ within the animal's 99% range using the animal's specific RSF model coefficients as β. From this we calculated the predicted number of relocations N_{pred} necessary to estimate β given $\text{Var}[R(X_\beta)]$ (i.e. $N_{0.05,0.05}(\beta)$; Equation (3)).

For comparison, we resampled N_{sam} relocations with replacement from the animal's dataset, where $N_{sam} = \left\lfloor \frac{iN_{total}}{50} \right\rfloor$, $i \in \{1, \ldots, 25\}$, and N_{total} is the total number of relocations recorded for that animal. This unconventional sequence was selected because (i) it produced a compara-

ble number of observed values of N to that in the M validation procedure (25 observed N vs. 29 pairings of M_{pred} and M_{obs}) while (ii) keeping the increments small enough to retain detail given that estimates of N can be orders of magnitude larger than those of M. We generated 4000 samples of $N_{sam,i}$ relocations and fit an RSF to each individual sample (i.e. 4000 RSFs assuming $N_{sam,i}$ relocations). We retained all originally generated available locations in each RSF so as to maintain a constant availability kernel between RSFs with different relocations. We then calculated the mean selection coefficient $\bar{\beta}$ and its 95% empirical interval at a given N_{sam}. The value of N_{sam} at which the empirical interval no longer contains 0 is the observed minimum N necessary to reject $H_0 : \beta = 0$ at significance level $p \leq 0.05$, N_{obs}, and should correspond to N_{pred}.

Results

The equations (3, 8) at the basis of our methods provide analytically predicted values for the number of relocations N and the number of animals M required to paramaterize an RSF. Simple 1-to-1 plots of N_{pred} vs. N_{obs} and M_{pred} vs. M_{obs} across all 10 species revealed remarkable agreement between observation and prediction (Fig. 3). Interestingly, 1 outlier was identified for N and 1 for M. Visual inspection of the data revealed that these outliers occurred alongside availability samples within individual RSFs that did not properly describe the true spatial integral of resource availability (i.e. $\int_\Omega A(\mathbf{x}')W(\mathbf{x}')d\mathbf{x}'$; Equation (1)). That is, the 1:1 used/available sampling protocol undersampled the available space. Thus, N_{pred} and M_{pred} can be sensitive to insufficient spatial sampling of availability, and care should be taken to to avoid such undersampling before applying these methods.

Given this, we then asked, what is the role of the definition of availability (sensu Johnson 1980) in shaping these relationships? Our original calculations of N_{pred} and M_{pred} used individual availability (i.e. each animal has its own available resources within its unique 99% KDE). We repeated our calculations of N_{pred} and M_{pred}, and bootstrap estimation of N_{obs} and M_{obs}, using 2 additional availability definitions that varied the spatial extent of availability for a given animal:

(i) within the entire collection of 99% KDEs (i.e. animals have access to resources within all KDEs equally), and (ii) within the entire site (i.e. animals have access to all resources within the study site, including those outside of 99% KDEs). This mimics the problem of sufficiently sampling availability described above, but now availability is driven by conceptual or ecological definitions rather than by the sampling protocol itself. Similar consistency in $\hat{\beta}$ was observed across M within a given definition of availability, but the sign and magnitude of $\hat{\beta}$ varied with availability from individual- to site-level (Fig. 4). Despite the change in sign and magnitude, Equation (8) is able to calculate M_{pred} consistent with observation across availability definitions. By inclusion, given that N_{pred} is a component of M_{pred} (see Equation (5)), we also observe that Equation (3) is consistent with observation across availability definitions.

Lastly we asked, what are the primary drivers of N_{pred} and M_{pred} as estimated by Equations (3, 8)? A key outcome of our method is that this question can be answered analytically, by simply inspecting Equations (3, 8). Equation (3) shows that N_{pred} is inversely correlated to both $\text{Var}[R(X_\beta)]$ and β^2, indicating that as either landscape variation or selection strength increase, so must N_{pred}. Similarly, because β^2 is contained in the denominator of Equation (8), M_{pred} must decrease with increasing selection strength. To demonstrate this graphically, we plotted log-log regressions of N_{pred} and M_{pred} against $\text{Var}[R(X_\beta)]$ and $|\beta|$, respectively, using data from all 10 species to evaluate whether these analytical predictions bear out under real data scenarios (Fig. 5). Per the analytical predictions, both N_{pred} and M_{pred} declined as their respective predictors (landscape variation or habitat selection strength) increased. It is also worth noting that inclusion of both predictors within the same log-log regression (i.e. M_{pred} as a function of both $\text{Var}[R(X_\beta)]$ and $|\beta|$) returned $R^2 = 1$, as expected given that N_{pred} and M_{pred} are determined only by $\text{Var}[R(X_\beta)]$ and β.

Discussion

Conventional wisdom regarding sample size in RSA holds that a sample size of $M \geq 30$ animals tagged is necessary for consistent and reliable inference (Leban *et al.* 2001; Hebblewhite & Haydon 2010). Convention also holds that more complex landscapes (i.e. those with higher landscape variance $\mathrm{Var}[R(X_\beta)]$) require more relocations per animal N to characterize selection (Wisz *et al.* 2008). Our analytical models and validation procedures return a contrasting set of results, contradicting conventional wisdom. First, we found that M_{pred} was often (but not always) substantially less than 30, and this prediction strongly agreed with observation based on resampling of GPS-based telemetry across a variety of ecologically contrasting species (Figs. 3, S1-S20). Strikingly, our analytical results show conclusively that M can only decline with increasing absolute magnitude of β (Equation (8)), indicating the most biologically relevant effects (i.e. those with the greatest $|\beta|$) can often be estimated with only a few animals (Fig. 5). This reveals important ethical and budgetary implications for wildlife studies. For example, consider the mule deer dataset containing 106 tagged individuals (Table S1). Our findings show that the strongest effects on the utilization distribution (i.e. selection for temperature, evergreen forest, and shrublands) may be estimated with fewer than 20 animals (Fig. S20), i.e. 80% fewer animals than were used. This means that, using a conservative estimate of US$2,450 for each GPS collar and data fees (K. L. Monteith, pers. obs.), if the sole aim of the study were to identify the relevant resource drivers of animal distributions as in typical RSF studies, this project would have overspent by $210,700 (excluding researcher/technician effort, which has significant cost in itself). Compared to the popular approach of tagging as many animals as possible and constructing phenomenological models to identify ecological mechanisms post hoc (colloquially referred to as "collar-and-foller"; Dunn 2004; Fieberg & Johnson 2015), our analytical results suggest researchers start with efforts aimed at constructing a priori hypotheses and associated models, then use our Equations (3, 8) to estimate the number of animals and locations per animal required for the study aims (Johnson *et al.* 2015).

Second, N_{pred} (the number of relocations per individual) also strongly agreed with observation, with both predicted and observed N in the 1000s or larger (Fig. 3). This agrees with findings that within-replicate sample sizes should generally be large (e.g. Wisz *et al.* 2008); however, our analytical expressions also conclusively demonstrate that N is directly calculable (Equation (3)) and as with M is expected to generally decline with increasing $\mathrm{Var}[R(X_\beta)]$ and β. These conclusions for both M and N are not only analytically proven but are additionally supported by real data bearing out the analytical predictions (Figs. 3–5). As such, our findings demonstrate that not only are M and N imminently calculable given a known landscape and some expectation of β, but the expected trends in M and N with respect to landscape complexity and the strength of animal preference are precisely opposite those predicted by conventional wisdom and previous studies.

Why are our results contrary to so much of the preceding literature? One possibility could lie in the "golden rule" of sample size, i.e. that $M \geq 30$ is required for a sample size sufficient to invoke the Central Limit Theorem and assume a roughly normal distribution of possible sample means (Aho 2014, p. 154), or to ignore non-normality because a model structure is somehow "robust" to non-normality (e.g. Hector 2015, p. 48). This is reinforced by an absence of mathematical attention to the sample size question. Previous studies have used simulation or empirical analyses to explore sample size sufficiency within particular species or systems (e.g. Leban *et al.* 2001; Loe *et al.* 2012; Sequeira *et al.* 2019), leading to conclusions that are quite specific to a given study but then are widely adopted as inferring pattern across all systems. By defining the problem mathematically (i.e. at what values of M and N do we reject the null hypothesis $100(1-\alpha)\%$ of the time at significance p?), we instead arrive at general analytical solutions that then may be tested with simulations and empirical analyses that are specifically designed for those solutions, rather than relying on intuitive but incorrect assumptions about the relationships between landscape variation relative to selection strength and RSA sample size sufficiency.

Our calculations show that the required M and N for a given study are dependent entirely on $|\beta|$ and $\mathrm{Var}[R(X_\beta)]$. The latter can be directly calculated given a landscape and an expectation

for β, but selecting an appropriate expected β is a critical step in estimating M and N. For *a priori* planning this could be accomplished using expert knowledge and previous literature; however, there may be no conceivable prior expectation of β in some RSA exercises. In such a case, one may elect to perform for example a sensitivity analysis given a range of β to select conservative estimates of M and N. Further, observe that β is often affected by a variety of ecological phenomena, including resource availability, competitor density, and seasonal effects (Mysterud & Ims 1998; McLoughlin *et al.* 2010; van Beest *et al.* 2016; Paolini *et al.* 2018). This implies that Equations (3 & 8) estimating N and M respectively are in fact hierarchical with dependencies not only on landscape variance (i.e. $\text{Var}[R(X_\beta)]$) but also landscape composition and structure as they determine β. In scenarios where we are uncertain about possible values of β, we may construct informed models suggesting likely values of β given an expectation for how the animal should behave as resource availability changes (e.g. generalized functional response models; Matthiopoulos *et al.* 2011). Such a hierarchical approach "borrows" information from the functional response model to provide a more ecologically informed range of possible β for a sensitivity analysis (Hobbs & Hooten 2015).

Our results also provide new insight into the importance of sufficient spatial sampling of availability. There was 1 outlier in the 1-to-1 comparison of N_{pred} and N_{obs}, and 1 in that of M_{pred} and M_{obs} (Fig. 3). These occurred because the 99% range of the animals under observation was so large, and the underlying landscape rasters so finely grained, that our 1:1 use/availability sample did not accurately portray the spatial integral of availability $\int_\Omega A(\mathbf{x}')W(\mathbf{x}')d\mathbf{x}'$ (Equation (1)). This caused M_{pred} and N_{pred} to be based on a different, incomplete availability set compared to the fitted RSFs. This highlights an unexpected but critical conclusion: the sampling intensity for availability in RSF-styled models should be only as large as necessary to correctly characterize the availability integral. Previous RSF-styled studies (including SSF) have almost exclusively sampled availability as we did here using ratios (i.e. 1:1, 1:10, 1:100, etc.; e.g. Boyce & McDonald 1999; Fortin *et al.* 2005; Street *et al.* 2016). This encourages either sampling at an intensity insufficient to approximate the spatial integral (as occurred here for outlying points in Fig. 3), or at

too great an intensity leading to overinflated sample sizes and biased standard errors, confidence intervals, and p-values. Both scenarios may affect inference, but despite these issues no general rule has been promoted for availability sampling in RSA. Based on our findings, we propose that this rule should be regular (non-random) sampling at a spatial interval equal to the resolution of the underlying landscape data such that every possible location within the availability boundary is considered. This would produce an availability observation for every raster pixel and thus overlap between used and available locations. Although it is suggested that such overlap is to be avoided (e.g. Wisz *et al.* 2008), logically a used location must also be available otherwise it cannot be selected, and removing used locations from availability can potentially omit important effects from the availability sample. Our equations indicate that this overlap is required by the mathematics of resource selection.

This finding reinforces that defining resource availability at the scale of the estimated model is a critical first step in planning a RSA. Our multi-scale analysis of mule deer produced remarkably different estimates for M at each of the three definitions of availability (site-wide, population-wide, and individual availability; Fig. 4), indicating that failure to properly define the available space can lead to incorrect estimates of both M and N. This is not a new finding; the importance of properly defining what is available for an animal to select is a long-standing issue in RSA research (e.g. Johnson 1980; Boyce & McDonald 1999; Fortin *et al.* 2005). However, the difficulty of calculating M and N for planning a RSA study increases with the biological scale of the intended model. Site-wide availability assumes all animals have access to resources on the entire landscape and is similar in concept to first-order selection (i.e. where the species is located; Johnson 1980), but availability may be sampled as a regular grid across the entire site. Population-wide availability refines the scale toward second-order selection (i.e. where animals situate their home ranges), but accurately defining a perimeter for the likely population range *a priori* within which to sample availability is non-trivial. This becomes even more difficult under individual availability; how can we anticipate the size and placement of individual home ranges? A feasible solution may be to delineate population boundaries and within this delineation generate random ranges

with area determined by the literature and expert knowledge. This would enable calculation of an average theoretical availability for any animal in the study site with appropriate standard error. This could then be used to produce an average prediction for M and N, and associated confidence limits, across the average home range composition.

We approached this analysis with the specific intention of evaluating how many GPS-tagged animals M are needed for RSF estimation, but there are many RSF applications that do not seek M or require GPS-tagging (e.g. plant distributions). For example, RSAs estimated for rare species will typically lack sufficient data for individual-based estimation of the utilization distribution $u(x)$ such that M is irrelevant and only N need be evaluated. RSAs can be sensitive to small sample sizes (Wisz *et al.* 2008), yet they often generate accurate predictions for rare species with small datasets (McCune 2016), suggesting that for some rare species smaller N is sufficient to achieve a robust model. Our findings permit evaluation of this. Consider a hypothetical scenario where RSA is conducted for a rare species with 100 observations and β is recorded. Here, Equations (3–4) could be used to calculate N_{pred} as a *post hoc* metric of confidence assuming β is the true population/species-level average selection coefficient. If $N_{pred} \leq 100$, then one could trust the outcome of the RSA; conversely, $N_{pred} > 100$ would indicate additional data collection is necessary. Where that is not possible, one could systematically adjust z_α and $z_{p/2}$ (Equation (3)) to determine the percent confidence interval that rejects the null hypothesis $H_0 : \beta = 0$ and establish a degree of confidence for model outcomes. Although there are issues with this approach (e.g. individual variation is ignored), this is a limitation of small datasets and not the equations identified here. Similarly, although we performed validation using GPS-based datasets, Equation (3) is agnostic to how data are collected and may also be applied to sessile organisms. Provided we can plausibly accept that β is roughly true and individual variation is either minimal or accommodated by the population-level β (presumably what has been estimated), our equations may be easily extended to evaluate most any RSA-based study.

We must emphasize that although M may only decline with increasing β, Equation (3) allows for a turning point to occur such that N initially decreases with $|\beta|$ but eventually increases at

very large $|\beta|$ (see Supplemental Information, Equation (25)). When selection strength is particularly strong, smaller sample sizes make it much more likely to obtain perfect separation between used and unused resources. In such a case one must collect more data to observe the animal not using a resource unit it should strongly prefer (or in the case of negative selection, to observe it using a resource it should strongly avoid). Practically, this means that sampling intensity for RSA is a greater concern for specialist organisms than generalists because specialists should exhibit typically larger $|\beta|$ for preferred/avoided resource units than generalists. Although the equations identified here allow us to directly calculate N for any landscape and expected selection strength, we should generally expect that specialists will require larger N for precise RSA estimation.

The equations identified here explicitly evaluate the compatibility of a dataset with a given hypothetical model (i.e. β). Calculating their solutions across gradients of N and M reveals how the number of data points (relocations) and number of replicates (animals) affect determination of compatibility. Rather than the values of N and M required to achieve statistical significance, we instead suggest these be used to determine the relevant sample sizes necessary to achieve "consistent" results, i.e. if we increase sampling intensity would we see substantial change in estimated coefficients? From this perspective, we conclude that the number of animals M required to consistently estimate the most biologically relevant effects in an RSA can be well below commonly touted sample size thresholds (i.e. $M \geq 30$), particularly when selection strength is strong (Fig. 3, 5). Moreover, the number of required relocations N can also be quite small but tends toward larger sample sizes when landscape variation is small. The sufficiency of samples sizes M and N is dependent entirely on the strength of selection ($|\beta|$) and landscape variation with respect to selection strength ($\text{Var}[R(X_\beta)]$). Rather than simply reporting sample sizes in RSA studies, researchers should pay explicit attention to the effect their sample size has on their findings. Regardless of study organism, ecosystem, or scenario, our equations may be equally applied to *any* RSF-based study to evaluate the consistency of expected outcomes given a dataset of a particular size. This will partially address the so-called "replicability crisis" by explicitly characterizing the consistency of model outputs in relation to sample sizes and effect

sizes, thereby increasing reader (and reviewer) confidence in such studies. Similarly, editors and reviewers should abandon preconceived notions of what makes a sufficient sample size in RSA in favor of evaluating the sensitivity of findings to sample size based on the mathematical rules identified here, for it is also feasible (and indeed demonstrable) that consistent findings can be achieved with as few as $N = 100$ relocations per animal and $M = 2$ animals (Fig. 3). Because M and N can be easily calculated provided knowledge of ecological and landscape effects, we argue that such calculations should henceforth be a mandatory component for all RSA studies.

Acknowledgments

We thank the Movement Ecology Special Interest Group of The British Ecological Society for valuable discussion regarding this topic, in particular Marie Auger-Méthé. GMS thanks the Mississippi Agricultural and Forestry Experiment Station (MAFES); the Forest and Wildlife Research Center (FWRC); the United States Department of Agriculture National Institute of Food and Agriculture (USDA NIFA); and the Mississippi Department of Wildlife, Fisheries, and Parks (MDWFP) for supporting this research and associated data collection. JRP thanks the School of Mathematics and Statistics at the University of Sheffield for granting him study leave which has helped enable the research presented here. CMP, LJN, and EV respectfully acknowledge that Riding Mountain National Park is the traditional homeland of the Anishinabe People and the Métis Nation, within Treaty 2 territory and at the crossroads of Treaties 1 and 4. Contributions of SD and BKS were partially supported by the Mississippi State University Extension Service (MSUES), FWRC, and MDWFP. Contributions of JCB, OER, PES, GD, DAK, and DAB were partially supported by USDA Animal and Plant Health Inspection Service (APHIS), Wildlife Services (WS), National Wildlife Research Center (NWRC), and U.S. Department of Energy (DOE) through Cooperative Agreement number DE-FC09-07SR22506 with the University of Georgia Research Foundation. Contributions of ARR and JMF were supported by the Ontario Ministry of Natural Resources and Forestry (OMNRF). Contributions of KLM and SPD were supported

by Wyoming Game and Fish Department (WGFD), Bureau of Land Management (BLM), Muley Fanatic Foundation, Boone and Crockett Club, Wyoming Wildlife and Natural Resources Trust, Knobloch Family Foundation, Wyoming Animal Damage Management Board, Wyoming Governor's Big Game License Coalition, Bowhunters of Wyoming, Wyoming Outfitters and Guides Association, United States Forest Service (USFS), and United States Fish and Wildlife Service (USFWS). Contributions of CMP, LJN, and EV were supported primarily by Parks Canada Agency (Riding Mountain National Park of Canada) and the Natural Science and Engineering Research Council of Canada (NSERC). Contributions of FMvB, LTB, and NMS were supported by the AUFF Starting Grant (AUFF-F-2016-FLS-8-16).

Author Contributions

GMS conceived and directed the project and developed the validation and resampling framework. JRP derived the analytic expressions, and GMS and JRP conducted the data analyses in collaboration with LB. GMS, JRP, and LB wrote the manuscript. Remaining authors collected and contributed data and contributed equally to edits and revisions.

Data Accessibility

RSF and landscape data will be uploaded to a permanent repository following formal acceptance of this manuscript for publication.

References

Aho, K.A. (2014). *Foundational and Applied Statistics for Biologists using R*. CRC Press, Boca Raton, FL.

van Beest, F.M., McLoughlin, P.D., Mysterud, A. & Brook, R.K. (2016). Functional responses in habitat selection are density dependent in a large herbivore. *Ecography*, 39, 515–523.

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

Dunn, W.C. (2004). More suggestions for raising the bar for conservation - a response to anderson et al. *Wildlife Society Bulletin*, 32, 594–597.

Elith, J. & Leathwick, J.R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40, 677–697.

Fieberg, J. & Johnson, D.H. (2015). MMI: multimodel inference or models with management implications? *Journal of Wildlife Management*, 79, 708–718.

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

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

Guillera-Arroita, G. & Lahoz-Monfort, J.J. (2012). Designing studies to detect differences in species occupancy: power analysis under imperfect detection. *Methods in Ecology & Evolution*, 3, 860–869.

Hebblewhite, M. & Haydon, D.T. (2010). Distinguishing technology from biology: a critical review of the use of GPS telemetry data in ecology. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365, 2303–2312.

Hector, A. (2015). *The New Statistics with R: An Introduction for Biologists*. Oxford University Press, Oxford, UK.

Hobbs, N.T. & Hooten, M.B. (2015). *Bayesian Models: A Statistical Primer for Ecologists*. Princeton University Press, Princeton, NJ.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer, New York, NY.

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

Johnson, P.C.D., Barry, S.J.E., Ferguson, H.M. & Müller, P. (2015). Power analysis for generalized linear mixed models in ecology and evolution. *Methods in Ecology & Evolution*, 6, 133–142.

Leban, F.A., Wisdom, M.J., Garton, E.O., Johnson, B.K. & Kie, J.G. (2001). Effect of sample size on the performance of resource selection analyses. In: *Radio Tracking and Wildlife Populations* (eds. Millspaugh, J.J. & Marzluff, J.M.). Academic Press, New York, NY, chap. 11, pp. 291–307.

Loe, L.E., Bonenfant, C., Meisingset, E.L. & Mysterud, A. (2012). Effects of spatial scale and sample size in GPS-based species distribution models: are the best models trivial for red deer management? *European Journal of Wildlife Research*, 58, 195–203.

Manly, B.F.J., McDonald, L.L., Thomas, D.L., McDonald, T.L. & Erickson, W.P. (2002). *Resource Selection by Animals: Statistical Design and Analysis for Field Studies*. 2nd edn. Kluwar Academic Publishers, Dordrecht, The Netherlands.

Matthiopoulos, J., Hebblewhite, M., Aarts, G. & Fieberg, J. (2011). Generalized functional responses for species distributions. *Ecology*, 92, 583–589.

McCune, J.L. (2016). Species distribution models predict rare species occurrences despite significant effects of landscape context. *Journal of Applied Ecology*, 53, 1871–1879.

McLoughlin, P.D., Morris, D.W., Fortin, D., Vander Wal, E. & Contasti, A.L. (2010). Considering ecological dynamics in resource selection functions. *Journal of Animal Ecology*, 79, 4–12.

Mysterud, A. & Ims, R.A. (1998). Functional responses in habitat use: availability influences relative use in trade-off situations. *Ecology*, 79, 1435–1441.

Paolini, K.E., Strickland, B.K., Tegt, J.L., VerCauteren, K.C. & Street, G.M. (2018). Seasonal variation in preference dictates space use in an invasive generalist. *PLoS One*, 13, e0199078.

Potts, J.R., Auger-Méthé, M., Mokross, K. & Lewis, M.A. (2014). A generalized residual technique for analysing complex movement models using earth mover's distance. *Methods in Ecology and Evolution*, 5, 1012–1022.

Renner, I.W. & Warton, D.I. (2013). Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics*, 69, 274–281.

Schlather, M., Malinowski, A., Oesting, M., Boecker, D., Strokorb, K., Engelke, S. *et al.* (2016). Randomfields: Simulation and analysis of random fields, r package. *Webpage http://CRAN. R-project. org/package= RandomFields*.

Sequeira, A.M.M., Heupel, M.R., Lea, M.A., Eguíluz, V.M., Guarte, C.M., Meekan, M.G. *et al.* (2019). The importance of sample size in marine megafauna tagging studies. *Ecological Applications*, In Press, e01947.

Street, G.M., Fieberg, J., Rodgers, A.R., Carstensen, M., Moen, R., Moore, S.A., Windels, S.K. & Forester, J.D. (2016). Habitat functional response mitigates reduced foraging opportunity: implications for animal fitness and space use. *Landscape Ecology*, 31, 1939–1953.

Strickland, M.D. & McDonald, L.L. (2006). Introduction to the special section on resource selection. *Journal of Wildlife Management*, 70, 321–323.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A. & NCEAS Predicting

Species Distributions Working Group (2008). Effects of sample size on the performance of species distribution models. *Diversity & Distributions*, 14, 763–773.

Figures

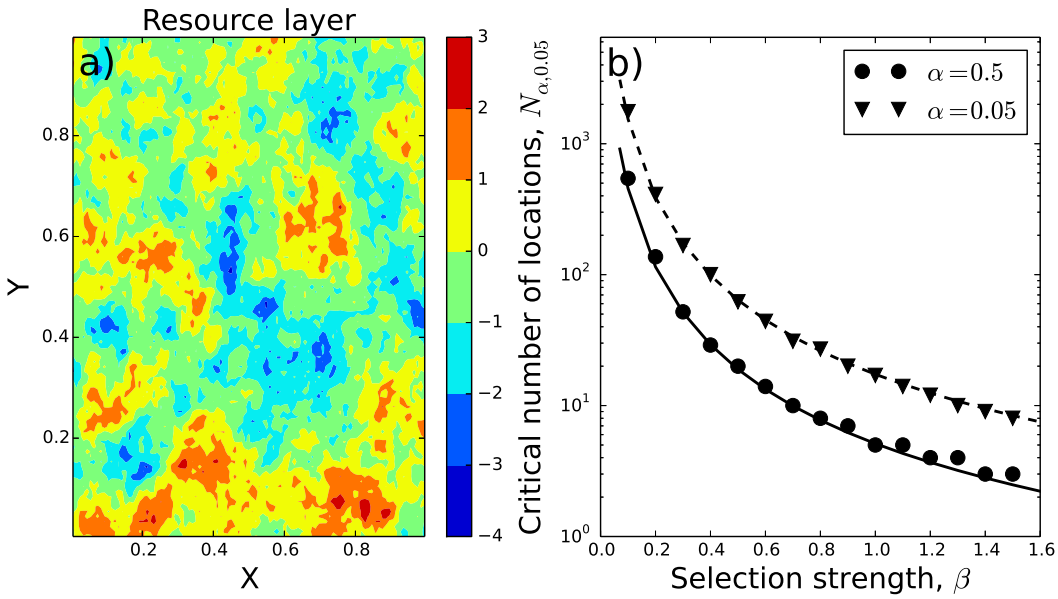

Figure 1: **Performance of analytic expression on simulated data.** Panel (a) shows a simulated resource layer, $R(\mathbf{x})$, which was used to construct the utilisation distribution (Equation 2) from which the simulated animal locations were samples. The circles (resp. triangles) in Panel (b) show the empirically-derived values of $N_{0.5,0.05}(\beta)$ (resp. $N_{0.05,0.05}(\beta)$), the minimum number of samples required so that there is a 50% chance (resp. 95% chance) of rejecting the null hypothesis that $\beta = 0$ at a significance level of $p = 0.05$. The solid line (resp. dashed line) in Panel (b) shows the corresponding analytic approximations given by Equation (3) and the remarkable agreement with the empirically-derived values.

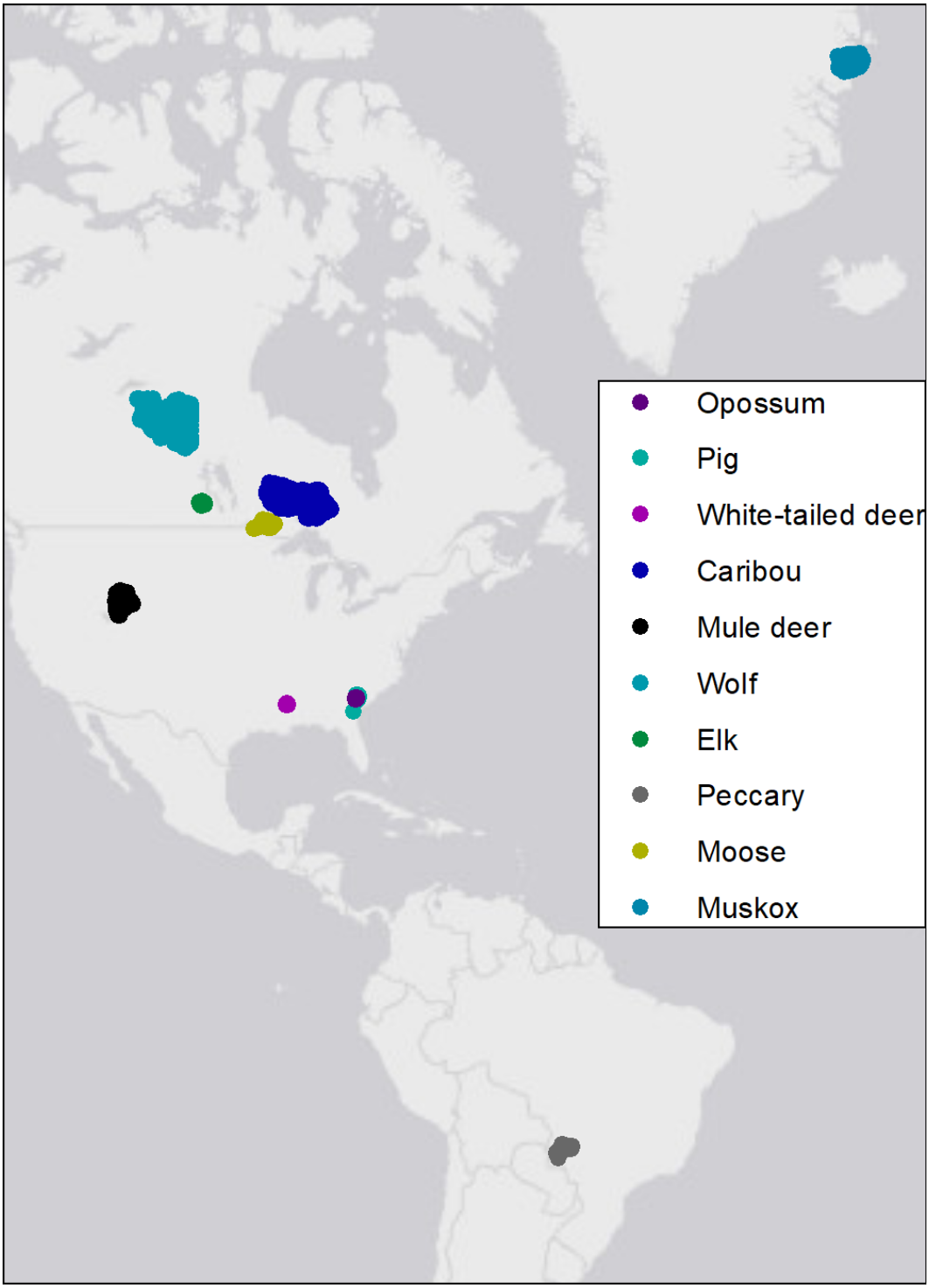

Figure 2: **Data distribution.** Geographic locations of GPS datasets (5,678,623 GPS relocations) across 511 individually collared members of 10 species.

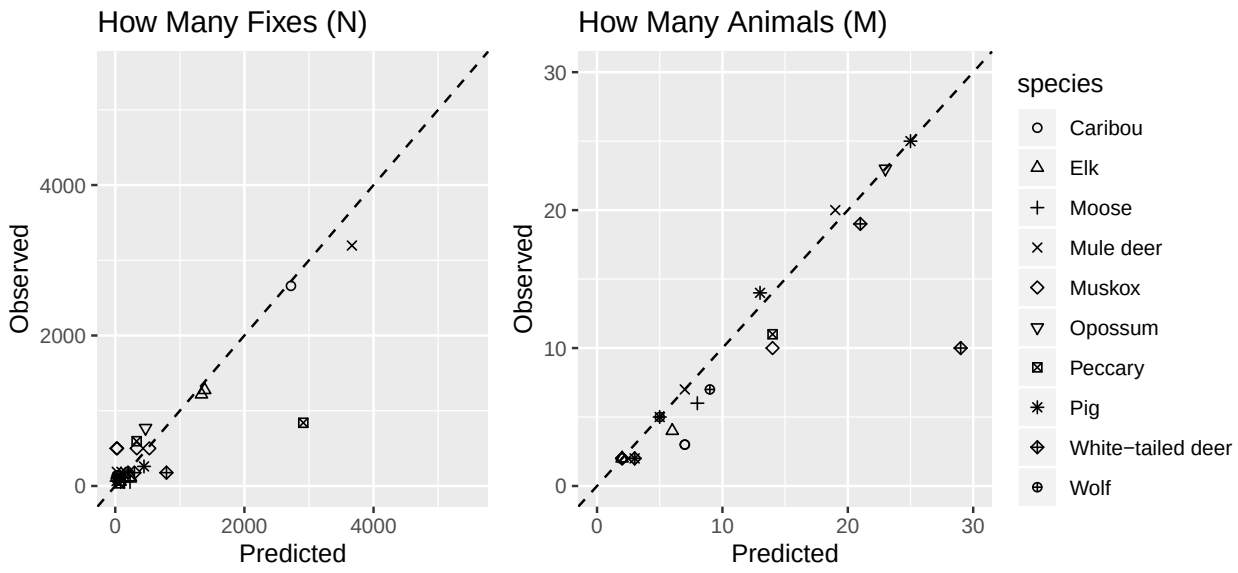

Figure 3: **1-to-1 comparison of predicted and observed M and N.** Three outliers are observed for N and one for M due to mismatch between sampled and true availability within the animals' 99% ranges. Dashed lines are those with gradient 1 crossing through the origin.

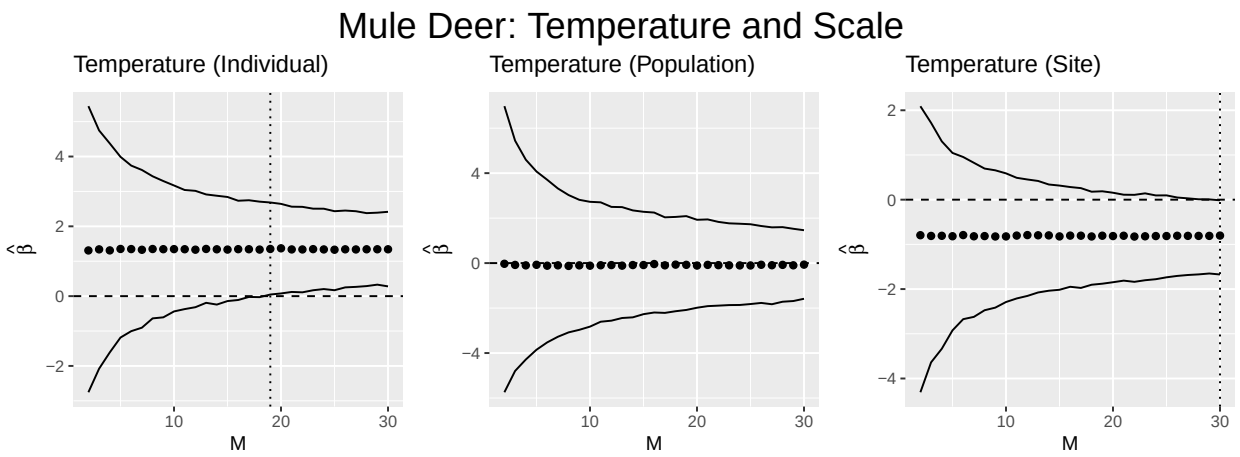

Figure 4: **Comparison of predicted M across orders of availability.** M_{pred} (vertical dotted line) changes depending on whether availability for the RSF is defined at the scale of the individual (each animal has its own available locations within its own 99% KDE), population (all animals have equal access to resources within all animal's 99% KDEs), or site (all animals have equal access to resources across the entire site). If no vertical dotted line occurs, then $M_{pred} > 30$.

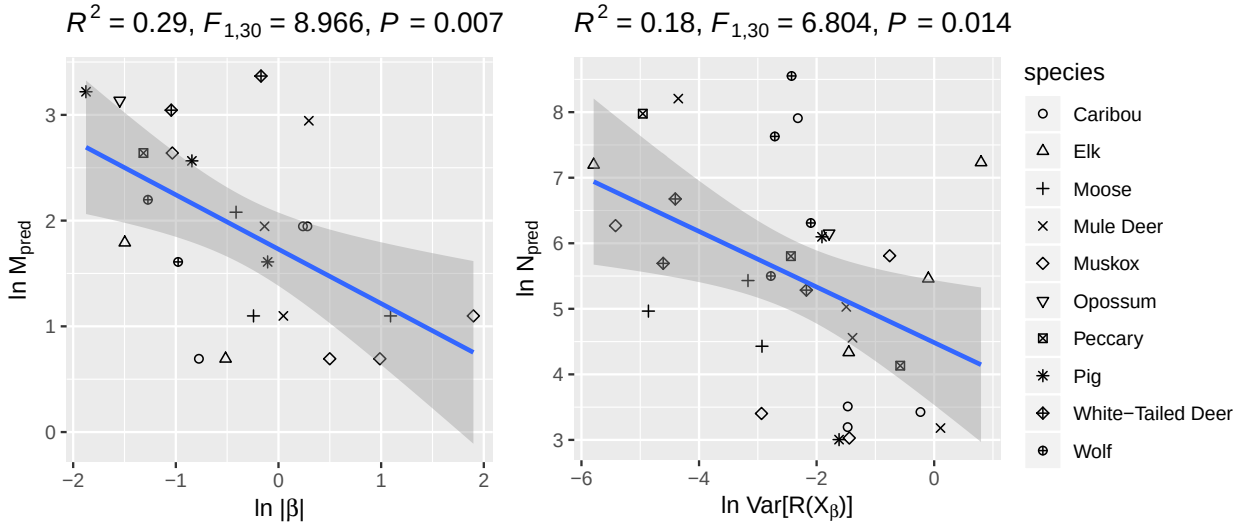

Figure 5: **Log-log regressions of predicted M vs. $|\beta|$ and predicted N vs. Var$[R(X_\beta)]$.** The predicted number of animals necessary M_{pred} declines with increasing absolute magnitude of selection (i.e. stronger effects require fewer animals to estimate), and the predicted number of relocations N_{pred} declines with increasing landscape complexity.