Received: 22 September 2020 | Revised: 24 March 2021 | Accepted: 23 June 2021

DOI: 10.1111/ddi.13384

METHOD

Diversity and Distributions WILEY

Generalized nonlinear models can solve the prediction problem for data from species-stratified use-availability designs

Nels G. Johnson[1] | Matthew R. Williams[2] | Erin C. Riordan[3,4]

[1]USDA Forest Service, Pacific Southwest Research Station, Albany, CA, USA

[2]National Science Foundation, National Center for Science and Engineering Statistics, Alexandria, VA, USA

[3]Laboratory of Tree-Ring Research, University of Arizona, Tucson, AZ, USA

[4]Riverside-Corona Resource Conservation District, Riverside, CA, USA

Correspondence
Nels Gordon Johnson, USDA Forest Service, Pacific Southwest Research Station, 800 Buchanan St, Albany, CA 94710, USA.
Email: nels.johnson@usda.gov

Editor: José Brito

Abstract

Aim: Habitat suitability modelling methods for presence-only species data are limited in their ability for making true predictions and are therefore often misused in ecological applications. A use-availability design—also known as a case–control design with contaminated controls—combines presence-only species data with a background sample of covariates where the species presence/absence is unknown. Assuming a log link function for the true probability of presence/absence, the use-availability data then can be analysed as a logistic regression model with a biased estimate of the intercept. Due to the biased intercept, the model is unable to make true predictions. Instead, ranking the "pseudo-predictions" from the model with biased intercept provides a viable alternative for making predictive inference in single-species models. We show the ranks are no longer conserved across species when such a single-species model is extended to multiple species, limiting predictive inference.

Innovation: By assuming a logit link function for the true probabilities of the presence/absence data, the resulting model allows for predictive inference even when extended to multiple species. We provide theoretical details justifying both fully Bayesian and large background sample asymptotic Bayesian generalized nonlinear model approaches.

Main conclusions: We illustrate how multiple species can be analysed using these approaches in R and Stan software using presence-only data for foundational shrubland taxa occurring in California, USA. Predictive inference highlights differences in habitat suitability rankings among individual species and among infraspecific taxa within a single species, improving the application of habitat suitability models for ecological restoration of southern California shrublands.

KEYWORDS

case–control, contaminated controls, habitat selection, logistic regression, species distribution models

1 | INTRODUCTION

Habitat suitability and species distribution modelling are broadly about quantifying where in space species/organisms are present versus absent and describing relationships between habitat data and the presence/absence status of a species/organism at a particular location. Statistical models for species presence/absence are models for binary response data, like logistic regression or some more general variant of it like generalized additive models (Wood, 2017) or classification trees (Breiman et al., 1984). Sometimes, the species presence/absence values are aggregated over some larger space to produce species counts that could be analysed using a Poisson log-linear model (McCullagh & Nelder, 1989) or some more general variant.

Frequently, species presence/absence suffers from some form of measurement error. For example, a bird identified by its song may go undetected because it did not sing while the study plot was being measured. The species is present, but marked as absent. Alternatively, its home range might not overlap exactly with the study plot so it goes undetected because it was someplace else during measurement. The species is absent and marked as absent, but it might make more sense in some projects to mark the species as present.

Both these scenarios could be considered issues with *prospective* studies. That is, studies where the species presence/absence status is unknown in each study plot and measurements are taken to determine their status. It is also common to find *retrospective* studies of species distributions, where species presence/absence values already known and records are sampled conditional on the presence/absence values of the species. This is known as a case–control design, where *cases* are presence values and *controls* are absence values (Hosmer et al., 2013).

Data from case–control studies can also be subject to measurement error. For example, records housed in herbaria, museum and databases may only contain locations where species were found to be present and no location records where species were found to be absent. This is known as presence-only data. That is, data where species presence/absence values are only available when the species is present. If the study design allowed for a census of all organisms of the species of interest, then presence-only data are equivalent to presence/absence data as any study plot missing the species status is by deduction equivalent to a record of species absence. When a census does not occur, we call location records *contaminated controls* when the habitat data are available, but the species/organism presence/absence status is unknown.

In the ecological literature, species distribution data in a retrospective study collected according to a case–control design with contaminated controls are called a use-availability design, where the *use* sample is a sample of cases and the *availability* sample is a sample of contaminated controls. Sometimes the availability sample is also called a *background* sample or *quadrature points*. A number of the technical details for analysing use-availability designs for single species are provided by Lancaster and Imbens (1996). A good review of the limitations of use-availability design for single species and how to address these limitations are provided by Keating and Cherry (2004).

The goal of this work is to layout an updated framework for analysing data collected according to use-availability designs for both single- and multiple-species models. To do this, we will update the model of Lancaster and Imbens (1996) so that it can be implemented as a generalized nonlinear model in both maximum likelihood and Bayesian frameworks. We will extend the updated model from one to multiple species for use-availability designs stratified (or blocked) by species. Along the way, we will update some of the inferential limitations described by Keating and Cherry (2004) of models for data from single-species use-availability designs. In particular, we will focus on limitation of estimating and predicting the probability of species presence/absence. Most notably, we will show a set of modelling choices that ensures cross species comparisons of estimated and predicted probabilities are interpretable, while highlighting another that leads to uninterpretable cross species comparisons.

The multiple-species case we consider should be referred to as a multiple-species use-availability design stratified (or blocked) by species, which is to say a list of species of interest is generated, then a separate use-availability sampling design is done for each species in series, and the model is for the combined data. For example, there may be a database with records for multiple species and a use-availability design is used to collect records for each species one at a time.

The remainder of the paper is organized as follows: First, we will describe the technical details of the use-availability design, some common modelling choices made, how they generalize from single to multiple species and their limitations. Next, we will discuss how to implement use-availability models as a generalized nonlinear model in software. We will then illustrate these methods on a real dataset from Riordan et al. (2018) of plant data collected in California, USA. We conclude with a discussion.

2 | METHODS

2.1 | Technical formulation of the use-availability design and model

The multiple-species case we are considering is a use-availability design stratified (or blocked) by species. Which is to say it is a series of single-species use-availability sampling designs implemented sequentially for each species of interest.

In the single-species case, we would consider a species s on a bounded landscape of interest (e.g. the state of California) represented by the union of a large finite number of point locations where an organism of species s could exist. Note: this implies the approach is a discrete-space approximation to a continuous-space problem. At each point-location i, there is an associated vector of habitat data $\mathbf{x}_{si} = (x_{si1}, x_{si2}, \ldots, x_{siK})$, composed of K habitat variables/covariates, and species presence/absence data y_{si}, such that $y_{si} = 1$ when the

species s is present, and $y_{si} = 0$ when the species is absent. We assume at each point location the species presence/absence was determined by a Bernoulli distribution with a probability of presence $P(y_{si} = 1 \mid \mathbf{x}_{si}) = p_{si}$ that is conditional on the habitat data. For the multiple-species case, we would have these same assumptions hold true for species $s = 1, 2, \ldots, S$.

2.1.1 | Sampling scheme and use-availability likelihood

In an ideal world, we would have a random sample of point locations stratified by species. That is, a random sample of $(y_{si}, \mathbf{x}_{si})$ for each of the S species where we observe true species presence/absence. We could then proceed with analysis using some method Bernoulli response data, like logistic regression. We would then proceed with inference about $\beta_{s,0}, \boldsymbol{\beta}_s$ and p_{si} as is pertinent to the project at hand.

Under use-availability sampling of point locations stratified by species, we take a random sample of size n_{s1} of point locations where species s is present (or assume the data collected is a random sample), called use data, and a random sample of size n_{s0} of point locations from the whole population of habit data (i.e., y_{si} could be 1 or 0), called availability data. The total sample size is $N_s = n_{s1} + n_{s0}$. Let $C_{si} = 1$ when point-location i is in our use sample for species s and let $C_{si} = 0$ when point-location i is in our availability sample for species s. Mathematically, the use sample is a sample from the distribution of the habitat given the species is present $f_{X_s \mid y_s = 1}(\mathbf{x}_s \mid y_s = 1)$ which by Bayes rule can be rewritten as $P(y_s = 1 \mid \mathbf{x}_s) f_X(\mathbf{x}_s) / P(y_s = 1)$. The availability sample is a sample from the marginal distribution of the habitat $f_X(\mathbf{x})$. Using these quantities, Lancaster and Imbens (1996) show the likelihood for use-availability sampling has the following form:

$$L_{C,X}(C, X \mid \beta_{1:S,0}, \beta_{1:S}, h_{1:S}, q_{1:S}) = \prod_{s=1}^{S} \prod_{i=1}^{N} \left[h_s \frac{p_{si} f_X(x_{si})}{q_s} \right]^{C_{si}} \left[(1 - h_s) f_X(x_{si}) \right]^{1 - C_{si}}, \tag{1}$$

where $q_s = P(y_{si} = 1)$ and $h_s = P(C_{si} = 1)$. Let us define $\psi_s = q_s(1 - h_s)h_s^{-1}$, $P(C_{si} = 1 \mid \mathbf{x}_{si}) = \tilde{p}_{si} = p_{si}(p_{si} + \psi_s)^{-1}$ and $\tilde{q}_s = q_s(q_s + \psi_s)^{-1} = h_s$. Appendix A contains a justification for parameterization in terms of ψ_s. For convenience, we will decompose the likelihood into its parts:

$$\begin{aligned} L_{C,X}(\mathbf{C}, \mathbf{X} \mid \beta_{1:S,0}, \boldsymbol{\beta}_{1:S}, \psi_{1:S}, q_{1:S}) &= L_{C|X}(\mathbf{C} \mid \mathbf{X}, \beta_{1:S,0}, \boldsymbol{\beta}_{1:S}, \psi_{1:S}) \\ &\quad \times L_X(\mathbf{X} \mid \beta_{1:S,0}, \boldsymbol{\beta}_{1:S}, \psi_{1:S}, q_{1:S} \text{ or } h_{1:S}), \end{aligned} \tag{2}$$

where

$$\begin{aligned} L_{C|X}(\mathbf{C} \mid \mathbf{X}, \beta_{1:S,0}, \boldsymbol{\beta}_{1:S}, \psi_{1:S}) &= \prod_{s=1}^{S} \prod_{i=1}^{N} \tilde{p}_{si}^{C_{si}} (1 - \tilde{p}_{si})^{1 - C_{si}}, \\ L_X(\mathbf{X} \mid \beta_{1:S,0}, \boldsymbol{\beta}_{1:S}, \psi_{1:S}, q_{1:S} \text{ or } h_{1:S}) &= \prod_{s=1}^{S} \prod_{i=1}^{N} \frac{1 - \tilde{q}_s}{1 - \tilde{p}_{si}} f_X(\mathbf{x}_{si}) \propto \prod_{s=1}^{S} \prod_{i=1}^{N} \frac{1 - \tilde{q}_s}{1 - \tilde{p}_{si}}, \end{aligned} \tag{3}$$

and where the likelihood $L_{C|X}$ has the form of the joint likelihood for Bernoulli distributed data, conditional on the habitat covariates, while the likelihood of L_X has form dependent on the distribution of the habitat covariates f_X.

2.1.2 | Practical details of estimation and inference

Convention is to use the conditional likelihood $L_{C|X}$ for inference as Bernoulli likelihoods are well studied with lots of tools for inference. Indeed, Lancaster and Imbens (1996) show that when we set $(1 - h_s)h_s^{-1} = n_{s0}n_{s1}^{-1}$, generalized method of moments estimates and constrained maximum likelihood estimates from $L_{C|X}$ will be the same as from $L_{C,X}$. For large N, the methods of moments and maximum likelihood estimators from $L_{C|X}$ are asymptotically equivalent. They are asymptotically unbiased and follow a normal distribution where the asymptotic covariance is the sandwich estimator from $L_{C|X}$. If the full likelihood was used, the asymptotic covariance would be the inverse Fisher information from $L_{C,X}$. The sandwich covariance produces larger variance estimates than the inverse Fisher information, which is the cost to using the conditional instead of the full likelihood.

In many practical settings, obtaining a large sample size N effectively means obtaining a large availability sample sizes n_{s0} for all S species. For example, when habitat data are generated from GIS layers of some landscape dataset. We show in Appendix B that for large availability samples, the joint likelihood $L_{C,X}$ is well approximated by $L_{C|X}$. As a result, the sandwich estimator of the covariance is unnecessary under this condition when using a maximum likelihood approach because the sandwich covariance collapses to the inverse Fisher information. Additionally, these derivations focus on a form $L_{C|X}$ parameterized in terms of ψ, as in Equation (3). Computer optimizers only need to solve two equations instead of the three of Lancaster and Imbens (1996) because it does not need to keep tract of the constraint.

An alternative to an asymptotic maximum likelihood approach to estimation and inference is taking a Bayesian approach. We will propose two Bayesian approaches which are easily implemented in R computer programming language (R Core Team, 2019). The first is an asymptotic Bayesian approach for large availability samples that, like the maximum likelihood approach, uses the conditional likelihood $L_{C|X}$ parameterized in terms of ψ instead of the full likelihood $L_{C,X}$. The second is a fully Bayesian approach that uses the full joint likelihood of $L_{C,X}$, including L_X. The fully Bayesian approach may be more appropriate for small-scale use-availability studies, rather than large, landscape-scale use-availability studies which motivated this work. The details of these Bayesian approaches are substantial. We will discuss the asymptotic Bayesian approach more in later sections and both Bayesian approaches in Appendix C.

2.2 | The prediction problem in use-availability designs

Thus far, we have not discussed the form of p_{si} when in use-availability designs. However, choice of p_{si} is key to solving the prediction model. Under large availability samples, the model using $L_{C|X}$ for inference can be written as follows:

$$\begin{aligned} C_{si} \mid \mathbf{x}_{si} &\sim \text{Bernoulli}\left(\tilde{p}_{si}\right), \\ \tilde{p}_{si} &= \frac{p_{si}}{p_{si} + \psi_s}. \end{aligned} \tag{4}$$

This is a nonlinear model. However, we can make it linearizable by choosing p_{si} to be the exponential probability function:

$$\begin{aligned} C_{si}|\mathbf{x}_{si} &\sim \text{Bernoulli}\left(\tilde{p}_{si}\right), \\ \tilde{p}_{si} &= \frac{\exp\left(\beta_{s,0}+\mathbf{x}_{si}\boldsymbol{\beta}_s\right)}{\exp\left(\beta_{s,0}+\mathbf{x}_{si}\boldsymbol{\beta}_s\right)+\psi_s} \\ &= \text{logistic}\left(-\log\psi_s+\beta_{s,0}+\mathbf{x}_{si}\boldsymbol{\beta}_s\right). \end{aligned} \tag{5}$$

This approach is equivalent to logistic regression of C_{si} on $\mathbf{x}_{si}$ where only a biased intercept $\beta^*_{s,0} = -\log\psi_s + \beta_{s,0}$ is identifiable. This is a very popular approach in practice because logistic regression is very easy to do in most statistical software languages, but it has a major issue. Specifically, without being able to identify $\beta_{s,0}$, it is difficult to return exact inference on p_{si} or other functions of $\beta_{s,0}$ without very strong prior assumptions about the value of q_s. This is known as the prediction problem in use-availability designs–biased estimates of p_{si} when the probability function is exponential. Keating and Cherry (2004) provide a nice review for single-species case.

One option to partially solve the prediction problem is to rank habitat. Because $\tilde{p}_{si}$ is a monotonic function of p_{si}, the ranks of $\tilde{p}_{si}$ will conserve the ordering of the ranks of p_{si}. However, this breaks down when multiple species are considered as each set of pseudo-probabilities $\tilde{p}_{si}$ is biased by a different amount $-\log\psi_s$ relative to the true probabilities p_{si} for each species. This is illustrated in Figure A1. Thus, the ranks of the pseudo-probabilities are only meaningful within species and not across.

Instead, we propose treating the use-availability model as the generalized nonlinear model that it is. By carefully picking the form of p_{si} such that all the model parameters are identifiable, we can recover inference on all model parameters and functions of them like p_{si}. This approach was suggested by Keating and Cherry (2004), but now software has advanced to the point that implementing generalized nonlinear models is much easier. In the next section, we will outline one such approach.

2.3 | A generalized nonlinear model for use-availability designs

Consider now the case where $p_{si} = \text{logistic}(\beta_{s,0} + \mathbf{x}_{si}\boldsymbol{\beta}_s)$, then a generalized nonlinear model that can be used easily in software has the following form:

$$\begin{aligned} C_{si} &\sim \text{Bernoulli}\left(\tilde{p}_{si}\right), \\ \tilde{p}_{si} &= \left[1+\exp\left(\alpha_{s,0}\right)+\exp\left(\alpha_{s,0}-\eta_{si}\right)\right]^{-1}, \\ \alpha_{s,0} &= \log\psi_s, \\ \eta_{si} &= \beta_{s,0}+\mathbf{x}'_{si}\boldsymbol{\beta}_s. \end{aligned} \tag{6}$$

Introducing η_{si}, the linear predictor of the original logistic model, should simplify nonlinear model parameter specification in a lot of software packages. It also clarifies how to extend the model to include predictors that are nonlinear (with known or unknown forms): change the right-hand side of the equation to whatever form is relevant to your research question. Parameterizing $\log\psi_s$ as $\alpha_{s,0}$ is useful because $\alpha_{s,0}$ is an unbounded quantity, whereas ψ_s is bounded below by 0. Unbounded quantities tend to have more stable estimates when fitting nonlinear models as it is impossible during optimization to either accidentally explore an area of the likelihood that is out of bounds or find an optimum there. It may be tempting to let $\log\psi_s$ also have covariate information on the right side of the equation, for example $\mathbf{z}'_{si}\alpha_s$. We caution against it without careful consideration as its interpretation implies the sampling design is not a use-availability design as we have outlined. A full exploration of the implications is beyond the scope of this work, but we do provide some additional details in Appendix D. For example, it could be appropriate to include covariate information such as date of sample or any further stratification such as repeatedly measured organisms within species.

2.4 | Illustration of the prediction problem on plant data from California

We use the data of Riordan et al. (2018, 2021). It is of common dominant shrubs and subshrubs in southern California chaparral, coastal sage scrub and alluvial scrub plant communities. For a full description of sampling protocols, data cleaning and data sources, see their work. In this dataset, there are multiple varieties and subspecies (infraspecies) from the same species. We treat each of these infraspecies as a separate 'species' for purposes of modelling. Taxonomic information is available in Table A2. Their use data were collected from herbarium records and field surveys.

The environmental covariate data are 30-year averages of climate data from 1951–1980 at 270 m resolution from the California Basin Characterization Model (Flint et al., 2013), a regional water-balance model that simulates hydrological responses to climate across the California hydrologic region. Variables names, descriptions, means and standard deviations are in Table A3. Availability data were randomly sampled from an area within 100 km of known species or infraspecies occurrences.

All analyses were done in R version 3.5.3 R Core Team (2019). We used the *brm* function in the BRMS package version 2.9.0 to generate samples from the posterior distribution. We ran four chains in parallel on four cores each for 4,000 iterations with 500 iterations of warm-up (aka burn-in). We set the max tree depth to 15 (default is 10) as recommended by the function to reduce some minor computational warnings during the warm-up phase of the NUTS algorithm. Some example pseudocode is available in Appendix E in Figure A2.

We placed normal priors on $\beta_{s,0}$ and $\beta_{s,0}$ with mean zero and standard deviation 5 and 2.5, respectively, and scaled the covariates to have mean 0 and standard deviation 1. This is a slightly more informative version of the normal prior in Ghosh et al. (2018). We opted for the more informative prior as the resulting posterior had better mixing in the NUTS sampler, reducing computational times from about 95 hr to about 40 hr on a Windows 10 PC with an Intel(R) Xenon(R) CPU E3-1245 v5 @ 3.50 GHz 3.50 GHz processor with

64 GB of RAM. We placed a normal priors with mean −2.05 and standard deviation 3.19 on each $\alpha_{s,0}$. As described in Table A1, this is approximately equivalent to placing normal priors with mean 0 and standard deviation 5 on the log-odds of both h_s and q_s.

Figures were made using the *ggplot* function in the GGPLOT2 package version 3.1.1 (Wickham, 2016).

Some comparisons are made to Model (5). Its parameters were estimated using the *glm* function in base R by fitting the model as a generalized linear model for binomial response data with logistic link function.

3 | RESULTS

Posterior means of the species presence/absence probabilities p_{si}^{logis} are plotted in Figure 1. From this figure, we can see that *Arctostaphylos glandulosa* subsp. *cushingiana* and *Eriogonum fasciculatum* var. *fasciculatum* appear to be rare everywhere. Also, *Acmispon glaber* var. *glaber*, *Arctostaphylos glandulosa* subsp. *glandulosa*, *Eriodictyon crassifolium* var. *crassifolium* and *Eriogonum fasciculatum* var. *foliolosum* appear to have high probability of being found in the coastal areas. *Eriogonum fasciculatum* var. *polifolium* is the only infraspecies to have high probability of being found in the desert region of southeast California.

Plots of the presence/absence pseudo-probabilities $\tilde{p}_{si}^{\text{exp}}$ as well as their within and across species scaled ranks are in Figures A4–A6 of Appendix F.2 along with a discussion highlighting the limitations of psuedo-probabilities and their ranks which was theoretically shown in a previous section. We use these computed ranks in Figures 2 and 3 to show how much inappropriately ranking pseudo-probabilities can lead to increased over- and under-ranking of habitat use when compared to the posterior means of p_{si}^{logis}. Figure 2 contains the difference of ranks between $\tilde{p}_{si}^{\text{exp}}$ and p_{si}^{logis} scaled by the max absolute difference in rank across all species. A value of 0.4 means $\tilde{p}_{si}^{\text{exp}}$ over-ranks p_{si}^{logis} by 40% and a value of – 0.4 it under-ranks by 40%. We can see that there is a lot of over-ranking going on for *Eriogonum fasciculatum*. This same process was done within species and plotted in Figure 3. These two figures have the colour-range scaled in the same way, so it is clearer when the scaled rank differences are closer to the true values for each species the amount of grey space increases. However, some infraspecies such as *Eriodictyon crassifolium* var. *crassifolium* still had issues. This could be because the Bayesian approach is akin to penalized likelihood approach thanks to the prior distribution, whereas the generalized linear model implemented by the *glm* function is a regular (i.e., non-penalized) likelihood approach. However, for such large sample sizes prior smoothing should be limited.

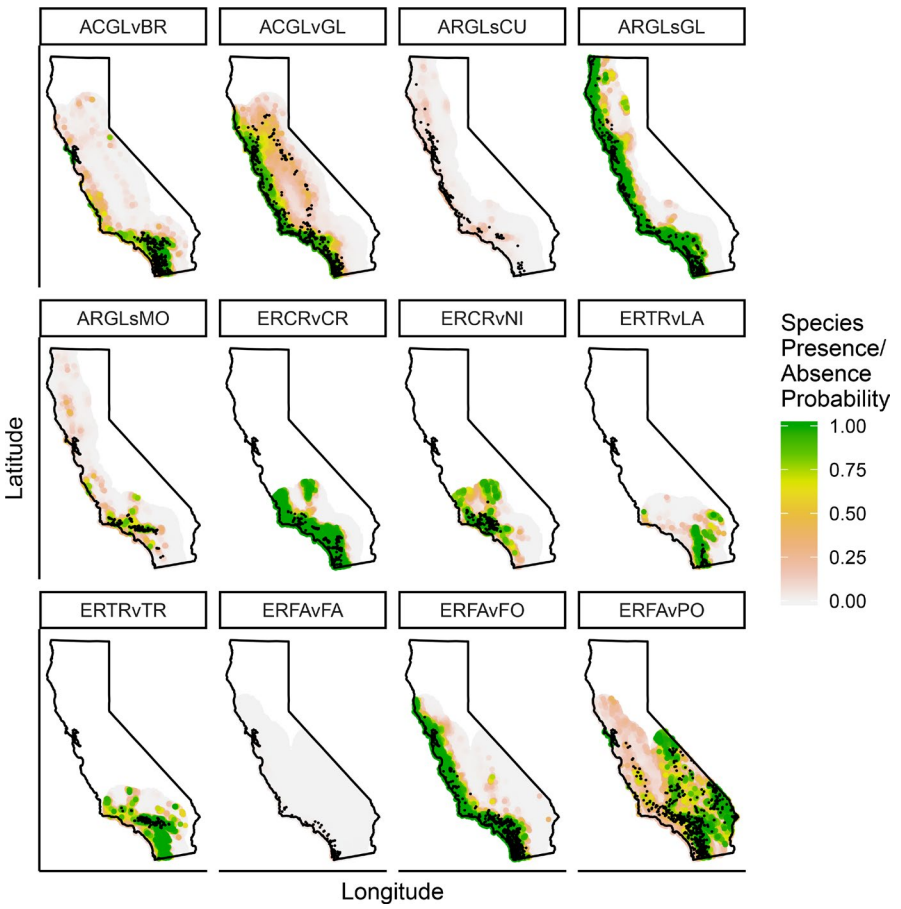

FIGURE 1 Posterior mean species habitat use probabilities from Model (6): p_{si}^{logis}. Black circles are observed species presences

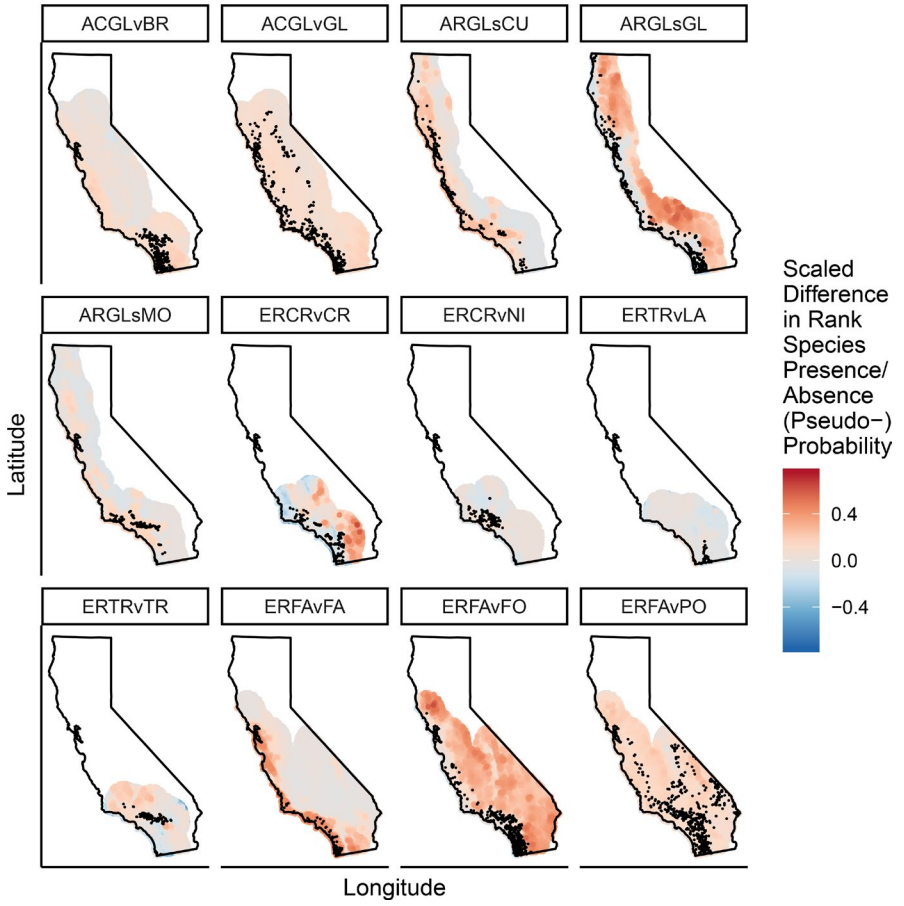

FIGURE 2 Difference of rank species habitat use pseudo-probability and probability from Models (5) and (6), respectively, scaled by the maximum absolute observed difference. Black circles are observed species presences. Red indicates over-ranking, blue indicates under-ranking, and grey indicates minimal over-/under-ranking

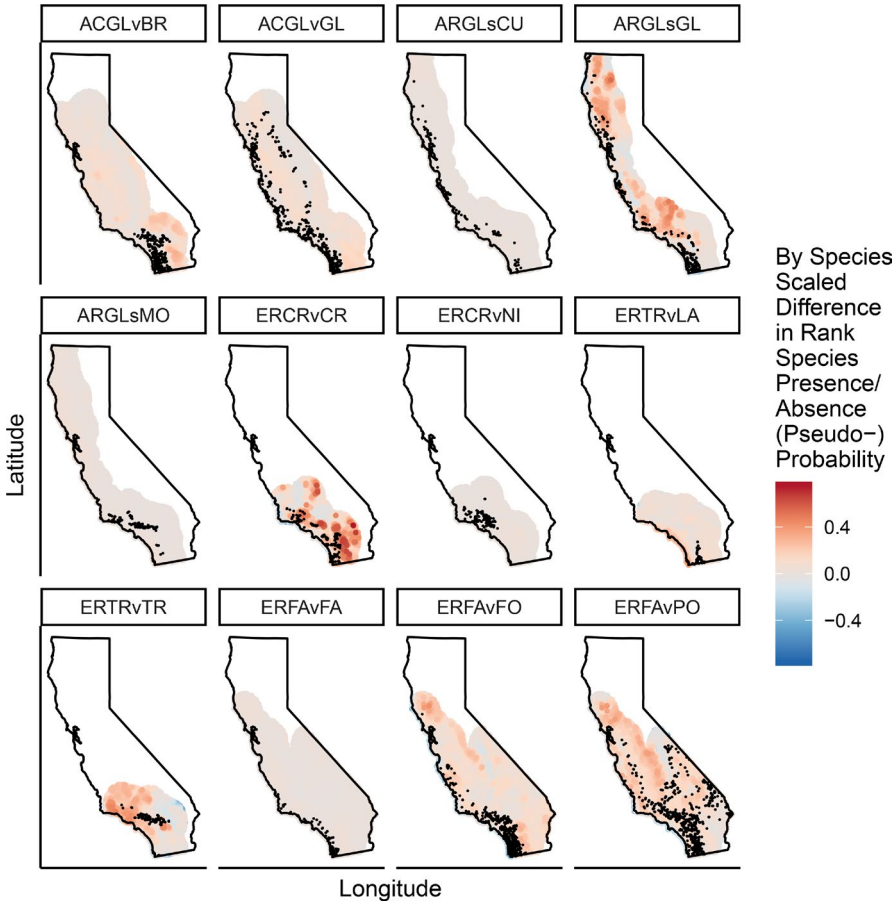

FIGURE 3 Within species difference of rank species habitat use pseudo-probability and probability from Models (5) and (6), respectively, scaled by the within species maximum absolute observed difference. Black circles are observed species presences. Red indicates over-ranking, blue indicates under-ranking, and grey indicates minimal over-/under-ranking

Additional results can be found in Appendix F. These results include forest plots of the covariate effects in Figure A3 and Bayes factors in Table A4.

4 | DISCUSSION

Our major contributions in this work are as follows: (a) identifying that some models for use-availability designs as described in Lancaster and Imbens (1996) can be implemented as generalized nonlinear models for binary response data and doing so allows us to make direct inference on presence/absence probabilities, (b) generalizing these models from one to multiple species when the sampling design is stratified by species, (c) identifying that ranked pseudo-probabilities from models for data collected from use-availability designs for multiple species does not conserve the order of the ranked probabilities from the presence/absence model and therefore can not be used for cross species habitat ranking, (d) developing an unrestricted parameterization of use-availability models to use in maximum likelihood and Bayesian inference, (e) developing approximate and fully Bayesian approaches to models for data from use-availability designs, (f) providing some guidance for prior selection when taking a Bayesian approach, and (g) showing how this can all be done with widely available software for statistical analysis.

We used our approach to highlight important differences in habitat suitability ranking between and within foundational shrub species. Inference on predicted habitat probabilities could be used to assess how species distributions will change under different land management scenarios, climate change scenarios, and other natural or interventionist changes to land and climate. It can also be used to select habitat variables that are important to explaining or predicting habitat use probabilities, as well as compare how the impact of said variables is different across species.

There are some pitfalls we have yet to mention. First is that this approach really relies on having informative habitat data. If none of the habitat data contain any information about y_{si}, then $p_{si} = q_s$ and the likelihood is degenerate. A second pitfall is having availability samples that are too small to contain some species presence values for rare species. When the availability sample is really a sample of controls $f_{X|Y=0}(\mathbf{x}_s \mid y_s = 0)$, rather than contaminated controls $f_X(\mathbf{x})$, then the derivations of the joint likelihood $L_{C,X}$ are not the same and the conditional likelihood $L_{C|X}$ becomes misspecified.

While the approaches we have described are for multiple-species use-availability designs, species is an arbitrary label and any stratification or blocking variable could be used its place. In our application, we used infraspecies. For organisms that move, the stratification variable could be organism ID in a single-species model. Depending on how sampling is done, it could be appropriate to stratify by database if multiple databases are used to generate the presence-only samples. It is important for the user to pay close attention to how the data were gathered when making these decisions.

ACKNOWLEDGMENTS

We would like to thank Jim Baldwin (USDA Forest Service, retired), Andy Hoegh (Montana State University), Inyoung Kim (Virginia Tech), Haiganoush Preisler (USDA Forest Service, retired) and Natalie van Doorn (USDA Forest Service) for their comments while developing this manuscript and Arlee Montalvo (Riverside-Corona Resource Conservation District, retired) and Jan Beyers (USDA Forest Service, retired) for contributing the California plant data.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/ddi.13384.

DATA AVAILABILITY STATEMENT

Data and code are archived at https://doi.org/10.2737/RDS-2021-0060. Please cite Riordan et al. (2021) when using the data.

ORCID

Nels G. Johnson https://orcid.org/0000-0002-8500-0194

REFERENCES

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411.

Flint, L. E., Flint, A. L., Thorne, J., & Boynton, R. (2013). Fine-scale hydrologic modeling for regional landscape applications: The California Basin Characterization Model development and performance. *Ecological Processes*, *2*, 1–21.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383.

Ghosh, J., Li, Y., & Mitra, R. (2018). On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*, *13*(2), 359–383.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons Inc.

Jepson Flora Project. (2017). *Jepson eFlora [Database]*. Retrieved from http://ucjeps.berkeley.edu/eflora/

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

Lancaster, T., & Imbens, G. (1996). Case-control studies with contaminated controls. *Journal of Econometrics*, *71*(1–2), 145–160.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall/CRC.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Riordan, E. C., Montalvo, A. M., & Beyers, J. L. (2018). *Using species distribution models with climate change scenarios to aid ecological restoration decisionmaking for southern California shrublands. Research Paper PSW-RP-270*. U.S. Department of Agriculture, Forest Service, Pacific Southwest Research Station.

Riordan, E. C., Montalvo, A. M., Beyers, J. L., Johnson, N. G., & Williams, M. R. (2021). *Southern California shrub and subshrub use-availability data*. Forest Service Research Data Archive.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.

Wood, S. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman and Hall/CRC.

BIOSKETCH

Nels Gordon Johnson is a mathematical statistician serving in the role of station statistician at the USDA Forest Service Pacific Southwest Research Station. There, he provides statistical support for research projects that span a diverse area of application including forestry, ecology, wildlife biology, natural resource management and social sciences. His statistical areas of interest include the design and analysis of landscape and species monitoring studies, design and analysis of studies involving measurement error and missing data, and Bayesian methods and computation.

Matthew Richard Williams is a mathematical statistician with the National Center for Science and Engineering Statistics at the National Science Foundation. Matt works in the Statistics and Methods Program at NCSES and provides statistical consulting and research support to NCSES, NSF, and the federal statistical system. Mattâ€™s current research interests include Bayesian analysis of complex surveys and data collections, formal privacy methods such as differential privacy, and combining multiple data sources for the use of evidence building and evaluation. Prior to working at NCSES, Matt worked for the Substance Abuse and Mental Health Services Administration, providing sampling, estimation, weighting and methodological support for the National Survey on Drug Use and Health. Matt was also a researcher for the National Agricultural Statistics Service and has consulted with the National Aeronautics and Space Administration. Matt has collaborated on numerous international projects related to agriculture and public health.

Erin Coulter Riordan is an applied ecologist working at the intersection of plants, people and climate. She uses quantitative tools, including species distribution models, to understand how plants respond to climate with the goal of promoting resilience in complex socio-ecological systems. Her interests include restoration of ecological and cultural landscapes, conservation of crop wild relatives and climate change adaptation in arid land food systems.

How to cite this article: Johnson, N. G., Williams, M. R., & Riordan, E. C. (2021). Generalized nonlinear models can solve the prediction problem for data from species-stratified use-availability designs. *Diversity and Distributions*, 27, 2077–2092. https://doi.org/10.1111/ddi.13384

APPENDIX A

UNRESTRICTED MAXIMUM LIKELIHOOD ESTIMATION OF MODEL PARAMETERS FOR DATA FROM USE-AVAILABILITY DESIGNS

This section of the appendix and Appendices B and C form a kind of paper within a paper as they contain detailed justification for the methods we used in the main body of the manuscript. We will reintroduce the model from a more technical perspective, provide an alternative parameterization of the model that eases implementation in some software programs and theoretically justify these methods for data with large availability/background samples.

A.1 | Use-availability design, model and likelihood

Consider a binary response variable y where $y_i = 1$ when an organism of some species is present in location i and $y_i = 0$ when the species is absent. At each location i, there are also habitat data $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$. Our goal is inference on $f(y_i \mid \mathbf{x}_i, \boldsymbol{\beta})$ which follows a Bernoulli distribution with probability parameter $P(y_i = 1 \mid \mathbf{x}_i, \boldsymbol{\beta}) = p(\mathbf{x}_i, \boldsymbol{\beta})$, where p is a known (i.e., chosen) probability function and $\boldsymbol{\beta}$ is vector parameters (e.g. an intercept and K linear regression coefficients). Suppose we have complete habitat data for all i, but incomplete data on y_i. In particular, we only have data where $y_i = 1$, also known as presence-only data. In these situations, we can use a use-availability design to make inference on $p(\mathbf{x}_i, \boldsymbol{\beta})$ and its model parameters. A use-availability design is a sequence of Bernoulli trials such that when $C = 1$ a sample is taken from $f(\mathbf{x} \mid y = 1)$, called the "use," "presence-only" or "case" sample, and when $C = 0$, a sample is taken from $f(\mathbf{x})$, called the "availability," "background," or "contaminated control" sample. Note, this implies $f(\mathbf{x} \mid C = 0) = f(\mathbf{x})$. We will use this result later in a derivation. The technical details for such a model were originally described by Lancaster and Imbens (1996). The single observation likelihood for a use-availability design is as follows:

$$L_{C,X}(C_i, \mathbf{x}_i \mid y_i) = [f(\mathbf{x} \mid y_i = 1) P(C_i = 1)]^{C_i} [f(\mathbf{x}_i) P(C_i = 0)]^{1-C_i}.$$

Let us define $p(\mathbf{x}_i, \boldsymbol{\beta}) = P(y_i = 1 \mid \mathbf{x}_i)$, $h = P(C_i = 1)$ and $q = P(y_i = 1)$. We can rearrange the likelihood and then decompose it to have the following form:

$$\begin{aligned} L_{C,X}(C_i, \mathbf{x}_i \mid y_i, \boldsymbol{\beta}, q, h) &= [f(\mathbf{x} \mid y_i = 1, \boldsymbol{\beta}) h]^{C_i} [f(\mathbf{x}_i)(1-h)]^{1-C_i} \\ &= \left[\frac{p(\mathbf{x}_i, \boldsymbol{\beta})}{p(\mathbf{x}_i, \boldsymbol{\beta}) + \frac{q(1-h)}{h}}\right]^{C_i} \left[1 - \frac{p(\mathbf{x}_i, \boldsymbol{\beta})}{p(\mathbf{x}_i, \boldsymbol{\beta}) + \frac{q(1-h)}{h}}\right]^{1-C_i} \\ &\times \left[\frac{p(\mathbf{x}_i, \boldsymbol{\beta}) + \frac{q(1-h)}{h}}{q + \frac{q(1-h)}{h}}\right] f(\mathbf{x}_i) \\ &= L_{C \mid X}(C_i \mid \mathbf{x}_i, y_i, \boldsymbol{\beta}, q, h) L_X(\mathbf{x}_i \mid \boldsymbol{\beta}, q, h) \end{aligned}$$

TABLE A1 Approximate distribution of $\alpha_{s,0} = -\log[1+\exp(-\gamma_{s,q})] - \gamma_{s,h}$ from 10 million Monte Carlo samples of $\gamma_{s,q}$ and $\gamma_{s,h}$. The estimated Kullback–Leibler divergence (KLD) over given range between the actual Monte Carlo distribution and the approximation

Prior for $\gamma_{s,q}$ and $\gamma_{s,h}$	c	Approximate prior for $\alpha_{s,0}$	Range	KLD (*bits*)
Cauchy(0, c)	2.5	Cauchy(− 1.53, 3.59)	(−25,25)	2.61e-4
	5	Cauchy(− 2.76, 7.23)	(−75,75)	1.22e-4
	10	Cauchy(− 5.35, 14.5)	(−100,100)	7.81e-5
$T_{df=7}(0, c)$	2.5	$T_{df=7}$(− 1.53, 3.59)	(−15,15)	2.03e-4
	5	$T_{df=7}$(− 2.37, 5.80)	(−30,30)	1.20e-4
	10	$T_{df=7}$(− 4.55, 9.83)	(−50,50)	6.59e-5
Normal(0, c)	2.5	Normal(− 1.23, 2.86)	(−10,10)	1.16e-4
	5	Normal(− 2.05, 3.19)	(−20,20)	7.92e-5
	10	Normal(− 4.05, 11.55)	(−50,50)	6.06e-5

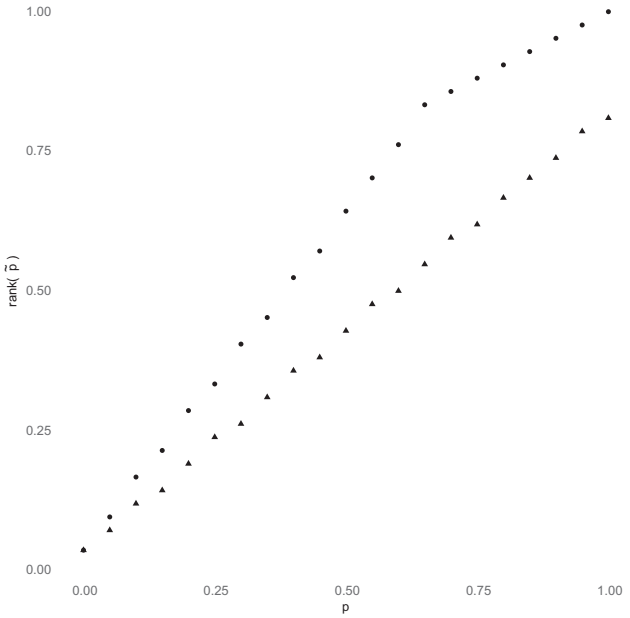

FIGURE A1 Scatter plot illustrating how the rank of $\tilde{p}_{si}$ is not guaranteed to conserve the ordering of p_{si} for two generic species denoted by triangles and circles. The circles have $\psi_s = 0.11$ and the triangles have $\psi_s = 0.17$. The ranks have been scaled to be between 0 and 1

where

$$L_{C|X}(C_i|\mathbf{x}_i, y_i, \boldsymbol{\beta}, q, h) = \tilde{p}(\mathbf{x}_i, \boldsymbol{\beta}, q, h)^{C_i}\ \ 1-\tilde{p}\left(\mathbf{x}_i, \boldsymbol{\beta}, q, h\right)\Big]^{1-C_i},$$

$$L_X(\mathbf{x}_i|\boldsymbol{\beta}, q, h) = \frac{p\left(\mathbf{x}_i, \boldsymbol{\beta}\right) + \frac{q(1-h)}{h}}{q + \frac{q(1-h)}{h}} f\left(\mathbf{x}_i\right),$$

$$\tilde{p}\left(\mathbf{x}_i, \boldsymbol{\beta}, q, h\right) = \frac{p\left(\mathbf{x}_i, \boldsymbol{\beta}\right)}{p\left(\mathbf{x}_i, \boldsymbol{\beta}\right) + \frac{q(1-h)}{h}}.$$

The likelihood $L_{C|X}$ corresponds to a Bernoulli likelihood of C_i conditional on $\mathbf{x}_i$ with parameter $P(C_i = 1 \mid \mathbf{x}_i, \boldsymbol{\beta}, q, h) = \tilde{p}(\mathbf{x}_i, \boldsymbol{\beta}, q, h)$, the likelihood of L_X appears to be determined by the form of $p(\mathbf{x}_i, \boldsymbol{\beta})$, the values of q and h, and the distribution $f(\mathbf{x}_i)$. Lancaster and Imbens (1996) considered the implications of when L_X is the likelihood of multinomial distribution. Such a strong assumption is unnecessary. All the information about $\boldsymbol{\beta}$ is contained within the conditional distribution of $Y \mid X$, and all the information about h and q is the distribution of C and Y, respectively. This implies that $f_X(\mathbf{x}_i)$ is a proportionality constant in L_X.

In the next subsection, we will walk through an updated version of the technique of Lancaster and Imbens (1996) for constrained maximum likelihood estimation. After that, we will provide an alternative parameterization of $L_{C|X}$ that is unconstrained while still following the same constraint as in Lancaster and Imbens (1996).

A.2 | Constrained maximum likelihood inference on model parameters

In this section, we will describe the approach of Lancaster and Imbens (1996) for constrained maximum likelihood estimation of $(\boldsymbol{\beta}, q, h)$from the likelihood $L_{C|X}$. They show that as h is the probability associated with a sequence of Bernoulli trials, independent of $(y, \mathbf{x})$, it can be maximized separately and plugged into $L_{C|X}$when that function is maximized for $(\boldsymbol{\beta}, q)$. Specifically, the maximum likelihood estimator for h is $\hat{h} = N^{-1}\sum_{i=1}^{N} C_i = n_1(n_1+n_0)^{-1}$, where n_1 is the size of the use sample and $n_0$is the size of the availability sample. Additionally, the maximum likelihood estimator of q must satisfy the following restriction $q = \int p(\mathbf{x}, \boldsymbol{\beta})\, \mathrm{d}F(\mathbf{x})$. This restriction is equivalent to requiring $h = \int \tilde{p}(\mathbf{x}, \boldsymbol{\beta}, q, h)\, \mathrm{d}F(\mathbf{x})$, which in practice would require $\hat{h} = N^{-1}\sum_{i=1}^{N} \tilde{p}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}, \hat{q}, \hat{h})$. Rearranged, it has the following form:

$$N^{-1}\sum_{i=1}^{N}\left[C_i - \tilde{p}\left(\mathbf{x}_i, \hat{\boldsymbol{\beta}}, \hat{q}, \hat{h}\right)\right] = 0.$$

The log likelihood would be:

$$\mathscr{L}_{C|X} = \log\prod_{i=1}^{N}\left[L_1(C_i|\mathbf{x}_i, y_i, \boldsymbol{\beta}, q, h)\right]$$
$$= \sum_{i=1}^{N}\left\{C_i \log p\left(\mathbf{x}_i, \boldsymbol{\beta}\right) - C_i \log\frac{q(1-h)}{h} - \log\left[p\left(\mathbf{x}_i, \boldsymbol{\beta}\right) + \frac{q(1-h)}{h}\right]\right\} + N\log\frac{q(1-h)}{h}$$

And the set of equations we would need to solve to find maximum likelihood estimates are the following:

```
# Load brms package into R
library(brms)

# Set the working directory to where your data is stored
setwd('your_working_directory')

# Import your data from a .csv file into R
your_data <- read.csv('your_data.csv', header = TRUE)

# Specify prior distributions
prior_list <- c(prior(normal(-2.12, 5.76), nlpar = "alpha"),
                prior(normal(0, 2.5), nlpar = "eta"),
                prior(normal(0, 5), nlpar = "beta0"))

# Use the brm function to implement a Bayesian generalized nonlinear model
brms.out <- brm(bf(C ~ (1 + exp(alpha) + exp(alpha - beta0 - eta))^(-1),
                       alpha ~ 0 + species, beta0 ~ 0 + species,
                       eta ~ 0 + species:(aet + bio4 + bio15 + bio18 +
                         bio19 + cwd + djf + jja),
                   family = bernoulli("identity"), nl = TRUE),
                data = your_data, inits = "random", warmup = 1000,
                chains = 4, iter = 4000, thin = 1, cores = 4,
                control = list(max_treedepth = 15), sample_prior = "yes",
                prior = prior_list)
```

FIGURE A2 Example R code for implementing Model (6) using the *brm* function in the BRMS package on the California plant data when environmental covariates have been centred and scaled to mean 0 and standard deviation 1

$$\frac{\partial \mathcal{L}_{C|X}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \frac{p'(\mathbf{x}_i, \boldsymbol{\beta})}{p(\mathbf{x}_i, \boldsymbol{\beta})} \left[C_i - \tilde{p}(\mathbf{x}_i, \boldsymbol{\beta}, q, h)\right] = 0$$
$$\frac{\partial \mathcal{L}_{C|X}}{\partial q} = -\frac{1}{q} \sum_{i=1}^{N} \left[C_i - \tilde{p}(\mathbf{x}_i, \boldsymbol{\beta}, q, h)\right] = 0$$
$$\frac{\partial \mathcal{L}_{C|X}}{\partial h} = \frac{1}{h(1-h)} \sum_{i=1}^{N} \left[C_i - \tilde{p}(\mathbf{x}_i, \boldsymbol{\beta}, q, h)\right] = 0$$

After plugging in $\hat{h}$ into these three equations, we can see that both $\partial \mathcal{L}_{C|X}/\partial q$ and $\partial \mathcal{L}_{C|X}/\partial h$ are proportional to the constraint needed to ensure values $\hat{\boldsymbol{\beta}}$ and $\hat{q}$ are appropriately constrained. Therefore, any values of $(\hat{\boldsymbol{\beta}}, \hat{q})$ that solve this set of equations with $\hat{h}$ plugged into h will be guaranteed to have the appropriate constraint and maximize the likelihood.

A.3 | An alternative way to parameterize the *L*

Suppose we parameterize $L_{C,X}$ so that $\psi = q(1-h)h^{-1}$ and L_X so that $\tilde{q} = q(q+\psi)^{-1} = h$. The likelihood then has the following form:

$$L_{C,X}(C_i, \mathbf{x}_i \mid y_i, \boldsymbol{\beta}, q, \psi) = \prod_{i=1}^{N} L_{C|X}(C_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \psi) L_X(\mathbf{x}_i \mid \boldsymbol{\beta}, \psi, q \text{ or } h)$$

where its components have form:

$$L_{C|X}(C_i|\mathbf{x}_i, \boldsymbol{\beta}, \psi) = \left[\frac{p(\mathbf{x}_i, \boldsymbol{\beta})}{p(\mathbf{x}_i, \boldsymbol{\beta}) + \psi}\right]^{C_i} \left[1 - \frac{p(\mathbf{x}_i, \boldsymbol{\beta})}{p(\mathbf{x}_i, \boldsymbol{\beta}) + \psi}\right]^{1-C_i}$$
$$L_X(\mathbf{x}_i|\boldsymbol{\beta}, \psi, q \text{ or } h) = \frac{1-\tilde{q}}{1-\tilde{p}(\mathbf{x}_i, \boldsymbol{\beta}, \psi)} f_X(\mathbf{x}_i) \propto \frac{1-\tilde{q}}{1-\tilde{p}(\mathbf{x}_i, \boldsymbol{\beta}, \psi)} = \frac{1-h}{1-\tilde{p}(\mathbf{x}_i, \boldsymbol{\beta}, \psi)}$$

Are the values $\hat{\boldsymbol{\beta}}$ and $\hat{\psi}$ that maximize $L_{C|X}$ appropriately constrained? As before, they need to satisfy the following restriction:

$$N^{-1} \sum_{i=1}^{N} \left[C_i - \tilde{p}(\mathbf{x}, \boldsymbol{\beta}, \psi)\right] = 0$$

The log likelihood $\mathcal{L}_{C|X}$ is as follows:

$$L_{C|X} = \log \prod_{i=1}^{N} L_{C|X}(C_i|\mathbf{x}_i, y_i, \boldsymbol{\beta}, \psi)$$
$$= \sum_{i=1}^{N} \left\{C_i \log p(\mathbf{x}_i, \boldsymbol{\beta}) - C_i \log \psi - \log\left[p(\mathbf{x}_i, \boldsymbol{\beta}) + \psi\right]\right\} + N \log \psi$$

And the set of equations needed to find the values $\hat{\boldsymbol{\beta}}$ and $\hat{\psi}$ that maximize L_1 are as follows:

$$\frac{\partial \mathcal{L}_{C|X}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \frac{p'(\mathbf{x}_i, \boldsymbol{\beta})}{p(\mathbf{x}_i, \boldsymbol{\beta})} \left[C_i - \tilde{p}(\mathbf{x}_i, \boldsymbol{\beta}, \psi)\right] = 0$$
$$\frac{\partial \mathcal{L}_{C|X}}{\partial \psi} = -\frac{\psi'}{\psi} \sum_{i=1}^{N} \left[C_i - \tilde{p}(\mathbf{x}_i, \boldsymbol{\beta}, \psi)\right] = 0$$

The second equation $\partial \mathcal{L}_{C|X}/\partial \psi$ is proportional to the restriction necessary to guarantee that any $\hat{\boldsymbol{\beta}}$ and $\hat{\psi}$ that maximize $L_{C|X}$ are appropriately constrained. This approach absolves us needing to use plug-in estimator $\hat{h}$. Thus, it allows us to perform constrained maximum likelihood with an unconstrained likelihood. It is equivalent to maximum likelihood estimation of parameters from a generalized nonlinear model with Bernoulli distributed response C_i, nonlinear mean/probability function $\tilde{p}(\mathbf{x}_i, \boldsymbol{\beta}, \psi)$ and identity link function. It is still possible to find $\hat{q}$ from this model using the definition of its restricted form, $\hat{q} = n_0^{-1} \sum_{i=1}^{N} (1 - C_i) p(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$.

APPENDIX B

APPROXIMATE LIKELIHOOD WHEN AVAILABILITY SAMPLE SIZE N_0 IS LARGE

Lancaster and Imbens (1996) showed that the asymptotic distribution of maximum likelihood estimators $(\hat{\boldsymbol{\beta}}, \hat{q}, \hat{h})$ from $L_{C|X}$ is asymptotically

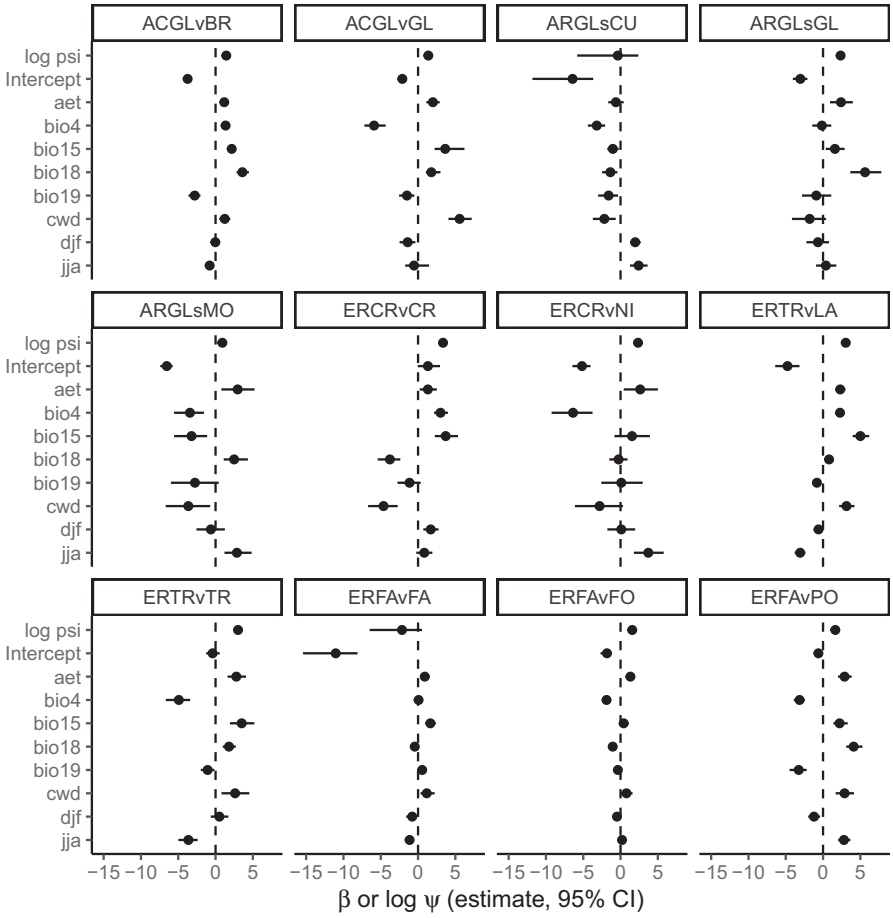

FIGURE A3 Forest plot of posterior means and 95% credible intervals for $\log \psi_s$, $\beta_{s,0}$ (denoted Intercept), and $\boldsymbol{\beta}_s$ for each species in the California plant data

unbiased and normally distributed with a sandwich estimator for the covariance as N goes to infinity. We are interested in a slightly different asymptotic result: where the size of the availability sample $n_0 = N - n_1$ goes to infinity. When n_0 goes to infinity, ψ goes to infinity. Thus, it is sufficient to check the behaviour of the log likelihood $\mathscr{L}_{C,X}$ at the latter limit. First, we will focus on approximating the limit of $\mathscr{L}_X$. Next, we will use this result to show $\mathscr{L}_{C,X}$ is approximately a Poisson likelihood.

Proposition 1. *The log likelihood* $\mathscr{L}_X(\mathbf{X} \mid \mathbf{p}, \psi, q)$ *is approximately 0 as* $\psi \to \infty$, *provided* $|p_i| < |\psi|$ *for all* i *and* $|q| < |\psi|$.

Proof. Let $\mathscr{L}_X(\mathbf{C}, \mathbf{X} \mid \mathbf{p}, \psi)$ denote the joint log likelihood of $\mathbf{X}$ from data collected according to a use-availability design with arbitrary probability function (e.g. logistic). It can be written as follows:

$$\begin{aligned}\mathscr{L}_X(\mathbf{X}|\mathbf{p}, \psi) &\propto \sum_{i=1}^{N} \log \frac{p_i + \psi}{q + \psi} \\ &= \sum_{i=1}^{N} \left[\log\left(p_i + \psi\right) - \log\left(q + \psi\right)\right].\end{aligned}$$

From the Taylor expansions, we get $\log(p_i + \psi) = \log \psi + p_i \psi^{-1} + O(\psi^{-2})$ and $\log(q + \psi) = \log\psi + q\psi^{-1} + O(\psi^{-2})$ for $|p_i| < |\psi|$ and for $|q| < |\psi|$, which leads us to:

$$\begin{aligned}\sum_{i=1}^{N} \left[\log\left(p_i + \psi\right) - \log\left(q + \psi\right)\right] &= \sum_{i=1}^{N} \left[\log\psi + \frac{p_i}{\psi} + O\left(\psi^{-2}\right) - \log\psi - \frac{q}{\psi} + O\left(\psi^{-2}\right)\right] \\ &\approx \frac{1}{\psi} \sum_{i=1}^{N} \left(p_i - q\right)\end{aligned}$$

Let us define some sample averages of p_i: $\bar{p} = N^{-1} \sum_{i=1}^{N} p_i$, $\bar{p}_1 = n_1^{-1} \sum_{i=1}^{N} p_i^{C_i}$ and $\bar{p}_0 = n_0^{-1} \sum_{i=1}^{N} p_i^{1-C_i}$. We then have the following:

$$\begin{aligned}\frac{1}{\psi} \sum_{i=1}^{N} \left(p_i - q\right) &= \frac{1}{\psi} \sum_{i=1}^{N} \left(p_i - \bar{p} + \bar{p} - q\right) \\ &= \frac{N}{\psi} \left(\bar{p} - q\right) \\ &= \frac{N}{\psi} \left[\frac{n_1 \bar{p}_1 + \left(N - n_1\right) \bar{p}_0}{N} - q\right]\end{aligned}$$

Now, we plug-in $\psi = q(N - n_1)n_1^{-1}$, to make explicit how ψ goes to infinity as N goes to infinity:

$$\begin{aligned}\frac{N}{\psi} \left[\frac{n_1 \bar{p}_1 + \left(N - n_1\right) \bar{p}_0}{N} - q\right] &= \frac{N}{q} \frac{n_1}{N - n_1} \left[\frac{n_1 \bar{p}_1 + \left(N - n_1\right) \bar{p}_0}{N} - q\right] \\ &= \frac{N}{q} \frac{n_1}{N - n_1} \frac{n_1 \bar{p}_1}{N} + \frac{N}{q} \frac{n_1}{N - n_1} \frac{\left(N - n_1\right) \bar{p}_0}{N} - \frac{N}{q} \frac{n_1}{N - n_1} q \\ &= \frac{n_1 \bar{p}_1}{q} \frac{n_1}{N - n_1} + \frac{n_1 \bar{p}_0}{q} - \frac{n_1 N}{N - n_1}\end{aligned}$$

From here, we take the limit as N goes to infinity for fixed n_1 to get:

$$\lim_{N \to \infty} \frac{n_1 \bar{p}_1}{q} \frac{n_1}{N - n_1} + \frac{n_1 \bar{p}_0}{q} - \frac{n_1 N}{N - n_1} = n_1 \frac{\bar{p}_0}{q} - n_1$$

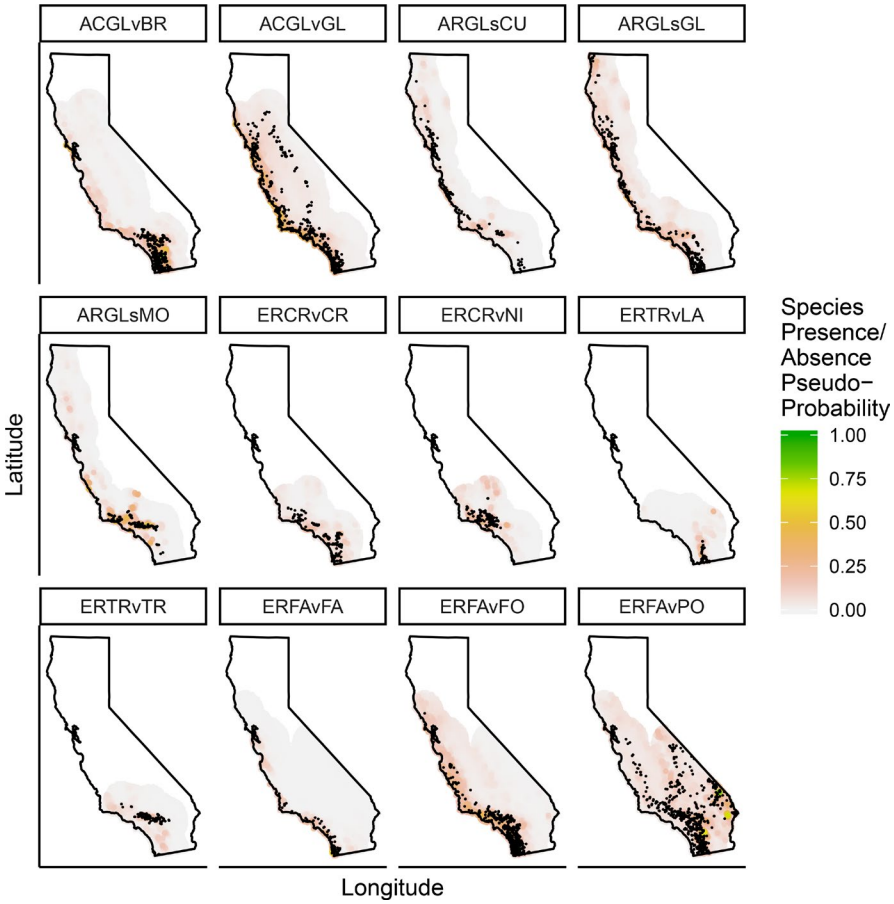

FIGURE A4 Species habitat use pseudo-probabilities from Model (5): $\tilde{p}_{si}^{\exp}$. Black circles are observed species presences

By the law of large numbers $\bar{p}_0 \to q$, therefore:

$$\begin{aligned} n_1 \frac{\bar{p}_0}{q} - n_1 &\to n_1 \frac{q}{q} - n_1 \\ &= n_1 - n_1 \\ &= 0. \end{aligned}$$

With this result, we can justify approximating $\mathscr{L}_{C,X}$ by $\mathscr{L}_{C|X}$ for large availability samples. We can create a similar approximation to $\mathscr{L}_{C|X}$ by a Poisson likelihood for large availability samples.

Proposition 2. *A conditional Poisson likelihood* $\mathscr{L}_{C|X,Poi}(\mathbf{C}\,|\,\boldsymbol{X}, \lambda)$, *where the Poisson intensities are* $\lambda_i = p_i\psi^{-1}$, *approximates the Bernoulli conditional likelihood* $\mathscr{L}_{C|X,Bern}(\mathbf{C}\,|\,\boldsymbol{X}, \boldsymbol{p}, \psi)$ *as* $\psi \to \infty$, *provided* $|\,p_i\,| < |\,\psi\,|$ *for all i.*

Proof. Let $\mathscr{L}_{C|X,Bern}(\mathbf{C}, \mathbf{X}\,|\,\mathbf{p}, \psi)$ denote the joint likelihood of $\mathbf{C}$ conditional on $\mathbf{X}$ from data collected according to a use-availability design with arbitrary probability function (e.g. logistic). It can be written as follows:

$$\begin{aligned} \mathscr{L}_{C|X,Bern}(\mathbf{C}, \mathbf{X}|\mathbf{p}, \psi) &= \sum_{i=1}^{N} \left[C_i \log \frac{p_i}{p_i + \psi} + (1 - C_i) \log \frac{\psi}{p_i + \psi} \right] \\ &= \sum_{i=1}^{N} \left(C_i \log \frac{p_i}{\psi} + \log \frac{\psi}{p_i + \psi} \right). \end{aligned}$$

From the Taylor expansions $\log[\psi(p_i + \psi)^{-1}] = -p_i\psi^{-1} + O(\psi^{-2})$ for $|\,p_i\,| < |\,\psi\,|$, we get:

$$\begin{aligned} \sum_{i=1}^{N} C_i \log \frac{p_i}{\psi} + \log \frac{\psi}{p_i + \psi} \Big) &= \sum_{i=1}^{N} \left[C_i \log \frac{p_i}{\psi} - \frac{p_i}{\psi} + O\left(\psi^{-2}\right) \right] \\ &\approx \sum_{i=1}^{N} \left(C_i \log \frac{p_i}{\psi} - \frac{p_i}{\psi} \right) \\ &= \mathscr{L}_{C|X,Poi}\left(\mathbf{C}|\mathbf{X}, \lambda = \mathbf{p}\psi^{-1}\right). \end{aligned}$$

Finally, these two propositions can be combined to get an approximation for $\mathscr{L}_{C,X}$ by $\mathscr{L}_{C|X,Poi}$ for large availability samples.

Proposition 3. *The Poisson likelihood* $\mathscr{L}_{C|X,Poi}\left(\boldsymbol{C}|\boldsymbol{X}, \lambda = \boldsymbol{p}\psi^{-1}\right)$ *approximates the joint likelihood* $\mathscr{L}_{C,X}(\mathbf{C}, \boldsymbol{X}\,|\,\boldsymbol{p}, \psi)$ *as* $\psi \to \infty$, *provided* $|\,p_i\,| < |\,\psi\,|$ *for all i.*

Proof. Follows directly from Proposition (1) and Proposition (2), as

$$\mathscr{L}_{C,X}(\mathbf{C}, \mathbf{X}\,|\,\mathbf{p}, \psi) = \mathscr{L}_{C|X}(\mathbf{C}\,|\,\mathbf{X}, \mathbf{p}, \psi) + \mathscr{L}_{X}(\mathbf{X}\,|\,\mathbf{p}, \psi).$$

The implication of these propositions is that generalized nonlinear models for Bernoulli or Poisson distributed binary response data can be used for large background samples as an alternative to methods that involve the full likelihood $L_{C,X}$.

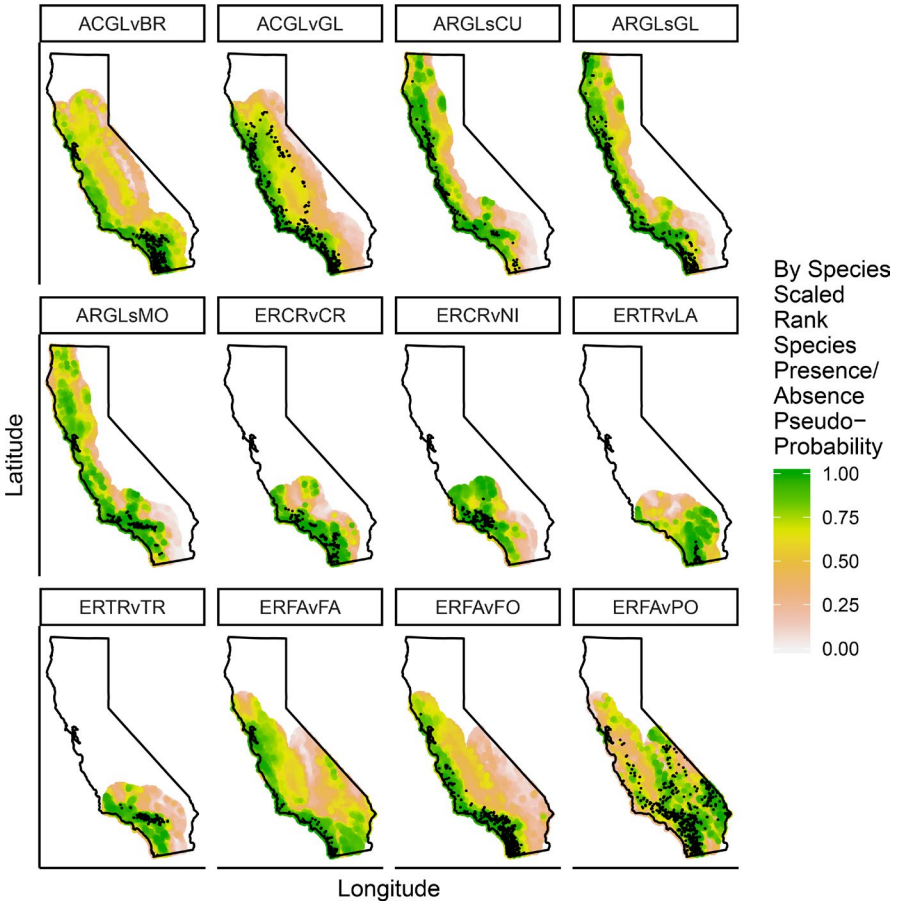

FIGURE A5 Scaled rank species habitat use pseudo-probabilities from Model (5): $\tilde{p}_{si}^{\exp}$. Black circles are observed species presences

APPENDIX C

BAYESIAN INFERENCE

A fully Bayesian approach would use the full likelihood $L_{C,X}$, as both $L_{C|X}$ and L_X contain information about the parameters β, h and q. When the size of the background sample is large, Proposition (1) allows us to take an approximate Bayesian approach that only relies on $L_{C|X}$. This approximate Bayesian approach can be framed as a Bayesian generalized nonlinear model and implemented such as the BRMS package (Bürkner, 2017, 2018) in R (R Core Team, 2019). The generalized nonlinear model approach is discussed in the main body of the text. In the next subsections will give some brief details of how to implement the fully Bayesian approach and propose some possible prior distribution for ψ to be used in the approximate Bayesian analysis.

C.1 | A Markov chain Monte Carlo algorithm for fully Bayesian inference

The fully Bayesian approach uses the full joint likelihood $L_{C,X}$. As noted previously, the likelihood is overparameterized. So instead of sampling from the conditional posterior of q during MCMC, we set q to its constrained value at each iteration. A rough outline of a Metropolis-type algorithm for the fully Bayesian approach is as follows:

Algorithm 1 *First choose initial conditions* $h^{(0)}$, $q^{(0)}$ *and* $\boldsymbol{\beta}^{(0)}$, *and number of MCMC iterations T. Then, iterate the following*:

- FOR $t = 1, 2, \ldots, T$:
  - Generate $h^{(t)} \sim f\left(h^{(t-1)}|\mathbf{C}\right)$
  - Generate $\boldsymbol{\beta}^{(t)} \sim f\left(\boldsymbol{\beta}^{(t-1)}|\mathbf{C}, \mathbf{X}, h^{(t)}, q^{(t-1)}\right)$
  - Set $q^{(t)} = n_0^{-1} \sum_{i=1}^{N} (1 - C_i) p\left(\mathbf{x}_i, \boldsymbol{\beta}^{(t)}\right)$
- END,

where the conditional posterior distributions are defined as follows:

$$
\begin{aligned}
f(h|\mathbf{C}) \quad & \propto L_{C,X}(\mathbf{C}, \mathbf{X}|\beta, q, h) f(h) \\
& \propto \left[\prod_{i=1}^{N} \text{Bernoulli}\left(C_i; h\right)\right] f(h), \\
f(\beta|\mathbf{C}, \mathbf{X}, h, q) \quad & \propto L_{C,X}(\mathbf{C}, \mathbf{X}|\beta, q, h) f(\beta) \\
& \propto \left\{\prod_{i=1}^{N} \text{Bernoulli}\left[C_i; \tilde{p}\left(\mathbf{x}_i, \beta, q, h\right)\right] \frac{1-h}{1-\tilde{p}\left(\mathbf{x}_i \beta, q, h\right)}\right\} f(\beta),
\end{aligned}
$$

and where $f(h)$ and $f(\beta)$ are the prior distribution on h and β, respectively.

C.2 | Prior selection in Bayesian generalized nonlinear model approaches to use-availability models

We are unaware of any work on weakly informative default prior selection in Bayesian use-availability models. We can use work on weakly informative default prior selection in logistic regression models to guide prior selection of $\beta_{s,0}$ and $\boldsymbol{\beta}_s$ as they are logistic regression model parameters. For instance, Gelman et al. (2008) recommend a Cauchy(0, 10) for $\beta_{s,0}$ and Cauchy(0, 2.5) for $\boldsymbol{\beta}_s$ as default priors when each environmental covariate has been centred and scaled to mean

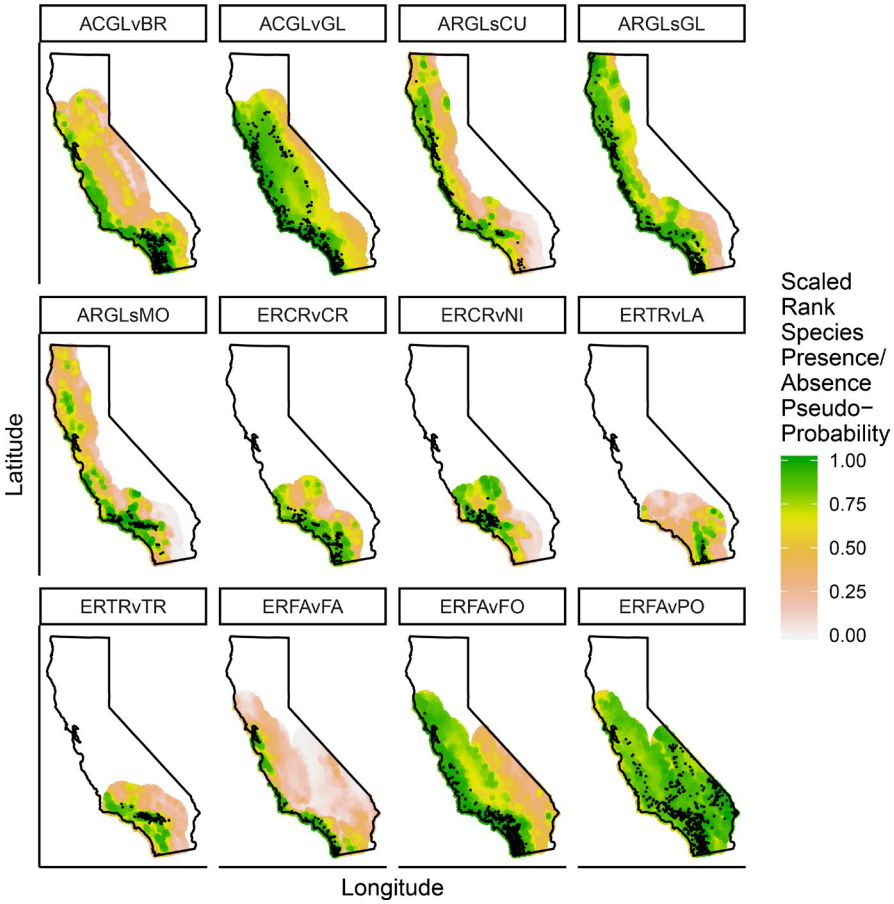

FIGURE A6 By species scaled rank species habitat use pseudo-probabilities from Model (5): $\tilde{p}_{si}^{\exp}$. Black circles are observed species presences

TABLE A2 Information on infraspecies included in the use-availability model. Taxonomy follows Jepson eFlora (Jepson Flora Project, 2017)

Scientific name	Common name	Figure label
Acmispon glaber var. *brevialatus* (Ottley) Brouillet	deerweed, California broom	ACGLvBR
Acmispon glaber (Vogel) Brouillet var. *glaber*	deerweed, California broom	ACGLvGL
Arctostaphylos glandulosa subsp. *cushingiana* (Eastw.) J.E. Keeley, M.C. Vasey & V.T. Parker	Cushing's manzanita	ARGLsCU
Arctostaphylos glandulosa Eastw. subsp. *glandulosa*	Eastwood's manzanita	ARGLsGL
Arctostaphylos glandulosa subsp. *mollis* (J.E. Adams) P.V. Wells	Transverse Range manzanita	ARGLsMO
Eriodictyon crassifolium Benth. var. *crassifolium*	thickleaf yerba santa	ERCRvCR
Eriodictyon crassifolium var. *nigrescens* Brand	bicolored yerba santa	ERCRvNI
Eriodictyon trichocalyx var. *lanatum* (Brand) Jeps.	San Diego yerba santa	ERTRvLA
Eriodictyon trichocalyx A. Heller var. *trichocalyx*	hairy yerba santa	ERTRvTR
Eriogonum fasciculatum Benth. var. *fasciculatum*	coastal California buckwheat	ERFAvFA
Eriogonum fasciculatum var. *foliolosum* (Nutt.) Abrams	leafy California buckwheat	ERFAvFO
Eriogonum fasciculatum var. *polifolium* (Benth.) Torr. & A. Gray	Mojave desert California buckwheat	ERFAvPO

zero and standard deviation 0.5. However, Ghosh et al. (2018) show that Cauchy priors have tails too heavy to address some data issues in logistic regression (such as perfect separation). They recommend either the normal or, to maintain some of the robustness of the Cauchy prior, the non-standardized T with 7 degrees of freedom, each with the same guidelines for scaling discussed in Gelman et al. (2008). Weakly informative default priors for $\log\psi_s = \alpha_{s,0}$ is an open problem, as it is the sum of a log-odds and a log probability. In this section, we propose some prior distributions for $\alpha_{s,0}$ motivated from the ideas of Gelman et al. (2008) and Ghosh et al. (2018).

Suppose we follow the suggestion of Gelman et al. (2008) that some of their recommendations for logistic regression may also be appropriate for parameters on the log scale, such as $\log\psi_s = \alpha_{s,0}$. Their recommendation is to choose a $\text{Cauchy}(0, 10)$ prior for $\alpha_{0,s}$, as

TABLE A3 Names, descriptions, means and standard deviations of environmental variables included in the use-availability model for the California plan data

Variable name	Variable description	Mean (*SD*)
aet	Actual evapotranspiration	269.579 (116.608)
bio4	Standard deviation of the monthly mean temperatures	615.590 (135.127)
bio15	Standard deviation of the monthly precipitation estimates expressed as a percentage of the annual mean of those estimates	82.2179 (15.5021)
bio18	Precipitation of the warmest quarter	20.3406 (14.8064)
bio19	Precipitation of the coldest quarter	252.618 (213.123)
cwd	Climatic water deficit	1,018.99 (262.621)
djf	Minimum temperature averaged over the coldest months (December, January, and February)	2.20037 (3.43438)
jja	Maximum temperature averaged over the hottest months (June, July, and August)	31.5515 (4.82089)

TABLE A4 Difference of β_s between the varieties species *Acmispon glaber*: var. *brevialatus*–var. *glaber*. Posterior means and 95% credible intervals for the differences, and Bayes Factors and the associated posterior probability comparing the difference is zero versus non-zero (i.e., the effect of the covariate is the same for both species versus otherwise). Bayes factors above 1 give evidence for the difference being zero and Bayes factors below 1 give evidence that the difference is not zero. Posterior probabilities above 0.5 give evidence for the difference being zero and posterior probabilities below 0.5 give evidence for the difference being non-zero

Variable	Estimate	95%CI	Bayes Factor	Posterior Probability
aet	1.56	(0.49,2.67)	0.13	0.12
bio4	−1.27	(−2.19,−0.37)	0.16	0.14
bio15	1.80	(0.70,3.04)	0.04	0.03
bio18	5.15	(4.06,6.42)	0.00	0.00
bio19	−2.90	(−4.17,−1.75)	0.00	0.00
cwd	2.09	(0.67,3.58)	0.09	0.8
djf	−0.74	(−1.63,0.20)	2.07	0.67
jja	2.60	(1.70,3.55)	0.00	0.00

it is a kind of intercept term, akin to $\beta_{s,0}$. Some other choices of distribution would be the non-standardized *t*, normal and Laplace with a similar scaling rule. However, we are not sure any of these choices should be centred on zero if a symmetric distribution is used. The quantity $(1-h_s)h_s^{-1}$ is largely determined by the sampling scheme with expected value $n_{s,0}n_{s,1}^{-1}$ where $n_{s,0}$ and $n_{s,1}$ are the sizes of the availability and use samples, respectively, for species *s*. This implies a kind of practical upper bound on what we would expect $\alpha_{0,s}$ to be as $\log q_s$ is strictly negative. As a result, some choices of symmetric real valued prior distributions centred at zero (e.g. Cauchy(0, 10) may end up putting too much prior probability on support above $\log n_{s,0} - \log n_{s,1}$.

To remedy this, we attempted to derive the implied prior distribution on $\alpha_{0,s}$ given the rules recommended by Gelman et al. (2008): that the log-odds should follow a Cauchy$(0,c)$ distribution. We re-parameterized $\alpha_{0,s}$ such that q_s and h_s were a function of their log-odds $\gamma_{s,q}$ and $\gamma_{s,h}$, respectively, to get the following:

$$\begin{aligned}\gamma_{s,q} &= \log\frac{q_s}{1-q_s}\\ \gamma_{s,h} &= \log\frac{h_s}{1-h_s}\\ \alpha_{s,0} &= -\log\left[1+\exp\left(-\gamma_{s,q}\right)\right]-\gamma_{s,h}\\ \gamma_{s,q},\gamma_{s,h} &\sim \text{Cauchy}\,(0,c)\end{aligned} \tag{7}$$

The prior distribution of $\alpha_{s,0}$ is not closed form when $\gamma_{s,q},\gamma_{s,h} \sim$ Cauchy$(0,c)$ nor is it for Normal$(0,c)$ or $\mathrm{T}_{\mathrm{df}=7}(0,c)$. We could parameterize Model (6) in terms of $\gamma_{s,q}$ and $\gamma_{s,h}$ and let the Hamiltonian Monte Carlo algorithm used by the *brm* function perform the integration for us. This would mean the chains for $\gamma_{s,q}$ and $\gamma_{s,h}$ would not necessarily converge, but the chain for the computed quantity $\alpha_{0,s}$ would. A faster alternative is to do Monte Carlo integration upfront to find an approximate prior distribution in terms of one we already know a bit about. When $\gamma_{s,q}$ and $\gamma_{s,h}$ are Cauchy, normal or $\mathrm{T}_{\mathrm{df}=7}(0,c)$, then the $\alpha_{s,0}$ is approximately Cauchy, normal or $\mathrm{T}_{\mathrm{df}=7}(0,c)$, respectively. See Table A1 for the approximate priors based on these distributions with values for $c = 2.5, 5, 10$. Approximation based on 10 million Monte Carlo samples.

APPENDIX D

OTHER FORMS OF ψ_s

As software easily allows for more complicated specification of $\log\psi_s$ than we have considered, we want to briefly discuss the implications of allowing $\log\psi_s$ to be a function of covariate information beyond species-specific intercept $\alpha_{s,0}$. Or rather, what alterations could we to make to the sampling scheme of a use-availability design for it to

make sense that $\log\psi_s$ should be a function of covariate information beyond species. As ψ_s is already composed of three quantities, q_s, h_s and $(1 - h_s)$, we will first look at each individually then seek to combine when appropriate. We should note that fully developing these concepts is beyond the scope of this paper, but are of interest.

Lancaster and Imbens (1996) show that the distribution of the availability sample is simply $f(\mathbf{x}_{si})$, while the distribution of the use sample is $f(\mathbf{x}_{si} \mid y_{si} = 1)$ which according to Bayes rule is $P(y_{si} = 1 \mid \mathbf{x}_{si}) f(\mathbf{x}_{si}) / P(y_{si} = 1) = f(\mathbf{x}_{si}) p_{si} q_s^{-1}$. From here, we can see that q_s makes it into the model implicitly as a quantity that has been marginalized over $\mathbf{x}_{si}$. Therefore, it is by definition a quantity that should not be a function of environmental covariate information that is integrated over in this manner. Besides species, other design induced stratification such as temporal effects, block effects or organism specific effects if an organism within species was measured repeatedly (as might be the case for mobile species) could be included if appropriate under the study design.

As for h_s, an often overlooked assumption of use-availability models is that for each species s the C_s's are a sequence of Bernoulli trials that determine whether to take a sample from the use data $f(\mathbf{x}_s \mid y_s = 1)$ or the availability data $f(\mathbf{x}_s)$. We have dropped the i subscripts for a reason—because the locations i are what is being sampled in these two scenarios. Thus, one cannot use spatially dependent environmental covariate information $\mathbf{x}$ to determine h_s as the spatial locations are determined after the value of C_s. Any covariate information used to determine h_s would need to be linked to the spatial location sampled after the sampling occurs and cannot be spatially located a priori. The list of possibilities overlaps quite a bit with q_s: stratification by species, organism within species, or block, or temporal effects, possibly others. However, h_s cannot be a function of treatment effects as they are not by definition a stratified variable (if they were, they would be called blocks).

Putting this all together, we can have our design induced covariate information and the associated location sampled $\mathbf{z}_{si} = (z_{si1}, z_{si2}, \ldots, z_{si1})$. We can define the following relationship:

$$\log\psi_{si} = \log\left[q_{si}\frac{1 - h_{si}}{h_{si}}\right] = \alpha_{s,0} + \mathbf{z}_{si}\boldsymbol{\alpha}_s, \tag{8}$$

where $\alpha_{s,0}$ is the species intercept and $\boldsymbol{\alpha}_s = (\alpha_{s1}, \alpha_{s2}, \ldots, \alpha_{sK_2})$ are the species-specific covariates for the other effects.

APPENDIX E

EXAMPLE GENERALIZED NONLINEAR MODEL CODE

Example code we used to implement Model (6) on the California plant data using the *brm* function in the BRMS package is available in Figure A2.

APPENDIX F

ADDITIONAL RESULTS FOR THE CALIFORNIA PLANT DATA

In this section, we include some additional information about the analysis of plant data beyond prediction.

F.1 | Forest plots of species-specific habitat effects

Pairwise comparisons of environmental covariate across species were computed using the hypothesis function in BRMS package. In order for the *hypothesis* function to return Bayes factors, the sample_prior = "yes" option has to be specified.

One way to compare parameter estimates across species is to use forest plots of the posterior means and 95% credible intervals. The forest plots of $\alpha_{s,0}$, $\beta_{s,0}$ and $\boldsymbol{\beta}_s$ for each of the infraspecies are in Figure A3. Alternatively, credible intervals and Bayes factors could be computed for post hoc comparisons as shown in Table A4 for the species *Acmispon glaber*. In this example, the Bayes factors are comparing the hypothesis that the two varieties of *Acmispon glaber* share the same covariate effect versus the hypothesis that each of these varieties has its own effect. In this case, Bayes factors above 1 favour the first hypothesis and Bayes factors below 1 favour the second hypothesis. The Bayes factors can be converted to probabilities where values above 0.5 favour the first hypothesis and values below 0.5 favour the second. The only covariate for *Acmispon glaber* where the evidence favours no difference is minimum winter temperature (djf) with a posterior probability of 0.67. The evidence for the remaining covariates favours the hypothesis of variety specific covariate effects. For three of these covariates, precipitation of the warmest quarter (bio18), precipitation of the coldest quarter (bio19) and maximum summer temperature (jja), the estimated probability that the varieties share the same effect is 0.00.

F.2 | Pseudo habitat use probabilities and their ranks

Figure A4 contains the estimated pseudo-probabilities $\tilde{p}_{si}^{\exp}$ for Model (5). Notice they do a very poor job spanning the full range of probabilities 0 to 1. This is due to the biased estimate for the intercept. In single species, we could rank and scale the pseudo-probabilities by the max rank so they are between 0 and 1 as in Figure A5. Notice that for taxa such as ***Eriogonum fasciculatum*** var. ***polifolium*** the scaled ranks are high over the entire region. This is an artefact of inappropriately ranking across all taxa when the ranks are not comparable across taxa. Instead, it is better to both rank pseudo-probabilities and scale the ranks within species as in Figure A6. This gives a more realistic picture of which habitat is suitable within taxa, but not across taxa.