



Research Paper

Cite this article: Schank CJ, Cove MV, Kelly MJ, Nielsen CK, O’Farrill G, Meyer N, Jordan CA, González-Maya JF, Lizcano DJ, Moreno R, Dobbins M, Montalvo V, Cruz Díaz JC, Pozo Montuy G, de la Torre JA, Brenes-Mora E, Wood MA, Gilbert J, Jetz W, and Miller JA (2019). A Sensitivity Analysis of the Application of Integrated Species Distribution Models to Mobile Species: A Case Study with the Endangered Baird’s Tapir. *Environmental Conservation* **46**: 184–192. doi: [10.1017/S0376892919000055](https://doi.org/10.1017/S0376892919000055)

Received: 1 February 2019

Revised: 18 April 2019

Accepted: 21 April 2019

First published online: 12 June 2019

Keywords:

species distribution model; occupancy model; effective sampling area; point process model; density; conservation; tapir; *Tapirus bairdii*

Author for correspondence: Cody J Schank, Email: codyschank@gmail.com

Thematic Section: Bringing Species and Ecosystems Together with Remote Sensing Tools to Develop New Biodiversity Metrics and Indicators

A Sensitivity Analysis of the Application of Integrated Species Distribution Models to Mobile Species: A Case Study with the Endangered Baird’s Tapir

Cody J Schank^{1,2}, Michael V Cove³, Marcella J Kelly⁴, Clayton K Nielsen⁵, Georgina O’Farrill⁶, Ninon Meyer^{7,8}, Christopher A Jordan^{2,9,10}, Jose F González-Maya¹¹, Diego J Lizcano^{12,13}, Ricardo Moreno^{8,14}, Michael Dobbins¹⁵, Victor Montalvo¹⁶, Juan Carlos Cruz Díaz^{16,17}, Gilberto Pozo Montuy¹⁸, J Antonio de la Torre^{19,20}, Esteban Brenes-Mora^{21,22}, Margot A Wood²³, Jessica Gilbert²⁴, Walter Jetz^{25,26} and Jennifer A Miller¹

¹Department of Geography and the Environment, The University of Texas at Austin, Austin, TX 78712, USA; ²Global Wildlife Conservation, Austin, TX, USA; ³Department of Applied Ecology, North Carolina State University, Raleigh, NC 27695, USA; ⁴Department of Fish and Wildlife Conservation, Virginia Tech, Blacksburg, VA 24061, USA; ⁵Department of Forestry and Cooperative Wildlife Research Laboratory, Southern Illinois University, Carbondale, IL 62901-6504, USA; ⁶Department of Ecology and Evolutionary Biology, University of Toronto, 25 Harbord Street, Toronto, Ontario, M5S 3G5, Canada; ⁷El Colegio de la Frontera Sur, Departamento de Conservación de la Biodiversidad, Lerma, Campeche, Mexico; ⁸Fundación Yaguara-Panama, Ciudad del Saber, Panama; ⁹Panthera, New York, NY, USA; ¹⁰Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI, USA; ¹¹Proyecto de Conservación de Aguas y Tierras, ProCAT Colombia/Internacional, Bogotá, Colombia; ¹²Departamento Central de Investigación, Universidad Laica Eloy Alfaro de Manabí, Manta, Ecuador; ¹³The Nature Conservancy, Bogotá, Colombia; ¹⁴Smithsonian Tropical Research Institute, Balboa, Panama; ¹⁵Department of Geography, University of Florida, Gainesville, FL, USA; ¹⁶Instituto Internacional en Conservación y Manejo de Vida Silvestre, Universidad Nacional, Heredia 3000-1350, Costa Rica; ¹⁷Department of Environmental Conservation, University of Massachusetts Amherst, MA, 01003, USA; ¹⁸Conservación de la Biodiversidad del Usumacinta A.C., Emiliano Zapata, Tabasco, C.P. 86990, Mexico; ¹⁹Instituto de Ecología, UNAM, Laboratorio de Ecología y Conservación de Vertebrados Terrestres, Ap. Postal 70-275, C.P. 04510 Ciudad Universitaria, Mexico; ²⁰Bioconciencia A.C., Ciudad de México, Mexico; ²¹Nai Conservation, San José, Costa Rica; ²²Escuela de Biología, Universidad de Costa Rica, Ciudad Universitaria, San José 2060, Costa Rica; ²³Conservation International, Arlington, VA, USA; ²⁴Department of Wildlife and Fisheries Sciences, Texas A&M University, College Station, TX, USA; ²⁵Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA and ²⁶Department of Life Sciences, Imperial College London, Silwood Park Campus, Ascot, Berkshire, UK

Summary

Species distribution models (SDMs) are statistical tools used to develop continuous predictions of species occurrence. ‘Integrated SDMs’ (ISDMs) are an elaboration of this approach with potential advantages that allow for the dual use of opportunistically collected presence-only data and site-occupancy data from planned surveys. These models also account for survey bias and imperfect detection through the use of a hierarchical modelling framework that separately estimates the species–environment response and detection process. This is particularly helpful for conservation applications and predictions for rare species, where data are often limited and prediction errors may have significant management consequences. Despite this potential importance, ISDMs remain largely untested under a variety of scenarios. We performed an exploration of key modelling decisions and assumptions on an ISDM using the endangered Baird’s tapir (*Tapirus bairdii*) as a test species. We found that site area had the strongest effect on the magnitude of population estimates and underlying intensity surface and was driven by estimates of model intercepts. Selecting a site area that accounted for the individual movements of the species within an average home range led to population estimates that coincided with expert estimates. ISDMs that do not account for the individual movements of species will likely lead to less accurate estimates of species intensity (number of individuals per unit area) and thus overall population estimates. This bias could be severe and highly detrimental to conservation actions if uninformed ISDMs are used to estimate global populations of threatened and data-deficient species, particularly those that lack natural history

and movement information. However, the ISDM was consistently the most accurate model compared to other approaches, which demonstrates the importance of this new modelling framework and the ability to combine opportunistic data with systematic survey data. Thus, we recommend researchers use ISDMs with conservative movement information when estimating population sizes of rare and data-deficient species. ISDMs could be improved by using a similar parameterization to spatial capture–recapture models that explicitly incorporate animal movement as a model parameter, which would further remove the need for spatial subsampling prior to implementation.

Introduction

Species distribution models (SDMs) are a widely applied and rapidly developing statistical tool used in the study of wildlife, with new methods regularly proposed as solutions to various challenges encountered during modelling (Elith & Leathwick 2009, Franklin 2010). A deficiency of most SDMs is the failure to account for imperfect detection – the possibility that a species may go undetected even when it is present (Lahoz-Monfort et al. 2014). Occupancy models, a similar but distinct field of research from SDMs, account for this scenario by separating the species–environment response from that of the detection process through the use of a hierarchical modelling framework (MacKenzie et al. 2003). Another challenge for most SDMs is how to appropriately use presence-only (PO) data, which are often the most common type of data used in SDMs due to their ease of collection. This type of data is sometimes also referred to as presence-background (PB) for the class of models that combine PO data with the background environment in order to estimate species–environment responses. Recently, the challenge of using PO data in SDMs has been addressed through the use of point process models (Warton & Shepherd 2010, Renner et al. 2015). Integrated SDMs (ISDMs) represent a new development that uses both of these approaches, combining opportunistic (e.g., PO) and higher-quality site-occupancy (SO) data in the same model (Dorazio 2014, Fithian et al. 2015, Koshkina et al. 2017). ISDMs have potential as useful tools, but they require further investigation (i.e., sensitivity analyses), as there are few applied examples to follow.

Data simulation is a powerful tool used to answer questions about how models react to various user decisions (Zurell et al. 2010, Miller 2014). However, the design of simulated studies sometimes assumes data conditions that are unrealistic for many rare or cryptic species. The assumptions of the simulations used in the two studies that introduced ISDMs (Dorazio 2014, Koshkina et al. 2017) include a larger and more even sample than is typically available for most species. Simulation studies that do not mirror reality are especially problematic for a species like the endangered Baird’s tapir (*Tapirus bairdii*), which is wide ranging and relatively rare, leading to wide gaps in spatial coverage of high-quality presence data. Schank et al. (2017) applied an ISDM to c. 800 PO observations and 1600 camera trap detection histories (created from SO data) for Baird’s tapir. This research estimated a total population size of c. 200,000 individuals for the species, more than an order of magnitude higher than expert estimates, which range from 3000 mature adults to 16,500 total individuals (Medici et al. 2005, García et al. 2016). There are a variety of reasons that could explain this discrepancy, including violations of model assumptions and the sensitivity of the model to various modelling decisions. We focus on two assumptions: independence between sites and population closure. As with most statistical models, occupancy models require independent observations (MacKenzie et al. 2006). In this case, observations would not be independent if the same individual was detected at more than one site during the same observation period. In order to avoid this possibility, Schank et al. (2017) used a spatial subsampling procedure to enforce a minimum distance between sites, as many

of the sites were so close together that independence would be violated. Occupancy models also assume population closure, which states that no immigration or emigration of individuals from the site occurs during the sampling period (MacKenzie et al. 2006). Violation of the closure assumption can originate from a sampling period that is too long (Rota et al. 2009).

Our research here investigated the effect of user decisions on model outputs and population estimates when using ISDMs, focusing on how issues of spatial and temporal scale relate to the model assumptions above. Specifically, we investigated the effect of different settings for site area, subsampling radius and season length using data from our prior Baird’s tapir analysis (Schank et al. 2017). ISDMs have great potential as useful tools for conservation; however, researchers using these tools need clear recommendations for how to apply them, particularly when making conservation and management decisions for threatened and data-deficient species. The results of this research shed light on how these models can be applied appropriately to such species of conservation concern.

Methods

The complete sensitivity analysis covered three model formulations (PB, SO and integrated), four site area sizes and three season lengths, with 100 spatially subsampled iterations – a total of 3600 models. Custom R code was adapted from Dorazio (2014) and Royle and Dorazio (2008) to run the models.

Model Descriptions

With ISDMs, two separate models are formulated to accommodate the two types of data used (PB and SO), both based on a Poisson point process model. In these models, $\lambda(s)$ is the expected intensity (number of individuals per unit area) at location s . In the context of the SDM, $\lambda(s)$ is formulated as a log-linear function of unknown parameters and location-specific regressors $x(s)$:

$$\log(\lambda(s)) = \beta_0 + \beta'x(s)$$

The general class of models used here are hierarchical models, which have separate levels for abundance (the process of interest) and detection (the nuisance process). Though they share the same SDM based on a point process model, the two types of data use different model formulations to account for imperfect detection, including that which results from spatially biased survey effort. With opportunistic data, spatial bias and imperfect detection are incorporated through an independent thinning of the point process. This thinned point process is the product of the original point process and $p_{pb}(s)$, the probability that the site is surveyed and the species is detected. $p_{pb}(s)$ is formulated as a logistic function of unknown parameters and location-specific regressors $w_{pb}(s)$:

$$\log \text{it}(p_{pb}(s)) + \alpha_{0,pb} + \alpha_{pb}'w_{pb}(s)$$

With data from planned surveys (SO), imperfect detection is modelled following a zero-inflated binomial distribution (Koshkina et al. 2017). Under this model, the presence or absence of the species at a site i follows a Bernoulli distribution. In this case, the detection histories at each site, y_i , have non-detections (i.e., zeros) due to both species absence and imperfect detectability – the fact that an individual may go undetected even when present (MacKenzie et al. 2003). This relationship is modelled as a Binomial distribution with J trials and the probability of success (i.e., species detection) equal to the product of z_i (the occupancy state, $z_i = I(N_i > 0)$) and p_{so} , the probability of detection at the site. As with detectability in the PB model, $p_{so}(s)$ is formulated as a logistic function of unknown parameters and location-specific regressors $w_{so}(s)$:

$$\log \text{it}(p_{so}(s)) = \alpha_{0,so} + \alpha_{so}' w_{so}(s)$$

In ISDMs, the PB and SO models are estimated simultaneously, such that one set of parameters for the SDM is created (i.e., the β values), while separate detectability parameters are estimated (i.e., the α values) for the two models.

Model Convergence and Parameter Identifiability

We determined two levels of model convergence. First, the *optim* function in R was used to estimate model parameters from the model likelihood. Occasionally, this function failed to return an optimized set of parameters. Next, if estimates were returned from this function, we determined whether or not they were correctly identified using the reciprocal of the condition number. This number is the ratio of the smallest to the largest eigenvalues in the Fisher information matrix and can be used to determine whether the parameters of the SDM are identifiable (Dorazio 2014). The reciprocal of the condition number falls between 0 and 1, with values near zero indicating ill conditioning (Golub & Van Loan 2012). If this number fell below a certain threshold (in this case 1×10^{-6}), the results for that model were not used in the subsequent analysis.

Presence Data

The species presence data used here consisted of 784 PO observations compiled from 11 data sources and 1595 camera trap detection histories from 19 sources. These data came from a multinational collaboration examining Baird's tapir occurrence and distribution (for more details about the compilation and processing of the data, see Schank et al. 2017).

Spatial Subsampling

We processed the presence data prior to model fitting using a random spatial subsampling procedure to help preserve independence among sites. The algorithm began by randomly choosing one observation point and removing any other observation points within a given radius. We added the chosen observation to the subset and repeated the steps until no observations were left in the original data. A similar type of subsampling is sometimes used to remove survey bias in observation data (Beck et al. 2014). However, this grid-based approach can lead to samples that remain close in space if they fall just across a boundary in an adjacent grid cell.

The effect of this procedure was to enforce a minimum distance between sampling points. This minimum distance was matched to the site area to ensure that no site contained more than one data

point. For example, we used a subsampling radius of 5657 m for a site area of 16 km² (the diagonal length of a square that size). This process was repeated 100 times for each radius to capture the variability introduced by the randomness of the sampling. Model parameters were then averaged across iterations. A set of models also were fit on the complete (non-sampled) presence data to investigate the effect of violating the independence assumption.

Predictor Variables

Environmental variables were grouped into five classes: climate, land cover, anthropogenic, topographic and sampling variables (i.e., the variables used in the detection process). Climate variables at 1-km resolution were downloaded from CHELSA (Karger et al. 2017) and included annual precipitation, maximum temperature of the warmest month, temperature seasonality and precipitation seasonality. Land cover variables consisted of percentage tree cover for the year 2000 at 30-m resolution (Hansen et al. 2013), distance to/within protected areas (IUCN & UNEP-WCMC 2014) and mean enhanced vegetation index (EVI) from MODIS for years 2000–2015 downloaded using Google Earth Engine. Anthropogenic variables included forest loss between 2000 and 2014 (Hansen et al. 2013), road density (Eugster & Schlesinger 2010) and density of fires between 2001 and 2014 (NASA 2017). Anthropogenic variables were then converted to focal averages using a moving circular window and a 10-km radius (centred on each 1-km pixel in the study area) to account for the fact that humans are mobile and presence in one area means access is likely within a reasonable distance (Barber et al. 2014). Slope was calculated from 90-m resolution elevation data downloaded from the 'raster' package in R (Hijmans et al. 2016).

Sampling variables (i.e., those used in the detectability process) for the PB data included binary indicators for forest (Arino et al. 2012) and protected status (IUCN & UNEP-WCMC 2014) and distance to roads (Eugster & Schlesinger 2010), while sampling variables for SO data were the same tree cover and distance to/within protected areas used in the land cover group, as well as distance to roads and maximum slope. With the PB data, these variables were meant to capture sampling bias in the PO data, which we believe heavily favours forested and protected areas that are reasonably accessible by road. We included a quadratic term for distance to roads, as there could be optimal locations that are far enough from roads to minimize anthropogenic factors, but close enough to facilitate sampling. With SO data, the sampling variables were chosen as variables that might influence the detectability of the species. For example, tapir detectability might decrease as distance from protected areas increases due to likely increased levels of hunting outside of protected areas and thus increased response by the species to avoid humans (de la Torre et al. 2017, Ferregueti et al. 2017).

All variables were scaled to have a mean of zero and a standard deviation of one, except distance to/within protected areas (zero represents the border of the protected area). The models also incorporated quadratic terms for all climate variables, EVI and distance to road to account for their suspected non-monotonic relationships with tapir occurrence (aided by single-variable response curves created in the early stages of the modelling process).

We resampled all environmental variables to four spatial resolutions: 1, 2, 4 and 8 km. These resolutions correspond to site areas of 1, 4, 16 and 64 km². Estimates of home range size for Baird's tapir vary from 1.25 km² reported in Costa Rica (Foerster & Vaughan 2002) to 8–10 km² in Nicaragua (Jordan et al. 2019), while estimates of maximum distance travelled range up to

10.5 km in Mexico from camera trap data on a marked individual over 4 years (Reyna-Hurtado et al. 2016). From this information, it is possible that an individual could be detected at more than one site, but the likelihood of this is probably small, especially for the larger site areas used in this analysis.

Season Length

When creating camera trap detection histories, researchers can adjust their data structure by defining the length of each sampling occasion and the number of sampling occasions to use in a discretized season. Since camera traps operate continuously, there is some flexibility in determining the sampling occasion and season length. These decisions can be made (and adjusted) after the data are collected and will determine the balance of detections and non-detections. It is also important to consider the behaviour of the target species. For sampling occasion length, one suggestion is to select a length of time during which an individual will visit all or most of its home range, and GPS telemetry data suggest Baird's tapirs cycle through their home ranges about once every 10–12 days (Jordan 2015). For this reason, we used a sampling occasion length of 10 days.

With season length (i.e., number of sampling occasions), it is important to consider the assumption of closure and to choose a length of time during which immigration/emigration at a site is unlikely. As the season length increases, it becomes more likely that the assumption of closure will be violated. On the other hand, it is crucial to include enough sampling occasions to estimate detectability reliably. Some recommendations suggest three as the minimum number of occasions to use, though this number should be higher for species with low detectability (MacKenzie & Royle 2005). Baird's tapir is unsurprisingly a species with a low detection probability (range of 0.2–0.3) (Cove et al. 2014, Jordan 2015). Considering a detectability in this range, there is an approximately 4–13% chance a present individual will go undetected after nine sampling occasions. With the SO data, we tested season lengths of 30, 60 and 90 days (three, six and nine samples). The PB data contain only one sample, as there are not repeat observations at each site for this dataset.

Accuracy Assessment

We used two PO accuracy measures to assess the spatial predictions of each model: the Boyce Index (Boyce et al. 2002) and the minimum predicted area (MPA) (Engler et al. 2004). We did not use detection/non-detection data with accuracy measures that require presence-absence data given the difficulty of properly defining absences (Lobo et al. 2010) and given the bias of these measures when test data are missing from large portions of the study area (Bean et al. 2012). After the spatial subsampling step, the retained PO data were randomly split following a 75/25 training/testing ratio (Fielding & Bell 1997). For both accuracy measures, intensity was converted to occupancy, ψ , which ranges from 0 to 1, using the formula from Dorazio (2014), where N is the number of individuals in the spatial unit, C :

$$Pr(N(C) > 0) = \psi = 1 - \exp(-\mu(C))$$

$$\mu(C) = \int_C \lambda(s) ds$$

To calculate the Boyce Index, we partitioned the occupancy surface into bins (i.e., $0.0 < \psi < 0.1, \dots, 0.9 < \psi < 1.0$) and calculated the

percentage of test data occurring in each bin (P_i). We then compared the proportion of the area covered by the bin with respect to the study area (E_i). Finally, we converted these two measures to a ratio: $F_i = P_i/E_i$. If the model correctly predicts low-suitability areas, the low-suitability classes should contain fewer test points than expected by chance (i.e., $F_i < 1$) and the graph of F_i versus average suitability of each bin should be monotonically increasing. The Boyce Index is the correlation between the average suitability of each bin and its respective F_i , with values greater than zero indicating a model whose predictions are consistent with the test data and negative values indicating an incorrect model. The continuous version of this measure uses overlapping bins (Hirzel et al. 2006).

The MPA is the smallest possible area covered by a thresholded prediction map that contains at least 90% of the test PO points. The smaller the MPA, the more parsimonious the model and the less likely there are to be errors of commission in the predictions (Rupprecht et al. 2011).

Results

The ISDM converged (with estimated standard errors) in more than 97% of model iterations, while the PB model had low convergence rates across site areas and the SO model exhibited a sharp drop-off in convergence at 16 km² and above (See Supplementary Material, available online). Both measures of model accuracy showed that the ISDM was the most accurate framework, a relationship that was consistent across site area and number of samples (Fig. 1). Focusing on the ISDM, estimates of total population decreased exponentially as site area increased (Fig. 2). The number of samples used and whether the data had been subsampled had much smaller effects on population estimates.

The decrease in population estimates across site areas was driven by estimates of model intercepts, primarily β_0 , while the coefficients representing species–environment relationships remained relatively stable (Table 1). Annual precipitation (+), tree cover (+) and road density (–) were the three most important environmental variables in the model. Temperature seasonality (–), precipitation seasonality (+), maximum temperature of the warmest month (+), EVI (–), forest loss (–) and maximum slope (–) also appeared as significant environmental predictors. Annual precipitation was the only environmental variable with a clearly significant quadratic term. In the PB detectability process, presence in a protected area (+) was the most important variable. Distance to roads (–) was significant in both detectability components (PB and SO). Maximum slope (+) was also a significant variable in the SO detectability process.

Discussion

In the SO model, convergence decreased for larger site areas, possibly due to reduced sample sizes following the subsampling step (mean sample sizes: 1 km², 663; 4 km², 370; 16 km², 182; 64 km², 93). The ISDM was able to maintain convergence at these larger site areas possibly because of the added information from the PB data. However, there was a detectable decline in convergence with shorter season lengths at these larger site areas. Clearly, sample size is affected by both number of sites and number of repeated observations at those sites (MacKenzie & Royle 2005). In addition to higher rates of convergence, the ISDM was consistently the most accurate model. Taken together, these results demonstrate the importance of this new modelling framework. The ability to

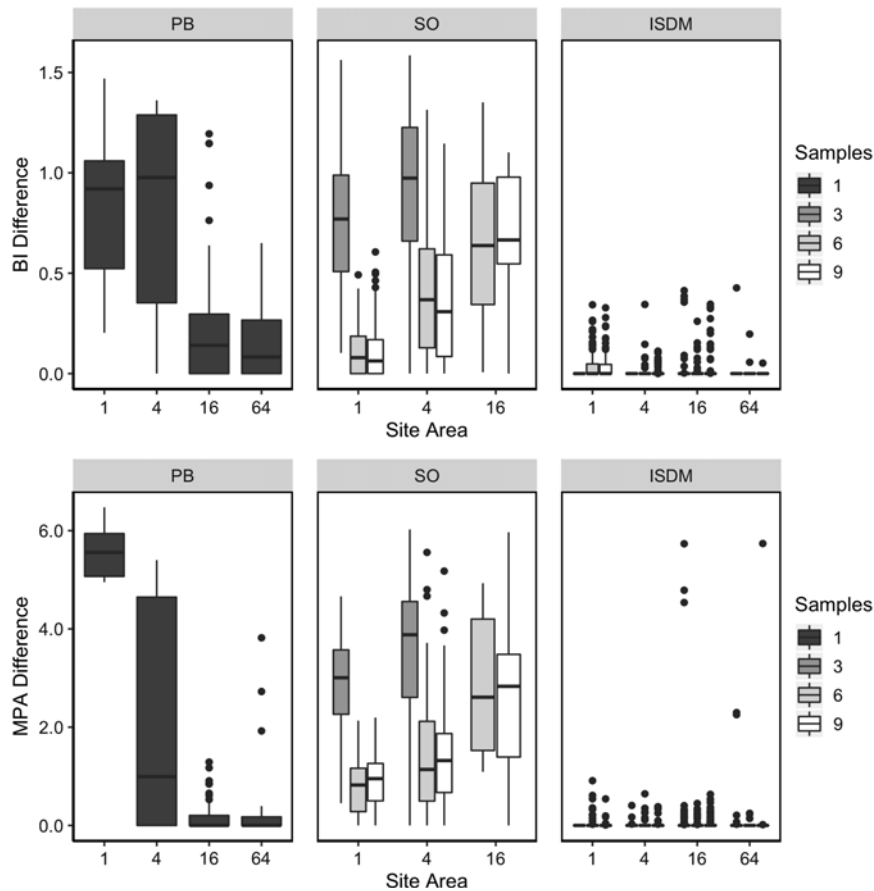


Fig. 1. Model accuracy using Boyce Index (BI) and minimum predicted area (MPA). The difference is calculated from the maximum (i.e., most accurate) BI and from the minimum (i.e., most accurate) MPA within each combination of model settings. The greater the difference (from zero), the less accurate the result. MPA differences have been rescaled to positive numbers and are represented in units of 100,000 km². ISDM = integrated species distribution model; PB = presence-background; SO = site occupancy.

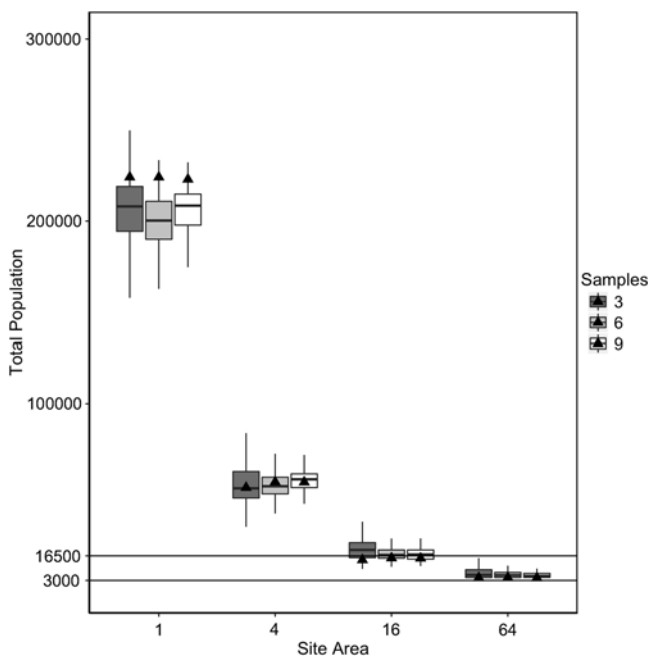


Fig. 2. Population estimates from the integrated species distribution model framework. Horizontal lines are placed at expert population estimates for the species of 3000 (current Red List assessment: Garcia et al. 2017) and 16,500 (Population Viability Assessment Report: Medici et al. 2005). Black triangles are the estimates for models run on the complete (not subsampled) set of presence data.

combine two types of presence data in the same model leads to better results.

In the ISDM, many of the species–environment relationships exhibited the expected outcome for this species (e.g., a preference for forest and avoidance of humans; Cove et al. 2014, Jordan et al. 2016). However, there are two results that contradicted expectations. First, EVI had a negative relationship with tapir intensity, as also seen in Schank et al. (2017). Tapir intensity should be positively associated with increasing vegetation (higher EVI) because vegetation is both a food source and it provides cover (Brooks et al. 1997, Pettorelli et al. 2011). This outcome could be explained if secondary forest is associated with higher EVI values, as there is evidence that tapir prefer these areas (Foerster & Vaughan 2002). In fact, subsequent modelling efforts that included an interaction term between forest cover and EVI provide evidence of this relationship (see chapter 4 in Schank 2018).

Also surprising was the positive relationship with maximum slope in the detectability of the SO data. This variable was included in this part of the model as it was suspected to either have a negative effect on sampling effort, because steep terrain is harder to sample, or on actual detectability of tapirs due to the same constraints, as difficult terrain is an impediment to wildlife movement as well (Bailey et al. 1996, Mair & Ruete 2016).

Clearly, the most important factor driving estimated population (and underlying magnitude of intensity) was the assumption about the size of our sampling unit, which we refer to as ‘site area’. The number of sampling occasions and whether or not the

Table 1. Coefficient estimates and standard errors for the integrated species distribution model (samples = 6) averaged across 100 model iterations fit on randomly subsampled presence data.

Coefficient	Site area			
	1	4	16	64
beta0	-1.403 (0.214)	-2.660 (0.258)	-3.800 (0.416)	-4.654 (0.812)
temp_seasonality	-0.527 (0.092)	-0.458 (0.109)	-0.438 (0.131)	-0.338 (0.163)
precip_seasonality	0.448 (0.134)	0.382 (0.160)	0.394 (0.195)	0.302 (0.240)
max_temp_warmest_month	0.291 (0.129)	0.236 (0.156)	0.204 (0.191)	0.115 (0.244)
annual_precip	1.448 (0.252)	1.398 (0.298)	1.428 (0.356)	1.410 (0.430)
temp_seasonality_sq	-0.651 (0.090)	-0.557 (0.105)	-0.410 (0.123)	-0.285 (0.148)
precip_seasonality_sq	-0.180 (0.089)	-0.164 (0.108)	-0.147 (0.130)	-0.129 (0.160)
max_temp_warmest_month_sq	0.066 (0.026)	0.050 (0.032)	0.023 (0.041)	-0.006 (0.055)
annual_precip_sq	-1.045 (0.202)	-1.100 (0.247)	-1.213 (0.312)	-1.225 (0.379)
treecover2000	1.692 (0.166)	1.568 (0.193)	1.496 (0.223)	1.309 (0.270)
distancePA	-0.005 (0.103)	-0.041 (0.119)	-0.099 (0.141)	-0.095 (0.172)
EVI	-0.554 (0.111)	-0.567 (0.140)	-0.649 (0.168)	-0.600 (0.229)
EVI_sq	-0.025 (0.051)	-0.090 (0.071)	-0.090 (0.078)	-0.176 (0.119)
forestloss_focal	-0.217 (0.056)	-0.179 (0.066)	-0.104 (0.078)	-0.040 (0.095)
road_length_focal	-1.147 (0.201)	-1.185 (0.244)	-1.314 (0.322)	-1.426 (0.411)
fire_density_focal	0.061 (0.091)	0.035 (0.112)	-0.032 (0.142)	-0.031 (0.163)
max_slope	-0.395 (0.080)	-0.357 (0.095)	-0.312 (0.117)	-0.215 (0.146)
alpha0.pb	-7.353 (0.243)	-6.310 (0.297)	-5.405 (0.465)	-5.050 (0.884)
alpha0.so	-1.326 (0.335)	-1.839 (0.526)	-2.316 (0.751)	-1.913 (0.858)
pb.forest	0.043 (0.183)	0.164 (0.224)	0.177 (0.271)	0.440 (0.350)
pb.protected	1.614 (0.180)	1.462 (0.203)	1.280 (0.235)	1.186 (0.281)
pb.road_distance	-0.792 (0.107)	-0.751 (0.127)	-0.721 (0.157)	-0.744 (0.207)
pb.road_distance_sq	0.146 (0.024)	0.137 (0.030)	0.125 (0.042)	0.116 (0.056)
so.treecover2000	0.301 (0.298)	0.751 (0.469)	0.954 (0.652)	0.348 (0.726)
so.distancePA	-0.043 (0.178)	-0.010 (0.232)	-0.162 (0.354)	-0.193 (0.468)
so.road_distance	-0.565 (0.222)	-0.811 (0.298)	-1.309 (0.469)	-1.590 (0.798)
so.road_distance_sq	0.103 (0.082)	0.166 (0.108)	0.211 (0.201)	-0.080 (0.607)
so.max_slope	0.316 (0.087)	0.307 (0.110)	0.344 (0.168)	0.412 (0.224)

presence data had been subsampled to preserve site independence had much smaller effects, with no discernible patterns. The models developed in Dorazio (2014) and Koshkina et al. (2017) explicitly incorporate site area in a way that is different from in traditional occupancy models. This likely explains the behaviour of total population changing proportional to the area of a grid cell used in the analysis. In the traditional formulation of an occupancy model (see panel 3.8 in Royle & Dorazio 2008), site area is not included anywhere in the model likelihood. The important question that remains is: what exactly constitutes a site? For sessile species or species that have small home ranges relative to the survey method, the site is easily defined as the area covered during the survey (i.e., a quadrat). However, when the species is relatively mobile, with a home range that is much bigger than the area covered in the survey, the concept of the site is less straightforward (Efford & Dawson 2012).

For example, when using camera traps, the cone of detection (i.e., the area in which a species can trigger the camera) is often very small in relation to the movements of the target species, which are typically large and mobile. The effective sampling area (ESA) is the area that contains the activity centres of any individuals that could come into contact with this cone of detection (Fig. 3) (White, 1982). This area should be approximately equal to the average home range of the species. Original estimates of the home range for Baird’s tapirs were *c.* 1 km² (Foerster & Vaughan 2002), which is surprisingly small for a species of its size. More recent estimates put the home range size closer to 10 km² (Jordan et al. 2019). The differences in reported estimates of home range size could be due to differences in the methodology used and differences in topography and the availability of resources, specifically regarding the availability of water. In mountainous sites with complex

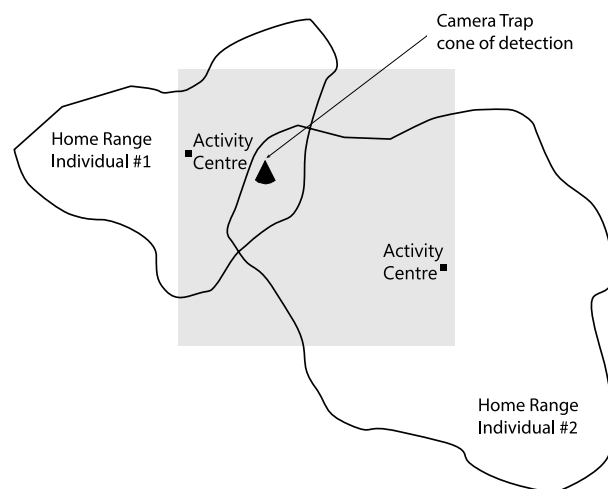


Fig. 3. Effective sampling area. A simplified diagram of two individuals with overlapping home ranges and a camera trap in the area of their intersection. The effective sampling area is equal to the area that incorporates the activity centres of all individuals detected at the camera.

topography and permanent availability of quality water throughout the year, the home range could be much smaller than in flat sites with very marked seasonality (Botello et al. 2017). Interestingly, using a site area of 16 km² provides a total population estimate that is within the range of expert estimates for the species (Fig. 2).

In Schank et al. (2017), the model was implemented using a site area of 1 km². To contextualize the population estimates from that model, the results were compared to multiple independent studies that focused on the estimated abundance of Baird’s tapir (Naranjo-

Piñera 1995, Gonzalez-Maya et al. 2012, Carbajal-Borges et al. 2014, Mejía-Correa et al. 2014, Botello et al. 2017). The estimates from those studies were similar to the estimates using the ISDM, which provided a conflicting story, as the total population estimates were thought to be overestimates by at least an order of magnitude. Most of these independent studies used capture–recapture methods and did control for the ESA; however, they used the old estimate of home range size from Foerster and Vaughan (2002). Clearly, accounting for the size of a species' home range and the variability in those estimates has a huge effect on abundance and population estimates using the ISDM (Fig. 2).

In fact, some authors have called into question the ability to use capture–recapture on a species like Baird's tapir, which does not have obvious and distinct markings by which individuals can be identified (Foster & Harmsen 2012). In the case of Baird's tapirs, sexually immature sub-adults lose their juvenile pelage before 1 year of age and develop very quickly (CA Jordan, pers. comm. 2018). In addition to making individual adults difficult to identify, this makes older juveniles effectively indistinguishable from mature adults in camera traps. This means that sexually immature individuals have likely been included in prior population estimates using capture–recapture methods, and this puts into doubt whether those studies accurately estimate the effective population size.

In addition to problems with misidentification of individuals, capture–recapture can overestimate species abundance due to the *ad hoc* correction of ESA (Noss et al. 2012), which can also lead to large errors in population estimates due to extrapolation (Foster & Harmsen 2012). These critiques recommend the use of the newer spatial capture–recapture (Royle et al. 2013), which explicitly accounts for species movement using an additional scaling parameter in the model.

Conclusion


Our research has demonstrated the potential connection between ISDMs and the ESA. Yet, the methods used in this research include *ad hoc* procedures that should be replaced by formal incorporation into the statistical model. In our models, accounting for the ESA is done in a way that matches site area with the best information about average movements for the species. Spatial capture–recapture provides an example for properly scaling the model to incorporate animal movement (Royle et al. 2013). Rather than approximating this effect through the selection of an appropriate site area, it would be better to combine concepts from spatial capture–recapture with the ISDMs used here.

Second, sometimes additional *ad hoc* steps must be taken in order to 'fix' the data. In this case, we used the spatial subsampling approach to avoid duplicate observations of the same individual at more than one site. Here, spatial capture–recapture can provide some guidance as well. These models require that sites are close enough to ensure that individuals are observed at more than one site, and they use this information to help estimate the spatial scalar of movement for the species. Thus, combining concepts from spatial capture–recapture with ISDMs may allow for the use of all data possible, although some alterations may be necessary for sites that have data covering more than one season (as the tapir data used here do).

A significant contribution of this research is the linkage between ESA and estimating abundance using ISDMs (or any other SDM/occupancy model). It is unclear why the discussion of ESA is almost entirely tied to capture–recapture models that use marked individuals. However, there is at least one study that

addresses this issue as it relates to occupancy models (Efford & Dawson 2012). The issue created by incorrectly accounting for ESA only becomes apparent in a small number of situations: when studying mobile species, estimating their abundance and extrapolating these estimates to produce population estimates. With Baird's tapir, these steps made it clear that something could be incorrect in our model. While it is possible to hypothesize multiple reasons for this disparity (see the conclusion section in Schank et al. 2017), the most straightforward answer is that ESA was not properly accounted for.

The incorporation of ESA into SDMs and occupancy models could use additional research. Failure to account for this properly could lead to inaccurate estimation of occupancy or abundance. However, as is seen in this research, species–environment relationships might remain the same. In order to improve our understanding, a future study using a detailed simulation (incorporating the movement of individuals) is needed. Future modelling efforts for this species should also explore unexpected species–environment relationships in more detail (e.g., negative associations between species presence and EVI and positive associations between species detectability and slope). It is possible that there is some interaction with other variables that caused these unexpected results. Finally, the reciprocal of the condition number used to determine parameter identifiability is new to species distribution modelling and should be investigated further.

Author ORCIDs.  Margot A. Wood, 0000-0001-6004-7378

Supplementary Material. For supplementary material accompanying this paper, visit <http://www.journals.cambridge.org/ENC>.

Author Contributions. CJS conceived the research; CJS, MVC, MJK, CKN, GO, NM, CAJ, JFG-M, DJL, RM, MD, VM, JCCD, GPM, JAT, EB-M, MAW and JG collected the data; CJS analysed the data; CJS, MVC and JAM led the writing. All authors contributed to writing the manuscript.

Financial Support. Funding for CJS during the research and writing of this manuscript was provided by the Donald D Harrington Fellowship through the Graduate School at the University of Texas at Austin. GPM: National Commission of Protected Natural Areas of Mexico and Biosphere Reserve Selva El Ocote Direction by the PROCER program 2014–2016. RR-H: El Colegio de la Frontera Sur in Mexico provided time and money to collect some data. NM and RM: funding from MWH, Ministerio de Ambiente de Panamá, GEMAS/Fondo Darién and Fundación Natura. JF: funding from PANTHERA, GEMAS/Fondo Darién and Fundación Natura, with additional support from the McIntire-Stennis Cooperative Forestry Research Program, Department of Forestry and Cooperative Wildlife Research Laboratory at Southern Illinois University. Permits from Ministerio de Ambiente de Panamá. Logistical support from PeaceCorps – Panamá and Azuero Earth Project. EB-M: Zoological Society of London and American Society of Mammalogists for providing funding. Sistema Nacional de Áreas de Conservación of Costa Rica for logistical support. JAT and MR: Conservation Program of Endangered Species (PROCER-Mexico) of the National Commission of Protected Areas (CONANP-Mexico) and the Natural Resources Protection Area La Fraileskana.

Conflict of Interest. None.

Ethical Standards. None.

Acknowledgements. Angelica Diaz-Pulido, Andrew Carver, Carolina Saenz Bolaños, Celso Poot, Eduardo Carrillo Jimenez, Eduardo Mendoza, Francisco Botello, Jessica Fort, Joel Saenz, Manolo García, Manuel Spinola, Marina Rivero, Nereyda Estrada, Oscar Godínez-Gómez, Rafael Reyna-Hurtado, Niall McCann, Raquel Leonardo and Sebastian Mejia are acknowledged for providing data.



References

- Arino O, Perez JJR, Kalogirou V, Bontemps S, Defourny P, Van Bogaert E (2012) Global land cover map for 2009 (GlobCover 2009). PANGAEA [www document]. URL <https://doi.pangaea.de/10.1594/PANGAEA.787668>.
- Bailey DW, Gross JE, Laca EA (1996) Mechanisms that result in large herbivore grazing distribution patterns. *Journal of Range Management* 49: 386–400.
- Barber CP, Cochrane MA, Souza Jr CM, Laurance WF (2014) Roads, deforestation, and the mitigating effect of protected areas in the Amazon. *Biological Conservation* 177, 203–209.
- Bean WT, Stafford R, Brashares JS (2012) The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. *Ecography* 35: 250–258.
- Beck J, Böller M, Erhardt A, Schwanghart W (2014) Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics* 19: 10–15.
- Botello F, Romero-Calderón AG, Sánchez-Hernández J, Hernández O, López-Villegas G, Sánchez-Cordero V (2017) Densidad poblacional del tapir centroamericano (*Tapirella bairdii*) en bosque mesófilo de montaña en Totontepec Villa de Morelos, Oaxaca, México. *Revista Mexicana de Biodiversidad* 88: 918–923.
- Boyce MS, Vernier PR, Nielsen SE, Schmiegelow FKA (2002) Evaluating resource selection functions. *Ecological Modelling* 157: 281–300.
- Brooks DM, Bodmer RE, Matola S (1997) *Tapir Action Plan*. Campo Grande, Brazil: IUCN/SSC Tapir Specialist Group.
- Carbajal-Borges JP, Godínez-Gómez O, Mendoza E (2014) Density, abundance and activity patterns of the endangered *Tapirus bairdii* in one of its last strongholds in southern Mexico. *Tropical Conservation Science* 7: 100–114.
- Cove MV, Pardo Vargas LE, de la Cruz JC, Spínola RM, Jackson VL, Saézn JC, Chassot O (2014) Factors influencing the occurrence of the endangered Baird's tapir *Tapirus bairdii*: potential flagship species for a Costa Rican biological corridor. *Oryx* 48: 402–409.
- de la Torre JA, Rivero M, Camacho G, 'lvarez-Márquez LA (2017) Assessing occupancy and habitat connectivity for Baird's tapir to establish conservation priorities in the Sierra Madre de Chiapas, Mexico. *Journal for Nature Conservation* 41: 16–25.
- Dorazio RM (2014) Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography* 23: 1472–1484.
- Efford MG, Dawson DK (2012) Occupancy in continuous habitat. *Ecosphere* 3: 1–15.
- Elith J, Leathwick JR (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics* 40: 677–697.
- Engler R, Guisan A, Rechsteiner L (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology* 41: 263–274.
- Eugster MJA, Schlesinger T (2010) osmar: OpenStreetMap and R. *R-Journal* [www document]. URL <http://osmar.r-forge.r-project.org/RJpreprint.pdf>.
- Ferreguetti ÁC, Tomás WM, Bergallo HG (2017) Density, occupancy, and detectability of lowland tapirs, *Tapirus terrestris*, in Vale Natural Reserve, southeastern Brazil. *Journal of Mammalogy* 98: 114–123.
- Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* 24: 38–49.
- Fithian W, Elith J, Hastie T, Keith DA (2015) Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution* 6: 424–438.
- Foerster CR, Vaughan C (2002) Home range, habitat use, and activity of Baird's tapir in Costa Rica. *Biotropica* 34: 423–437.
- Foster RJ, Harmsen BJ (2012) A critique of density estimation from camera-trap data. *Journal of Wildlife Management* 76: 224–236.
- Franklin J (2010) *Mapping Species Distributions: Spatial Inference and Prediction*. Cambridge, UK: Cambridge University Press.
- García M, Jordan CA, O'Farrill G, Poot C, Meyer N, Estrada N, ... Ruiz-Galeano M (2016) *Tapirus bairdii*. The IUCN Red List of Threatened Species [www document]. URL <http://dx.doi.org/10.2305/IUCN.UK.2016-1.RLTS.T21471A45173340.en>.
- Golub GH, Van Loan CF (2012) *Matrix Computations*. Baltimore, MD, USA: Johns Hopkins University Press.
- González-Maya JF, Schipper J, Polidoro B, Hoepker A, Zárrate-Charry D, Belant JL (2012) Baird's tapir density in high elevation forests of the Talamanca region of Costa Rica. *Integrative Zoology* 7: 381–388.
- Hansen MC, Potapov PV, Moore R, Hancher M, Turubanova SA, Tyukavina A, ... Townshend JRG (2013) High-resolution global maps of 21st-century forest cover change. *Science* 342: 850–853.
- Hijmans RJ, van Etten J, Cheng J, Mattiuzzi M, Sumner M, Greenberg JA, ... Wueest R (2016) Package 'raster' [www document]. URL <http://healthstat.snu.ac.kr/CRAN/web/packages/raster/raster.pdf>.
- Hirzel AH, Le Lay G, Helfer V, Randin C, Guisan A (2006) Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling* 199: 142–152.
- Jordan CA (2015) The dynamics of wildlife and environmental knowledge in a bioculturally diverse coupled natural and human system in the Caribbean region of Nicaragua. PhD thesis. East Lansing, MI, USA: Michigan State University.
- Jordan CAJ, Hoover B, Dans AJ, Schank C, Miller JA (2019) The impact of Hurricane Otto on Baird's tapir movement in Nicaragua's Indio Maíz Biological Reserve. In: *Movement Ecology of Neotropical Forest Mammals*, eds R Reyna-Hurtado, CA Chapman, pp. 5–20. New York, NY, USA: Springer.
- Jordan CA, Schank CJ, Urquhart GR, Dans AJ (2016) Terrestrial mammal occupancy in the context of widespread forest loss and a proposed interoceanic canal in Nicaragua's decreasingly remote south Caribbean region. *PLoS One* 11: e0151372.
- Karger DN, Conrad O, Böhrer J, Kawohl T, Keft H, Soria-Auza RW, ... Kessler M (2017) Climatologies at high resolution for the earth's land surface areas. *Scientific Data* 4: 170122
- Koshkina V, Wang Y, Gordon A, Dorazio RM, White M, Stone L (2017) Integrated species distribution models: combining presence-background data and site-occupancy data with imperfect detection. *Methods in Ecology and Evolution* 8: 420–430.
- Lahoz-Monfort JJ, Guillera-Arroita G, Wintle BA (2014) Imperfect detection impacts the performance of species distribution models. *Global Ecology and Biogeography* 23: 504–515.
- Lobo JM, Jiménez-Valverde A, Hortal J (2010) The uncertain nature of absences and their importance in species distribution modelling. *Ecography* 33: 103–114.
- MacKenzie DI, Nichols JD, Hines JE, Knutson MG, Franklin AB (2003) Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology* 84: 2200–2207.
- MacKenzie DI, Nichols JD, Royle JA, Pollock KH, Bailey LL, Hines JE (2006) *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*. Amsterdam, The Netherlands: Elsevier.
- MacKenzie DI, Royle JA (2005) Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology* 42: 1105–1114.
- Mair L, Ruete A (2016) Explaining spatial variation in the recording effort of citizen science data across multiple taxa. *PLoS One* 11: e0147796.
- Medici EP, Carrillo L, Montenegro OL, Miller PS, Carbonell F, Chassot O, ... Mendoza A (2005) *Baird's Tapir (Tapirus bairdii) Conservation Workshop Population and Habitat Viability Assessment (PHVA)*. Campo Grande, Brazil: IUCN/SSC Tapir Specialist Group.
- Mejía-Correa S, Díaz-Martínez A, Molina R (2014) Densidad y hábitos alimentarios de la danta *Tapirus bairdii* en el Parque Nacional Natural Los Katios, Colombia. *Tapir Conservation* 23: 16–23.
- Miller JA (2014) Virtual species distribution models: using simulated data to evaluate aspects of model performance. *Progress in Physical Geography* 38: 117–128.
- Naranjo-Piñera E (1995) Abundancia y uso de hábitat del tapir (*Tapirus bairdii*) en un bosque tropical húmedo de Costa Rica. *Vida Silvestre Neotropical* 4: 20–31.
- NASA (2017) MODIS Collection 6 NRT Hotspot/Active Fire Detections MCD14DL [www document]. URL <https://doi.org/10.5067/FIRMS/MODIS/MCD14DL.NRT.006>.
- Noss AJ, Gardner B, Maffei L, Cuéllar E, Montaña R, Romero-Muñoz A, ... O'Connell AF (2012) Comparison of density estimation methods for mammal populations with camera traps in the Kaa-Iya del Gran Chaco landscape. *Animal Conservation* 15: 527–535.

- Pettorelli N, Ryan S, Mueller T, Bunnefeld N (2011) The normalized difference vegetation index (NDVI): unforeseen successes in animal ecology. *Climate Research* 46: 15–27.
- Renner IW, Elith J, Baddeley A, Fithian W, Hastie T, Phillips SJ, . . . Warton DI (2015) Point process models for presence-only analysis. *Methods in Ecology and Evolution* 6: 366–379.
- Reyna-Hurtado R, Sanvicente-López M, Pérez-Flores J, Carrillo-Reyna N, Calmé S (2016) Insights into the multiannual home range of a Baird's tapir (*Tapirus bairdii*) in the Maya Forest. *THERYA* 7: 271–276.
- Rota CT, Fletcher Jr RJ, Dorazio RM (2009) Occupancy estimation and the closure assumption. *Journal of Applied Ecology* 46: 1173–1181.
- Royle JA, Chandler RB, Sollmann R, Gardner B (2013) *Spatial Capture-Recapture*. Amsterdam, The Netherlands: Elsevier Science.
- Royle JA, Dorazio RM (2008) *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*. Amsterdam, The Netherlands: Elsevier Science.
- Rupprecht F, Oldeland J, Finckh M (2011) Modelling potential distribution of the threatened tree species *Juniperus oxycedrus*: how to evaluate the predictions of different modelling approaches? *Journal of Vegetation Science* 22: 647–659.
- Schank CJ (2018) Investigation of novel methods to predict the distribution, abundance, and connectivity of rare species: a case study for the conservation of Baird's tapir. Doctoral dissertation. Austin, TX, USA: University of Texas at Austin.
- Schank CJ, Cove MV, Kelly MJ, Mendoza E, O'Farrill G, Reyna-Hurtado R, . . . Miller JA (2017) Using a novel model approach to assess the distribution and conservation status of the endangered Baird's tapir. *Diversity and Distributions* 23: 1459–1471.
- UNEP-WCMC (2014) The World Database on Protected Areas [data set] [www document]. URL <https://protectedplanet.net>.
- Warton DI, Shepherd LC (2010) Poisson point process models solve the 'pseudo-absence problem' for presence-only data in ecology. *Annals of Applied Statistics* 4: 1383–1402.
- White GC (1982) *Capture-Recapture and Removal Methods for Sampling Closed Populations*. Los Alamos, NM, USA: Los Alamos National Laboratory.
- Zurell D, Berger U, Cabral JS, Jeltsch F, Meynard CN, Münkemüller T, . . . Grimm V (2010) The virtual ecologist approach: simulating data and observers. *Oikos* 119: 622–635.