GfÖ
GfÖ Ecological Society of Germany, Austria and Switzerland
Basic and Applied Ecology 67 (2023) 61–69
Basic and Applied Ecology
www.elsevier.com/locate/baae

RESEARCH PAPER

"Mixed" occupancy designs: When do additional single-visit data improve the inferences from standard multi-visit models?

Gesa von Hirschheydt[a,*], Silvia Stofer[a], Marc Kéry[b]

[a]*Swiss Federal Research Institute for Forest, Snow and Landscape Research WSL, 8903 Birmensdorf, Switzerland*
[b]*Swiss Ornithological Institute, 6204 Sempach, Switzerland*

Received 27 May 2022; accepted 29 January 2023
Available online 31 January 2023

Abstract

Estimating occupancy while accounting for imperfect detection typically requires repeated surveys at sampling units. However, *mixed* sampling designs are very common, where only a subset of sites is visited repeatedly, while the remainder are visited only once, providing single-visit (SV) data. It is unclear whether SV data contribute to parameter estimates. Consequently, they have often been discarded in occupancy analyses. We conducted two simulation studies to understand the degree to which SV data contribute information to the estimation of occupancy and detection probability. In Simulation 1, we simulated detection/non-detection data under different scenarios of repeated sampling and varying magnitudes of occupancy and detection probabilities. In Simulation 2, we included continuous covariates, to see whether these could enhance the information content of SV data. To each simulated data set, we fitted models containing between 0 and 5000 SV sites and compared the standard errors of the occupancy and detection estimates. We found that SV data always contributed some information to the estimation of both occupancy and detection in a mixed design. Their relative contribution was greatest when > 2 visits were conducted at the repeated-visit sites, and for species with higher detection probabilities. These results suggest that SV data are valuable when combined with repeated-visit data and lead to more precise estimates than when repeated-visit data are used alone. Including suitable continuous covariates into the analysis of the simulated data increased the contribution of SV data even more. This suggests that, in a mixed design, occupancy estimation could be optimized by measuring and modelling continuous covariates that explain at least some heterogeneity in occupancy and detection amongst sites. Thus, we recommend that for mixed-design data all the available information be used in a joint model to obtain the most precise detection-corrected occupancy estimates.

Keywords: Occupancy models; Occupancy design; Imperfect detection; Repeated surveys; Single surveys; Data simulation

Introduction

Species distributions and the factors driving them have long fascinated ecologists. Species distributions are often expressed by occupancy probability, which is the probability with which a species occurs at a site (or any spatial unit), given the value of measured environmental and other variables at the site. Assessment of this probability is challenging due to the presence of measurement errors, the dominant one being the failure to detect individuals that are present, leading to false negatives. MacKenzie et al. (2002) and Tyre et al. (2003) proposed a simple, intuitive modelling

Abbreviations: Standard error, (SE); Single-visit, (SV); Multi-visit, (MV)
*Corresponding author.
E-mail address: gesa.vonhirschheydt@wsl.ch (G. von Hirschheydt).

https://doi.org/10.1016/j.baae.2023.01.003

framework that enables estimation of and correction for false-negative sampling errors. The error is represented as a parameter for detection probability, i.e., the probability that a species is detected during a survey given that it occurs at the site. The standard occupancy model requires a data set with multiple (i.e., replicated) observations for at least some of the sites, in the form of binary detections and non-detections. If the occupancy status of a site is constant across all observations (i.e., the assumption that the population is closed is not violated), the model permits estimation of the probability of occurrence and of detection separately. The former is of direct interest in a species distribution model, while the latter is typically treated as a nuisance parameter that must be accounted for to avoid bias in the primary estimation target. However, conducting multiple surveys is usually costly. It may be necessary to employ several field technicians simultaneously or to visit a site on different occasions. This requires additional resources that may not always be available.

The trade-off between a need for repeated visits for reliable estimation of detection probability and the desire to cover as many sites as possible is sometimes solved with a *mixed design*, where only a subset of all sites is surveyed multiple times while the remainder (often the vast majority) are visited only once. Contrary to integrated occupancy models, which combine data from different sources or sampling methods (e.g., Koshkina et al., 2017; Miller et al., 2019), under the mixed design all data are collected with the same method. Sites to be surveyed repeatedly may be chosen randomly (or according to some strata) amongst all sites or in practice often also haphazardly. For instance, in and around Switzerland alone, several monitoring programs use this strategy, including major contributors to the national Biodiversity Monitoring Switzerland (www.biodiversitymonitoring.ch), the Swiss National Forest Inventory (www.lfi.ch), and the Global Observation Research Initiative in Alpine Environments GLORIA (www.gloria.ac.at), and without a doubt, there are countless others. The primary goal of these schemes is often not to obtain detection-corrected occupancy estimates but presumably to cover as much heterogeneity of the landscape as possible in order to detect large-scale patterns or community changes over time. In many such schemes, repeated-visit data have not been used to estimate (or account for) species-specific detection probabilities, but merely to assess the reproducibility of the measurements (Nilsson & Nilsson, 1983). If it is deemed satisfactory, the real parameter of interest, i.e., occupancy or abundance, is then typically assessed using only the data from the first surveys (i.e., SV data only) and resulting estimates remain uncorrected for detection errors. However, whenever detection errors occur, estimates from these procedures will be biased (Guillera-Arroita, 2017; Guillera-Arroita et al., 2014; Kéry & Schmidt, 2008; Lahoz-Monfort et al., 2014; Ruiz-Gutiérrez & Zipkin, 2011). Instead of conducting separate analyses of data reproducibility and species distribution, we suggest that both can and should be achieved simultaneously by simply fitting an occupancy model (MacKenzie et al., 2002; Tyre et al., 2003) to the mixed data. Surprisingly though, there is a dearth of research on the efficiency of occupancy models that combine multi-visit (MV) data and single-visit (SV) data in a single model. MacKenzie and Royle (2005) investigated whether surveying a small number of sites with equal number of visits (standard design) was more efficient in terms of precision of the occupancy estimator than surveying a larger number of sites, but only some of them repeatedly and the rest only once (mixed design). They found that the standard design, with identical replication, is almost always more efficient for a given total number of surveys. However, they did not address the question of whether it pays, in terms of estimator precision, to add SV data into an analysis of an otherwise MV-only data set. Our aim with this study was to identify whether, how much, and under which conditions the addition of such SV data in a mixed design improves estimator precision. Additionally, we wanted to see whether adding continuous covariates to the model would further improve precision, since they have been shown to aid estimation in the case in which only SV data are available (Lele et al., 2012).

We addressed these questions using simulation, so that truth was known (Chapter 4 in Kéry & Royle, 2016). We conducted two simulation studies in which we generated detection/non-detection data under a mixed design and for widely varying scenarios defined by the number of SV sites as well as the magnitude of the probability of occupancy and detection. In Simulation 1, we focused on the effects of the number of MV sites and the number of SV sites in the simplest possible model without covariates. In Simulation 2, we investigated whether the utility of SV data can be enhanced by incorporation of continuous covariates.

Materials and methods

Data simulation

We used function simOcc in the R package AHMbook (Kéry et al., 2021) to simulate detection/non-detection data sets under a wide range of conditions and with or without the effects of a continuous covariate in either the occupancy or the detection portion of the model. Function simOcc first generates true presence/absence z at M sites based on a defined probability of occupancy Ψ, where Ψ can vary with environmental covariates in the form of a logistic regression. After generating true presence/absence data z, the function simulates detection/non-detection data y for J visits to each site, with a probability p of detecting the species during a visit to an occupied site, and a probability of 0 of detecting it at an unoccupied site. Variation in sampling conditions that may affect p can again be modelled with a logistic regression. We used this scheme to generate data in our two simulation studies.

Simulation 1

Here, we investigated under which conditions SV data contribute any information to the estimation of Ψ and p in a mixed design when no covariates are included in a model. We compared three schemes of repeated sampling:

- *Case2×150* with 2 visits each to 150 sites (i.e., number of visits $J = 2$, number of sites $M = 150$)
- *Case2×300* with 2 visits each to 300 sites
- *Case4×150* with 4 visits each to 150 sites.

We chose these numbers to reflect designs from moderately small to medium sample sizes. We further varied conditions by selecting a gradient for occupancy probability Ψ and detection probability p between 0.1 and 0.9 in steps of 0.02, and by considering five levels for the number of SV sites S added to the multi-visit (MV) data: 0, 150, 500, 1000, and 5000. For each combination of J, M, Ψ, and p, we initially simulated 1000 data sets with J visits and a total number of $N = 5000 + M$ sites. We then defined the first M sites of each data set to be the repeated-visit sites and turned observations from visits $J \geq 2$ in all remaining sites into NAs, so they became SV data. The full SV portion was then subset five-ways to produce the five levels of factor S, i.e., for each of the 1000 original data sets, we created a total of five variants corresponding to the five levels of the factor S. When a simulated data set contained either only detections or not a single detection amongst all MV sites, it was discarded (to avoid boundary estimates of the probability parameters in the model) and replaced by a new data set.

Simulation 2

In this set of simulations, we evaluated whether continuous covariates in occupancy or detection affect the degree to which SV sites contribute usable information in an occupancy model. We based all simulations on *Case2×150* above, i.e., where we assumed 2 visits each to 150 sites (i.e., $J = 2$, $M = 150$) and varied Ψ, p, and S in the same way as in Simulation 1. With covariates, Ψ and p here represent the intercepts expressed on the probability scale. Data under each parameter combination were simulated under four different covariate settings:

- *CovNull* without any covariates (identical to *Case2×150*)
- *CovOcc* with one continuous site-specific covariate for occupancy and none for detection
- *CovDet* with one continuous visit-specific covariate for detection and none for occupancy
- *CovBoth* with one continuous site-specific covariate for occupancy and one visit-specific covariate for detection.

Each covariate was randomly drawn from a standard normal distribution and was linked to the respective probability via a logistic regression model. The logit-scale effect of the occupancy covariate was simulated as -1, while the effect of the detection covariate was set to 1. We simulated 1000 data sets for each scenario and parameter combination.

Analysis of simulated data

Simulations and analyses were run in R (version 4.1.1; R Core Team, 2021) and static occupancy models (MacKenzie et al., 2002; Tyre et al., 2003) were fitted by maximum likelihood using function occu in the R package unmarked (Fiske & Chandler, 2011). We identified numerical failures in model fitting by the presence of either missing (NA), unreasonably large (> 3 on the logit scale) or unreasonably small (< 0.005) estimated standard errors ($\widehat{\text{SE}}$). Beyond a standard error of 3, confidence intervals of probabilities cover essentially the full range of values from 0 to 1, while standard errors below 0.005 unrealistically suggest near-perfect estimation and were always associated with boundary estimates (either occupancy or detection was estimated at 0 or 1). For data sets where the simulated detection and occupancy probabilities were low, the proportion of models with such numerical failures was substantial (up to 89%). These cases were ignored in the description of our results below, but we tally their frequency in Appendix 1. All analyses are based on $\widehat{\text{SE}}$s associated with the estimates on the original logit scale.

In Simulation 2, each data set was analysed with an occupancy model with identical covariate structure as in the data-generating model. To assess the amount of information contributed by the SV data in an occupancy fit, we analysed the magnitude of the $\widehat{\text{SE}}$, on the logit scale, and the rate of change in $\widehat{\text{SE}}$ as we went from 0 SV sites to 5000 SV sites added to a given number of MV site data.

For each combination of Ψ and p, we used $\widehat{\text{SE}}$ of the logit-scale $\widehat{\Psi}$ and $\widehat{p}$ to fit a generalized linear mixed model across all 1000 simulations using the function lmer in the R package lme4 (version 1.1–29; Bates et al., 2015) to investigate how logit-scale $\widehat{\text{SE}}$ changes with increasing number of SV sites:

$$\widehat{\text{SE}}(\widehat{\Psi}_{k,I}) = \gamma_0 + \gamma_1 * I_s + \delta_k + \varepsilon_{k,l}$$

where $\widehat{\text{SE}}(\widehat{\Psi}_{k,I})$ is the estimated standard error associated with the maximum likelihood estimate of Ψ for data set k and S factor level l, γ_0 is the intercept, γ_1 is the coefficient for the factor level I of S, δ_k is the random effect associated with the $k = 1\ldots1000$ simulated data sets, and $\varepsilon_{k,l}$ is the residual. We then conducted the analogous analysis also for $\widehat{\text{SE}}(\widehat{p}_{k,I})$. Note that we regressed the estimated SEs on the factor levels [0, 1, 2, 3, 4] of variable S instead of directly using the numbers [0, 150, 500, 1000, 5000]. Our reason for this is that we wanted a simple indicator for the magnitude of the change, and when plotting the SEs against different versions of the number of SV sites (e.g., raw numbers, log-transformed numbers, factor levels), the relationship to levels of factor S was most nearly linear. In this regression, the estimated intercept $\widehat{\gamma}_0$ represents the $\widehat{\text{SE}}$ of an estimate when

$S = 0$, i.e., when the model is fit to data with repeated visits only.

We assessed the contribution of SV data to estimator precision in two ways. The first one was to identify that part of the examined parameter space (if any) of occupancy and detection probability for which the precision of the estimated parameters improved when SV data were added, i.e., where $\widehat{\mathrm{SE}}$ become smaller. The second was to compare the magnitude of that improvement between different sampling schemes and covariate structures, wherein we define improvement as a negative estimated slope $\hat{\gamma}_1$ in the regression on the level of S described above.

In this paper, we focus on the precision of occupancy and detection estimators. The effect of SV data on the precision of covariate coefficients is discussed in Appendix 2. In order to keep within the scope of the paper, we only briefly discuss the accuracy of the estimates and refer the reader to the appendices for figures and a short discussion of the effect of SV data on estimator bias (Appendix 3) and root mean squared error (Appendix 4).

Results

Simulation 1

Of the 1000 simulated data sets generated for each parameter combination, 907 on average resulted in valid estimates (range: 111 - 1000; see Appendix 1 for the number of non-valid estimates). We found that the addition of SV data was always beneficial in terms of precision of the occupancy and detection estimates for all sampling schemes examined in Simulation 1 (Figs. 1 and 2). Regression slopes $\hat{\gamma}_1$ of $\widehat{\mathrm{SE}}(\hat{\Psi})$ against SV sites were consistently negative, i.e., the precision of estimates improved, for all combinations of Ψ and p and for all sampling schemes of the MV data (Fig. 1A). This improvement was greatest when Ψ was small and p was large.

For $\widehat{\mathrm{SE}}(\hat{p})$ too, the addition of SV data in an occupancy model always paid in terms of estimator precision: regression slopes were consistently negative for all combinations of Ψ and p (Fig. 1B). However, unlike for $\widehat{\mathrm{SE}}(\hat{\Psi})$, the contribution of SV data was greatest when both Ψ and p were

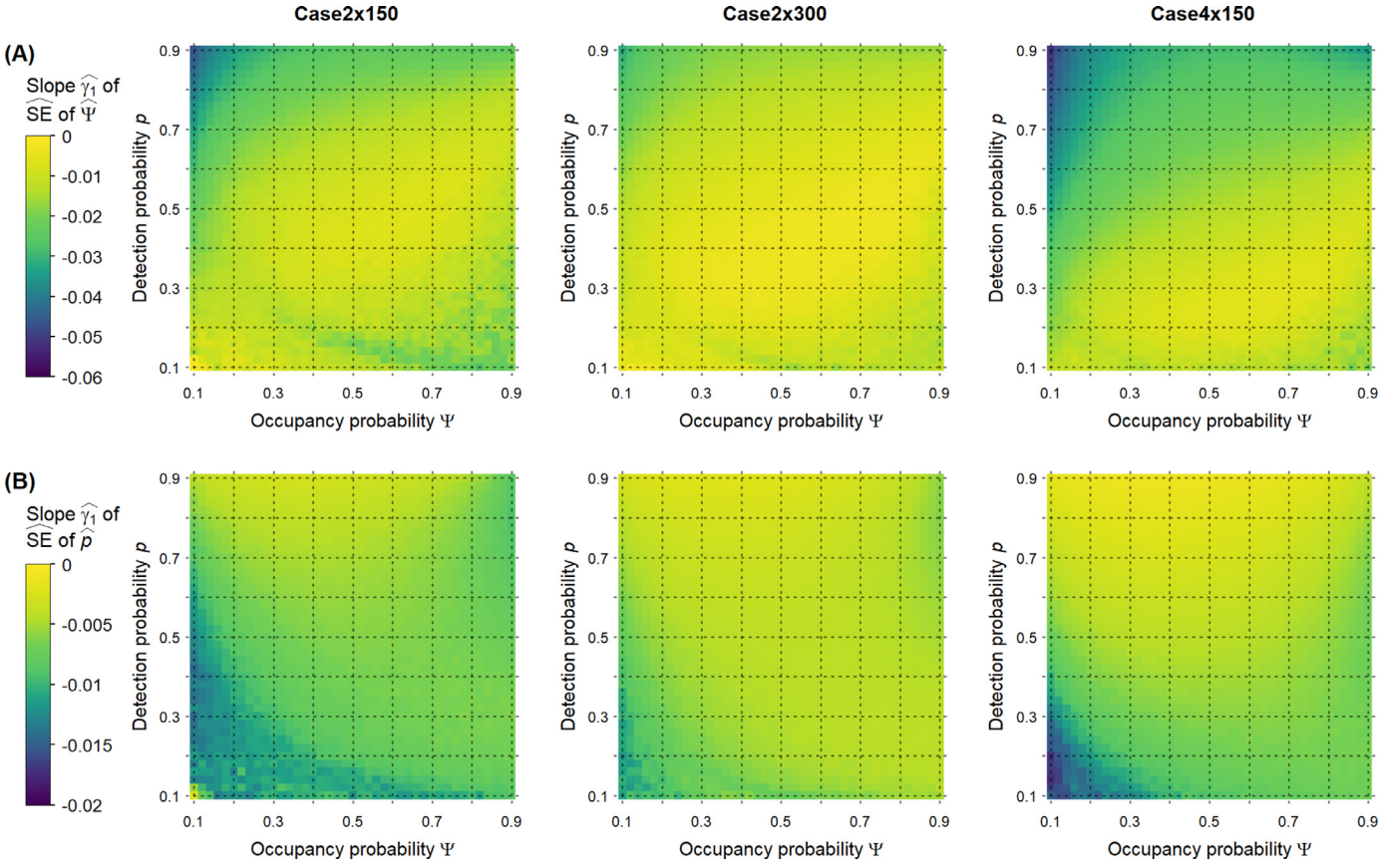

Fig. 1. Heatmaps showing slope $\hat{\gamma}_1$ of linear regressions of the estimated standard errors of $\hat{\Psi}$ (A) and of $\hat{p}$ (B) against the number of single-visit sites I_S in relation to true occupancy Ψ and detection probability p (along the axes). Columns represent the different cases of repeated sampling with (left) 2 visits each to 150 sites (in addition to 0–5000 SV sites), (middle) 2 visits each to 300 sites, and (right) 4 visits each to 150 sites. Note that a negative slope indicates an improvement of estimator precision with increasing numbers of single-visit sites added.

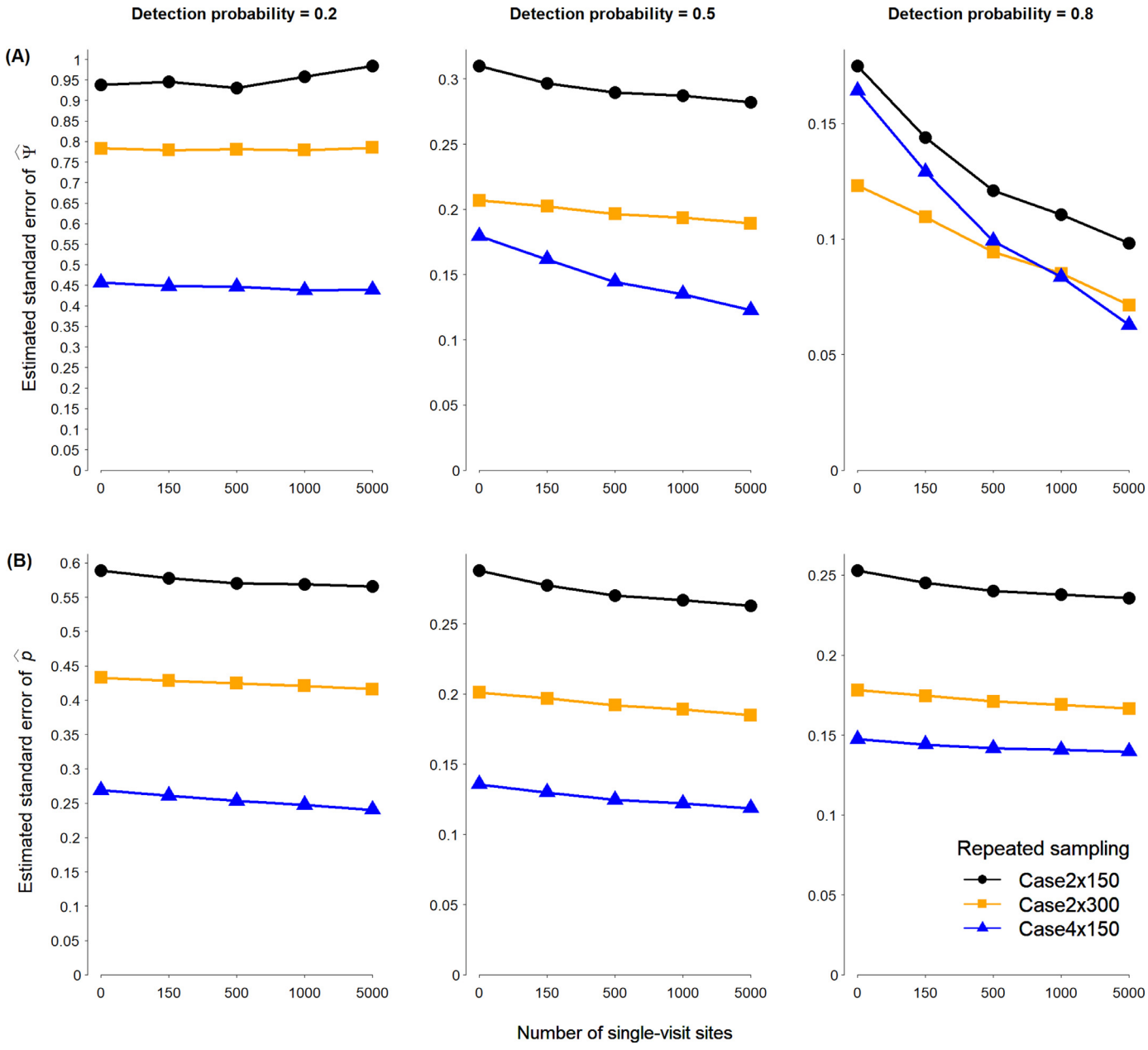

Fig. 2. Plots showing (on the y axis) the average magnitude across simulations of the estimated standard error of $\hat{\Psi}$ (A) and of $\hat{p}$ (B) as more single-visit sites (on the x axis) are added to the analysis (Simulation 1). Columns represent different combinations of data-generating rates where Ψ is kept constant at 0.5 and p is set to 0.2 (left), 0.5 (middle), and 0.8 (right). Note that the scale of the y-axes varies, but ticks are drawn consistently at intervals of 0.05.

low. Doubling the number of repeated-visit sites (*Case2×300*) reduced the relative contribution of additional SV data compared to *Case2×150* for the entire parameter space. Doubling the number of visits (*Case4×150*) on the other hand increased the contribution of SV data on the estimation of $\hat{\Psi}$, especially when the detection probability was high, as illustrated by the more negative slopes of $\widehat{\mathrm{SE}}(\hat{\Psi})$ with additional SV data (Figs. 1 and 2). Both *Case2×300* and *Case4×150* show considerably lower absolute $\widehat{\mathrm{SE}}$ compared to *Case2×150* (see the intercepts in Fig. 2) due to the greater information content of the MV data.

Overall, the $\widehat{\mathrm{SE}}$ of $\hat{\Psi}$ improved a lot more than the $\widehat{\mathrm{SE}}$ of $\hat{p}$ for the same number of additional SV sites. For example, for a moderately common species ($\Psi = 0.5$) which is easily detected ($p = 0.8$) and a repeated sampling scheme of *Case2×150*, the addition of data from only 500 SV sites reduced the $\widehat{\mathrm{SE}}(\hat{\Psi})$ from 0.175 to 0.121 (Fig. 2A), which represents a reduction of 31%. For the same settings, the $\widehat{\mathrm{SE}}(\hat{p})$ decreased only by 5%, from 0.253 to 0.240. Even for combinations of Ψ and p for which the contribution of SV data on $\widehat{\mathrm{SE}}(\hat{p})$ is greater, adding 500 SV sites never reduced the $\widehat{\mathrm{SE}}(\hat{p})$ by more than 14% (results not shown).

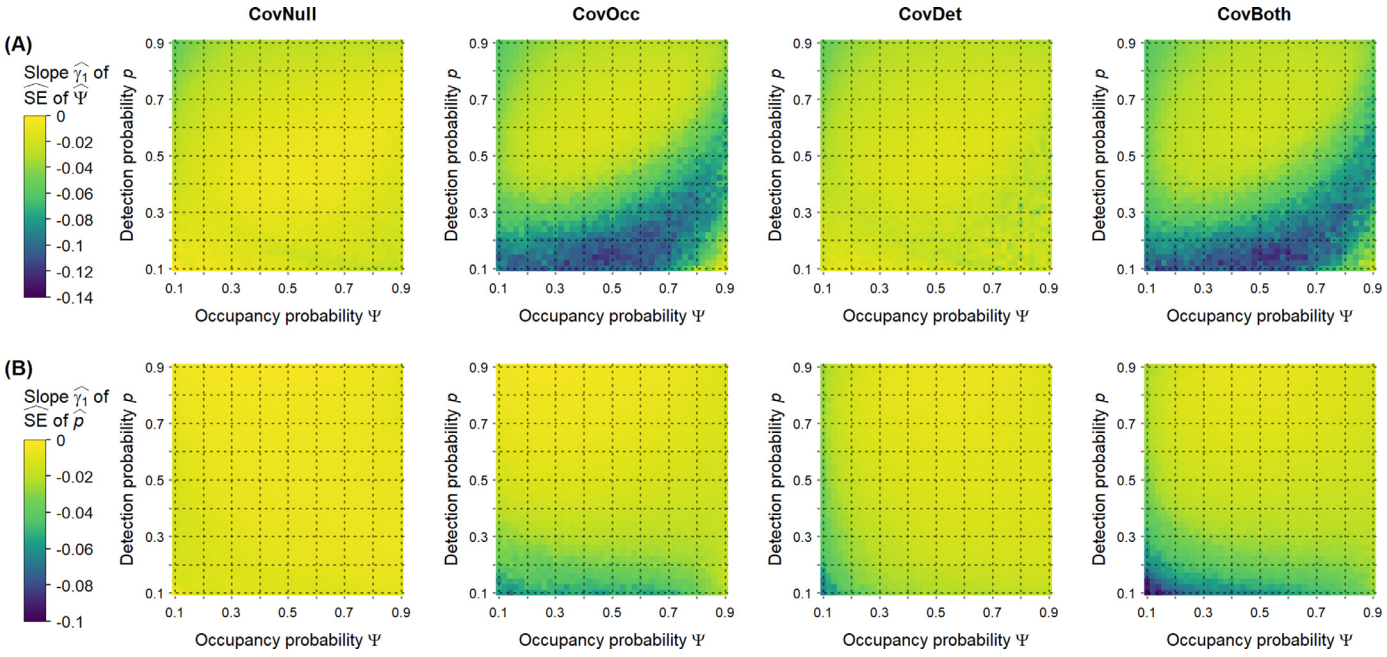

Fig. 3. Heatmaps showing slope $\hat{\gamma}_1$ of linear regressions of the estimated standard errors of $\hat{\Psi}$ (A) and of $\hat{p}$ (B) against the number of single-visit sites I_S in relation to true occupancy Ψ and detection probability p (along the axes). Columns represent the different covariate settings: *CovNull* is the intercept-only model and is identical with *Case2×150* in Simulation 1, *CovOcc* has one covariate for occupancy and none for detection, *CovDet* has one covariate for detection and none for occupancy, and *CovBoth* has one continuous covariate each. Colour indicates the magnitude of the regression slope for each combination of values of Ψ and p in the parameter space for 1000 simulated data sets each. Note that the colour scales differ from those in Fig. 1. All models include 150 sites visited twice.

In terms of accuracy, both estimates showed negligible deviations from the truth for species with $p \geq 0.5$ for *Case2×150*. Below this threshold, occupancy was slightly under- and detection probability slightly overestimated which confirms previous findings (e.g., MacKenzie et al., 2002). The magnitude of the bias decreased with additional MV sites or greater number of repeated visits, but was not affected by the number of SV sites (Appendix 3).

Simulation 2

On average, 905 simulations per parameter combination resulted in valid estimates (range: 111–1000). We found that the inclusion of a site-specific occupancy covariate (in scenarios *CovOcc* and *CovBoth*) strongly enhanced the contribution of SV data to the estimation of occupancy (Fig. 3A) but did only slightly do so in the estimation of detection (Fig. 3B). The steepest slope, i.e., the greatest improvement in $\widehat{SE}$, was found for species with low detection probability and moderate occupancy probability (Fig. 4). Precision of the covariate coefficients improved likewise, with the occupancy covariate showing a pattern similar to that of the occupancy estimate, and the detection covariate showing a pattern similar to that of the detection estimate (Appendix 2).

Adding a visit-specific covariate of detection into the model (*CovDet*) slightly increased the usable information content of SV data for the estimation of detection, but only for species with very low Ψ and p. However, it had little effect on the occupancy estimate.

In terms of accuracy, the coefficient of the occupancy covariate showed some bias when detection probability was low in combination with very low or high values of occupancy probability (i.e., near 0.1 or 0.9; Appendix 3). The coefficient of the detection covariate showed a slight positive value for all parameter combinations. For both covariates, accuracy improved substantially when SV data were added to the data set (Appendix 3).

Discussion

We explored what we call "mixed sampling designs" in an occupancy modelling framework. That is, where one portion of the sites is sampled multiple times, as in the standard occupancy design (MacKenzie et al., 2002; Tyre et al., 2003), and the other portion is sampled only once. Across the scenarios examined, we found that the addition of SV data always improved estimator precision for both occupancy and detection probability when compared to the analysis of data from the MV sites alone. However, the magnitude of this gain in precision varied, and depended on the magnitude of species occupancy and detection, as well as on the type of covariate included. In general, estimates of occupancy $\widehat{\Psi}$ benefited more from adding SV data than did estimates of detection probability $\widehat{p}$.

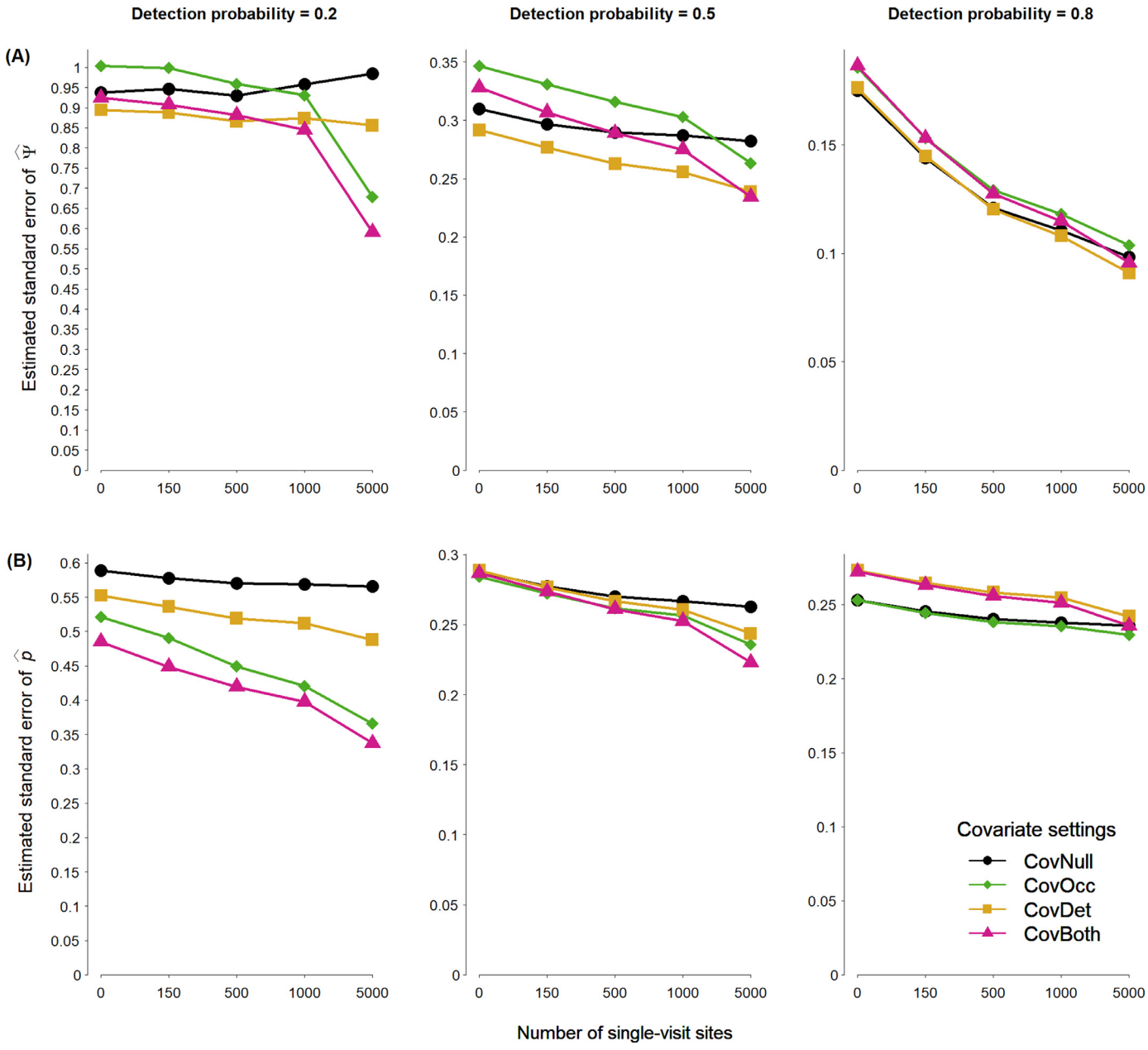

Fig. 4. Plots showing the average magnitude across simulations of the estimated standard error of $\hat{\Psi}$ (A) and of $\hat{p}$ (B) against the number of single-visit sites and with or without covariates in the model (Simulation 2). Columns represent different combinations of data-generating rates where Ψ is kept constant at 0.5 and p is set to 0.2 (left), 0.5 (middle), and 0.8 (right). The four lines illustrate different covariate settings. Note that the scale of the y-axes varies, but ticks are drawn consistently at intervals of 0.05.

Different repeated-sampling schemes

The standard occupancy design uses the MV portion of the data to provide information on detection probability. Expanding the starting data set from 150 sites visited twice (*Case2×150*) with additional visits or additional MV sites should therefore improve overall estimator precision. As expected, we found that doubling the number of sites with two visits (*Case2×300*) improved estimates in terms of precision, but it reduced the relative improvement of precision that was observed when SV data were added to the analysis. Doubling the number of visits (*Case4×150*) strongly improved precision of the estimates, especially when detection probability was low, as has previously been reported (MacKenzie & Royle, 2005; Reich, 2020). An interesting and new finding from our study is that a larger number of visits in the repeated portion of the data also has benefits for the relative contribution of SV data: when MV sites are visited four times, additional SV data carry relatively more information (and hence precision is improved relatively more) than when MV sites are visited only twice. In other words, greater precision in the detection estimate obtained through greater

number of visits allows the model to make better use of the information on occupancy contributed by the SV data.

Different covariate structures

Covariates may carry valuable additional information on the probability that a site is occupied or not. One may therefore expect that the inclusion of one, or better two, continuous covariates (i.e., at least one at site level and another at the visit level) would make it easier for the model to utilize additional SV data. Our simulations clearly confirmed this for the estimation of occupancy: Incorporating a covariate for occupancy strongly improved the contribution of SV data to estimation of $\widehat{\Psi}$ for species with low to moderate detection probabilities. We further found that covariates also improved estimates of detection $\widehat{p}$, but here the improvement was less pronounced and restricted to cases where detection probabilities were low. As standard occupancy models require repeated visits to estimate detection probability, it may seem counterintuitive that SV data should improve estimates of detection at all. In fact, MacKenzie et al. (2003) write "repeated surveys may be restricted to a subsample of sites in order to collect sufficient information for estimating detection probabilities, which can then be applied to those sites only visited once" suggesting that the detection estimate is informed only by the MV data. Our results show, however, that SV data can actually improve estimates of detection probabilities, especially when the model includes an occupancy covariate and when overall detection probability is low. This means that by adding information about the occupancy status of a site, an occupancy covariate indirectly contributes information about the detection probability. Finding suitable occupancy covariates should be relatively easy. Potential covariates may be elevation, yearly mean temperatures or precipitation, vegetation density, proximity to water or to human settlements.

Adding a detection covariate that varies at the visit level had little effect on the contribution of SV data to the estimation of $\widehat{\Psi}$, but slightly improved their contribution for $\widehat{p}$. As with the occupancy covariate, this effect was more prominent when detection probability was low, but the overall improvement was smaller than for the occupancy covariate. Not unexpectedly, the greatest benefits were obtained when the model contained one unique (or "private") continuous covariate each for occupancy and detection. Lele et al. (2012) and Sólymos et al. (2012) used continuous covariates to estimate detection probability (separately from occupancy) from data of SV sites alone. In contrast to a design with purely SV data, a mixed design does not *require* the use of continuous covariates to guarantee parameter identifiability. Therefore, it may also be more robust to assumption violations compared to a model fit to SV-only data (Knape & Korner–Nievergelt, 2015). Our results show that especially analyses of difficult-to-detect species, i.e., species with a low detection probability, can be greatly improved when adequate covariates are included. Examples of possible detection covariates that may vary between visits are date, climatic conditions during sampling (temperature, rainfall, wind, cloud cover, etc.), or some continuous measure of observer experience such as the proportion of species successfully identified in a test. We note that categorical covariates such as observer identity should not be expected to be informative in this regard (Lele et al., 2012). We did not assess the effect of detection covariates that vary at site level, but we would expect it to be of similar magnitude.

Implications for survey design

Several studies have evaluated the performance of various sampling designs in occupancy studies and tried to identify optimal strategies for a constant total survey effort (e.g., Guillera-Arroita, 2017; Guillera-Arroita et al., 2010; MacKenzie & Royle, 2005; Reich, 2020). We emphasize that our goal was *not* to show that mixed designs are particularly powerful at estimating occupancy and detection rates when compared with other designs. In fact, ever since MacKenzie and Royle (2005) it has been known that a mixed design, which they called "double sampling", is rarely the ideal solution when the aim is to obtain a precise estimate of occupancy with a fixed number of surveys. Rather, the aim of our study was to provide guidance for an analysis in the common situation *when a survey has already been conducted*, and when both SV and MV data are available. Especially in vegetation studies (vascular plants, mosses, fungi, and lichens), there are numerous data sets that have such a mixed structure, both from past surveys and from ongoing monitoring programs and often, separate analyses are conducted of the SV and the MV data, perhaps because the mixed data does not seem ideal for any joint analysis. Our results show, however, that the MV and SV data from such studies can be analysed jointly and that this will improve the estimates of both occupancy and detection probability, even if the design was not optimized for the purpose of correcting for imperfect detection in the most efficient way. Thus, our take-home message is this: if you have additional single-visit data in the analysis of standard (i.e., repeated-visit) occupancy data, then use them all in a single occupancy model.

Data Availability

All R code used in this study to simulate, analyse, and visualize the data is available on https://doi.org/10.5281/zenodo.7272029.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are very grateful to Jim Baldwin, Ken Kellner, and Andy Royle for statistical advice. We also thank Christoph Scheidegger and Stefan Ekman for feedback on project ideas and manuscript, and Oliver Hawlitschek, Pius Korner, and Wesley Hochachka, and three anonymous reviewers for carefully reviewing earlier drafts of the manuscript. This work was supported by the Federal Office for the Environment (FOEN; project number 00.5147.PZ/0DF268D21).

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.baae.2023.01.003.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. doi:10.18637/jss.v067.i01.

Fiske, I., & Chandler, R. (2011). unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*, *43*(10), 1–23. doi:10.18637/jss.v043.i10.

Guillera-Arroita, G. (2017). Modelling of species distributions, range dynamics and communities under imperfect detection: Advances, challenges and opportunities. *Ecography*, *40*(2), 281–295. doi:10.1111/ecog.02445.

Guillera-Arroita, G., Lahoz-Monfort, J. J., MacKenzie, D. I., Wintle, B. A., & McCarthy, M. A. (2014). Ignoring imperfect detection in biological surveys is dangerous: A response to 'fitting and interpreting occupancy models'. *PloS one*, *9*(7), e99571. doi:10.1371/journal.pone.0099571.

Guillera-Arroita, G., Ridout, M. S., & Morgan, B. J. T. (2010). Design of occupancy studies with imperfect detection. *Methods in Ecology and Evolution*, *1*(2), 131–139. doi:10.1111/j.2041-210X.2010.00017.x.

Kéry, M., Royle, A., & Meredith, M. (2021). AHMbook: Functions and data for the book 'Applied hierarchical modeling in ecology' vols 1 and 2 *(R package version 0.2.3)*. https://CRAN.R-project.org/package=AHMbook.

Kéry, M., & Royle, J. A. (2016). *Applied hierarchical modeling in ecology: Analysis of distribution, abundance and species richness in R and BUGS*. Elsevier/AP.

Kéry, M., & Schmidt, B. (2008). Imperfect detection and its consequences for monitoring for conservation. *Community Ecology*, *9*(2), 207–216. doi:10.1556/ComEc.9.2008.2.10.

Knape, J., & Korner-Nievergelt, F. (2015). Estimates from non-replicated population surveys rely on critical assumptions. *Methods in Ecology and Evolution*, *6*(3), 298–306. doi:10.1111/2041-210X.12329.

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R. M., 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*(4), 420–430. doi:10.1111/2041-210X.12738.

Lahoz-Monfort, J. J., Guillera-Arroita, G., & Wintle, B. A. (2014). Imperfect detection impacts the performance of species distribution models. *Global Ecology and Biogeography*, *23*(4), 504–515. doi:10.1111/geb.12138.

Lele, S. R., Moreno, M., & Bayne, E. (2012). Dealing with detection error in site occupancy surveys: What can we do with a single survey? *Journal of Plant Ecology*, *5*(1), 22–31. doi:10.1093/jpe/rtr042.

MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G., & Franklin, A. B. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, *84*(8), 2200–2207. doi:10.1890/02-3090.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J. A., & Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, *83*(8), 2248–2255. doi:10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2.

MacKenzie, D. I., & Royle, J. A. (2005). Designing occupancy studies: General advice and allocating survey effort. *Journal of Applied Ecology*, *42*(6), 1105–1114. doi:10.1111/j.1365-2664.2005.01098.x.

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

Nilsson, S. G., & Nilsson, I. N. (1983). Are estimated species turnover rates on islands largely sampling errors? *The American Naturalist*, *121*(4), 595–597. doi:10.1086/284087.

R Core Team. (2021). R: A language and environment for statistical computing *(4.1.2)*. R Foundation for Statistical Computing. https://www.R-project.org/.

Reich, H. T. (2020). Optimal sampling design and the accuracy of occupancy models. *Biometrics*, *76*(3), 1017–1027. doi:10.1111/biom.13203.

Ruiz-Gutiérrez, V., & Zipkin, E. F. (2011). Detection biases yield misleading patterns of species persistence and colonization in fragmented landscapes. *Ecosphere*, *2*(5), 1–14. doi:10.1890/ES10-00207.1.

Sólymos, P., Lele, S., & Bayne, E. (2012). Conditional likelihood approach for analyzing single visit abundance survey data in the presence of zero inflation and detection error. *Environmetrics*, *23*(2), 197–205. doi:10.1002/env.1149.

Tyre, A. J., Tenhumberg, B., Field, S. A., Niejalke, D., Parris, K., & Possingham, H. P. (2003). Improving precision and reducing bias in biological surveys: Estimating false-negative error rates. *Ecological Applications*, *13*(6), 1790–1801. doi:10.1890/02-5078.