Generalized sample verification models to estimate ecological state variables with detection-nondetection data while accounting for imperfect detection and false positive errors

John D. J. Clare[1*], Benjamin Zuckerberg[1], and Philip A. Townsend[1]

[1]Department of Forest and Wildlife Ecology, University of Wisconsin – Madison, Madison, Wisconsin

[*]jclare2@wisc.edu; 1630 Linden Drive, Madison, Wisconsin 53706

Abstract

Spatially indexed repeated detection-nondetection data is widely collected in ecological studies in order to estimate parameters such as species distribution, relative abundance, density, richness, or phenology while accounting for imperfect detection. Given growing evidence that false positive error is also present within most data, more recent model development has focused on also explicitly accounting for this error type. To date, however, most modeling efforts have improved occupancy estimation. We describe a generalizable structure for using verified samples to estimate and account for false positive error in detection-nondetection data that can be flexibly implemented within many existing model types. We use simulation to demonstrate that estimators for relative abundance, arrival time, and density exhibit bias under realistic levels of false positive prevalence, and that our modified estimators improve performance. As ecologists increasingly use expedient but potentially error-prone techniques to classify growing volumes of data, properly accounting for misclassification will be critical for sound ecological inference. The generalized model structure presented here provides ecologists possessing even a small amount of verified data a means to correct for false positive error and estimate several state variables more accurately.

Introduction

Binary detection non-detection data is widely used in ecology because it directly describes many variables of interest, can be used to derive many other variables, and is more reliably collected and more easily modeled than continuous, count, or categorical types. Repeated detection-nondetection data has a long history of use within ecology for the purposes of accounting for zero-inflation, first within capture-recapture studies (e.g., Otis et al. 1978). The logistical difficulties associated with marking or repeatedly identifying individual organisms across large

spatial or temporal domains are much greater than repeatedly identifying species or other organism states across those domains. Unsurprisingly, there has been rapid recent adoption of space or site-structured models that record species or other phenomena repeatedly at specific locations to estimate parameters associated with species distribution, relative abundance, density, or phenology (MacKenzie et a. 2002, Royle and Nichols 2003, Roth et al. 2014, Ramsey et al. 2015) or associated dynamics while still accounting for the imperfect detection that motivated capture-recapture studies.

However, a growing body of evidence suggests that binary data are collected with both false negative observation error and false positive error. Across a variety of protocols, false positives interspecifically range from nearly negligible to constituting 20% of observations or more (Simons et al. 2007, McClintock et al. 2010, Swanson et al. 2016, Norouzzadeh et al. 2018). This has motivated model extensions to account for false positive error as well as false negative error in binary data using a variety of sampling techniques (e.g., Miller et al. 2011, Chambert et al. 2015). Simulation and empirical studies indicate that ignoring false-positive error can lead to severely biased inference regarding occupancy or occupancy dynamics, and that using false positive extensions provides more reliable inference (e.g., Miller et al. 2015). However, efforts to address false positives have largely been confined to occupancy estimation even though any parameter that can be estimated with binary data is likely to be sensitive to un-assumed error.

Here, we address this issue by first reformulating the sample-verification false positive model for occurrence presented by Chambert et al. (2015) to make it more easily extensible to other models. We use simulation to demonstrate that estimators using repeated binary data to

estimate relative abundance, density, and species arrival are sensitive to false positive error, and to show that our model extensions permit more rigorous inference.

Methods

Chambert et al. (2015) assume that an investigator has recorded repeated detection-nondetection data $\boldsymbol{y}$ of some species (or other phenomena of interest) at i = 1, 2, …R locations over j = 1, 2…T discrete sampling intervals, where $y_{i,j}$ = 1 if species is observed and 0 otherwise. The purpose of the sampling is to estimate the binary occurrence state within a finite sample of sites (z_i) or the population level probability of occurrence ψ, assuming $z_i \sim$ Bernoulli (ψ). Within some number of sampling intervals at specific sites, observations ($v_{i,j}$) have been verified as either containing only true positives ($v_{i,j} = 1$), only false positives ($v_{i,j} = 2$), both ($v_{i,j} = 3$), or no observations ($v_{i,j} = 4$). These observations might include images, audio or video recordings, or physical samples (hair or scat) that can subsequently be confirmed via expert evaluation, laboratory analysis, or other means. Data from confirmed samples $v_{i,j} \sim$ Categorical ($\mathbf{\Omega}$), where the elements of $\mathbf{\Omega}$ are conditional on whether the species occurs at site i or not. If $z_i = 1$ (species occurs), then $\mathbf{\Omega} = [\{(s_1 \times (1 - s_0)\}\ \{s_0 \times (1 - s_1)\}\ \{s_1 \times s_0\}\ \{(1 - s_1) \times (1 - s_0)\}]$. If $z_i = 0$, the only possible outcomes are a false positive detection or no detection, and $\mathbf{\Omega} = [\{0\}\ \{s_0\}\ \{0\}\ \{1 - s_0\}]$. Here s_1 and s_0 represent the probabilities that a sampling interval contains > 0 true positive or false positive observations conditional upon the occurrence state. A similar conditional statement for unconfirmed data is $y_{i,j}|z_i \sim$ Bernoulli ($z_i \times p_{11} + (1 - z_i) \times p_{10}$). If present, a species can either be truly or falsely detected with probability $p_{11} = s_1 + s_0 - (s_1 \times s_0)$, and if not it can only be falsely detected with probability $p_{10} = s_0$.

Our initial description includes two alterations to the sampling protocols described by Chambert et al. (2015) that we have previously shown to be valid (Clare et al. *in review).*

Chambert et al. (2015) describe $\boldsymbol{v}$ as including all sampling intervals at a subset of sites k such that $\boldsymbol{v}$ and $\boldsymbol{y}$ are completely distinct and have different indexing. We imagine that it is more likely that samples will be verified across any number of sites or intervals such that y and v share consistent indexing: the only critical constraint is that specific samples cannot be included within the likelihoods for both $\boldsymbol{v}$ and $\boldsymbol{y}$. We also envision that most verification efforts will primarily focus on intervals with > 0 observations (either true or false positive). That is, investigators may never effectively evaluate sampling intervals with no detections (where $v_{i,j} = 4$), but Ω_4 must remain within the likelihood for $\boldsymbol{v}$ because confirmed data share parameters with non-confirmed data where nondetections are possible and may constitute the majority of outcomes.

Two further alterations make the sample verification model easier to generalize across other model structures. First, we define s_1 as a derived parameter reflecting a combination of state and observation processes that describe the *unconditional* probability of true detection (Royle and Dorazio 2008), which will consequently vary (at least) across locations i. For an occupancy model, $s_{1,i} = z_i \times p$, where p is the *conditional* probability of truly detecting a present species (MacKenzie et al. 2002). As before, $\boldsymbol{\Omega_i} = [\{(s_{1,i} \times (1 - s_0)\} \{s_0 \times (1 - s_{1,i})\} \{ s_{1,i} \times s_0\} \{(1 - s_{1,i}) \times (1 - s_0)\}]$ and $v_{i,j} \sim$ Categorical ($\boldsymbol{\Omega_i}$). The difference is that $s_{1,i}$ here is derived from a function of other parameters, and in the original treatment it is equivalent to the p of MacKenzie et al. (2002). Secondly, we do not derive conditional parameters p_{11} and p_{10}, but instead consider $y_{i,j} \sim$ Bernoulli $(1 - \Omega_{4,i})$: the species is detected truly, falsely, or both, or is not detected. The hierarchical likelihood is then

$$z_i \sim \text{Bernoulli}\ (\psi)$$

$$s_{1,i} = z_i \times p$$

$$\mathbf{\Omega}_i = [\{(s_{1,i} \times (1 - s_0)\}\ \{s_0 \times (1 - s_{1,i})\}\ \{\ s_{1,i} \times s_0\}\ \{(1 - s_{1,i}) \times (1 - s_0)\}]$$

$$v_{i,j} \sim \text{Categorical}\ (\mathbf{\Omega}_i)$$

$$y_{i,j} \sim \text{Bernoulli}\ (1 - \Omega_{4,i}\ )$$

We emphasize that this is primarily a reorganization of derived parameters within Chambert et al.'s (2015) model and the different parameterizations produce almost exactly the same results (Figure S1). The sole difference in the likelihood is that the original description treats $y_{i,j} = 1$ as constituting either a true or false positive detection, whereas here $y_{i,j} = 1$ may also reflect a mixture of true and false observations (Figure 1). Conditioning $\mathbf{\Omega}$ upon the underlying state is straightforward for models with two states, but burdensome for models with many possible states or if the range of states has unknown dimension (i.e., population size). The general structure of our reformulation instead conditions $s_{1,i}$ upon the underlying ecological state, and allows extension to several models dependent on binary data simply by redefining $s_{1,i}$ as equivalent to the applicable unconditional probability of (true) detection.

We use three models as examples. First, consider an occupancy model designed to estimate the timing of some ephemeral phenomena such as migration arrival following Roth et al. (2014). The model is exactly the same as presented above, except that organisms can only be truly detected during sampling intervals after they have arrived and occupied sites. Let arrival time at site i be denoted as x_i and assume that $x_i \sim$ Poisson (φ). To simplify presentation, we define x_i in terms of sampling intervals j rather than specific dates. A hierarchical description is:

$$z_i \sim \text{Bernoulli}(\psi)$$

$$x_i \sim \text{Poisson}\ (\varphi)$$

$$s_{1,i,j} = z_i \times p \times I(j > x_i)$$

$$\mathbf{\Omega_{i,j}} = [\{(s_{1,i,j} \times (1 - s_0)\}\ \{s_0 \times (1 - s_{1,i,j})\}\ \{\ s_{1,i,j} \times s_0\}\ \{(1 - s_{1,i,j}) \times (1 - s_0)\}]$$

The likelihoods for confirmed and unconfirmed observations are then $v_{i,j} \sim$ Categorical ($\mathbf{\Omega_{i,j}}$) and $y_{i,j} \sim$ Bernoulli $(1 - \Omega_{4,i,j})$.

In the model presented by Royle and Nichols (2003), the unconditional probability of detection $s_{1,i} = 1 - (1 - r)^{N_i}$, where r is the probability of detecting an individual during a sampling interval, while $N_i \sim$ Poisson (λ) and denotes the abundance of a species at site i. The false positive extension can be described hierarchically as:

$$N_i \sim \text{Poisson}\ (\lambda)$$

$$s_{1,i} = 1 - (1 - r)^{N_i}$$

$$\mathbf{\Omega_i} = [\{(s_{1,i} \times (1 - s_0)\}\ \{s_0 \times (1 - s_{1,i})\}\ \{\ s_{1,i} \times s_0\}\ \{(1 - s_{1,i}) \times (1 - s_0)\}]$$

The statements for $v_{i,j}$ and $y_{i,j}$ follow the occupancy description.

As a final example, the spatially explicit variant of Royle and Nichols' (2003) model (Ramsey et al. 2015) uses z_i to denote whether individuals i = 1,2…M exist within a geographic space $||S||$ with probability ψ. The state variable of interest, population size N in $||S||$is estimated as $\widehat{N} = \sum_{i=1}^{M} z_i$, and population density is derived as $\widehat{N}/Area_{||S||}$. Individuals have distinct activity centers located within $||S||$and the coordinates of these activity centers are denoted as s_i; individuals are detected at any of j detectors on given sampling occasions k with probability $p_{i,j.}$ The unconditional probability of detection is a function of whether an individual exists, the distance between an individual's latent activity center and the location of the detector, $d_{i,j}$, and the parameters g_0 and σ that respectively relate to the probability of individual detection at $d_{i,j} = 0$

and the rate at which individual encounter probability decays. Individuals are not distinguished, so these parameters are inferred by marginalizing across the latent individual encounter histories at a specific detector such that the unconditional probability of detection $s_{1,j} = 1 - \prod_{i=1}^{M}(1 - p_{i,j})$.

$$z_i \sim \text{Bernoulli}\ (\psi)$$

$$p_{i,j} = g_0(-d_{i,j}/2\sigma^2) \times z_i$$

$$s_{1,j} = 1 - \prod_{i=1}^{M}(1 - p_{i,j})$$

$$\mathbf{\Omega_j} = [\{(s_{1,j} \times (1 - s_0)\}\ \{s_0 \times (1 - s_{1,j})\}\ \{\ s_{1,j} \times s_0\}\ \{(1 - s_{1,j}) \times (1 - s_0)\}]$$

Here, $v_{j,k} \sim$Categorical ($\mathbf{\Omega_j}$) and $y_{j,k} \sim$ Bernoulli $(1 - \Omega_{4,j})$.

Simulation Study

We use simulation to demonstrate both that false-positive error influences estimates of arrival time, relative abundance, and density, and that extensions to account for false positive error provide better performance.

Royle-Nichols Model

We considered 8 different simulation scenarios of interest while holding simulated sampling parameters constant: 200 sites, with 20 temporal replications each. For each scenario (Table 1) we generated 300 replicate datasets with site-specific abundances $N_{i,sim} \sim$ Poisson $(\lambda_{i,\ sim})$ and log $(\lambda_{i,sim}) = \beta_0 + \beta_1 X_{1,i,sim}$, where $X_{1,i,sim} \sim$ N (0, 1), $\beta_0 = 0$, and $\beta_1 = 0$ or 1. True detection data $y_{i,j,sim}$ was generated as Bernoulli $(p_{i,sim})$, where $p_{i,sim} = 1 - (1 - r_{i,sim})^{N_{i,sim}}$, logit $(r_{i,sim}) = \alpha_0 + \alpha_1 X_{2,i,sim}$, $X_{2,i,sim} \sim$ N (0, 1), $\alpha_0 = -1.73$, and $\alpha_1 = 0$ or 1.

For each simulated encounter history, we generated false-positive detections as occurring at random across all cells within a simulation. The probability of a false-positive detection within a cell was derived such that the number of false-positive detections was 5 or 10% of the number of true detections (with Binomial variance). These rates of false positive error are common across a variety of sampling situations (Simons et al. 2007, McClintock et al. 2010, Norouzzadeh et al. 2018) and have also been used as threshold definitions for accurate data (McShea et al. 2016, Swanson et al. 2016). Ten percent of positive detections were sampled to create the verified data $v_{i,j,sim}$ (across simulation replicates and scenarios, $\bar{x}$ = 74.95 verified detections, s = 15.22). Each of the 2400 generated datasets was used to fit both a standard Royle-Nichols model and our false-positive extension. We evaluated the performance of both estimators by calculating the mean error associated with the posterior mean of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, the relative bias of finite sample population size point estimates ($\widehat{N}^*$, derived as the posterior mode of $\sum_{i=1}^{R} \widehat{N}_i$), and the frequentist coverage of 95% CRI.

Phenology Model

Our subsequent simulation investigations were less thorough. To evaluate the sensitivity of the Roth et al. (2014) model for arrival we considered a single scenario with 200 sites and 20 sampling occasions (each treated as analogous to a 10-day sampling period). Parameterization of the simulated data was as follows: logit ($\psi_{i,sim}$) = $\beta_0 + \beta_1 X_{1,i,sim}$, $X_{1,i,sim} \sim N(0, 1)$, $\beta_0 = 0$, and $\beta_1 = 0.5$; logit ($p_{i,sim}$) = $\alpha_0 + \alpha_1 X_{2,i,sim}$, $X_{2,i,sim} \sim N(0, 1)$, $\alpha_0 = -2$, and $\alpha_1 = 0.5$; average arrival time φ was simulated as day 60. True observations $y_{i,j,sim}$ were generated as Bernoulli ($z_{i,sim} \times p_{i,sim} \times I(a_{i,j,sim})$), where $z_{i,sim} \sim$ Bernoulli ($\psi_{i,sim}$), and $I(a_{i,j,sim})$ is an indicator function associated with whether survey j falls after the site and simulation specific arrival time $a_{i,j,sim}$ itself is generated as Poisson (φ_{sim}). We simulated 300 replicate dataset, with false positive detections and a 10%

verification sample $\boldsymbol{v}$ created as before (summarizing the size of the verification sample across replicates, $\bar{x}$ = 20.57, s = 4.57).

We evaluated 5 scenarios using the simulated data. We fit a standard arrival model treating φ_{sim} as constant (1), a standard arrival model treating $\varphi_{i,sim}$ as a site-specific random effect distributed as N ($\mu_{\varphi,sim}$, $\sigma_{\varphi,sim}$) (2). We fit the second model because we have observed that sometimes random effects are operationally assumed to account for observation error. For scenario 3 we fit an arrival model incorporating false positives and treating φ_{sim} as constant. Scenario 4 also incorporated false positives and treated φ_{sim} as constant; here, we altered the verification sample to constitute 20% of detections within the first 10 sampling intervals ($\bar{x}$ = 26.79 verified samples, s = 5.34). As a fifth scenario, we used the extended model, and increased the size of the verification sample to 20% across all time periods ($\bar{x}$ = 40.82 verified samples, s = 7.22). We evaluated models on the basis of the relative bias and frequentist coverage of $\hat{\varphi}$ and finite sample estimate of the proportion of occupied sites ($\widehat{PAO}$, derived for each simulation as $\sum_{i=1}^{R} \hat{z}_i$), and the mean error and coverage associated with estimates of coefficients for occurrence.

Spatial Royle-Nichols Model

Because the spatially-explicit extension of the Royle-Nichols model is computationally intensive and typically used only to estimate one state variable of ecological interest (population size $\hat{N}$), we simulated only 100 replicates within a single scenario to demonstrate proof of concept. Fixed sampling parameters included a population size of 50 organisms; $||S||$ defined as a 20 × 20 unit square; detection parameters g_0 = 0.15 and σ = 0.5; 196 detector locations within a 14 × 14 square grid with 1 unit spacing, and 20 sampling intervals: only the location of individual

activity centers varied across simulation replicates. False positive observations and a verification sample were generated as before, although the lower number of simulated detections also resulted in a much smaller verification sample ($\bar{x}$ = 12.56 verified samples, s = 3.30). We compared the standard model and the false positive extension on the basis of relative bias and frequentist coverage of $\widehat{N}$.

Evaluating transferability of s_0

A potentially appealing property of generalizing the model as here is that the verification outcomes that primarily contribute to the estimation of s_0 and its underlying generating process are likely to be consistent regardless of how the unconditional probability of true detection is formulated. This suggests that if data or computational resources are lacking, one might be able to use an informative prior for s_0 given previous estimates of the parameter from a distinct (and more quickly fit) model. To briefly explore transferability, we fit standard occupancy models to the simulated data previously used to fit Royle-Nichols models, compared the congruency of s_0 estimates across model types, and refit an extended Royle-Nichols model to the simulated data using an informed prior distribution for s_0 derived from the posterior distribution of the occupancy estimator.

We fit models using JAGS (Plummer 2003) or Nimble (the spatial model, deAlpine et al. 2017) to perform Markov-Chain Monte Carlo simulation through R v 3.4 (R Core Team 2017). Simulation settings are detailed further within Appendix SI2.

Results

Across all model types, estimators incorporating false positive error performed better than standard implementations. The Royle-Nichols model exhibited positive-bias and permissive

coverage of finite-sample population size (relative bias ≥ 0.20, coverage ≤ 0.22) and the abundance intercept parameter (mean error ≤ –0.16, coverage ≤ 0.49) even with relatively low (5%) rates of false positive error. Overall performance was worse when site-specific abundance varied in relation to simulated covariates although associations were estimated more accurately than baseline prevalence (Table 1). False positive extensions were nearly unbiased and had approximately nominal coverage across all parameters or state parameters considered.

Similar results held for the other models considered. Regardless of whether expected arrival time was estimated as fixed or randomly varying across sites, phenological occupancy models ignoring false positive error were biased estimators of the time of arrival (relative bias = −0.44) and finite-sample occupancy (relative bias = 0.16, Table 2) and exhibited permissive coverage (0.07 for estimates of arrival date, and 0.20 for estimates of finite-sample occurrence). False positive extensions were less biased and had more nominal coverage (Table 2). However, results suggested trade-offs between verifying samples randomly across the survey duration or placing more focus on verifying samples during earlier sampling times around the time of arrival. Focusing verification efforts on earlier sampling times reduced bias and provided more nominal coverage of arrival date than verifying samples across all time periods (relative bias and coverage respectively 0.01 vs 0.04 and 0.96 vs. 0.85), but resulted in poorer estimation of finite-sample occupancy (relative bias and coverage probability respectively 0.18 vs. –0.08 and 0.33 vs. 0.80). When the verified sample was random but larger (scenario 5), bias was negligible and coverage approximately nominal for all parameters considered.

Spatially explicit estimators of population size were severely biased (relative bias = 0.82) and exhibited poor coverage (0.48) when false positive rates were 10%. Extended models exhibited better performance (relative bias = 0.24, coverage probability = 0.89). One particular

simulation appeared to produce non-identifiable data, as the posterior mode fell along the boundary of our data augmentation prior even after refitting models with a more diffuse prior (Figure 2): excluding this outlier, the relative bias for the standard and modified estimator was 0.80 and 0.21, respectively. In a few other simulation replicates, the upper boundary of the augmentation prior may have truncated posterior density (and the 95% CRI): because of this, we likely overestimate coverage probability slightly (particularly for the standard estimator, which appeared more prone to this issue).

Estimates of s_0 derived from occupancy models were correlated with—but greater than—estimates of s_0 derived from the Royle-Nichols model (Figure S5). Despite this discrepancy, using an informative prior distribution for false positive error within the Royle-Nichols model derived from an occupancy model's estimates of s_0 resulted in model performance that barely differed from when verification data was directly evaluated within the likelihood (Table 1, Figure S5.).

Discussion

Because repeated detection-nondetection data are relatively easily to collect, such data is extensively applied for monitoring and modeling purposes. As ecologists increasingly focus upon addressing broad-scale questions that require collecting or collating massive amounts of data (e.g., Sorrano and Schimel 2014), ease of collection plays an important role in study design. Modeling advancements are often dependent upon data availability and amount, and the parameters that can be estimated using detection-nondetection data continues to grow. While model complexity has grown in conjunction with increases in data amount, whether the data collected and used to fit these models are clean enough to permit accurate estimation is a continued limitation. Indeed, "big data" efforts often depend upon fast but potentially error-prone

data collection or processing methods such as algorithms or crowdsourcing (e.g., Swanson et al. 2016, Tabak et al. 2018). As the scale of inquiry grows – either the sample itself or predictive space – so too does the amount of potential absolute bias and the implications of biased estimation. Previous efforts have repeatedly shown that occupancy estimators are biased under realistic rates of false positive error (e.g., Miller et al. 2011, Chambert et al. 2015). Results here demonstrate that this bias is general to many estimators reliant upon repeated binary detection data. Bias associated with population size or the timing of arrival may be more problematic because these metrics both cover a wider range of potential values (bias can be more pronounced), and because they are more widely used to justify management decisions (e.g., timing of actions, quotas, recovery metrics) than occurrence or distribution.

Models are potentially sensitive to numerous violations of assumptions, but relative to more nuanced assumptions such as the form of a given parametric function, the assumption that an error type like false positives does not exist is particularly easy to evaluate, and as shown here, not prohibitively difficult to correct for. One cost of incorporating validated data to explicitly model false positive detections is to induce additional uncertainty associated with extra parameters and to require additional effort associated with verification. Results indicate that even very small number of verification samples (n = 15 – 20) can substantively improve inference even when false positives are scarce. The sample size of many of the simulated verification samples presented here is probably smaller than most investigators would prefer, particularly when fitting models with a great deal of intrinsic uncertainty given multiple latent variables, like the spatial Royle-Nichols model or a phenological occupancy model. Regardless, the size of the verified sample need not be prohibitively large to permit unbiased and reasonably precise estimation. Results here and elsewhere (Clare et al. *in* review) suggest that when s_0 is constant,

the sample verification model is generally unbiased when ~50 samples have been verified. Increasing the size of a verification sample beyond this may further result in reduced bias as investigators gain more power to approximate any underlying variability in s_0, but otherwise primarily increases estimator precision (Miller et al. 2015, Chambert et al. 2018).

A second cost of using false positive-extensions is that they require slightly more computational overhead. There are several ways to limit this cost. If there is no modeled temporal variation in true or false positive detections, verified sampling occasions could be aggregated across sites and treated as $\boldsymbol{v_i} \sim$ Multinomial ($\boldsymbol{\Omega_i}, \boldsymbol{k_i}$). Our results suggest that s_0 may be transferable enough across different model structures that investigators operating under stringent computational constraints could estimate s_0 using a simpler model and subsequently use an informed prior for more intensive analyses. The primary benefit of including verified data within the model likelihood rather than using an informed prior is that the verified data provide direct information about particular latent variables (e.g., z_i or N_i), but these are rarely of direct interest, and using an informed prior shrinks the dimensionality of the model matrix and may provide substantive increases in speed if the verification sample is large. Finally, the concepts here need not be implemented using MCMC simulation within a Bayesian framework. We describe the extensions hierarchically using a complete data likelihood because all models considered have been described in this fashion, but if a model can be fit using faster means such by maximizing a marginalized likelihood, so too can the extensions presented here.

Further extensions are possible and deserve more investigation. The degree to which false positive errors degrade dynamic or integrated extensions of the static models considered here (i.e., provide biased estimates of trends) is a subject of ongoing research, but previous work demonstrating that false positives induce biased estimates of occupancy dynamics (McClintock

et al. 2010b) and that integrated models are sensitive to misspecifications associated with any constituent data type (Zipkin et al. 2017) suggests that such extensions are likely to be similarly sensitive. False positive error may not always happen at random as in our simulations, and a natural way to deal with heterogeneity between sites or sampling intervals is via the use of random effects or covariates—e.g., logit ($s_{0,i}$) = $\boldsymbol{\beta X}$, where the vector or matrix $\boldsymbol{X}$ captures covariates associated with, for example, the occurrence of a similar looking species, or a metric associated with classification confidence. Additionally, sampling intervals may contain several distinct observations that are classified separately (e.g., recordings, images) but aggregated across a sampling interval for an analysis. In some cases, it may be easier to verify discrete observations rather than verify complete sets of observations within defined sampling intervals, and a larger number of observations within a sample suggests a greater probability that at least one observation is true (Chambert et al. 2015, 2018). One way to deal with this is to model a positive outcome within a sampling interval as arising from either > 0 true observations or all misclassified observations, where the probability that all observations are misclassified within a sampling interval $s_{0,i,j} = (1 - r_0)^{n_{i,j}}$, $n_{i,j}$ is the number of recorded observations within interval j at site i, and r_0 is the probability that a single observation is false positive (Chambert et al. 2015, Appendix SI3).

The sampling designs associated with verification effort may also deserve more attention. Here, a verification sample targeting earlier sampling intervals where simulated false positives were more prevalent than true positives permitted unbiased estimation of arrival time, but not overall occurrence, while a random verification sample across all sampling intervals provided better performance. For models in which observation is conditional upon multiple independent latent variables (e.g., arrival time and occurrence), different verification schemes may provide

more information about one parameter than another. Detections verified as true during earlier sampling periods provided more information about arrival time than detections verified at random, but focusing a verification effort upon earlier sampling periods appears to have biased the verification sample towards false positive detections prior to species arrival relative to true detections. In turn, this appears to have negatively biased estimates of the true probability of detection and led to a similar positive bias in the number of occupied sites as exhibited by models ignoring false positives. More generally, targeting the verification sample towards suspected false positives may require explicitly accounting for sampling bias within the verification effort.

Recent studies focusing on species distribution or demographic parameters generally acknowledge the existence of imperfect detection and use models that explicitly account for it. Explicitly accounting for imperfect detection while assuming no false positive error makes many of these models extremely sensitive to misclassification. As demonstrated here, accounting for false positives is both surmountable and important for making rigorous ecological inference across a broader class of models than previously recognized.

Acknowledgments

Support for this research was provided by NASA ESSF NNX16AO61H to JC and NASA Ecological Forecasting grant NNX14AC36G to PT and BZ.

References

Chambert, T., D.A.W. Miller, and J. D. Nichols. 2015. Modeling false positive detections in species occurrence data under different study designs. Ecology 96:332-339.

Chambert, T., J. H. Waddle, D.A.W. Miller, S. C. Walls, and J. D. Nichols. 2018. A new framework for analyzing automated acoustic species-detection data: occupancy estimation and optimization of recordings post-processing. Method ins Ecology and Evolution 9:560-570.

de Valpine, P., D. Turek, C. Paciorek, C. Anderson-Bergman, D. T. Lang, and R. Bodik. 2017. Programming with models: writing statistical algorithms for general model structures with NIMBLE. Journal of Computational and Graphical Statistics 26:403-413.

MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83:2248-2255.

McClintock, B. T., L. L. Bailey, K. H. Pollock, and T. R. Simons. 2010a. Experimental investigation of observation error in anuran call surveys. Journal of Wildlife Management 74:1882-1893.

McClintock B.T., L.L. Bailey, K. H. Pollock, and T. R. Simons. 2010b. Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections. Ecology 91: 2446–2454.

McShea, W. J., T. Forrester, R. Costello, Z. He, and R. Kays. 2016. Volunteer-run cameras as distributed sensors for macrosystem mammal research. Landscape Ecology 31:55-66.

Miller, D. A., J. D. Nichols, B. T. McClintock, E. H. Campbell Grant, L. L. Bailey, and L. A. Weir. 2011. Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. Ecology 92: 1422-1428.

Miller, D.A.W., L. L. Bailey, E. H. C. Grant, B. T. McClintock, L. A. Weir, and T. R. Simons. 2015. Performance of species occurrence estimators when basic assumptions are not met:

a test using field data where true occupancy status is known. Methods in Ecology and Evolution 6:557-565.

Norouzzadeh, M. S., A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, and J. Clune. 2018. Automatically identifying, counting, and describing wildlife animals in camera-trap images with deep learning. Proceedings of the National Academy of Sciences: 201719367.

Otis, D. L., K. P. Burnham, G. C. White, and D. R. Anderson. 1978. Statistical inference from capture data on closed animal populations. Wildlife Monographs 62:3-135.

Plummer, M. (2003). JAGS: a program for analysis of Bayesian graphical models using GIBBS sampling. Proceedings of the 3rd international workshop on distributed statistical computing.

Swanson, A., M. Kosmala, C. Lintott, and C. Packer. 2016. A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. Conservation Biology 30:520-531.

R Core Team. 2017. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Ramsey, D. S. L., P. A. Caley, and A. Robley. 2015. Estimating population density from presence-absence data using a spatially explicit model. Journal of Wildlife Management 79:491-499.

Roth, T., N. Strebel, and V. Amrhein. 2014. Estimating unbiased phenological trends by adapting site-occupancy models. Ecology 95:2144-2154.

Royle, J.A., and R. D. Dorazio. 2008. Hierarchical modeling and inference in ecology. Academic Press, London.

Royle, J. A., and J. D. Nichols. 2003. Estimating abundance from repeated presence-absence data or point counts. Ecology 84:777-790.

Simons, T. R., M.W. Alldredge, K. H. Pollock, and J. M. Wettroth. 2007. Experimental analysis of the auditory detection process on avian point counts. Auk 124:986-999.

Sorrano, P. A., and D. S. Schimel. 2014. Macrosystems ecology: big data, big ecology. Frontiers in Ecology and the Environment 12:3-3.

Swanson, A., M. Kosmala, C. Lintott, and C. Packer. 2016. A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. Conservation Biology 30:520-531.

Zipkin, E. F., S. Rossman, C. B. Yackulic, J. D. Wiens, J. T. Thorson, R. J. Davis, and E. H. Campbell Grant. Integrating count and detection-nondetection data to model population dynamics. Ecology 98:1640-1650.

Table 1. Simulation settings and model performance of the standard Royle-Nichols estimator and the false positive extension described herein either fit with verified data in the likelihood or using an informed prior. Dashes indicate value was not estimated.

		True Values					Mean Error					Relative Bias				Coverage
Estimator	Scenario	β_0	β_1	α_0	α_1	FP	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{N}^*$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{N}^*$
Standard	1	0	1	–1.73	1	0.1	0.71	–0.13	–0.57	–0.19	0.80	0	0.46	0.01	0.24	< 0.01
	2	0	0	–1.73	1	0.1	0.45	-	–0.38	–0.18	0.57	0.05	-	0.06	0.41	< 0.01
	3	0	1	–1.73	0	0.1	0.58	–0.11	–0.46	-	0.59	0	0.54	0.01	-	< 0.01
	4	0	0	–1.73	0	0.1	0.33	-	–0.27	-	0.37	0.11	-	0.14	-	0.01
	5	0	1	–1.73	1	0.05	0.43	–0.08	–0.36	–0.12	0.43	0.06	0.74	0.15	0.54	0.09
	6	0	0	–1.73	1	0.05	0.25	-	–0.21	–0.10	0.28	0.38	-	0.44	0.69	0.15
	7	0	1	–1.73	0	0.05	0.32	–0.06	–0.26	-	0.30	0.18	0.83	0.26	-	0.15
	8	0	0	–1.73	0	0.05	0.19	-	–0.16	-	0.20	0.49	-	0.54	-	0.22
False Positive Extension	1	0	1	–1.73	1	0.1	0.00	0.01	–0.01	0.01	< 0.01	0.95	0.92	0.94	0.95	0.94
	2	0	0	–1.73	1	0.1	–0.02	-	–0.11	0.01	–0.01	0.95	-	0.87	0.92	0.96
	3	0	1	–1.73	0	0.1	–0.02	0.02	–0.10	-	–0.02	0.96	0.95	0.91	-	0.95
	4	0	0	–1.73	0	0.1	–0.03	-	–0.08	-	–0.04	0.97	-	0.91	-	0.96
	5	0	1	–1.73	1	0.05	–0.02	< 0.01	<0.01	0.01	–0.02	0.97	0.94	0.96	0.94	0.97
	6	0	0	–1.73	1	0.05	–0.01	-	<0.01	< 0.01	–0.01	0.97	-	0.97	0.95	0.96
	7	0	1	–1.73	0	0.05	–0.01	0.01	<0.01	-	–0.01	0.94	0.96	0.93	-	0.95
	8	0	0	–1.73	0	0.05	0.00	-	<0.01	-	–0.01	0.95	-	0.91	-	0.93
	1[A]	0	1	–1.73	1	0.1	–0.06	0.03	<0.01	0.02	–0.03	0.95	0.92	0.95	0.94	0.95
	2[A]	0	0	–1.73	1	0.1	–0.04	-	–0.01	0.01	–0.03	0.96	-	0.98	0.92	0.97
	3[A]	0	1	–1.73	0	0.1	–0.06	0.02	0.02	-	–0.05	0.93	0.96	0.99	-	0.92
	4[A]	0	0	–1.73	0	0.1	–0.07	-	0.04	-	–0.07	0.90	-	0.98	-	0.87
	5[A]	0	1	–1.73	1	0.05	–0.04	0.01	0.01	0.01	–0.03	0.95	0.95	0.98	0.93	0.97
	6[A]	0	0	–1.73	1	0.05	–0.03	-	0.01	< 0.01	–0.03	0.97	-	0.99	0.95	0.95
	7[A]	0	1	–1.73	0	0.05	–0.07	0.02	0.04	-	–0.05	0.91	0.95	0.99	-	0.94
	8[A]	0	0	–1.73	0	0.05	–0.06	-	0.04	-	–0.06	0.94	-	0.98	-	0.92

[A]Model does not include verified data within the likelihood, but instead uses an informed prior for false positive error derived from posterior estimates of an occupancy model.

Table 2. Error, relative bias, and coverage of parameters associated with detection (α), occurrence (β), arrival time (φ), and the finite-sample proportion of occupied sites (PAO) for the 4 candidate phenological occupancy models described within the main text. The standard estimators (Models 1 and 2) overestimate overall occurrence, and underestimate arrival time and true detection probability. The modified estimators exhibit limited bias with respect to the timing of arrival, but also sub-optimally estimate overall occurrence. In particular, focusing the verification sample towards earlier time periods (Model 4) reduces the accuracy of occupancy estimation.

	Mean Error						Relative Bias						Coverage					
Model	β_0	β_1	α_0	α_1	φ	PAO	β_0	β_1	α_0	α_1	φ	PAO	β_0	β_1	α_0	α_1	φ	PAO
1	0.35	0.03	–0.23	–0.05	–0.44	0.16	0.6	0.93	0.35	0.88	0.07	0.21						
2	0.34	0.04	–0.23	–0.05	–0.44	0.16	0.61	0.93	0.36	0.87	0.07	0.21						
3	–0.17	> –0.01	0.17	–0.02	–0.04	-0.08	0.88	0.95	1	0.92	0.83	0.80						
4	0.4	0.03	–0.04	–0.04	–0.01	0.17	0.58	0.96	0.93	0.91	0.96	0.39						
5	< 0.01	0.05	0.02	< 0.01	0.01	< 0.01	0.94	0.96	0.96	0.97	0.92	0.97						

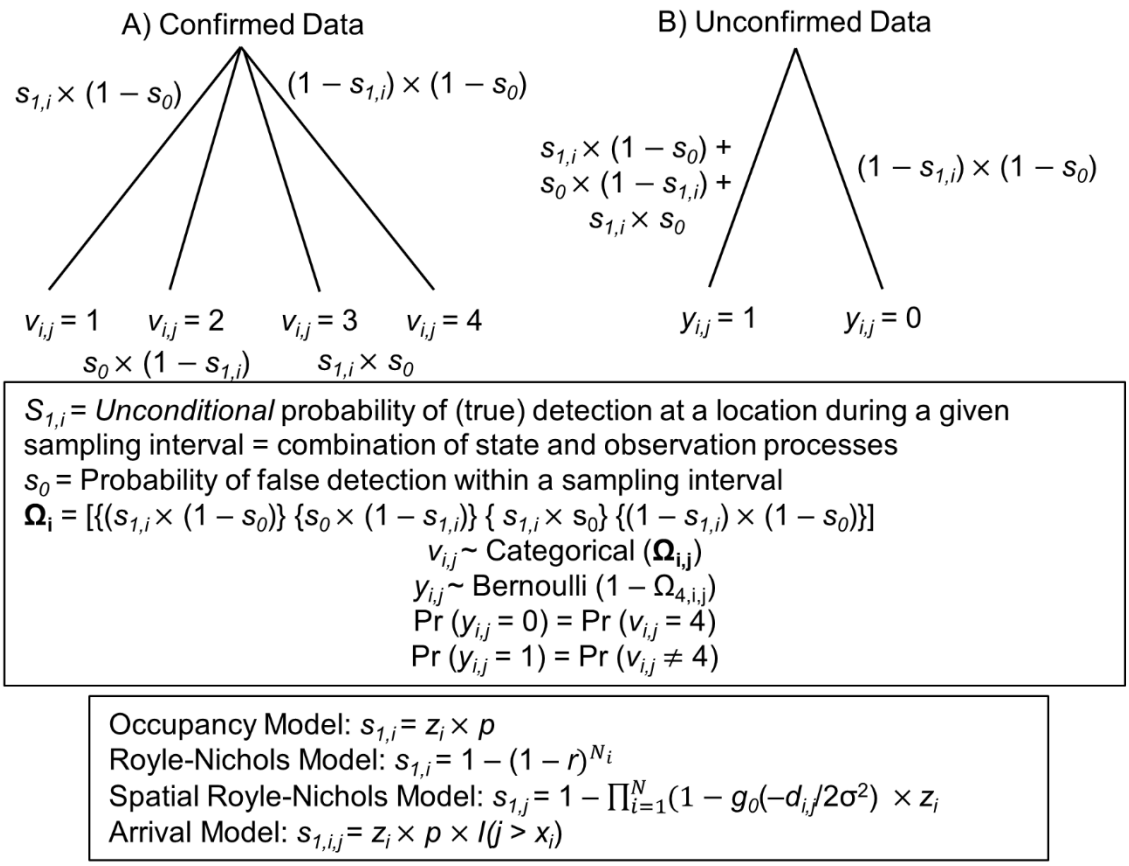

Figure 1. Schematic for the generalizable sample-verification model for false positive error presented here. Unconfirmed outcomes are positive if the species has either been truly or falsely detected, or negative if neither true nor false detection has occurred. Confirmed data can either result in a true detection only, a false detection only, both, or no detection. The probability of > 0 false detections within a verified sampling interval is denoted s_0, while the probability of > 0 true detections, $s_{1,i}$, is treated as a combination of functions associated with state and conditional observation models.

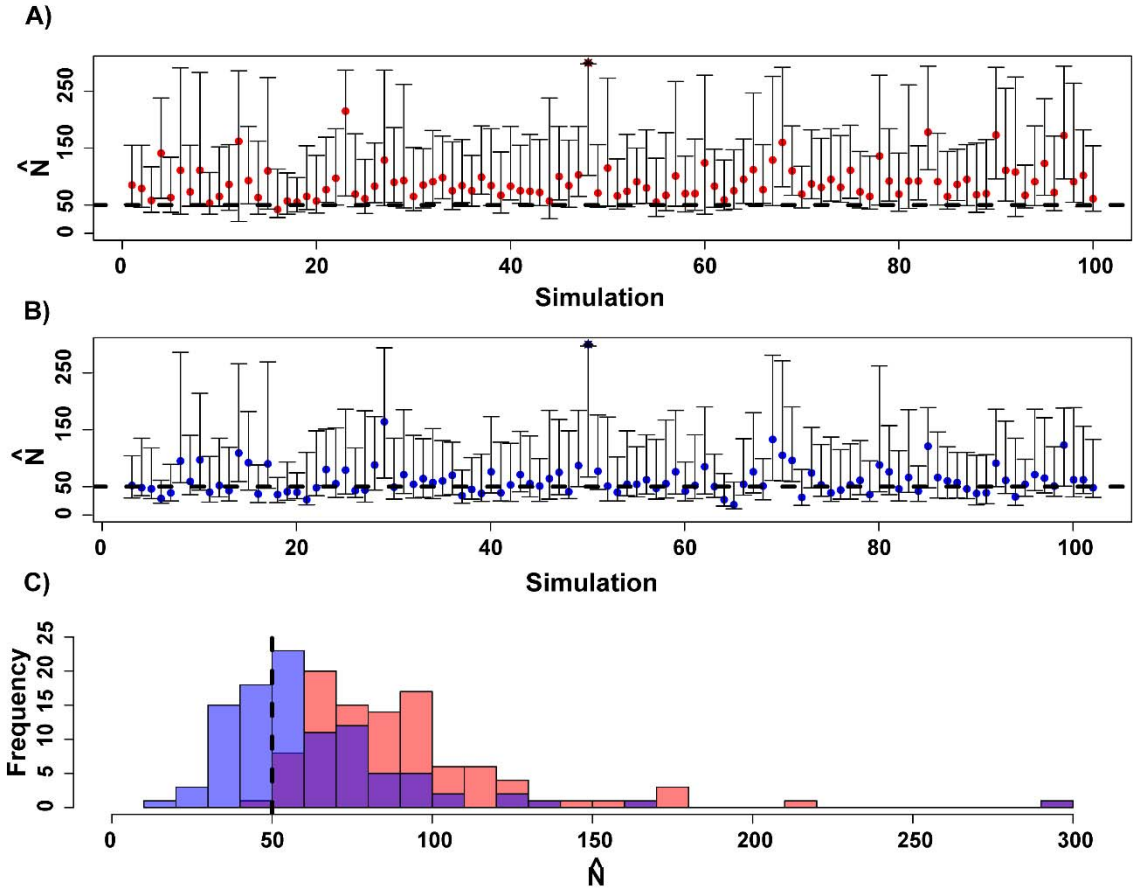

Figure 2. Simulation-specific point estimates and 95% CRI from the standard spatial Royle-Nichols model (A), an extension to account for false-positives (B), and the overall distribution of point estimates (C). Dashed lines denote true values. In A, B, and C, one point estimate falls right at the boundary, potentially indicating a non-identifiable dataset; other simulated replicates exhibit nonnegligible posterior density along the boundary such that the upper CRI abuts it. This is more common for the standard estimator, and one additional consequence of ignoring false positive error may be added computational cost associated with using a larger data augmentation prior for *N*.

Supporting Information S1.

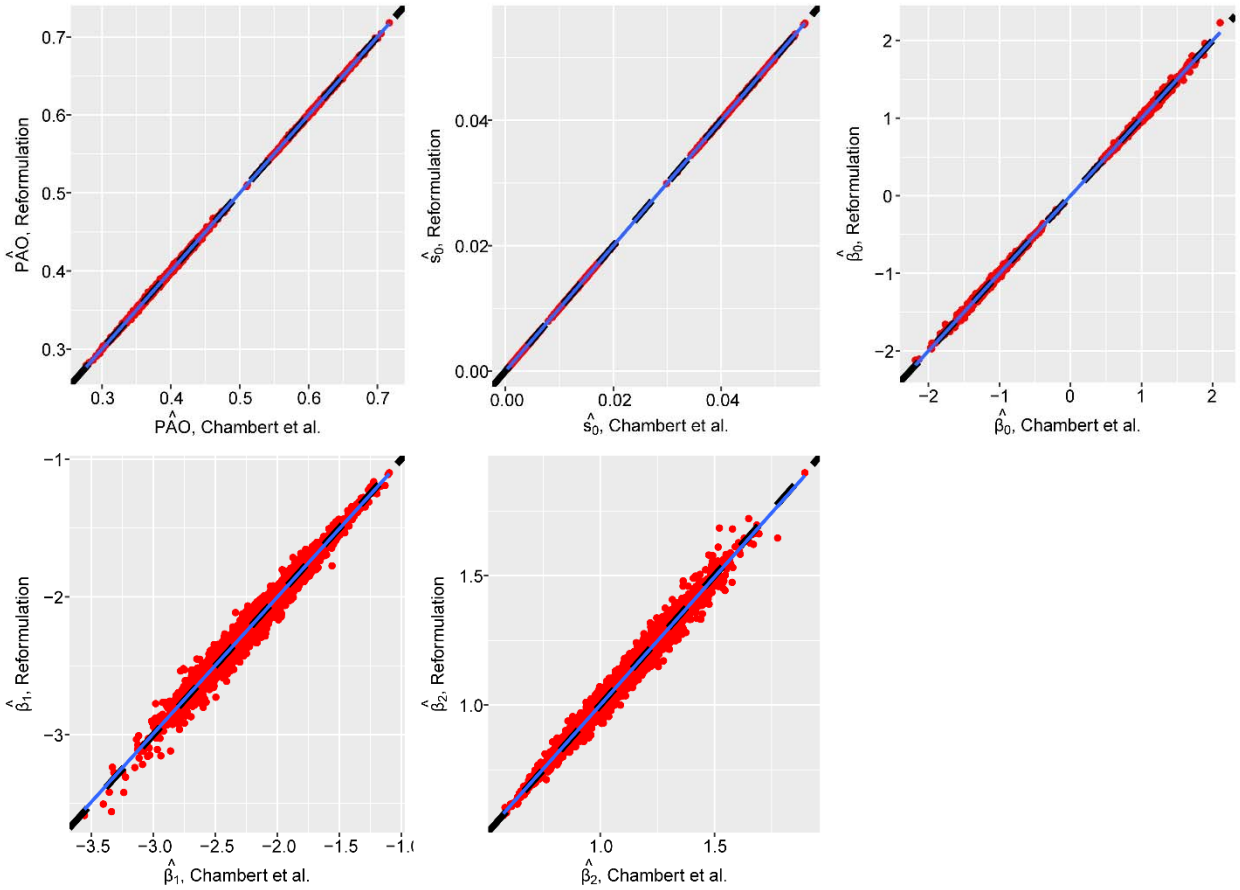

Figure S1**.** Point estimates (based upon the mode [top left] or mean of the posterior distribution) from Chambert et al.'s 2015 description of the sample verification model and our reformulation of the model are essentially equivalent: red dots indicate point estimates from specific simulate replicates, black dashed line is 1:1, blue line is fitted linear association. These estimates are derived from 18 simulation scenarios described in Clare et al. *in review*.

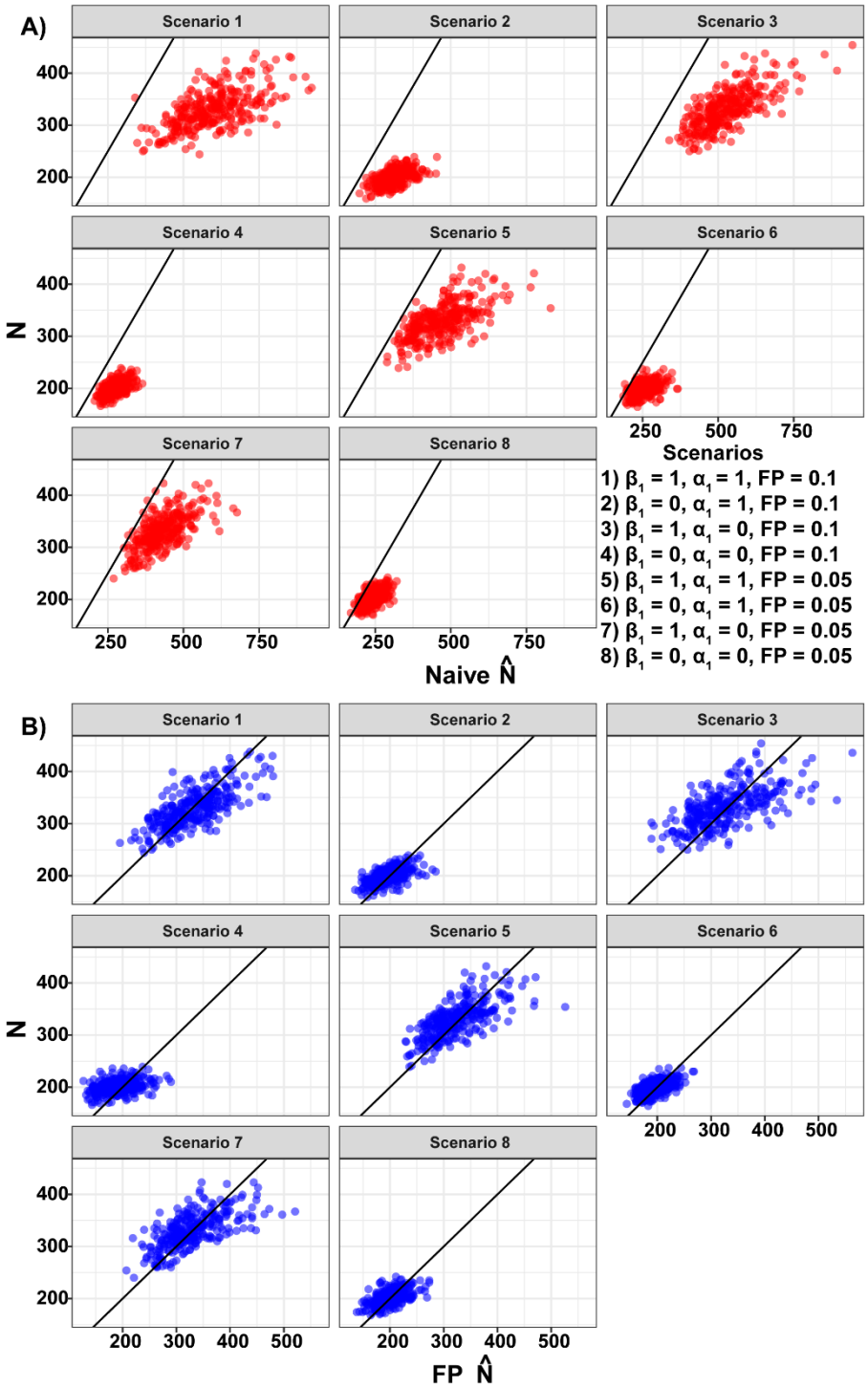

Figure S2. Abundance estimation using the Royle-Nichols model is biased under low to moderate rates of false positive error, particularly when abundance is modeled as spatially varying (A); in all scenarios, the false positive extension presented herein is nearly unbiased (B).

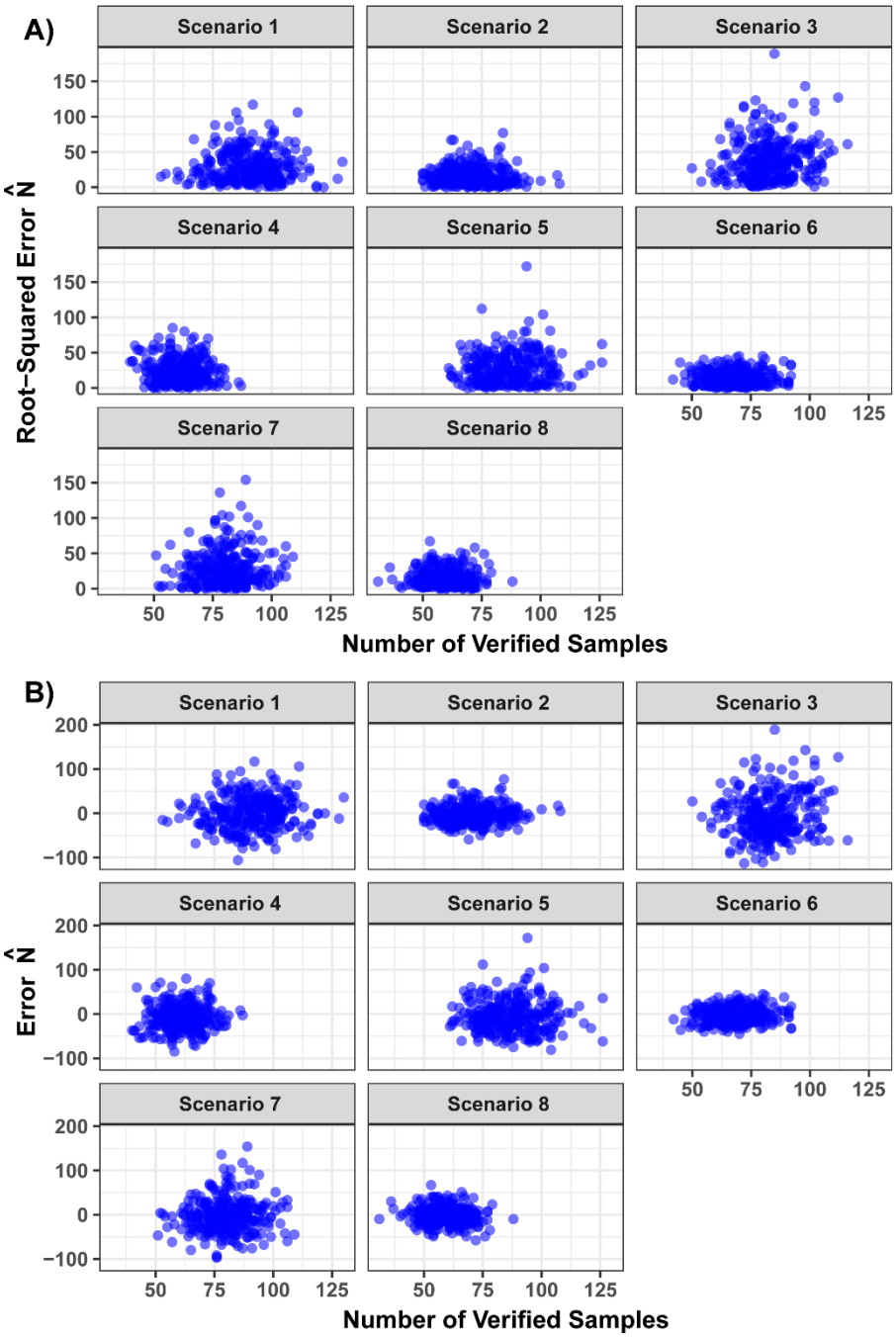

Figure S3. Root squared error (A) and bias (B) associated with the extended Royle-Nichols estimator exhibit little association with the size of the verification sample within the range considered: when false positive error is constant, gains in estimator precision may require much larger increases in sample size than considered here.

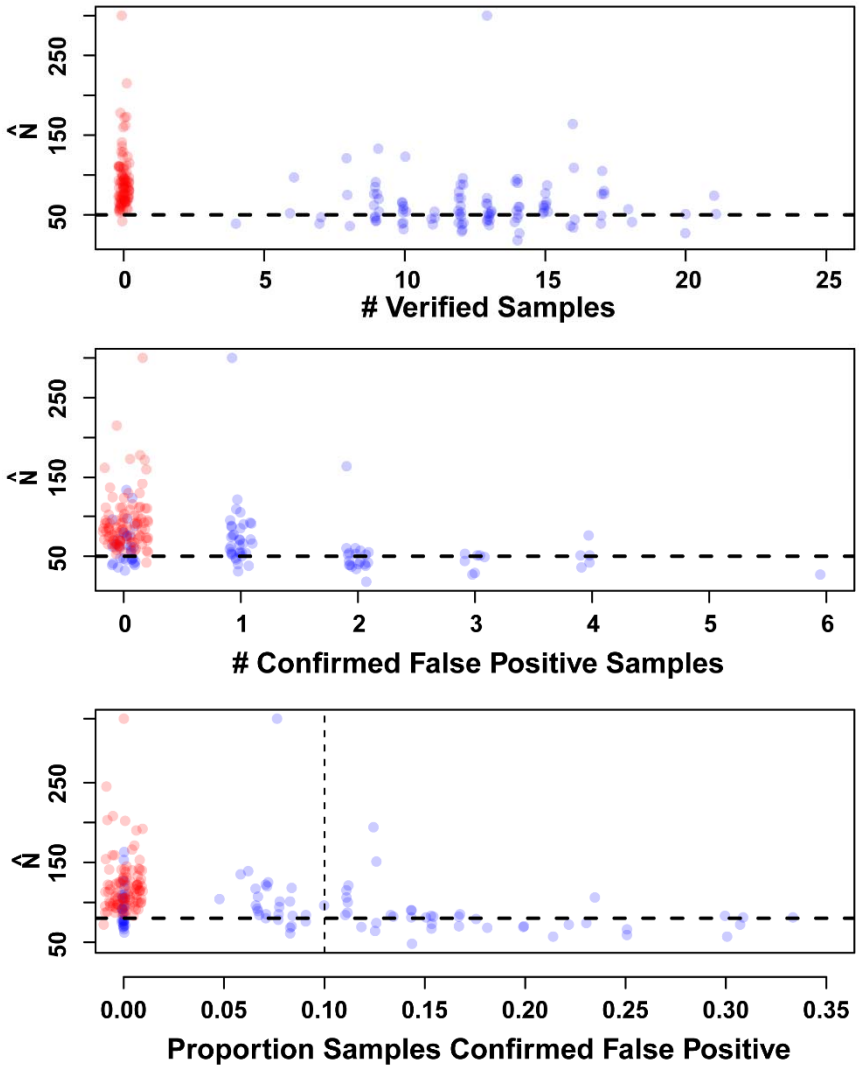

Figure S4. Point estimates from the extended spatial Royle-Nichols model (blue) exhibit sensitivity to the size of the verification sample and the empirical proportion of false positives within it appear more pronounced than for the non-spatial Royle-Nichols, likely because the simulated verification sample was very small. Regardless, the extended model exhibits better performance than the standard model (red, jittered).

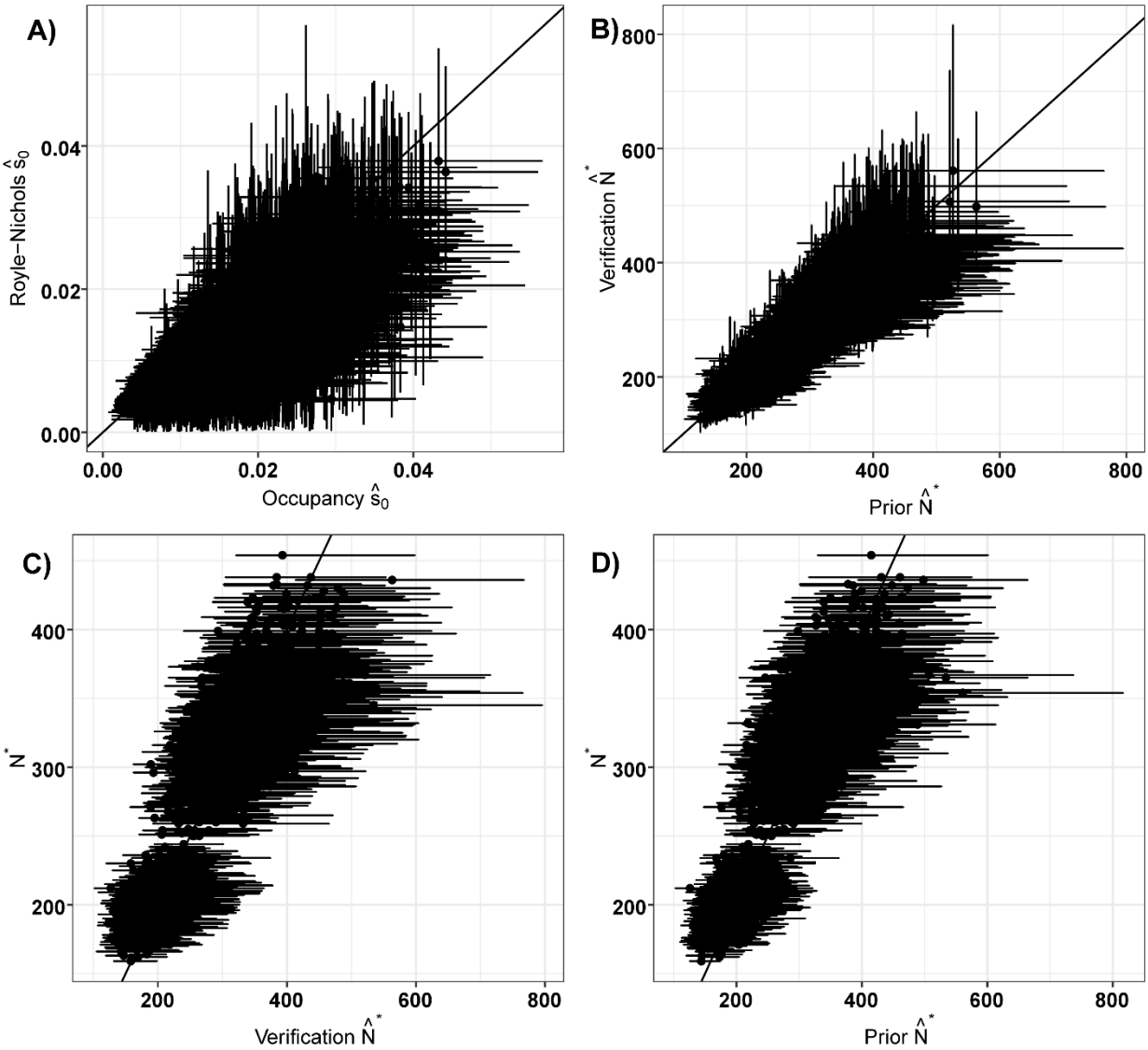

Figure S5. Associations between false positive probabilities estimated using the Royle-Nichols and occupancy estimator suggests that the occupancy estimator tends to produce slightly larger estimates of s_0 (A). However, estimates of population size using an informed prior derived from occupancy-based estimates are practically equivalent to estimates of population size when confirmed data is included within the likelihood (B), and both appear nearly unbiased (C, D).