Appendix S1

Matthew T. Farr, David S. Green, Kay E. Holekamp, and Elise F. Zipkin

Integrating distance sampling and presence-only data to estimate species abundance

Ecology

Model structure and assumptions

This appendix describes details on the model structure (Figure S1) and includes additional simulations to evaluate model assumptions. See our GitHub repository for the full code.

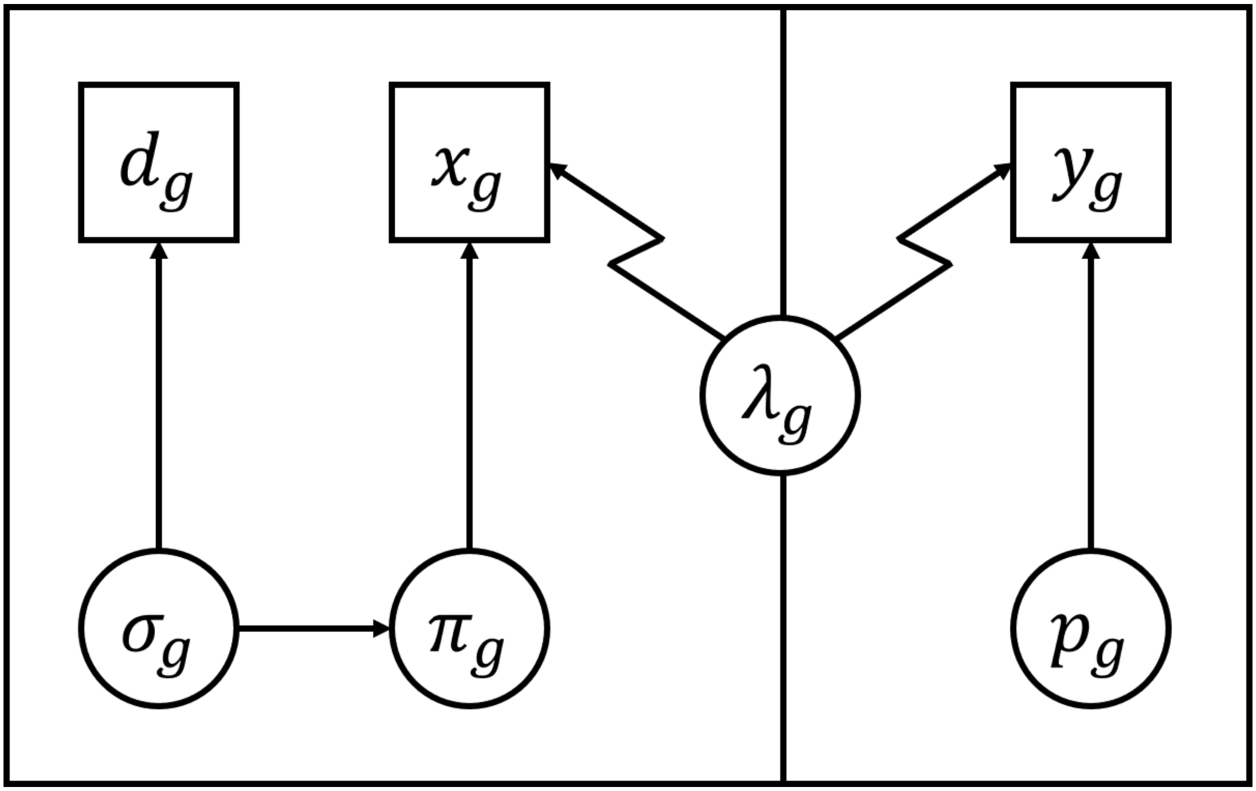

Figure S1. A directed acyclic graph for the integrated model showing links (arrows) between parameters (circles) and data sources (boxes). Both distance sampling and presence-only data are integrated b the biological process or intensity function λ_g. The left box indicates the distance sampling observation process with count data x_g, observed distances d_g, and estimated parameters σ_g (scale) and π_g (detection

probability) at pixel g. The presence-only observation process is within the right box with observed count of presences y_g and thinning rate or observation error p_g at pixel g. The jagged arrows indicates a potential change-of-support between scale for either distance sampling or presence-only data.

Section S1. Poisson thinning process

We assumed that presence-only and distance sampling data were a "thinned version" of true abundance, where each detection (i.e., data point) was observed with probability p (i.e., observation error from sampling bias and/or imperfect detection). We used a binomial-Poisson mixture for each observation process, which reduced to a Poisson thinning process. Letting Y = count data and N = true abundance, we then had: $Y \sim \mathit{binomial}(N, p)$ where $N \sim \mathit{Poisson}(\lambda)$. This hierarchical model reduced to: $Y \sim \mathit{Poisson}(p\lambda)$ through the following process (Casella & Berger 2002, pg. 163):

$$P(Y = y) = \sum_{n=0}^{\infty} P(Y = y, N = n) = \sum_{n=y}^{\infty} P(Y = y|N = n)P(N = n) \tag{1}$$

$$= \sum_{n=y}^{\infty} \binom{n}{y} p^y (1-p)^{n-y} \frac{\lambda^n e^{-\lambda}}{n!} \tag{2}$$

$$= e^{-\lambda} \sum_{n=y}^{\infty} \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y} \frac{\lambda^n}{n!} \tag{3}$$

$$= \frac{e^{-\lambda} p^y \lambda^y}{y!} \sum_{n=y}^{\infty} \frac{(1-p)^{n-y} \lambda^{n-y}}{(n-y)!} \tag{4}$$

$$= \frac{e^{-\lambda} p^y \lambda^y}{y!} \sum_{x=0}^{\infty} \frac{((1-p)\lambda)^x}{x!} \tag{5}$$

$$= \frac{e^{-\lambda} p^y \lambda^y}{y!} e^{(1-p)\lambda} \tag{6}$$

$$= \frac{e^{-p\lambda} (p\lambda)^y}{y!} \tag{7}$$

A further explanation of each equation is as follows:

1) Probability of count data given probability of true abundance.
2) Probability mass function of binomial-Poisson mixture.
3) Moved $e^{-\lambda}$ out of the series. Expanded binomial coefficient, $\binom{n}{y}$, to $\frac{n!}{y!(n-y)!}$.
4) Canceled out $\frac{n!}{n!}$. Split λ^n to $\lambda^y \lambda^{n-y}$. Moved $\frac{p^y \lambda^y}{y!}$ out of the series.
5) Switched $n - y$ to x. Combined $(1-p)^x \lambda^x$ to get $((1-p)\lambda)^x$.
6) Euler's number is the sum of infinite series, $e = \sum_{x=0}^{\infty} \frac{1}{x!}$, and $e^z = \sum_{x=0}^{\infty} \frac{z^x}{x!}$. Thus, $\sum_{x=0}^{\infty} \frac{((1-p)\lambda)^x}{x!} = e^{(1-p)\lambda}$.
7) Combined $e^{-\lambda}$ and $e^{(1-p)\lambda}$ to get $e^{-p\lambda}$. Combined $p^y \lambda^y$ to get $(p\lambda)^y$. This has now become equivalent to the probability mass function of a Poisson ($Y \sim \mathit{Poisson}(p\lambda)$) where expected value was $p\lambda$.

Section S2. Assumption of independence

For our simulations and case study, we described a model structure that assumed independence between presence-only and distance sampling data. In practice, this means each data type is linked to its own latent abundance that came from separate, independent draws (i.e., random variates) of a Poisson random variable with mean λ_g. We ran a set of simulations to assess the validity and flexibility of this assumption as there may be situations when this assumption is violated.

We compared the model assuming independence to a structure with dependencies between data types built into the model. The first version is the model described in the paper: $y_g \sim Poisson(\lambda_g p_g); x_g \sim Poisson(\lambda_g \pi_g)$. For each pixel g, y_g is presence-only data, λ_g is the intensity, p_g is the sampling bias for presence-only data, x_g is distance sampling data, π_g is the detection probability for distance sampling. This structure assumes independence between y_g and x_g. The second model assumes dependence between y_g and x_g by conditioning explicitly on latent abundance, N_g. The model structure is: $y_g \sim Binomial(N_g, p_g); x_g \sim Binomail(N_g, \pi_g); N_g \sim Poisson(\lambda_g)$. We ran 1000 simulations with varying amounts of distance sampling data (i.e., 0 - 20% of our study area). Figure S2 below shows the results of our simulations.

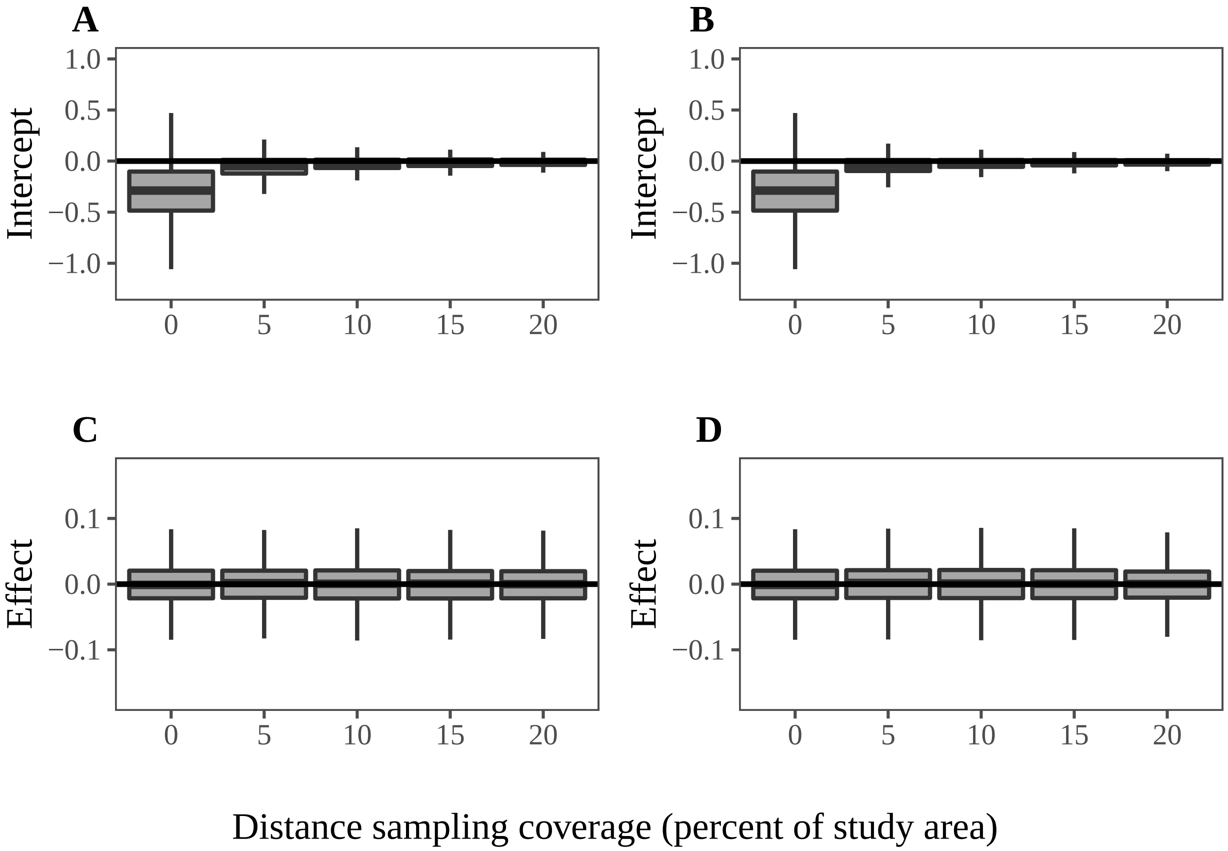

Figure S2. Intercept and effect parameter biases (estimated − truth) via box plots. The independent model structure, is on the left (panels A, C), which was reproduced from Figure 1C, E in the main text, and the dependent model structure is on the right (panels B, D). The panels show results for high amounts of presence-only data.

The models returned nearly identical results with respect to parameter estimates and uncertainty around these estimates (Figure S2). The frequency of false positive significance was also similar between structures where the 95% credible intervals captured the true value across simulations near 95% (Table S1). These results were also consistent for low quantities of presence-only data as well (results not shown). Based on these simulation results, we concluded that both models are sufficient for parameter estimation. We chose to present the independent version of the model in the main text as it is a more concise model structure and easier to program in R and JAGS. Though it may seem straight forward in this context, applications of the dependent model structures are generally more challenging as additional model components (e.g., change-of-support) or case-specific issues (e.g., data discrepancies between presence-only and distance sampling data) can make initializing latent abundance, N, difficult. However, we caution readers to test assumptions for each case-specific application.

Table S1. Percent of simulations where the true value was within the estimated 95% credible interval for both independent and dependent model structures across multiple distance sampling quantities and high quantities of presence-only data.

Structure	Distance sampling quantity	True Positive %
Independent	5%	93.9%
Dependent	5%	94.2%
Independent	10%	93.4%
Dependent	10%	93.9%
Independent	15%	94.1%
Dependent	15%	94.6%
Independent	20%	94%
Dependent	20%	94.4%

Section S3. Thinned Poisson assumption

Within our specified structure, p_g is the estimate of observation bias for presence-only data, which includes both information on detection probability and sampling bias (i.e., variation in sampling intensity). As such, p_g is the product of these two processes and cannot be parsed apart to decompose detection and sampling probabilities.

We thus developed a zero-inflated model (ZIF) to assess whether it is possible to parse apart each of these probabilities. The ZIF model was specified as follows: $y_g \sim Poisson(p_g z_g \lambda_g)$; $\lambda_g = 0$ accounts for true zeros (i.e., no individuals present); p_g is an estimate of detection probability and accounted for zeros due to imperfect detection; z_g is a latent parameter describing whether pixel g is either sampled ($z_g = 1$) or not sampled ($z_g = 0$): $z_g \sim Bernoulli(\psi_g)$ and ψ_g is the probability of pixel g being sampled and accounts for zeroes due to unsampled pixels.

We compared the non-ZIF model to the ZIF version assuming the data generating process of the ZIF model (as specified above). For both versions, we modeled λ_g with an intercept (λ_0) and an effect parameter (β_1) of a simulated ecological covariate. For the ZIF model, we modeled p_g with an intercept (p_0) and an effect parameter ($/alpha_1$) of a simulated detection covariate. For ψ_g, we used an intercept (ψ_0) and an effect parameter ($\omega 1$) of a simulated sampling bias/intensity covariate. For the original model, we modeled p_g with an intercept (p_0) and effect parameters (α_1, ω_1) of each the same detection and sampling intensity covariates. Figure S3 shows results from 500 simulations of each model.

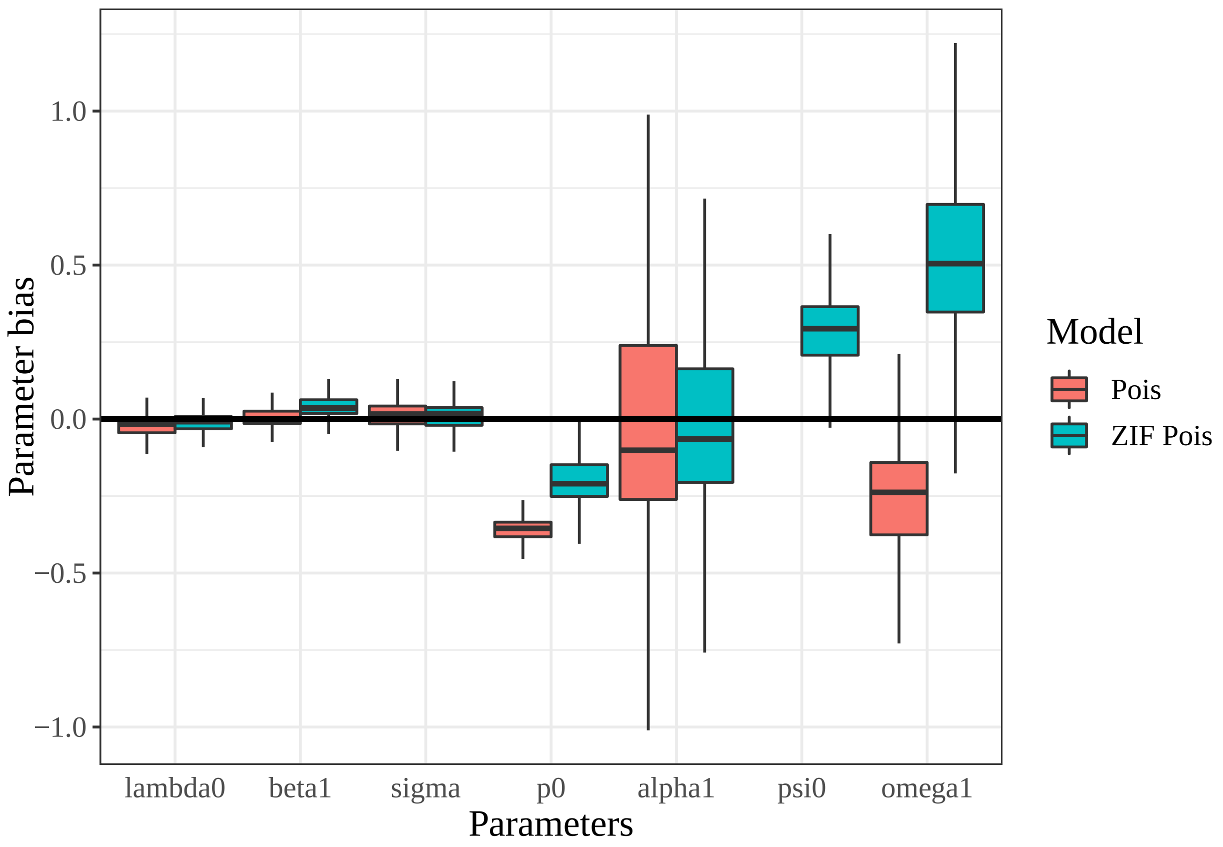

Figure S3. Parameter biases (estimated – truth) via box plots for each model structure (original is red and ZIF is blue).

We found that both models produced unbiased estimates of λ_g (including intercept [$\lambda 0$] and effect parameter [β_1]). The scale parameter (σ) of the distance function was also unbiased for both models. However, we found that both models produced slightly biased estimates of the observation parameters. Both the non-ZIF and ZIF model underestimated the intercept (p_0) of p_g and produced large uncertainty around the effect parameter (α_1). The ZIF model overestimated the intercept (ψ_0) and effect parameter (ω_1) on ψ_g. The non-ZIF model underestimated the effect of sampling intensity (ω_1) that was modeled on p_g. These simulation results demonstrated that parsing apart detection and sampling bias is not feasible with a ZIF model with standard quantities of data. The non-ZIF model is sufficient to estimate the biological process parameters as it can return the intercept on abundance and the effect of environmental covariates on abundance even when the data was simulated under the ZIF model assumptions. Estimates of the observation process parameters for presence-only data may be biased within an integrated framework and were also confounded since they are describing the product of sampling bias and imperfect detection. We thus caution readers on overly interpreting estimates of p_g including the intercept or effects parameters as they are the product of detection and sampling probabilities.

Literature Cited

Casella, G. & Berger, R.L. (2002) Statistical Inference, Thomson Learning, Boston.