Assessing the impacts of time to detection distribution assumptions on detection probability estimation

Adam Martin-Schwarze, Jarad Niemi, and Philip Dixon

Department of Statistics, Iowa State University, Ames, Iowa

September 20, 2016

Abstract

Abundance estimates from animal point-count surveys require accurate estimates of detection probabilities. The standard model for estimating detection from removal-sampled point-count surveys assumes that organisms at a survey site are detected at a constant rate; however, this assumption is often not justified. We consider a class of N-mixture models that allows for detection heterogeneity over time through a flexibly defined time-to-detection distribution (TTDD) and allows for fixed and random effects for both abundance and detection. Our model is thus a combination of survival time-to-event analysis with unknown-N, unknown-p abundance estimation. We specifically explore two-parameter families of TTDDs, e.g. gamma, that can additionally include a mixture component to model increased probability of detection in the initial observation period. We find that modeling a TTDD by using a two-parameter family is necessary when data have a chance of arising from a distribution of this nature. In addition, models with a mixture component can outperform non-mixture models even when the truth is non-mixture. Finally, we analyze an Overbird data set from the Chippewa National Forest using mixed effect models for both abundance and detection. We demonstrate that the effects of explanatory variables on abundance and detection are consistent across mixture TTDDs but that flexible TTDDs result in lower estimated probabilities of detection and therefore higher estimates of abundance.

Keywords: abundance; availability; hierarchical model; Markov chain Monte Carlo; N-mixture model; point counts; removal sampling; Stan; survival analysis

1 Introduction

Abundance estimates from animal point-count surveys require accurate estimates of detection probabilities. Removal sampling, where individuals are solely counted on their first capture, provides one established methodology for estimating detection probabilities (Farnsworth et al., 2002). A typical assumption in removal sampling is a constant detection rate throughout the observation period, but this assumption is often unjustified (Alldredge et al., 2007). In particular, animal behaviors such as intermittent singing in birds and frogs or diving in whales (Scott et al., 2005; Diefenbach et al., 2007; Reidy et al., 2011), differences in behavior across subgroups of animals (Otis et al., 1978; Farnsworth et al., 2005), observer impacts on animal behaviors (McShea and Rappole, 1997; Rosenstock et al., 2002; Alldredge et al., 2007), and variations in observer effort, e.g. saturation or lack of settling in period (Petit et al., 1995; Lee and Marsden, 2008; Johnson, 2008), can all lead to time-varying rates of detection.

In this manuscript, we develop a model for scenarios where detection rates are not constant over time. We consider the first time-to-detection as is done in survival analysis, defining a continuous random variable T for each individual's time to first detection with a probability density function (pdf) $f_T(t)$ and cumulative distribution function (cdf) $F_T(t)$. We refer to the distribution of T as a time-to-detection distribution (TTDD). One common strategy to deal with data that do not fit a constant-detection assumption is to model increased detection probability in the initial observation period via a mixture component (Farnsworth et al., 2002, 2005; Efford and Dawson, 2009; Etterson et al., 2009; Reidy et al., 2011), although this is not yet the standard (Sólymos et al., 2013; Amundson et al., 2014; Reidy et al., 2016). We consider the choice of whether to include a mixture component in conjunction with TTDDs with non-constant rates.

Unlike most survival analyses, the number of individuals N present at a survey is unknown and may be the primary quantity of interest. We embed the TTDD in a hierarchical framework for multinomial counts using an N-mixture model (Wyatt, 2002; Royle, 2004b). For our purposes, the N-mixture framework provides three clear benefits: 1) it handles counts within a flexible multinomial data framework (Royle and Dorazio, 2006) which accords with the interval-censored data collection that is customary in point-count surveys (Ralph et al., 1995), 2) the hierarchical structure readily lends itself to including abundance- and detection-related covariates and random effects (Dorazio et al., 2005; Etterson et al., 2009; Amundson et al., 2014), and 3) for a Bayesian analysis, we can sample the posterior joint distribution of N-mixture parameters straight-forwardly using Markov chain Monte Carlo (MCMC). The N-mixture framework models abundance as a latent variable with a Poisson or other discrete distribution and independently models detection probabilities. Several previous studies have employed the N-mixture framework to analyze removal sampled point-count data while assuming constant detection rates (Royle, 2004a; Dorazio et al., 2005; Etterson et al., 2009; Sólymos et al., 2013; Amundson et al., 2014; Reidy et al., 2016).

Framing a model in terms of time-to-detection leads to two practical differences vis-a-vis constant-detection models. First, in order to model covariate and random effects on detection, we perform mixed effects linear regression on the log of the rate parameter as in Sólymos et al. (2013), whereas most existing studies instead construct regression models on the logit of the equal-interval detection probability. The latter is not possible when detection rates are not constant. Second, because we can obtain interval-specific detection probabilities from the TTDD by partitioning its cdf, we can directly model the data according to their existing interval structure rather than subdividing the observation period into intervals of equal duration. Indeed our model fits exact time-to-detection data, whereas existing constant-detection removal models only approximate exact data by subdividing the observation interval into a large number of fine equal-duration intervals (Reidy et al., 2011; Amundson et al., 2014).

Section 2 provides a description of the interval-censored time-to-detection avian point count data under consideration. Section 3 introduces an N-mixture model with a generically defined TTDD for estimating abundance from removal-sampled point-count surveys. Section 4 provides three simulation studies to assess the impact of TTDD choice on estimated detection probability. Section 5 analyzes an Ovenbird data set under different TTDDs to determine the impact of this choice on estimated detection probability and therefore estimated abundance.

2 Interval-censored point counts

Our analysis is motivated by avian point-count surveys in Chippewa National Forest from 2008-2013 as part of the Minnesota Forest Breeding Bird Project (MNFB) (Hanowski et al., 1995). For our analysis, we focused on Ovenbird counts selected from one habitat type: sawtimber red pine stands with no recent logging activity. Each stand had up to four sites with sufficient geographical distance between sites to reduce or eliminate overlapping territories. The data included 65 sites and a total of 381 surveys with site specific variables including site age, stock density, and an indicator of select-/partial-cut logging during the 1990s.

Single-visit (per year) point-count surveys were conducted by trained observers at each site once annually (weather permitting). Fourteen different observers conducted surveys during the study period and 69% of surveys in our dataset involved observers in their first year at the MNFB. Survey durations were 10 minutes, with times to first detection censored into nine intervals: a two-minute interval followed by eight one-minute intervals. During each survey, the Julian date, time of day, and temperature were recorded.

While we focus on the estimation of detection probability in avian populations, the approach we describe is appropriate for point-count surveys of any species. The methodology allows the analysis of data with 1) recorded first (possible censored) detection of each individual, 2) site-specific explanatory variables, and 3) survey-specific explanatory variables.

3 Continuous time-to-detection N-mixture models

Before considering interval censoring and explanatory variables, we first present the scenario of exact time to detections with no explanatory variables. We then incorporate interval censoring and follow with inclusion of fixed and random effects for abundance and detection.

3.1 Exact time to detection

Suppose that, for each survey s ($s = 1, \dots, S$), N_s individuals are present. Imagine an observer could remain at the survey location until every individual is detected, recording the time to detection t_{sb} (for bird, $b = 1, \dots, N_s$) for each. Assuming detection times for all individuals at a survey are independent, identically distributed according to a common time-to-detection distribution (TTDD), we define T_{sb} as a random variable with cumulative distribution function (cdf) $F_T(t)$ and probability density function (pdf) $f_T(t)$. In reality, times to first detection are often truncated due to a finite survey length of C, meaning that each individual has a detection probability $p^{(det)} = F_T(C)$. The conditional distribution of observed detection times then has pdf $f_{T|det}(t) = f_T(t)/F_T(C)$ for $0 < t < C$, cdf $F_{T|det}(t) = \int_0^t f_{T|det}(x)dx$, and instantaneous detection rate, or hazard function, is $h(t) = f_T(t)/[1 - F_T(t)]$. We model the number of individuals at survey s for which t_{sb} is actually observed as $n_s^{(obs)} \overset{ind}{\sim} \text{Binomial}\left(N_s, p^{(det)}\right)$.

A common choice for TTDD is an exponential distribution, i.e. $T_{sb} \overset{ind}{\sim} \text{Exp}(\varphi)$, which imposes a constant first detection rate, i.e. $h(t) = \varphi$. Choosing another TTDD can allow for a systematic non-constant detection regime. For example, to model an observer effect where: (i) the observer's arrival suppresses or stimulates detectable cues, but (ii) organisms acclimate and gradually return to constant detection, a gamma TTDD could be appropriate. In addition to an exponential and gamma TTDD, we also consider Weibull and lognormal as these distributions are often used in survival analysis. To facilitate the later inclusion of fixed and random effects, we use the following rate-based parameterizations: $T \sim \text{Exp}(\varphi), E[T] = 1/\varphi$; $T \sim \text{Ga}(\alpha, \varphi), E[T] = \alpha/\varphi$; $T \sim \text{We}(\alpha, \varphi), E[T] = \Gamma(1 + 1/\alpha)/\varphi$; and $T \sim \text{LN}(\varphi, \alpha), E[T] = \exp(\alpha^2/2)/\varphi$. This parameterization of the lognormal relates to the standard (μ, σ^2) parameterization by $\varphi = \exp(-\mu)$ and $\alpha = \sigma$. The exponential distribution is a special case of both the gamma and Weibull distributions when $\alpha = 1$.

Using maximum likelihood, we can estimate the parameters in a TTDD from exact times to first detection and thus estimate $p^{(det)}$ and its uncertainty. With an estimate and uncertainty for $p^{(det)}$, we are in a scenario of a binomial model with unknown N_s and "known" p and thus can estimate N_s. In these models, point estimates of N_s are unstable unless $p^{(det)} > 0.4$ (Olkin et al., 1981). One approach to regularizing these estimates is to construct a hierarchical model for the site-specific abundance, e.g. $N_s \overset{ind}{\sim} \text{Po}(\lambda)$ (Raftery, 1988; Royle, 2004b). With this assumption, we can decompose N_s into observed and unobserved portions: $n^{(obs)} \sim Po(\lambda p^{(det)})$ and, independently, $n_s^{(unobs)} \sim \text{Po}\left(\lambda[1 - p^{(det)}]\right)$. Although alternative distributions could be considered, e.g. negative binomial, our experience with Ovenbird point counts suggests that, after accounting for appropriate explanatory variables, the resulting abundances are likely underdispersed rather than overdispersed, and thus we will use the Poisson assumption here.

3.2 Interval-censored times to detection

Due to the harried process of avian point counts, times to first detection are typically not recorded exactly, but are instead censored into I intervals. Let C_i for $i = 1, \ldots, I$ indicate the right endpoint of the ith interval then C_I is the total survey duration and, letting $C_0 = 0$, the ith interval is $(C_{i-1}, C_i]$. Let n_{si} be the number of individuals counted during interval i on survey s, $n_s^{(obs)} = \sum_{i=1}^{I} n_{si}$, and $\mathbf{n}_s = (n_{s1}, \ldots, n_{sI})$. Assuming independence amongst individuals and sites, we have $\mathbf{n}_s \overset{ind}{\sim} \text{Mult}\left(n_s^{(obs)}, \mathbf{p}_s\right)$, where $\mathbf{p}_s = (p_{s1}, \ldots, p_{sI})$ and is calculated from the TTDD: $p_{si} = F_{T|det}(C_i) - F_{T|det}(C_{i-1})$.

3.3 Detection heterogeneity across subgroups

It is common in avian point counts to observe increased detections in the first interval relative to an exponential distribution. This is often understood to reflect unmodeled detection heterogeneity across behavioral groups in the study population. Failure to account for such heterogeneity in the constant-detection scenario leads to negative bias in abundance estimates (Otis et al., 1978). To accommodate this empirical observation, many models of interval-censored removal times define a TTDD with a mixture component to increase the probability of observing individuals in the first interval (Farnsworth et al., 2002; Royle, 2004a; Farnsworth et al., 2005; Alldredge et al., 2007; Etterson et al., 2009; Reidy et al., 2011). We specify a mixture TTDD with mixing parameter $\gamma \in [0,1]$, a point-mass during the first observation interval, and a continuous-time detection distribution $F_T^{(M)}(t)$. The mixture TTDD cdf is defined: $F_T(t) = (1-\gamma) + \gamma F_T^{(M)}(t)$ for $t > 0$. If $\gamma = 1$, the non-mixture model is recovered.

3.4 Incorporating explanatory variables

As discussed in Section 2, explanatory variables are available for sites and for surveys. Generally, we suspect that site variables, e.g. habitat, will affect abundance and survey variables, e.g. time of day, will affect detection probability. Thus, we allow for incorporating explanatory variables on both the abundance and detection.

To incorporate explanatory variables on abundance, we model the expected survey abundance λ_s with log-linear mixed effects, i.e. $\log(\lambda_s) = \mathbf{X}_s^A \boldsymbol{\beta}^A + \mathbf{Z}_s^A \boldsymbol{\xi}^A$ where $\mathbf{X}_s^A$ are explanatory variables, $\boldsymbol{\beta}^A$ is a vector of fixed effects, $\mathbf{Z}_s^A$ specifies random effect levels, and $\xi_j^A \overset{ind}{\sim} N(0, \sigma^2_{A[j]})$ are random effects where $A[j]$ assigns the appropriate variance for the jth abundance random effect.

To incorporate explanatory variables on detection probability, we let the continuous portion of the TTDD depend on the explanatory variables through the now site-specific parameter φ_s. Specifically, we model $\log(\varphi_s) = \mathbf{X}_s^D \boldsymbol{\beta}^D + \mathbf{Z}_s^D \boldsymbol{\xi}^D$, where $\mathbf{X}_s^D$ are explanatory variables, $\boldsymbol{\beta}^D$ is a vector of fixed effects, $\mathbf{Z}_s^D$ specifies random effect levels, $\xi_j^D \overset{ind}{\sim} N(0, \sigma^2_{D[j]})$ are random effects where $D[j]$ assigns the appropriate variance for the jth detection random effect. For simplicity, we assume the shape parameter α as constant across sites.

3.5 Estimation

For ease of reference, the final full model is provided in equation (1) where the conditioning of the TTDD cdf on α and φ_s is made explicit.

$$\begin{aligned}
n_s^{(obs)} &\overset{ind}{\sim} \text{Po}(\lambda_s p_s^{(det)}) \\
\mathbf{n}_s &\overset{ind}{\sim} \text{Mult}(n_s^{(obs)}, \mathbf{p}_s); \qquad \mathbf{p}_s = (p_{s1}, \ldots, p_{sI}) \\
p_s^{(det)} &= F_T(C_I|\alpha, \varphi_s) \\
p_{s1} &= \left[(1-\gamma) + \gamma F_T^{(M)}(C_1|\alpha, \varphi_s)\right] / p_s^{(det)} \\
p_{si} &= \gamma \left[F_T^{(M)}(C_i|\alpha, \varphi_s) - F_T^{(M)}(C_{i-1}|\alpha, \varphi_s)\right] / p_s^{(det)} \\
\log(\lambda_s) &= \mathbf{X}_s^A \boldsymbol{\beta}^A + \mathbf{Z}_s^A \boldsymbol{\xi}^A; \qquad \xi_j^A \overset{ind}{\sim} N(0, \sigma^2_{A[j]}) \\
\log(\varphi_s) &= \mathbf{X}_s^D \boldsymbol{\beta}^D + \mathbf{Z}_s^D \boldsymbol{\xi}^D; \qquad \xi_j^D \overset{ind}{\sim} N(0, \sigma^2_{D[j]})
\end{aligned} \tag{1}$$

We adopt a Bayesian approach and therefore require a prior over the model parameters. To ease construction of a default prior for this model, we standardize all explanatory variables and then construct priors to be diffuse within a reasonable range of values. Normal prior mean and standard deviation (sd) for the abundance intercept was set at a median abundance of 3 birds per site and a 95% probability of 0-14 birds present (counted and uncounted). Normal prior mean and sd for the detection intercept were chosen so that, based on an intercept-only non-mixture model with $\alpha = 1$: (i) median prior detection probability was $p_s^{(det)} = 0.50$, and (ii) 95% of the prior detection probability was within $p_s^{(det)} \in (0.01, 1.0)$.

Normal priors for fixed effect parameters were centered at zero with standard deviations matching the appropriate intercept term. All standard deviations and α were given half-Cauchy priors with location 0 and scale 1 for the untruncated Cauchy, and the mixture parameter γ was assigned a Unif(0,1) prior in mixture models. All scalar parameters were assumed independent *a priori*.

We fit the models by MCMC sampling using the Bayesian statistical software Stan, implemented via the R package `rstan` version 2.8.0 (Stan Development Team, 2015). Stan model code and `rstan` code to run the models can be found in the Supplementary Materials. We discarded half of the iterations as warmup and then thinned by 10. We monitored convergence of the MCMC chains using Geweke z-score diagnostics (Geweke et al., 1991) and reran models if lack of convergence was indicated by a non-normal distribution of the z-scores or if the effective sample size for any parameter was below 1000. The number of iterations used depended on the model and is detailed later. For most models, we accepted Stan defaults for initial values; however, gamma and Weibull models sometimes failed to run unless care was taken in the specification of initial values.

4 Simulation studies

We conducted three simulation studies to explore the behavior of models with non-constant TTDDs. The first study compares mixture vs non-mixture models. The second study compares the TTDD families. In the first two studies, we utilized intercept only models to focus attention on the TTDD. For the third study, we included fixed and random effects for both abundance and detection and again compared the distribution families. In all simulation studies, we focus on accuracy in estimation of $p^{(det)}$ which then translates into estimation of abundance. For the following analyses we distinguish two categories of purely continuous TTDDs: peaked and nonpeaked. A peaked TTDD has a mode greater than zero (or C_1 for lognormal) while a non-peaked TTDD has a mode of zero (or less than C_1).

4.1 Mixture versus non-mixture TTDDs

To assess the need for incorporating a mixture component to increase the probability of detection in the initial interval as discussed in Section 3.2, we simulated 16 intercept-only datasets from each of 14 TTDDs: each combination of peaked/non-peaked, mixture/non-mixture, and exponential/gamma/Weibull/lognormal, where all exponential models are non-peaked. We chose the number of surveys (381) and true parameter values (Table S-1) to mimic values from the Ovenbird analysis (Section 5). In particular, we set parameters such that (i) the overall expected detection probability was 0.80, (ii) for mixture datasets, $\gamma = 0.65$ (meaning 35% of individuals were immediately detected), (iii) in nonpeaked models, 70% of *detected* individuals were observed during the first two minutes, and (iv) in peaked models, the detection mode for 'hard to detect' individuals occured at 5 minutes.

Each dataset was fit with two models: mixture and non-mixture version of the distribution family, e.g. exponential, used to simulate the data. For each dataset-analysis combination, we ran 60,000 iterations which showed no evidence of lack of convergence according to the Geweke diagnostic and reached over 1,000 effective samples for all parameters.

We summarized the analysis by calculating the average, across the simulations, of the posterior median of $p^{(det)}$ (Med. p), the average proportion of the posterior distribution of $p^{(det)}$ larger than the truth (Q(p)), and average coverage for 50% and 90% credible intervals. If the analyses are providing a reasonable estimate of $p^{(det)}$, we would expect Med. p to be around the true value of 0.8, Q(p) to be around 0.5 indicating that half of the posterior is above and below the true value of 0.8, and the coverage to be close to their credibility. Table 1 provides a summary of these quantities. When a mixture model is used to simulate the data (bottom half of the table), there is clearly a benefit to using a mixture model for inference. Using a non-mixture model for inference, the credible interval coverage is near zero for most models with the exponential model overestimating $p^{(det)}$ (Q(p) $\approx$ 1) and the other models underestimating. When a non-mixture model is used to simulate the data (top half of the table), there are no clearly discernable differences between the ability of a non-mixture or mixture model to capture $p^{(det)}$. For nonpeaked non-mixture data (top 4 lines), the mixture model is less biased for lognormal and Weibull scenarios (Q(p)$\approx$ 0.5), and CIs are 15-40% narrower (results not shown). These results support the general default use of a mixture model over a non-mixture model.

				Non-mixture model				Mixture model			
				Med. p	Q(p)	50%	90%	Med. p	Q(p)	50%	90%
TTDD used to simulate data	Non-mixture	Nonpk.	Gamma	0.76	0.41	0.75	0.88	0.84	0.66	0.56	0.94
			Lognormal	0.66	0.17	0.31	0.62	0.78	0.48	0.62	1.00
			Weibull	0.69	0.25	0.38	0.75	0.79	0.51	0.75	1.00
			Exponential	0.80	0.54	0.44	0.94	0.79	0.38	0.31	0.88
		Peaked	Gamma	0.80	0.55	0.50	1.00	0.82	0.70	0.38	0.88
			Lognormal	0.78	0.31	0.50	0.94	0.80	0.48	0.62	1.00
			Weibull	0.79	0.49	0.56	0.94	0.82	0.66	0.44	0.88
	Mixture	Nonpk.	Gamma	0.67	0.17	0.12	0.81	0.76	0.41	0.69	1.00
			Lognormal	0.56	0.04	0.00	0.31	0.72	0.30	0.44	0.88
			Weibull	0.51	0.02	0.00	0.06	0.71	0.32	0.44	1.00
			Exponential	0.96	1.00	0.00	0.00	0.77	0.37	0.38	0.94
		Peaked	Gamma	0.28	0.00	0.00	0.00	0.74	0.33	0.31	0.88
			Lognormal	0.22	0.00	0.00	0.00	0.76	0.36	0.38	0.94
			Weibull	0.22	0.00	0.00	0.00	0.70	0.29	0.56	0.94

Table 1: Summary of mixture vs. non-mixture model fits. In all cases, the inference model family matches the dataset family. Med p: average across simulations of the posterior median of $p^{(det)}$ (true value = 0.80). Q(p): average proportion of the posterior distribution of $p^{(det)}$ that is larger than the true value. 50% and 90% coverage is expressed as the proportion of 16 simulations for which the true value of $p^{(det)}$ lies within the appropriate credible interval.

4.2 Constant vs. non-constant detection mixture TTDDs

The previous section addressed model mis-specification in terms of the mixture component. Now we turn to model misspecification of the distribution family. We simulated 16 intercept-only datasets from the 7 different mixture TTDD models using the same settings as in the previous section, and we fit them with mixture models from each of exponential, gamma, lognormal, and Weibull families.

Figure 1 presents a representative example of posterior distributions for $p^{(det)}$ for data and models from the exponential and gamma mixture families. In this example, the posterior distribution under an exponential inference model accurately captures the true detection probability when the simulation model is an exponential, but overestimates (underestimates) the detection probability when the simulation model has a small (large) gamma shape parameter. In contrast, the posterior from a gamma family, which has the additional flexibility of the shape parameter, is able to accurately capture the truth in all scenarios with increased uncertainty.

To summarize these findings across the various simulation and inference models, Table 2 provides simulation-averaged Median estimates of $p^{(det)}$ (Med p), proportion of the $p^{(det)}$ posterior larger than the truth ($Q(p)$), and 50% and 90% coverage proportions. Each table quadrant assesses a model used for inference compared to the 7 different models used for simulation. The poorest estimation of $p^{(det)}$ occurs for the exponential inferential model when the model used for simulation has a peak, because the two parameters (rate and mixing parameter) do not provide enough flexibility for the mixture exponential distribution to adequately fit a TTDD with both an initial increase and a delayed mode. In this situation, the exponential model underestimates the actual detection probability. In contrast, if the simulated data are non-peaked, the exponential model typically overestimates the actual detection probability. However, exponential model estimates are both less biased and more precise when the data actually do derive from an exponential mechanism.

When comparing the different three-parameter TTDDs, model misspecification is not as serious an issue because these models can adequately account for the interval-censored times to detection. Nonetheless, it appears the gamma mixture model is better able to account for data from most non-peaked datasets, while gamma and Weibull mixture models have

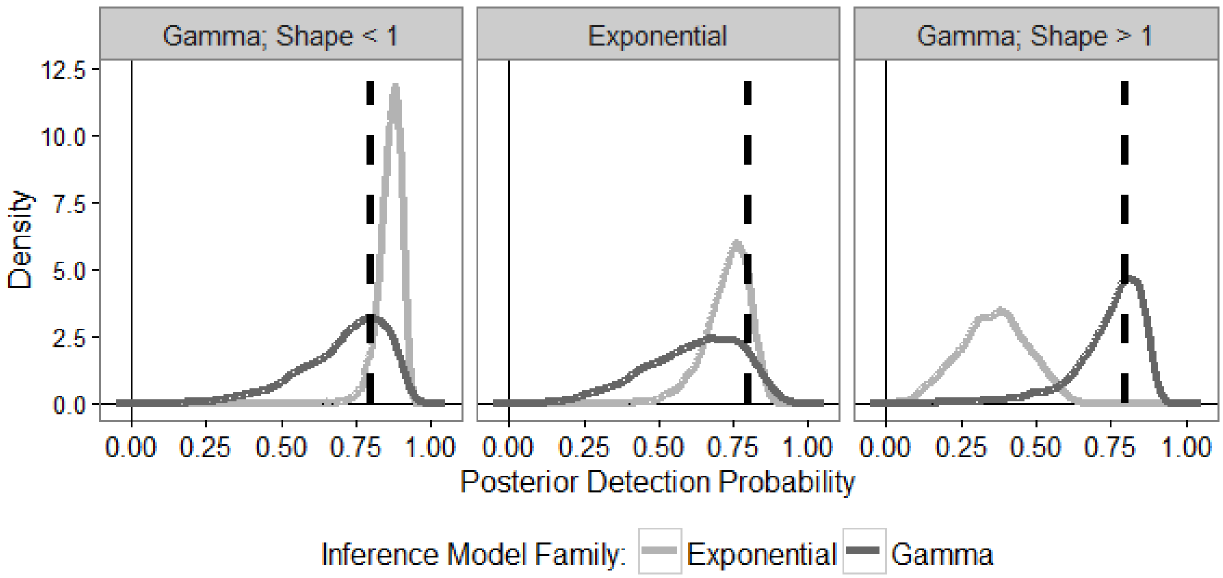

Figure 1: Representative example of the posterior distributions for $p^{(det)}$ from exponential (light solid line) and gamma (dark solid line) mixture models fit to data simulated from nonpeaked gamma (shape<1), exponential (shape=1), and peaked gamma (shape>1) mixture TTDDs, with true detection probabilities shown for comparison (dashed vertical line, $p^{(det)} = 0.8$).

comparable performance across peaked datasets. In the gamma quadrant of Table 2, the average posterior median is near the truth of 0.8, the average proportion of the posterior distribution of $p^{(det)}$ is near 0.5, and the credible intervals have coverage near their credibility. In contrast, when the lognormal model is used for inference, it performs worse across nearly all data types. However, gamma models required $\sim$ 10 times the computational time as did lognormal and Weibull models and almost 25 times that of exponential models.

4.3 Models including covariates and random effects

The previous sections studied effects of time to detection assumptions in the context of no explanatory variables. We now incorporate fixed and random effects for abundance and detection. We simulated data from each of the 7 mixture TTDDs and fit models from exponential, gamma, lognormal, and Weibull mixture models. We simulated data using the

			Exponential mixture model				Gamma mixture model			
			Med. p	Q(p)	50%	90%	Med. p	Q(p)	50%	90%
Data Mixture	Nonpk.	Gamma	0.88	0.91	0.06	0.69	0.76	0.41	0.69	1.00
		Lognormal	0.92	0.99	0.00	0.12	0.85	0.72	0.44	0.69
		Weibull	0.87	0.83	0.31	0.38	0.79	0.49	0.69	1.00
		Exponential	0.77	0.37	0.38	0.94	0.68	0.24	0.25	0.81
	Peaked	Gamma	0.36	0.00	0.00	0.00	0.74	0.33	0.31	0.88
		Lognormal	0.34	0.00	0.00	0.00	0.84	0.73	0.44	0.81
		Weibull	0.39	0.00	0.00	0.00	0.66	0.15	0.19	0.75

			Lognormal mixture model				Weibull mixture model			
			Med. p	Q(p)	50%	90%	Med. p	Q(p)	50%	90%
Data Mixture	Nonpk.	Gamma	0.65	0.16	0.19	0.81	0.67	0.24	0.31	1.00
		Lognormal	0.72	0.30	0.44	0.88	0.77	0.48	0.44	1.00
		Weibull	0.68	0.23	0.31	0.88	0.71	0.32	0.44	1.00
		Exponential	0.61	0.12	0.12	0.44	0.63	0.20	0.25	0.75
	Peaked	Gamma	0.65	0.13	0.06	0.56	0.79	0.52	0.69	0.88
		Lognormal	0.76	0.36	0.38	0.94	0.89	0.88	0.12	0.69
		Weibull	0.58	0.03	0.00	0.12	0.70	0.29	0.56	0.94

Table 2: Summary of mixture inference models of all families fit to mixture datasets. Med p: average across simulations of the posterior median of $p^{(det)}$ (true value = 0.80). Q(p): average proportion of the posterior distribution of $p^{(det)}$ that is larger than the true value. 50% and 90% coverage is expressed as the proportion of simulations for which the true value of $p^{(det)}$ lies within the appropriate credible interval.

median posterior parameter estimates obtained in the analysis of Ovenbird data in Section 5. Because the fitted Ovenbird models did not yield peaked distributions, we simulated peaked datasets by: (i) using the same intercepts, shape parameters, and mixing parameters as for peaked data in the previous simulations, (ii) using median covariate and random effects from the Ovenbird estimates, and (iii) scaling the detection intercept and random effect to achieve true detection probabilities ≈ 0.8 with a detection mode at 5 minutes, see Table S-2 for actual parameter values. Due to the computation time involved in estimating models with these fixed and random effects, we simulated each TTDD only once. To obtain reasonable convergence diagnostics and effective sample sizes, these analyses ranged from 250,000-375,000 iterations. Because of the difficulty in integrating random effects over all sites, approximate posterior distributions for the study-wide marginal $p^{(det)}$ were obtained by simulating data from each MCMC sample and calculating the proportion of simulated

Ovenbirds that were observed.

The results from this simulation are qualitatively similar to that Ovenbird analysis (Figure 2) and thus we only briefly review the results here and provide the corresponding figures and tables in the Supplementary Material. Patterns in posterior estimates of site-specific detection probabilities, $p_s^{(det)}$, with respect to mixture and family TTDD forms were the same as in the previous simulation studies – the inclusion of explanatory variables did not make models more robust to violations of mixture- and constant-detection assumptions (Figures S-1 and S-2). Posteriors for abundance fixed and random effects were the same regardless of what TTDD was assumed, see Section S-2.1 and figures therein. Posteriors for the mixing parameter γ and detection fixed and random effects were the same across gamma, lognormal, and Weibull mixture models but were narrower and location-shifted for the exponential mixture model.

5 Ovenbird analysis

We fit the Ovenbird dataset with exponential, gamma, lognormal, and Weibull mixture models. For the abundance half of our model, we used four covariates plus two random effects. The covariates were: (a) site age, (b) survey year, (c) an indicator of whether the site stock density was over 70%, and (d) an indicator of whether the site experienced select-/partial-cut logging during the 1990s. We associated random effects with each survey year and each stand. For the detection half of our model, we used covariates for: (a) Julian date, (b) time of day, (c) temperature, (d) an indicator of whether it is the observer's first year in the database, and (e) an interaction between (a) and (d) to approximate a new observer's learning curve. We associated random effects with each observer. Preliminary model fits did not support the inclusion of quadratic terms for any detection covariates. We centered and standardized all continuous covariates prior to fitting models. We ran chains 250,000-375,000 iterations; Geweke diagnostics showed no indication of lack of fit, and effecive sample sizes were over 1000 for all parameters.

Figure 2 presents posterior medians and credible intervals for model parameters, overall detection probability $p^{(det)}$, and the logarithm of the number of uncounted Ovenbirds. Estimates for the shape parameter α from the gamma and Weibull models are consistent with the data arising from an exponential distribution, although the uncertainty on this parameter remain relatively large.

Abundance covariate coefficient estimates were virtually the same across all models. The 95% credible intervals for two of the abundance parameters (site age and logging) do not contain zero, thereby suggesting notable effects. Select- and partial-cut logging events of the 1990s depressed local Ovenbird abundance during the study perior to roughly 25-50% of the abundance for unlogged sites. Credible intervals for site age coefficient indicate that each decade of age increases abundance from 1.5-13%. Credible intervals for detection parameters do not indicate significant effects, after adjusting for the other predictors, for any of the included predictors.

In spite of the similarity of effect parameter estimates, the posterior distributions for detection probability and uncounted abundance differ greatly between the exponential and non-exponential models. It is clear that the assumption of constant detection leads to much higher and more precise estimates of detection than would be obtained if we are unwilling to make that assumption.

6 Discussion

We formulated a model for interval-censored time to detection data that allows for non-constant detection rates. Our model adopts a time-to-event approach within a hierarchical N-mixture framework, and it allows times to first detection to be modeled according to flex-

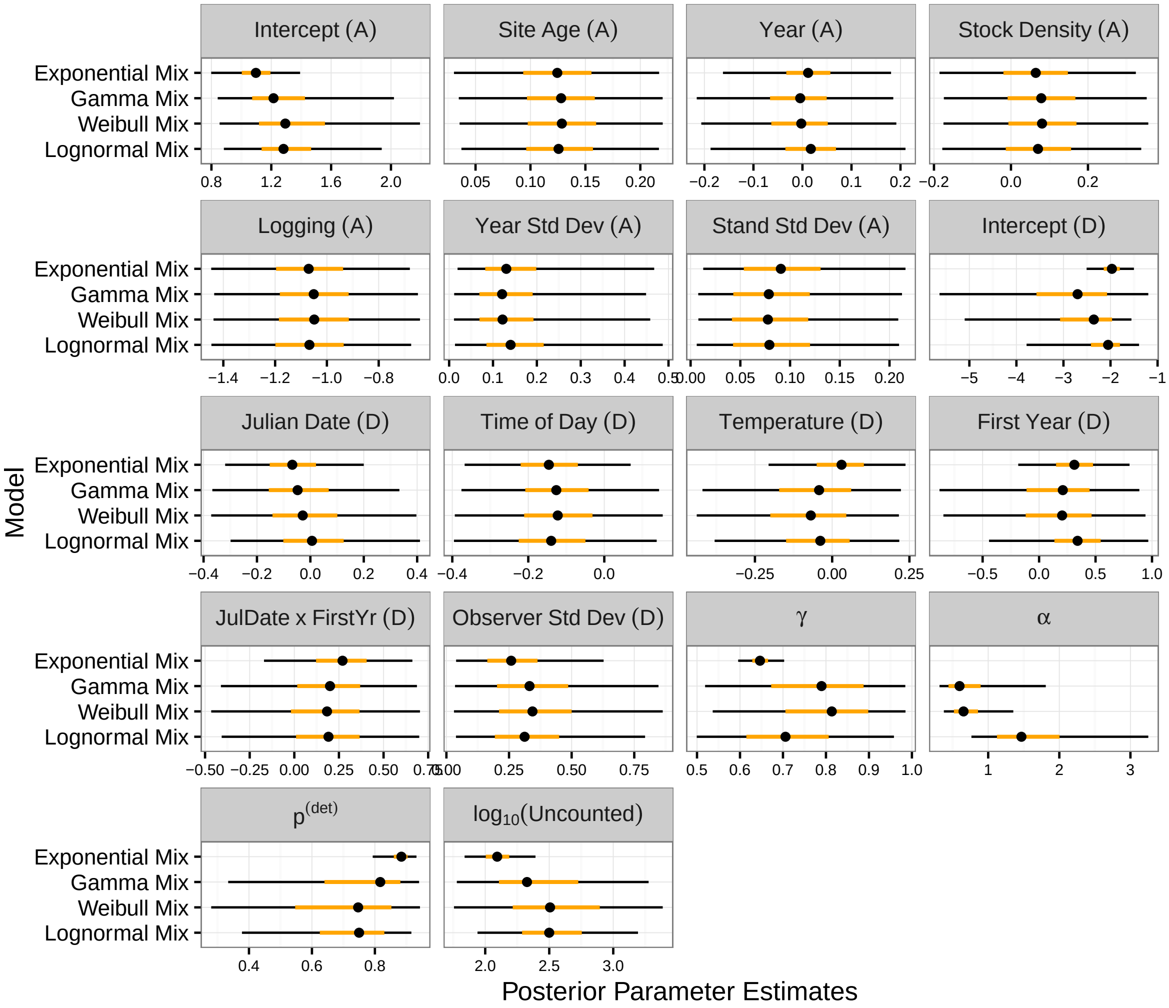

Figure 2: Posterior medians (black dots) with 50% (wider line) and 95% (narrowerline) credible intervals for the mixing parameter (γ), shape parameter (α) as well as abundance (A) and detection (D) fixed effects and random effect standard deviations. Posteriors are also available for the number of uncounted individuals and the overall probability of detection across all sites.

ibly defined TTDD families. Our results show that non-constant TTDDs can return reasonable estimates of detection probabilities across a variety of time-to-detection data patterns, whereas traditional constant-rate TTDDs return biased and overly precise estimates when data deviate from the constant-rate assumption, even when they include a mixture for heterogeneity across groups. Because the exponential TTDD is a special case of both gamma and Weibull TTDDs, we can interpret the differences in estimation between models as resulting from the information conveyed by the assumption of constant detection. Among the

TTDDs modeled in our study, the gamma mixture TTDD gave the most accurate estimates across data distributions, though the Weibull mixture TTDD was comparable for peaked datasets.

We have additionally demonstrated for non-constant models the utility of using a mixture TTDD formulation. Inference models with a mixture component are accurate whether the data have a mixture or not, whereas inference models without the mixture are badly biased when the data do feature a mixture. Mixture models even outperform non-mixture models when the data are non-mixture and nonpeaked.

If the estimation of effect parameters and the roles of explanatory variables are the primary interest, then our results suggest that the exact choice of TTDD may not be important. Abundance effect estimates are similar regardless of the chosen TTDD. Detection effect estimates, while conditional on the mixing parameter γ, are similar across all mixture non-exponential TTDDs. These findings may well not hold if the same covariate is modeled in both abundance and detection models (Kéry, 2008).

If the estimation of abundance is the primary interest, then the choice of TTDD has large consequences, and we may reasonably ask whether removal sampled point-count surveys are adequate for the purpose. The strategy behind removal sampling is to use the pattern of detections over time to estimate the proportion of individuals that would be detected if only the observation period lasted longer. As such, it is entirely based upon extrapolation and correctly estimating the probability in the tail of the TTDD based on recorded observations. An assumption of constant detection places constraints on the amount of uncertainty in the extrapolated tail but at the cost of potentially sizable bias. When we allow for non-constant detection, estimation of the tail probability becomes less certiain. In theory, estimation of the unsampled tail probability can be improved by conducting longer surveys, but the longer the survey lasts, the greater the risk that individuals enter/depart the study area or are double-counted, which violates the removal sampling assumption of a closed population

(Lee and Marsden, 2008; Reidy et al., 2011). Without extending the observation period, an alternative to removal sampling is to record complete detection records (all detections for every individual) instead of just the first (Alldredge et al., 2007); however, this may not be feasible in studies like MNFB where many species are observed simultaneously

Versions of time-varying models have been described for trap-based removal sampling and continuous-time capture-recapture. Time variation has been modeled through a non-constant hazard function (Schnute, 1983; Hwang and Chao, 2002), a randomly varying detection probability across trapping sessions (Wang and Loneragan, 1996), and constant detection probabilities that vary randomly from individual to individual (Mäntyniemi et al., 2005; Laplanche, 2010). Most of these approaches resulted marginally in a decreasing (nonpeaked) detection function over time. Their results generally echo what we have presented here. Schnute (1983) found that the equivalent of a mixture exponential adequately described their data. Wang and Loneragan (1996), Hwang and Chao (2002), and Mäntyniemi et al. (2005) all found constant-detection models to be flawed, producing underestimates of abundance and too-narrow error estimates; these resulted in inadequate coverage and also overstatement of effect significance.

Point-count survey data often include the recorded distance between observer and detected organism. Because our focus has been on modeling variations in detection rates during the survey period, we have not incorporated distance into our model. Consequently, our application of a TTDD represents an averaging across distance classes, which induces systematic bias in estimates of abundance (Efford and Dawson, 2009; Laake et al., 2011; Sólymos et al., 2013). To be consistent with the continuous time-to-event approach, distance can be incorporated into the detection model as an event-level modifier as is done in Borchers and Cox. This approach is distinct from earlier integrations of removal- and distance sampling, where distance has been treated as an interval-/survey-level modifier (Farnsworth et al., 2005; Amundson et al., 2014). The differences between these implementations may impact estimates of detection and abundance, especially in the presence of behavioral heterogeneity in availability rates across subgroups of the study population. This is an area of ongoing exploration.

We recommend that time-heterogeneous detection rates be explicitly modeled in analyses involving removal-sampled point-count survey data where estimation of detection probability or abundance is a primary objective. The assumption of constant detection, while computationally simple and reasonable as a null model, proves to be rather informative and can result in pronounced bias. Meanwhile, the causes of non-constant detection – i.e., observer effects on behavior and systematic variations in observer effort – are both plausible and not trivially discounted. It would be nice if the data itself could inform us whether constant detection is a reasonable assumption; however, our preliminary efforts to diagnose this assumption using deviance information criterion (DIC) and posterior predictive check statistics have led to weak and sometimes erroneous findings. Development of such a diagnostic tool would be useful, but given the limitations of first time-to-detection data, we are not confident a reliable tool could be easily developed. We believe that more informative data collection, such as complete time-to-detection histories and microphone arrays, offer more effective tools for time-to-event modeling going forward.

7 Supplementary Materials

The supplementary materials include supplementary figures as well as code to fit these models.

References

Alldredge, M. W., Pollock, K. H., Simons, T. R., Collazo, J. A., Shriner, S. A., and Johnson, D. (2007). Time-of-detection method for estimating abundance from point-count surveys. *The Auk* **124,** 653–664.

Amundson, C. L., Royle, J. A., and Handel, C. M. (2014). A hierarchical model combining distance sampling and time removal to estimate detection probability during avian point counts. *The Auk* **131,** 476–494.

Borchers, D. and Cox, M. J. (2016). Distance sampling: 2D or not 2D? `http://media.wix.com/ugd/343855_52e1d27058dc4062960f588903aa7e5e.pdf`. [Online; accessed 8-Sep-2016.

Diefenbach, D. R., Marshall, M. R., Mattice, J. A., Brauning, D. W., and Johnson, D. (2007). Incorporating availability for detection in estimates of bird abundance. *The Auk* **124,** 96–106.

Dorazio, R. M., Jelks, H. L., and Jordan, F. (2005). Improving removal-based estimates of abundance by sampling a population of spatially distinct subpopulations. *Biometrics* **61,** 1093–1101.

Efford, M. G. and Dawson, D. K. (2009). Effect of distance-related heterogeneity on population size estimates from point counts. *The Auk* **126,** 100–111.

Etterson, M. A., Niemi, G. J., and Danz, N. P. (2009). Estimating the effects of detection heterogeneity and overdispersion on trends estimated from avian point counts. *Ecological Applications* **19,** 2049–2066.

Farnsworth, G. L., Nichols, J. D., Sauer, J. R., Fancy, S. G., Pollock, K. H., Shriner, S. A., Simons, T. R., Ralph, C., and Rich, T. (2005). Statistical approaches to the analysis of point count data: a little extra information can go a long way. *USDA Forest Service General Technical Report PSW–GTR–191* pages 735–743.

Farnsworth, G. L., Pollock, K. H., Nichols, J. D., Simons, T. R., Hines, J. E., Sauer, J. R., and Brawn, J. (2002). A removal model for estimating detection probabilities from point-count surveys. *The Auk* **119,** 414–425.

Geweke, J. et al. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA.

Hanowski, J., Niemi, G. J., et al. (1995). Experimental design considerations for establishing an off-road, habitat specific bird monitoring program using point counts. *Monitoring bird populations by point counts. General Technical Report PSW-GTR-149. Pacific Southwest Research Station, Forest Service, US Department of Agriculture, Albany, CA* pages 145–150.

Hwang, W.-H. and Chao, A. (2002). Continuous-time capture-recapture models with covariates. *Statistica Sinica* pages 1115–1131.

Johnson, D. H. (2008). In defense of indices: the case of bird surveys. *The Journal of Wildlife Management* **72,** 857–868.

Kéry, M. (2008). Estimating abundance from bird counts: binomial mixture models uncover complex covariate relationships. *The Auk* **125,** 336–345.

Laake, J., Collier, B., Morrison, M., and Wilkins, R. (2011). Point-based mark-recapture distance sampling. *Journal of Agricultural, Biological, and Environmental Statistics* **16,** 389–408.

Laplanche, C. (2010). A hierarchical model to estimate fish abundance in alpine streams by using removal sampling data from multiple locations. *Biometrical Journal* **52,** 209–221.

Lee, D. C. and Marsden, S. J. (2008). Adjusting count period strategies to improve the accuracy of forest bird abundance estimates from point transect distance sampling surveys. *Ibis* **150,** 315–325.

Mäntyniemi, S., Romakkaniemi, A., and Arjas, E. (2005). Bayesian removal estimation of a population size under unequal catchability. *Canadian Journal of Fisheries and Aquatic Sciences* **62,** 291–300.

McShea, W. and Rappole, J. (1997). Variable song rates in three species of passerines and implications for estimating bird populations. *Journal of Field Ornithology* pages 367–375.

Olkin, I., Petkau, A. J., and Zidek, J. V. (1981). A comparison of n estimators for the binomial distribution. *Journal of the American Statistical Association* **76,** 637–642.

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife monographs* pages 3–135.

Petit, D. R., Petit, L. J., Saab, V. A., and Martin, T. E. (1995). Fixed-radius point counts in forests: factors influencing effectiveness and efficiency. Technical Report PSW-GTR-149, USDA Forest Service.

Raftery, A. E. (1988). Inference for the binomial n parameter: A hierarchical bayes approach. *Biometrika* **75,** 223–228.

Ralph, C. J., Droege, S., and Sauer, J. R. (1995). Managing and monitoring birds using point counts: Standards and applications. Technical Report PSW-GTR-149, USDA Forest Service.

Reidy, J. L., Thompson, F. R., and Bailey, J. (2011). Comparison of methods for estimating density of forest songbirds from point counts. *The Journal of Wildlife Management* **75,** 558–568.

Reidy, J. L., Thompson III, F. R., Amundson, C., and ODonnell, L. (2016). Landscape and local effects on occupancy and densities of an endangered wood-warbler in an urbanizing landscape. *Landscape Ecology* **31,** 365–382.

Rosenstock, S. S., Anderson, D. R., Giesen, K. M., Leukering, T., Carter, M. F., and Thompson III, F. (2002). Landbird counting techniques: current practices and an alternative. *The Auk* **119,** 46–53.

Royle, J. (2004a). Generalized estimators of avian abundance from count survey data. *Animal Biodiversity and Conservation* **27,** 375–386.

Royle, J. A. (2004b). N-mixture models for estimating population size from spatially replicated counts. *Biometrics* **60,** 108–115.

Royle, J. A. and Dorazio, R. M. (2006). Hierarchical models of animal abundance and occurrence. *Journal of Agricultural, Biological, and Environmental Statistics* **11,** 249–263.

Schnute, J. (1983). A new approach to estimating populations by the removal method. *Canadian Journal of Fisheries and Aquatic Sciences* **40,** 2153–2169.

Scott, T. A., Lee, P.-Y., Greene, G. C., McCallum, D. A., et al. (2005). Singing rate and detection probability: an example from the least bell's vireo (vireo belli pusillus). In *Proceedings of the Third International Partners in Flight Conference, US Department of Agriculture Forest Service, Pacific Southwest Research Station, General Technical Report PSW-GTR-191, Albany, CA, USA*, pages 845–853.

Sólymos, P., Matsuoka, S. M., Bayne, E. M., Lele, S. R., Fontaine, P., Cumming, S. G., Stralberg, D., Schmiegelow, F. K., and Song, S. J. (2013). Calibrating indices of avian density from non-standardized survey data: making the most of a messy situation. *Methods in Ecology and Evolution* **4,** 1047–1058.

Stan Development Team (2015). Rstan: the R interface to Stan, Version 2.8.0.

Wang, Y.-G. and Loneragan, N. R. (1996). An extravariation model for improving confidence intervals of population size estimates from removal data. *Canadian Journal of Fisheries and Aquatic Sciences* **53,** 2533–2539.

Wyatt, R. J. (2002). Estimating riverine fish population size from single-and multiple-pass removal sampling using a hierarchical model. *Canadian Journal of Fisheries and Aquatic Sciences* **59,** 695–706.