Biometrics ??, 1–23 DOI: ??
?? ??

Mark-Recapture with Multiple, Non-invasive Marks

Simon J Bonner*

Department of Statistics, University of Kentucky, Lexington, KY 40508, USA

**email:* simon.bonner@uky.edu

and

Jason Holmberg**

ECOCEAN, 1726 N Terrry Street, Portland, OR 97211, USA

**email:* jason@whaleshark.org

Summary: Non-invasive marks, including pigmentation patterns, acquired scars,and genetic markers, are often used to identify individuals in mark-recapture experiments. If animals in a population can be identified from multiple, non-invasive marks then some individuals may be counted twice in the observed data. Analyzing the observed histories without accounting for these errors will provide incorrect inference about the population dynamics. Previous approaches to this problem include modeling data from only one mark and combining estimators obtained from each mark separately assuming that they are independent. Motivated by the analysis of data from the ECOCEAN online whale shark (*Rhincodon typus*) catalog, we describe a Bayesian method to analyze data from multiple, non-invasive marks that is based on the latent-multinomial model of Link et al. (2010). Further to this, we describe a simplification of the Markov chain Monte Carlo algorithm of Link et al. (2010) that leads to more efficient computation. We present results from the analysis of the ECOCEAN whale shark data and from simulation studies comparing our method with the previous approaches.

Key words: Latent multinomial model; Mark-recapture; Multiple marks; Non-invasive marks; Photo-identification; Whale sharks

1. Introduction

Non-invasive marks (also called natural marks) include patterns in pigmentation, genetic markers, acquired scars, or other natural characertistics that allow researchers to identify individuals in a population without physical capture. Visible marks have long been used to identify individuals of some species that are hard to tag, particularly marine mammals, and non-invasive marks are now being used more widely. Yoshizaki et al. (2009) and Yoshizaki et al. (2011) reference studies including:

- studies based on photographs of large cats (cheetahs, snow leopards, and tigers),
- scar patterns on marine mammals (manatees and whales),
- skin patterns of reptiles and amphibians (snakes, crocodiles, and salamanders),
- and genetic marks in various species (bears, wombats, and whales).

The primary advantage of non-invasive marks over man-made marks is that they can often be observed from a distance or through the collection of secondary material (e.g. hair samples or scat). This means that individuals can be identified passively without physical contact. Further, many non-invasive marks allow every individual in the population to be identified from birth. However, mark-recapture data collected from non-invasive marks can present several modeling challenges. Previous statistical developments have considered that non-invasive marks may be misidentified at non-negligible rates (Lukacs and Burnham, 2005; Wright et al., 2009; Yoshizaki et al., 2011), that individuals' marks may change over time (Yoshizaki et al., 2009), and that some non-invasive marks (e.g scar patterns) may be restricted to a subset of the population (Da-Silva et al., 2003; Da-Silva, 2006). We consider the problem of modeling the demographics of a population from mark-recapture data when individuals have been identified from multiple, non-invasive marks.

The specific application we consider concerns modeling the aggregation of whale sharks (*Rhincodon typus*) in Ningaloo Marine Park (NMP), off the west coast of Australia. Whale

sharks aggregate at NMP each year between April and July. During this time, whale sharks are located by tour companies and photographs are taken by tourists and tour operators who upload their images to the online ECOCEAN whale shark library. Whale sharks can be identified by the unique pattern of spots on their flanks, and computer assisted methods are used to match photographs in the library. Matches are then used to generate capture histories which provide information about the timing of the sharks' arrival and departure from NMP and their survival across years (see Holmberg et al., 2009, for further details).

The challenge in modeling this data is that sharks may be photographed from either the left or the right side, but the spot patterns are not the same. This means that photographs from the two sides of a shark cannot be matched without further information. In particular, the spot patterns on the right and left can only be matched if the shark was photographed from both sides during one encounter or more. If this has not happened then photographs of the same shark taken from different sides on different occasions cannot be linked and the shark will contribute two separate histories to the observed data. Ignoring this problem and naively modeling the observed encounter histories will inflate the apparent number of sharks identified and create dependence between the encounter histories. This violates a key assumption of most mark-recapture models. One solution is to construct encounter histories based on photographs from either the left or right side alone, but this removes information from the data. As an alternative, Wilson et al. (1999, pg. 294) suggests combining inferences obtained from left- and right-side photographs of bottlenose dolphins by averaging separate point estimates and computing standard errors assuming that these estimates are independent. The bias of the combined estimate is the average of the biases of the individual estimates (the combined estimate is unbiased if the individual estimates are unbiased), but the assumption of independence is violated and standard errors will be underestimated. More recently, Madon et al. (2011) describes a method to estimate abundance from multiple marks

by adjusting the sufficient statistics required to compute the Jolly-Seber estimator, but we have concerns with this method. Though the observed counts underestimate some of the statistics and overestimate others, Madon et al. (2011) uses the same adjustment factor for all and constrains its value to be between 0 and 1. Simulations Madon et al. (2011) presents indicate a clear problem in that the coverage of confidence intervals is much lower than their nominal value, even when the population is large and the capture probability is close to 1. These issues are discussed further in Bonner (2013). We are also aware of methods similar to ours being developed concurrently by McClintock et al. (2013).

The primary contribution of our work is to provide a valid method of modeling a population's dynamics using data from multiple, non-invasive marks. We do so by constructing an explicit model of the observation process that allows for multiple marks and applying Bayesian methods of inference via Markov chain Monte Carlo (MCMC) sampling. Our model is a modification of the latent multinomial model (LMM) presented in Link et al. (2010) for modeling mark-recapture data based on genetic marks with non-negligible misidentification rates. Further to this, we provide a more efficient simplification of the MCMC algorithm of Link et al. (2010).

2. Data

Data for our analysis were obtained from the ECOCEAN on-line whale shark library (available at `www.whaleshark.org`). This library contains photographs of whale sharks taken by recreational divers and tour operators worldwide and submitted electronically. The library has been operational since 2003, and more than 41,000 photographs had been submitted by over 3,300 contributors as of January, 2013.

New photographs submitted to the library are matched against existing photographs using two computer algorithms (Arzoumanian et al., 2005; Van Tienhoven et al., 2007). Identities are based on the pattern of spots on the flank, believed to be unique, and the algorithms

operate independently using significantly different approaches to provide complementary coverage in evaluating matches. All matches generated by the computer algorithms are confirmed by two or more trained research staff to minimize the probability of false matches. Further details on the study site, the observation protocols, and the algorithms for matching photographs are provided in Holmberg et al. (2009).

We model only the data collected from the northern ecotourism zone of NMP during the 16 week period between April 1 and July 31, 2008. This period was divided into 8 capture occasions of 2 weeks each, and sharks may have been encountered multiple times during a single capture occasion. Five possible events may occur; on each occasion, a shark may:

1) not be encountered at all (event 0)
2) be photographed from the left only (event L),
3) be photographed from the right only (event R),
4) be photographed from both sides simultaneously on at least on encounter (event S), or
5) be photographed from both sides but never simultaneously (event B).

We will denote a generic encounter history made from these events by $\boldsymbol{\omega}$.

Problems with identification arise because the pattern of spots on the left and right flanks are not the same. It is only possible to match the skin patterns from the two sides of a shark if photographs of both sides were taken simultaneously during at least one capture occasion – i.e., there is at least one S in its encounter history. Otherwise, an individual photographed from both sides will contribute two encounter histories to the data set – one containing the observations of its right side and the other containing the observations of its left side.

Suppose, for example, that an individual's true encounter history is 00L0B0R0. This history is not observable because the two sides of the individual were never photographed simultaneously. Hence, the individual will contribute two observed histories to the data – 00L0L000 and 0000R0R0. Working backward, the observed histories 00L0L000 and 0000R0R0 may either

come from one individual encountered on three occasions or from two separate individuals each encountered on two two or more occasions.

For a study with T capture occasions there are $5^T - 1$ possible true capture histories (we condition on capture and ignore the zero history). Of these, $(5^T - 1) - (4^T - 1) + 2(2^T - 1)$ histories can be observed. These include the $(5^T - 1) - (4^T - 1)$ that contain at least one S, which we call simultaneous histories, the $2^T - 1$ histories that include only 0 and L, left-only histories, and the $2^T - 1$ histories that include only 0 and R, right-only histories. The remaining $(4^T - 1) - 2(2^T - 1)$ contain either L and R together and/or B but no S and cannot be observed. Individuals with these true histories contribute two observed histories to the data. When a left-only and right-only history, call them $\boldsymbol{\omega}_L$ and $\boldsymbol{\omega}_R$, combine to form a third history, $\boldsymbol{\omega}_C$, we say that $\boldsymbol{\omega}_L$ and $\boldsymbol{\omega}_R$ are the left and right parents of child $\boldsymbol{\omega}_C$.

3. Methods

3.1 *Latent Multinomial Model*

To account for uncertainty in the true encounter histories caused by multiple marks, we adapt the LMM model of Link et al. (2010). Suppose that individuals in the population can have one of K possible true histories which produce a total of $L \leqslant K$ possible observable histories. The genetic misidentification model of Link et al. (2010), for example, allows for three events on each capture occasion: individuals may be captured and identified correctly (1), captured and misidentified (2), or not captured (0). This produces $K = 3^T$ possible true histories but only $L = 2^T - 1$ observable histories. Following Link et al. (2010), we define $\boldsymbol{f}$ to be the L-vector of observed counts for the observable histories and $\boldsymbol{x}$ the latent K-vector of counts for the possible true histories. The LMM is based on two assumptions about these vectors. First, it assumes that each element of $\boldsymbol{f}$ is a known linear combination of the elements of $\boldsymbol{x}$. That is, there is a known $K \times L$ matrix $\boldsymbol{A}$ such that $\boldsymbol{f} = \boldsymbol{A}'\boldsymbol{x}$. This

limits the possible values of $\boldsymbol{x}$ given the observed value of $\boldsymbol{f}$, so we refer to it as the latent vector constraint. Second, the LMM assumes that $\boldsymbol{x}$ follows a multinomial distribution

$$\boldsymbol{x} \sim \text{Multinomial}(N, \boldsymbol{\pi}(\boldsymbol{\theta}))$$

with $N = \sum_{k=1}^{K} \boldsymbol{x}_k$ representing either the population or sample size (depending on whether the model conditions on first capture) and $\boldsymbol{\pi}(\boldsymbol{\theta})$, the cell probabilities dependent on parameter $\boldsymbol{\theta}$.

The specific model of $\boldsymbol{x}$ we have fit is an extension of the Link-Barker-Jolly-Seber (LBJS) model from Link and Barker (2005) modified to allow for multiple marks. We are primarily interested in the arrival and departure times of the sharks at NMP and so we condition on individuals being captured at least one time and ignore the zero history. In this case, N is the total number of individuals captured during the study. Note that unlike standard mark-recapture experiments the true value of N cannot be observed.

The key assumptions of our model are that all emigration from NMP is permanent, that the probability of remaining at NMP from one occasion to the next does not depend on how long an individual has been present (or any other factors), that encounters are independent between individuals and over time, that there are no losses on capture, and that the conditional probabilities of the events L, R, S, and B are constant. Under these conditions, the cell probability assigned to history $\boldsymbol{\omega}$ is:

$$\begin{aligned}\pi_{\boldsymbol{\omega}}(\boldsymbol{\theta}) =& \xi(a|\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{p}) \\ & \times \rho_{\omega_a} \prod_{t=a+1}^{b} \left[\phi_{t-1}(p_t \rho_{\omega_t})^{I(\omega_t \neq 0)}(1-p_t)^{I(\omega_t = 0)}\right] \\ & \times \chi(b|\boldsymbol{\phi}, \boldsymbol{p})\end{aligned}$$

where $a = \min\{t : \omega_t > 0\}$ and $b = \max\{t : \omega_t > 0\}$ denote the occasions of the first and last captures, and $I(\cdot)$ is the indicator function. The model is parameterized in terms of:

1) **Recruitment rates**: the number of individuals that enter the population between occasions t and $t+1$ per individual present on occasion t (γ_t), $t = 1, \ldots, T-1$,

2) **Survival probabilities**: the probability that an individual present on occasion t is also present on occasion $t+1$ (ϕ_t), $t = 1, \ldots, T-1$,

3) **Capture probabilities**: the probability that an individual present on occasion t is encountered once or more (p_t), $t = 1, \ldots, T$, and

4) **Event probabilities**: the conditional probability of event E given that an individual is encountered (ρ_E), $E \in \{L, R, S, B\}$.

The derived parameter $\xi(a|\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{p})$ models the probability that an individual is first captured on occasion a given that it is captured at least one time, and $\chi(b|\boldsymbol{\phi}, \boldsymbol{p})$ models the probability that an individual released on occasion b is not recaptured. Expressions for these parameters are provided in Appendix 6. Prior distributions for the model parameters were chosen to be non-informative whenever possible and are described in Appendix 6.

3.2 *Inference*

As Link et al. (2010) explains, maximum likelihood (ML) methods are hard to implement for the LMM. Although the likelihood function can be written down easily, it is difficult to compute. The distribution of $\boldsymbol{f}$ given N and $\boldsymbol{\theta}$ is a mixture of multinomial distributions, and its density is easily formulated by summing over all possible values of $\boldsymbol{x}$ that satisfy the latent vector constraint. Explicitly:

$$L(\boldsymbol{\theta}, N|\boldsymbol{f}) = \sum_{\{\boldsymbol{x}:\boldsymbol{A}'\boldsymbol{x}=\boldsymbol{f}\}} f(\boldsymbol{x}|N, \boldsymbol{\theta}).$$

However, there may be many values of $\boldsymbol{x}$ that satisfy these constraints (even for fixed N), and there is no simple way to identify them all. This makes it difficult to compute the sum directly and to apply ML inference. Instead, Link et al. (2010) applies Bayesian inference treating $\boldsymbol{x}$ as missing data and working with the joint posterior distribution of $\boldsymbol{x}$, N, and $\boldsymbol{\theta}$

given $\boldsymbol{f}$:

$$\pi(\boldsymbol{x}, N, \boldsymbol{\theta}|\boldsymbol{f}) \propto I(\boldsymbol{f} = \boldsymbol{A}'\boldsymbol{x}) f(\boldsymbol{x}|N, \boldsymbol{\theta})\pi(N, \boldsymbol{\theta}). \quad (1)$$

Inference is then obtained by sampling from this distribution via MCMC.

The MCMC algorithm that Link et al. (2010) presents is a variant of the Metropolis-within-Gibbs algorithm which alternately updates the values of $\boldsymbol{\theta}$ and $\boldsymbol{x}$ (note that N is fully defined by $\boldsymbol{x}$ and is treated as a derived parameter in the missing data approach). Updating the value of $\boldsymbol{\theta}$ given $\boldsymbol{x}$ is equivalent to a single MCMC iteration for the parameters of the underlying mark-recapture model and can be performed with standard methods. However, it is challenging to update $\boldsymbol{x}$ given $\boldsymbol{\theta}$ in an efficient way. If proposals are generated by making simple changes to $\boldsymbol{x}$, e.g. adding or subtracting from randomly selected elements, then they are unlikely to satisfy the latent vector constraint and will almost always be rejected. To avoid this problem, Link et al. (2010) suggests an algorithm that uses vectors from the null space of $\boldsymbol{A}'$ to generate proposals for $\boldsymbol{x}$ that always satisfy the latent vector constraint. Suppose that $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_R$ form a basis of null($\boldsymbol{A}'$). Given the current values of $\boldsymbol{\theta}$, $\boldsymbol{x}$, and N, call them $\boldsymbol{\theta}^{curr}$, $\boldsymbol{x}^{curr}$ and N^{curr}, the algorithm updates $\boldsymbol{x}$ and N by repeating the following two substeps for each $r = 1, \ldots, R$:

1) Generate proposals $\boldsymbol{x}^{prop}$ and N^{prop} by:

   i) sampling c_r from the discrete uniform distribution on $-D_r, \ldots, -1, 1, D_r$,

   ii) setting $\boldsymbol{x}^{prop} = \boldsymbol{x}^{curr} + c_r\boldsymbol{b}_r$, and

   iii) defining $N^{prop} = \sum_{r=1}^{K} x_r^{prop}$.

2) Compute the Metropolis-Hastings ratio:

$$\alpha(\boldsymbol{x}^{curr}, N^{curr}; \boldsymbol{x}^{prop}, N^{prop}) = \min\left\{1, \frac{f(\boldsymbol{x}^{prop}|N^{prop}, \boldsymbol{\theta}^{curr})\pi(N^{prop}, \boldsymbol{\theta}^{curr})}{f(\boldsymbol{x}^{curr}|N^{curr}, \boldsymbol{\theta}^{curr}))\pi(N^{curr}, \boldsymbol{\theta}^{curr})}\right\}$$

   and accept the proposals with probability $\alpha(\boldsymbol{x}^{curr}, \boldsymbol{x}^{prop})$.

The key to this algorithm is that $\boldsymbol{A}'\boldsymbol{b}_r = 0$ for each $r = 1, \ldots, R$ so that $\boldsymbol{A}'\boldsymbol{x}^{prop} = \boldsymbol{A}'\boldsymbol{x}^{curr} +$

$c_r \boldsymbol{A}'\boldsymbol{b}_r = \boldsymbol{f}$. This means that $\boldsymbol{x}^{prop}$ always satisfies the latent vector constraint (provided that $\boldsymbol{x}^{curr}$ also satisfies the constraint). The values $D_r \in Z$ are tuning parameters that need to be chosen interactively or before starting the chain.

Although this algorithm solves the problem of generating valid proposals for $\boldsymbol{x}$ and N, the computational cost grows exponentially with T. The dimension of null($\boldsymbol{A}'$) in the genetic misidentification problem considered by Link et al. (2010) is $R = 3^T - (2^T - 1)$. Each update of $\boldsymbol{x}$ requires 212 substeps if $T = 5$, 58,026 substeps if $T = 10$, and 3.5×10^9 substeps if $T = 20$.

The amount of computation grows even faster for the problem of multiple marks. Our model allows for $K = 5^T - 1$ possible true histories and $L = (5^T - 1) - (4^T - 1) + 2(2^T - 1)$ observable histories; the dimension of null($\boldsymbol{A}'$) is $r = (4^T - 1) - 2(2^T - 1)$. When $T = 8$, there are 390,624 possible true histories of which 325,599 are observable. The MCMC algorithm of Link et al. (2010) would require 65,025 substeps for each update of $\boldsymbol{x}$.

To show how the algorithm can be simplified we consider a toy example. Suppose that $T = 8$ and that only the six histories shown in the top of Table 1 are observed. These include two left-only, two right-only, and two simultaneous histories. Although there are 390,624 possible true histories with $T = 8$ entries, the vast majority of these are not compatible with the observed histories. In this example, only ten true histories are compatible with the observed data. These include the six observed histories plus the four extra histories formed by combining each left-only and each right-only history, shown in the bottom of Table 1. Any other true history would have produced an observed history not seen in the data.

[Table 1 about here.]

Modeling can now be conducted using only the six histories observed and the ten compatible true histories. Redefine $\boldsymbol{f}$ to be the vector of length 6 containing counts for the observed histories and $\boldsymbol{x}$ the vector of length 10 containing counts for the compatible true histories.

The latent vector constraints between $\boldsymbol{f}$ and $\boldsymbol{x}$ are determined by pairing each parent in the observed histories with its children in the compatible true histories. Specifically, the number of times a parent is observed must equal the sum of the counts from all of its children in the compatible true histories. In the toy example, the first observed history is a parent of the 1^{st}, 7^{th}, and 9^{th} compatible true histories. The corresponding constraint is $f_1 = x_1 + x_7 + x_9$. The remaining constraints are: $f_2 = x_2 + x_8 + x_{10}$, $f_3 = x_3 + x_7 + x_8$, $f_4 = x_4 + x_9 + x_{10}$, $f_5 = x_5$, and $f_6 = x_6$. One consequence is that $\boldsymbol{x}$ has only four free elements. New values of $\boldsymbol{x}$ can be sampled by updating only $x_7, \ldots, x_{10}$ in turn and adjusting the remaining counts accordingly. Further, the values of $x_7, \ldots, x_{10}$ are bounded by the observed counts. In the example, $0 \leqslant x_7 \leqslant \min(f_1, f_3)$, $0 \leqslant x_8 \leqslant \min(f_2, f_3)$, $0 \leqslant x_9 \leqslant \min(f_1, f_4)$ and $0 \leqslant x_{10} \leqslant \min(f_2, f_4)$. These bounds can be used to define proposal distributions that are free of tuning parameters.

Generally, let $L^{'}$ denote the number of unique histories observed and $K^{'}$ the number of compatible true histories. Explicitly, $L^{'} = L^{'}_L + L^{'}_R + L^{'}_S$ and $K^{'} = L^{'} + L^{'}_L L^{'}_R$ where $L^{'}_L$, $L^{'}_R$, and $L^{'}_S$ denote the numbers of left-only, right-only, and simultaneous histories observed. To describe the algorithm we need to know the order of the counts in $\boldsymbol{f}$ and $\boldsymbol{x}$. We order $\boldsymbol{f}$ so that the $L^{'}_L$ counts of the left-only histories come first, followed by the $L^{'}_R$ counts for the right-only histories, and finally by the $L^{'}_S$ counts for the simultaneous histories. We order $\boldsymbol{x}$ in the same way with the counts for the $L^{'}_L L^{'}_R$ extra, compatible true histories added at the end. For each of the extra histories let $l(k)$ and $r(k)$ be the indices of its left and right parents. In the toy example, $l(7) = 1$ and $r(7) = 3$. The latent vector constraints are then given by the constraints on the left-only histories:

$$f_j = x_j + \sum_{\{k:l(k)=j\}} x_k, \quad j = 1, \ldots, L^{'}_L,$$

the constraints on the right-only histories:

$$f_j = x_j + \sum_{\{k:r(k)=j\}} x_k, \quad j = L_L^{'} + 1, \ldots, L_L^{'} + L_R^{'},$$

and the constraints on the simultaneous histories:

$$f_j = x_j, \quad j = L_L^{'} + L_R^{'} + 1, \ldots, L^{'}.$$

These equations show that $\boldsymbol{x}$ is completely defined by the $L_L^{'}L_R^{'}$ elements $x_{L'+1}, \ldots, x_{K'}$ and that $x_k \leqslant \min(f_{l(k)}, f_{r(k)})$ for each $k = L^{'} + 1, \ldots, K^{'}$.

Updates to $\boldsymbol{x}^{curr}$ given $\boldsymbol{\theta}^{curr}$ can then be performed with the following algorithm. For each $k = L^{'} + 1, \ldots, K^{'}$:

1) Generate proposals $\boldsymbol{x}^{prop}$ and N^{prop} by:

   i) setting $\boldsymbol{x}^{prop} = \boldsymbol{x}^{curr}$,

   ii) sampling x_k^{prop} from $\{0, \ldots, \min(f_{l(k)}, f_{r(k)})\}$,

   iii) setting $x_{l(k)}^{prop} = x_{l(k)}^{curr} - (x_k^{prop} - x_k^{curr})$ and $x_{r(k)}^{prop} = x_{r(k)}^{curr} - (x_k^{prop} - x_k^{curr})$,

   iv) and defining $N^{prop} = \sum_{k=1}^{K^{'}} x_k^{prop}$.

2) Reject the proposals immediately if $x_{l(k)}^{prop} < 0$ or $x_{r(k)}^{prop} < 0$.

3) Otherwise, compute the Metropolis-Hastings ratio:

$$\alpha(\boldsymbol{x}^{curr}, \boldsymbol{x}^{prop}) = \min\left\{1, \frac{f(\boldsymbol{x}^{prop}|N^{prop}, \boldsymbol{\theta}^{curr})\pi(N^{prop}, \boldsymbol{\theta}^{curr})}{f(\boldsymbol{x}^{curr}|N^{curr}, \boldsymbol{\theta}^{curr}))\pi(N^{curr}, \boldsymbol{\theta}^{curr})}\right\}$$

   and accept $\boldsymbol{x}^{prop}$ and N^{prop} with probability $\alpha(\boldsymbol{x}^{curr}, \boldsymbol{x}^{prop})$

The advantage of this algorithm is that it uses only $L_L^{'}L_R^{'}$ steps to update $\boldsymbol{x}$. For the toy example with 6 observed histories, $\boldsymbol{x}$ can be updated in 4 steps. For the 2008 ECOCEAN whale shark data, $L_L^{'} = 27$ and $L_R^{'} = 24$ so the new algorithm requires only 648 substeps to update $\boldsymbol{x}$. This is much smaller than the 65,025 substeps required by the algorithm of Link et al. (2010).

We have implemented the MCMC sampling algorithm for fitting the multiple MARK model directly in `R` and using the JAGS interpreter for the BUGS language (Plummer, 2003, 2011;

Team, 2012). An `R` package providing functions to format the data and to fit these models is available from the website of the first author at `www.simon.bonners.ca/MultiMark`. In application to the 2008 ECOCEAN whale shark data, we ran three parallel chains with 10,000 burn-in iterations and 50,000 sampling iterations each. Convergence was monitored with the Gelman-Rubin-Brooks (GRB) diagnostic (Brooks and Gelman, 1998) as implemented in the `R` package CODA (Plummer et al., 2006).

4. Simulation Study

To assess the performance of the model presented in the previous section we conducted simulation studies under a variety of scenarios. Here we present the results from two simulation scenarios which illustrate our main results.

In our simulations, we compared the performance of the new model (the two-sided model) with two alternatives. First, we fit models using considering only the data from the left-side photographs (the one-sided model). Capture histories were constructed by combining all events that include a left-side photograph, namely L, S, and B, ignoring all right-side photographs. The models we fit to this data were equivalent to the LBJS model with prior distributions as given in Appendix 6. Second, we fit a Bayesian method of combining inferences from the two sides under the assumption of independence as in (Wilson et al., 1999) (combined inference). To do this, we fit separate models to the data from the left- and right-side photographs and averaged the values drawn on each iteration of the separate MCMC samplers prior to computing summary statistics. For example, let $\phi_t^{(k,L)}$ and $\phi_t^{(k,R)}$ represent the values of ϕ_t drawn on the k^{th} iterations of the MCMC samplers run separately for models of the the left- and right-side data. Let $\widehat{\text{Var}^{(L)}}(\phi_t)$ and $\widehat{\text{Var}^{(R)}}(\phi_t)$ be the posterior variances estimated from all iterations. Combined inference for ϕ_t was obtained by computing

the inverse variance weighted average of $\phi_t^{(k,L)}$ and $\phi_t^{(k,R)}$

$$\phi_t^{(k)} = \frac{(\widehat{\text{Var}^{(R)}}(\phi_t)\phi_t^{(k,L)} + \widehat{\text{Var}^{(L)}}(\phi_t)\phi_t^{(k,R)})}{\widehat{\text{Var}^{(L)}}(\phi_t) + \widehat{\text{Var}^{(R)}}(\phi_t)}$$

and then computing summary statistics from the new chain $\phi_t^{(1)}, \phi_t^{(2)}, \ldots$. Credible intervals can then be computed directly from the new chain without relying on normal approximations. The mean of the values in the new chain is exactly equal to the inverse-variance weighted average of means from the separate chains.

We expected that the new model would provide better inference than the two alternatives. In particular, we expected that credible intervals from the one-sided models would be wider than the corresponding intervals from the two-sided model. We also expected that credible intervals produced by combined inference would be narrower than the intervals from the two-sided model but would not achieve the nominal coverage probability.

In the first scenario, we generated data under the assumption that all events were equally likely given capture ($\rho_L = \rho_R = \rho_B = \rho_S = .25$). We set $T = 10$ and generated data by simulating true capture histories sequentially until 200 observed capture histories were produced (each true history contributing either 0, 1, or 2 histories to the observed data). Demographic parameters were simulated from the distributions:

$$\text{logit}(\phi_t) \sim N(\text{logit}(.80), .30), \quad \text{logit}(p_t) \sim N(\text{logit}(.80), .30), \quad \log(\gamma_t) \sim N(\log(.25), .30).$$

A total of 100 data sets were simulated and analyzed. The median number of true histories simulated before 200 observed histories were obtained was 164 (min=150,max=180), the median number of unique individuals observed was 138 (min=127,max=148), and the median number of captures per individual was 2 (min=1,max=10).

Table 2 presents statistics comparing the mean-squared error (MSE) of the posterior means and the mean width and estimated coverage probability of the 95% credible intervals obtained from the alternative models. The MSE of the two-sided model and the combined-inference were similar for all parameters and smaller than those of the one-sided model by between

10% and 25%. Credible intervals for both the one-sided and two-sided models achieved the nominal coverage rate for all parameters, but the credible intervals for the one-sided model were wider by approximately 10%. In comparison, the credible intervals from the combined inference were narrower than those of the two-sided model by 20% or more but failed to achieve the nominal coverage rate.

In the second scenario, we simulated data from the same model except that both marks were seen with probability one each time an individual was captured ($\rho_S = 1$). This represents the extreme situation in which there is complete dependence between the two marks and no uncertainty in the true capture histories. In this case, the one-sided and two-sided models produce identical results. The median number of histories simulated in the 100 data sets before 200 observed histories were obtained was 215 (min=204,max=227) and the median number of captures per observed individual was 2 (min=1,max=10).

Point estimates produced by the two models in this scenario were almost exactly equal and the MSE of the two models was indistinguishable (see Table 2). However, there were clear differences in the interval estimates. While the intervals produced by combined-inference were, on average, 30% narrower, the coverage of these intervals was well below the nominal value.

[Table 2 about here.]

5. Results

The data provided in the ECOCEAN whale shark library contained a total of 96 observed encounter histories for the 2008 study period. Of these, 27 histories (28%) were constructed from left-side photographs alone, 24 (25%) were constructed from right-side photographs alone, and 45 (47%) contained at least one encounter with photographs taken from both

sides simultaneously. Along with the model presented in Section 3, we computed inferences for $\boldsymbol{p}$, $\boldsymbol{f}$, and $\boldsymbol{\phi}$ from the alternative models described in Section 4.

Table 3 provides posterior summary statistics for the LBJS parameters model obtained from the two-sided model. Inferences about all parameters are relatively imprecise because of the relatively small number of individuals captured and the low capture probabilities, but the posterior means follow the expected patterns. Point estimates for the survival probability (the probability that a whale shark remains at NMP between occasions) are at or above .90 in the first two periods, below .70 in the last two periods, and about .80 in between. The posterior mean recruitment rate is very high in week two, suggesting that most of the sharks entered during this period, and lower thereafter. This table also provides summary statistics for the population growth rate, $\lambda_k = \phi_k + f_k$, $k = 1, \ldots, K-1$, computed as a derived parameter. Although the 95% credible intervals for λ_k cover 1.00 for all k, the point estimates are greater than 1.00 for the first two periods, close to 1.00 in the next three periods, and less than .75 in the last two periods. This suggests that the aggregation of whale sharks grew during the first two periods, remained almost steady during the next three periods, and declined during the last two periods. This supports the hypothesis that whale sharks aggregate at NMP to feed after the major coral spawn which occurred between April 9 and 12 in 2008 (Chalmers, 2008, pg. 33).

[Table 3 about here.]

Table 4 provides posterior summary statistics for the conditional event probabilities. These results show that sharks were photographed from both sides simultaneously most often ($\hat{\rho_S} = .45(.36, .54)$) and that the probabilities that an individual was photographed from either the left or right side only were similar ($\hat{\rho_L} = .29(.20, .38)$ versus $\hat{\rho_R} = .21(.13, .29)$).

The posterior mean of N, the number of unique sharks encountered during the 2008 season, was 88 with 95% credible interval (82,93). The full posterior distribution of N is shown

in Figure 1 and compared with the prior distribution of N generated by simulating data sets from the prior predictive distribution conditional on there being 96 observed capture histories and at least 72 true histories (the minimum number given that 24 of 96 observed histories included right-side photographs alone). Whereas the prior distribution of N is close to uniform, the posterior distribution is strongly peaked and concentrates 95% of its mass between 82 and 93. We conclude that between 3 (3.1%) and 14 (14.6%) of the sharks encountered during the 2008 season were photographed from both the left and right sides on separate occasions without ever being matched.

[Table 4 about here.]

[Figure 1 about here.]

Comparisons of the three chains starting from different initial values provided no evidence of convergence problems. Traceplots all indicated that the three chains converged within the burn-in period, GRB diagnostic values were all less than 1.02, and the estimated MCMC error was less than 2.6% of the posterior standard deviation for each parameter. Based on these results, we are confident that the chains were long enough to produce reliable summary statistics.

The plots in Figure 2 compare inferences for the survival, recruitment, and growth rates from the four alternative models. Posterior means from the four models are all very similar and the 95% credible intervals for all parameters overlap considerably. Comparison of the widths of the 95% credible intervals from the left- and right-side data alone showed that the two-sided model provided improved inference for most, but not all, parameters. On average, the 95% credible intervals for the recruitment rates produced by the two-sided model were 93% and 69% as wide as those produced from the left- and right-side data alone. The 95% credible intervals for the survival probabilities produced by the two-sided model were 78% as wide as those from the right-side data, on average, but 103% as wide as those from the left-

side data. This last result seems to be caused by issues with the upper bound on the survival probabilities as the 95% credible intervals for the logit transformed survival probabilities produced from the two-sided model were, on average, 90% and 89% as wide as those obtained from the left- and right-side data alone. Credible intervals produced via combined inference were on average 12% smaller than those obtained from the two-sided model; however, based on the results in the previous section, we believe that these intervals would not achieve the nominal coverage rate and do not reflect the variability of the parameters correctly.

[Figure 2 about here.]

6. Conclusion

The simulation results presented in Section 4 illustrate the main advantages of our model over the previous approaches to analyzing mark-recapture data with multiple, non-invasive marks. In general, estimates from our model will be more precise than estimates based on only one mark. In contrast, the apparent gain in precision from combining estimators computed separately for each mark under the assumption of independence is artificial and credible/confidence intervals computed by these methods will not achieve the nominal coverage rate. The effect is strongest when the probability that both marks are seen simultaneously is high and the separate estimators are highly dependent.

The disadvantage of combining data from multiple marks is that the model is more complex and computations take longer. A single chain of 60,000 iterations for the 2008 whale shark data implemented in native `R` code ran in 28.6 minutes on a Linux machine with a clock speed of 2.8 GHz. In comparison, a chain of the same length for the one-sided data finished in 6.2 minutes. Our algorithm is less complex than that of Link et al. (2010), but the amount of computation is still proportional to the square of the number of observed histories and the chains may take too long to run for some large data sets. We are exploring possible solutions

including developing more efficient methods of computation and approximating the posterior distribution.

Although we have described our model for two marks, it can easily be extended for data with any number of marks. We expect that including more marks will strengthen differences between our model, the one-sided model, and combined inference seen in the simulation study. The model can also be adapted easily to estimate the size of an open population. Following Link et al. (2010), one can include the null encounter history (vector of 0s) in the set of possible true histories. Then $\boldsymbol{x}$ would have length $K^{'} = L^{'} + L_L^{'} L_R^{'} + 1$ and $N = \sum_{k=1}^{K^{'}} x_k$ would denote the total population size. Because the observed histories do not restrict the number of individuals never encountered the constraints on $\boldsymbol{x}$ would not change. The only differences are that the MCMC algorithm presented in Section 3.2 would require one more substep to update the number of individuals never encountered and that the prior bound on N must be increased to allow for the unobserved individuals.

Non-invasive marks are especially useful for mark-recapture studies that rely on public data collection because they can often be observed without special equipment or physical interaction. So called citizen science projects involving "public participation in organized research efforts" (Dickinson and Bonney, 2012, pg. 1) play an important role in ecological monitoring. Large teams of volunteer researchers can cover large geographical areas and quickly collect large data sets. As examples of successful, large scale, citizen science projects in the United States, Dickinson and Bonney (2012) highlights the US Geological Survey's North American Breeding Bird Survey (BBS), the National Audubon Society's Christmas Bird Count, and projects of The Cornell Lab of Ornithology at Cornell University. The authors estimate that "200,000 people participate in [their] suite of bird monitoring projects each year" (Dickinson and Bonney, 2012, pg. 10).

One concern with many citizen science projects is the reliability of the data. Some general

issues concerning the accuracy and analysis of data from citizen science projects are discussed by Cooper et al. (2012). Though the ECOCEAN library does rely on reports from untrained observers, it differs from similar projects in that citizens provide no more than the raw data. Most importantly, the contributors do not identify the sharks they photograph. Instead, the submit their photographs to the library and matches suggested by the paired computer algorithms are all confirmed by trained researchers (see Section 2). Hence, the data does not depend on the ability of tourists or tour operators to identify spot patterns and matches can be reconfirmed at any time. Even the reported times that a photograph was taken can be confirmed from the digital timestamp. For these reasons, we are confident that errors in the data set are minimized and that the results provided in Section 5 accurately reflect the arrival and departure of sharks from NMP in 2008.

Although we are confident in our results, some of the assumptions of our model given in Section 3 may oversimplify the population's dynamics. Sharks may move temporarily to other areas of the reef and factors like age, sex, or fitness might affect the length of time that a shark remains at NMP. The objective of this research was to develop and illustrate a general method for modeling data with multiple marks, and we intend to explore more complicated models of the ECOCEAN data in further work. Changes in survival, fecundity, and capture over time or among individuals might be accounted for with covariates or random effects, and temporary emigration might be modeled with Pollock's robust design (Pollock, 1982). We also intend to model data from multiple years in order to assess changes in the population over time.

Acknowledgments

We thank Laura Cowen for providing comments on a previous draft of the manuscript. Matt Schofield also commented on drafts and provide valuable discussions during the development of our model. Support for this work was provided in part by the NSF-Kentucky EPSCOR

Grant (NSF Grant No. 0814194). We are aware that similar methods for analyzing data with multiple marks are being developed by Brett McClintock and Paul Conn and by Rachel Fewster, and recommend that our work be cited together.

References

Arzoumanian, Z., Holmberg, J., and Norman, B. (2005). An astronomical pattern-matching algorithm for computer-aided identification of whale sharks *Rhincodon typus*. *Journal of Applied Ecology* **42,** 999–1011.

Bonner, S. J. (2013). Response to: A new method for estimating animal abundance with two sources of data in capture-recapture studies. *Methods in Ecology and Evolution* In press.

Brooks, S. P. and Gelman, A. E. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7,** 434–455.

Chalmers, A. (2008). *Temporal and spatial variability in coral condition at Sandy Bay, Ningaloo*. PhD thesis, The University of Western Australia.

Cooper, C. B., Hochachka, W. M., and Dhondt, A. A. (2012). The opportunities and challenges of citizen science as a toolf for ecological research. In Dickinson, J. L. and Bonney, R., editors, *Citizen Science: Public Participation in Environmental Research*, pages 99–113. Comstock Pub. Associates.

Da-Silva, C. Q. (2006). Asymptotics for a population size estimator of a partially uncatchable population. *Annals of the Institute of Statistical Mathematics* **59,** 603–615.

Da-Silva, C. Q., Rodrigues, J., Leite, J. G., and Milan, L. A. (2003). Bayesian estimation of the size of a closed population using photo-id data with part of the population uncatchable. *Communications in Statistics - Simulation and Computation* **32,** 677–696.

Dickinson, J. L. and Bonney, R. (2012). Why citizen science? In Dickinson, J. L. and Bonney,

R., editors, *Citizen Science: Public Participation in Environmental Research*, pages 1–14. Comstock Pub. Associates.

Holmberg, J., Norman, B., and Arzoumanian, Z. (2009). Estimating population size, structure, and residency time for whale sharks *Rhincodon typus* through collaborative photo-identification. *Endangered Species Research* **7,** 39–53.

Link, W. A. and Barker, R. J. (2005). Modeling association among demographic parameters in analysis of open population capture-recapture data. *Biometrics* **61,** 46–54.

Link, W. A., Yoshizaki, J., Bailey, L. L., and Pollock, K. H. (2010). Uncovering a latent multinomial: Analysis of mark-recapture data with misidentification. *Biometrics* **66,** 178–185.

Lukacs, P. M. and Burnham, K. P. (2005). Estimating population size from DNA-based closed capture-recapture data incorporating genotyping error. *Journal of Wildlife Management* **69,** 396–403.

Madon, B., Gimenez, O., McArdle, B., Scott Baker, C., and Garrigue, C. (2011). A new method for estimating animal abundance with two sources of data in capture-recapture studies. *Methods in Ecology and Evolution* **2,** 390–400.

McClintock, B. T., Conn, P., Alonso, R., and Crooks, K. R. (2013). Integrated modeling of bilateral photo-identification data in mark-recapture analyses. *Ecology* In press.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, pages 1–10, Vienna, Austria.

Plummer, M. (2011). *rjags: Bayesian graphical models using MCMC*. R package version 3-5.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News* **6,** 7–11.

Pollock, K. H. (1982). A capture-recapture design robust to unequal probability of capture.

The Journal of Wildlife Management **46,** 752–757.

Team, R. D. C. (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Van Tienhoven, A., Den Hartog, J., Reijns, R., and Peddemors, V. (2007). A computer-aided program for pattern-matching of natural marks on the spotted raggedtooth shark *Carcharias taurus*. *Journal of Applied Ecology* **44,** 273–280.

Wilson, B., Hammond, P. S., and Thompson, P. M. (1999). Estimating size and assessing trends in a coastal bottlenose dolphin population. *Ecological Applications* **9,** 288–300.

Wright, J. A., Barker, R. J., Schofield, M. R., Frantz, A. C., Byrom, A. E., and Gleeson, D. M. (2009). Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples. *Biometrics* **65,** 833–40.

Yoshizaki, J., Brownie, C., Pollock, K. H., and Link, W. A. (2011). Modeling misidentification errors that result from use of genetic tags in capture-recapture studies. *Environmental and Ecological Statistics* **18,** 27–55.

Yoshizaki, J., Pollock, K. H., Brownie, C., and Webster, A. (2009). Modeling misidentification errors in capture-recapture studies using photographic identification of evolving marks. *Ecology* **90,** 3–9.

Received ??. Revised ??.

Accepted ??.

Appendix

Derived Parameters

As in Link and Barker (2005):

$$\xi(a|\boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{p}) = \frac{\kappa_a}{\sum_{t=1}^{T} \kappa_t}$$

where:

$$\begin{aligned}
\kappa_1 &= p_1 \\
\kappa_2 &= (\phi_1(1-p_1)+\gamma_1)p_2 \\
\kappa_{t+1} &= p_{t+1}\left(\frac{\kappa_t(1-p_t)\phi_t}{p_t} + \gamma_t \prod_{k=1}^{t-1}(\phi_k+\gamma_k)\right), \quad t = 2,\ldots,T-1.
\end{aligned}$$

Similarly:

$$\chi(t|\boldsymbol{\phi},\boldsymbol{p}) = (1-\phi_t) + \phi_t(1-p_{t+1})\chi(t+1|\boldsymbol{\phi},\boldsymbol{p}), \quad t = 1,\ldots,T-1$$

with $\chi(T|\boldsymbol{\phi},\boldsymbol{p}) = 1$.

Prior Distributions

Parameters in the model of the true histories were assigned the following prior distributions:

$$\begin{aligned}
\text{logit}(\phi_t) &\sim N(\mu_\phi, \sigma_\phi^2), \quad t = 1,\ldots,T-1 \\
\text{logit}(p_t) &\sim N(\mu_p, \sigma_p^2), \quad t = 1,\ldots,T \\
\log(\gamma_t) &\sim N(\mu_\gamma, \sigma_\gamma^2), \quad t = 1,\ldots,T-1 \\
(\rho_L,\rho_R,\rho_S,\rho_B) &\sim \text{Dirichlet}((1,1,1,1)^T) \\
N &\sim U\{0,\ldots,U_{\max}\}
\end{aligned}$$

The value $U_{\max}$ must be bigger than the true value of N. This can be achieved by setting $U_{\max} = \sum_{l=1}^{L} f_l$ when the model conditions on first capture.

Hyperparameters were assigned the prior distributions:

$$\begin{aligned}
\mu_\phi, \mu_p &\sim N(0,2) \\
\mu_\gamma &\sim N(0,.25) \\
\sigma_\phi, \sigma_p, \sigma_\gamma &\sim HT(3,.9)
\end{aligned}$$

Here $HT(\nu,\sigma)$ represents the half t-distribution with ν degrees of freedom and scale parameter σ. All prior distributions were assumed independent.

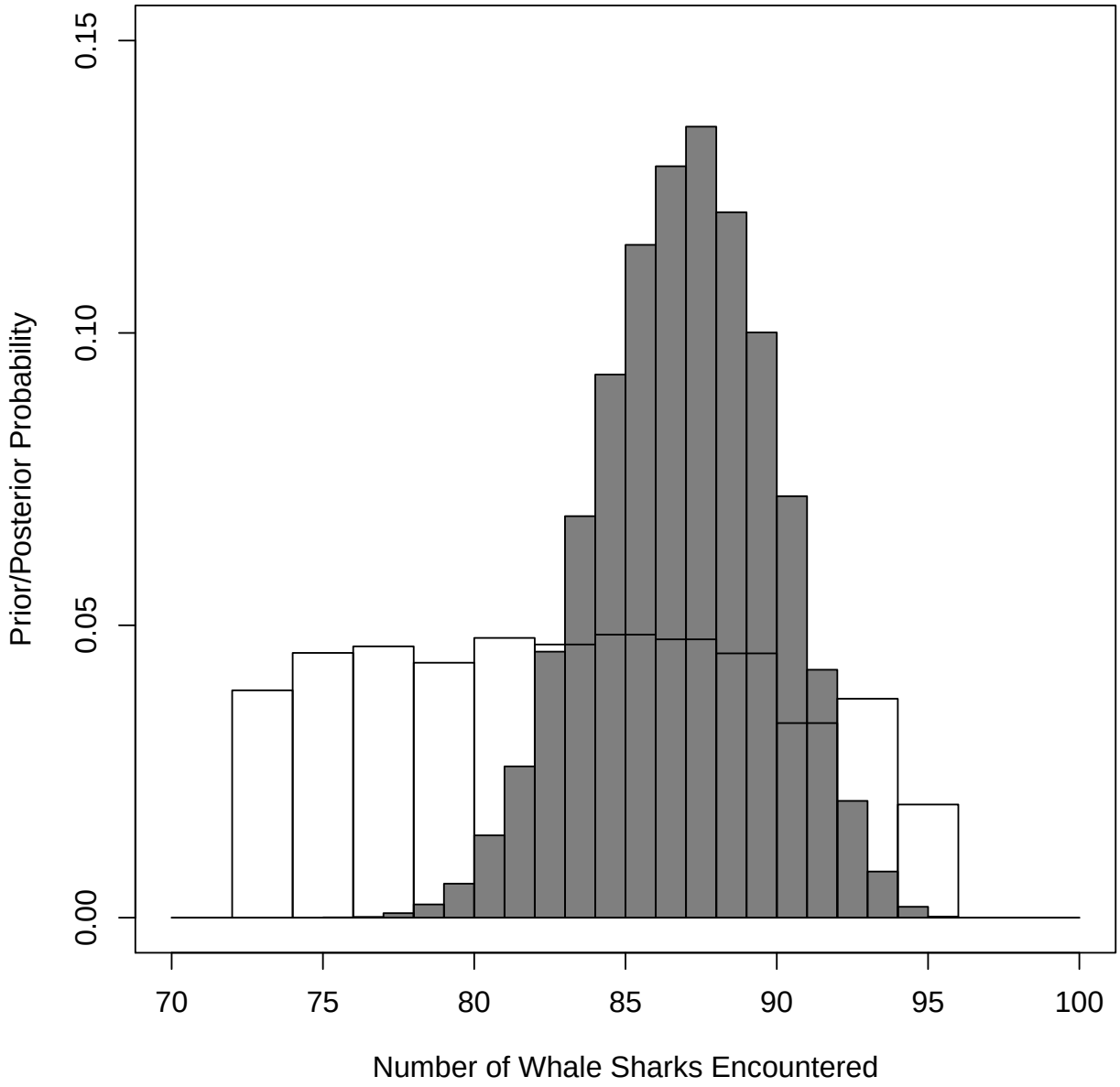

Figure 1. Comparison of the prior and posterior distribution of N. The prior distribution of N, conditional on there being 96 observed capture histories and at least 72 unique individuals, is shown by the histogram with white bars. The posterior distribution of N is shown by the histogram with grey bars.

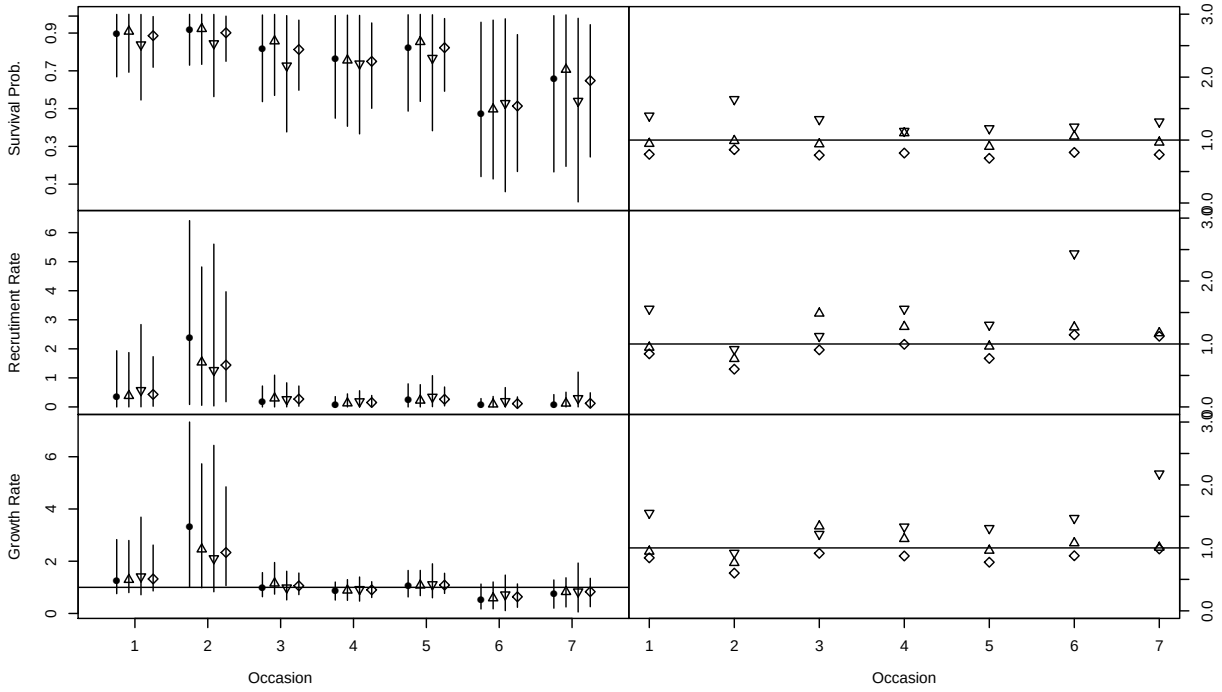

Figure 2. Comparison between the two-sided model and the three alternative models. The plots on the left-side of the figure compare the posterior means (points) and 95% credible intervals (vertical lines) of the survival probability (top), recruitment rate (middle), and population growth rate (bottom) obtained from the four models. The plots on the right side of the figure display the posterior standard deviations from the three alternative models relative to the posterior standard deviation from the two-sided model. Results from the two-sided model are represented by the circles, from the left-side photographs only by the upward pointing triangles, from the right-side photographs only by the downward pointing triangles, and from combined inference by the diamonds.

Table 1

Example of possible observed and true capture histories. Suppose that the data comprises the six observed histories given in the top of the table. The possible true histories that may have generated this data include these six plus the four additional histories in the bottom of the table.

	k	History
Observed	1	00L0L000
	2	0000L000
	3	00R00000
	4	000RR000
	5	00SBR000
	6	S0S00000
Unobserved	7	00B0L000
	8	00R0L000
	9	00LRB000
	10	000RB000

Table 2

Performance of the estimates from the two simulation scenarios. Each column of the table presents the MSE of the posterior mean relative to the MSE of the posterior mean of the one-sided model, and the median width and estimated coverage probability of the 95% credible intervals for the survival probability (ϕ), recruitment rate (f), and growth rate (λ) for one of the three models – one-sided (OS), two-sided (TS), or combined-inference (CI). The models are described in Section 4.

		Simulation 1			Simulation 2	
		OS	TS	CI	TS	CI
ϕ	MSE	1.00	.89	.87	1.00	1.00
	Width	.23	.20	.16	.17	.12
	Cover	.97	.96	.90	.95	.84
f	MSE	1.00	.88	.81	1.00	1.00
	Width	.35	.31	.24	.26	.18
	Cover	.97	.95	.90	.95	.84
λ	MSE	1.00	.88	.82	1.00	1.00
	Width	.41	.36	.29	.31	.22
	Cover	.98	.97	.95	.97	.87

Table 3

Posterior summary statistics for the demographic parameters ϕ_k, f_k, λ_k, and p_k obtained from the two-sided model. The columns of the table provide posterior means followed with equal-tailed 95% credible intervals.

Occ (k)	Survival (ϕ_k)	Recruitment (f_k)	Growth (λ_k)	Capture (p_k)
1	0.90(0.67,1.00)	0.36(0.00,1.93)	1.26(0.76,2.83)	0.23(0.08,0.43)
2	0.92(0.73,1.00)	2.40(0.08,6.41)	3.31(1.00,7.33)	0.19(0.05,0.33)
3	0.82(0.54,1.00)	0.17(0.00,0.72)	0.99(0.64,1.56)	0.26(0.15,0.43)
4	0.77(0.45,0.99)	0.09(0.00,0.36)	0.85(0.51,1.20)	0.22(0.13,0.34)
5	0.82(0.49,1.00)	0.23(0.00,0.79)	1.05(0.63,1.65)	0.22(0.12,0.36)
6	0.48(0.14,0.96)	0.06(0.00,0.29)	0.54(0.17,1.12)	0.25(0.14,0.42)
7	0.66(0.16,0.99)	0.09(0.00,0.42)	0.75(0.20,1.28)	0.20(0.06,0.37)
8	–	–	–	0.18(0.03,0.34)

Table 4
Posterior summary statistics for the conditional event probabilities.

Event (j)	Cond. Prob. (ρ_j)
1	0.29(0.20,0.38)
2	0.21(0.13,0.29)
3	0.45(0.36,0.54)
4	0.06(0.01,0.13)