bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Submitted to Statistical Science

Multivariate State Hidden Markov Models for Mark-Recapture Data

Devin S. Johnson, Jeff L. Laake, Sharon R. Melin, and Robert L. DeLong

National Marine Mammal Laboratory, Alaska Fisheries Science Center, NOAA

Abstract. State-based Cormack-Jolly-Seber (CJS) models have become an often used method for assessing states or conditions of free-ranging animals through time. Although originally envisioned to account for differences in survival and observation processes when animals are moving though various geographical strata, it has evolved to model vital rates in different life-history or diseased states. We further extend this useful class of models to the case of multivariate state data. Researchers can record values of several different states of interest; e.g., geographic location and reproductive state. Traditionally, these would be aggregated into one state with a single probability of state uncertainty. However, by modeling states as a multivariate vector, one can account for partial knowledge of the vector as well as dependence between the state variables in a parsimonious way. A hidden Markov model (HMM) formulation allows straightforward maximum likelihood inference. The proposed HMM models are demonstrated with a case study using data from a California sea lion vital rates study.

Key words and phrases: Capture-recapture, Cormack-Jolly-Seber, Hidden Markov Model, Multivariate, Partial observation, State uncertainty.

1. INTRODUCTION

The seminal papers by Cormack (1964), Jolly (1965), and Seber (1965) initiated 50 years of active development of capture-recapture theory and application. The Cormack-Jolly-Seber (CJS) model for survival estimation and the Jolly-Seber model for survival and abundance estimation and their extensions are still the most widely used capture-recapture models. They have been implemented in the computer software `MARK` (White and Burnham, 1999), `POPAN` (Arnason and Schwarz, 1999), `SURGE` (Lebreton et al., 1992), `M-SURGE` (Choquet et al., 2004), `E-SURGE` (Choquet, Rouan and Pradel, 2009), and, more recently, `marked` (Laake, Johnson and Conn, 2013) and `multimark` (McClintock, 2015).

Expansion of CJS models to account for movement between areas (multistate) was initiated by Darroch (1958) and Arnason (1973) and further developed by

National Marine Mammal Laboratory, Alaska Fisheries Science Center, NOAA National Marine Fisheries Service, Seattle, WA U.S.A. (e-mail: devin.johnson@noaa.gov)

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Hilborn (1990), Hestbeck, Nichols and Malecki (1991), Schwarz, Schweigert and Arnason (1993) and Brownie et al. (1993). Nichols et al. (1992) used the multistate model to describe survival with state transitions based on mass and this has been followed by numerous other examples using the multistate model (Lebreton and Pradel, 2002; Nichols and Kendall, 1995; White, Kendall and Barker, 2006). An important advance to the multistate CJS generalization was made by Kendall et al. (2004) who considered models with state uncertainty (unknown state) and Pradel (2005) who cast a multistate model with state uncertainty as a hidden Markov model (Zucchini and MacDonald, 2009). The state uncertainty models are often termed multievent models. In a further step, Kendall et al. (2012) combined a robust design with a multistate hidden Markov model to improve precision in the face of state uncertainty. We use the general term *state-based* to refer to all of these types of CJS models.

For traditional CJS analysis, the detection portion of the model is often considered to be a nuisance with little scientific interest and often determined largely by the resampling methods used by researchers to resight individuals. The achievement made by the development of the original CJS model was that it permitted estimation of survival even when individuals were not observed on every occasion. The state-based extensions allowed survival inference to account for time-varying heterogeneity or deviations from CJS model assumptions induced by states as well make inference to the state process itself (Lebreton and Pradel, 2002). With the recently developed extensions, one can often find state-based CJS analyses where survival is also considered a nuisance process and primary scientific interest lies with the transitions between states. See Gourlay-Larour et al. (2014) for a recent example. With this in mind, practicing ecologists are using the state-based CJS framework and the associated HMM formulation to analyze complex state transitions such as the combination of geographic location and reproductive or disease status. This presents the problem that a state may be only partially observed when an individual is resighted. For example, location will be known but reproductive status may not be observed. To further generalize the state-based CJS models, King and McCrea (2014) developed a closed form likelihood expression to handle partial knowledge of the state of an individual. However, a general framework for parametrizing complex state-based CJS models is not examined. Laake et al. (2014) made initial strides into multivariate state CJS models by proposing a bivariate state model to account for dependence in double-mark loss. Here we seek to augment the work of King and McCrea (2014) and Laake et al. (2014) to provide a general method to construct complex state-based CJS models.

In this paper, we propose a general modeling framework for multivariate state-based CJS models in which the state is defined by one or more discrete categorical variables and each variable maybe unknown when the animal is resighted. The modeling framework extends the state-based models in MARK by allowing any or all of the state variables to be uncertain rather than just a single uncertain state. We have implemented the model in the `marked` package (Laake, Johnson and Conn, 2013) for the R statistical environment (R Development Core Team, 2015) using a hidden Markov model for maximum likelihood estimation of the parameters which are cast in a log-linear framework (Cormack, 1989; Darroch and Ratcliff, 1980) to easily allow dependence among state variables. We provide an example using 18 years of annual resighting data collected on a single cohort

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of California sea lion (*Zalophus californianus*) pups that were both branded and double tagged on their flippers in 1996 at San Miguel Island off the west coast near Santa Barbara, California (Melin et al., 2011). The animal's state was defined based on the animal's location (San Miguel or Año Nuevo Islands) which were always known when the animal was resighted and the, possibly unknown, states of the two flipper tags (present or absent due to tag loss).

2. METHODS

To begin the description of multivariate state modeling of mark-recapture data, we must begin with the state and data structure. The general structure is based on the capture histories of individuals, indexed $i = 1, \ldots, I$, over capture occasions, indexed $j = 1, \ldots, J$. The state of an individual on a particular occasion is determined by the cross-classification of that individual to a cell in a multiway table where the margins represent state variables of interest. Specifically, the state, $\mathbf{s} = (s_1, \ldots, s_K)$ is a vector where the entries correspond to K classification criteria, e.g., location, reproductive status, or disease presence. Each state variable has a set of possible values $s_k \in \mathcal{S}_k = \{1, \ldots, m_k\}$, $k = 1, \ldots, K$. The space of the state vector, $\mathbf{s}$, is thus $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_K$. The number of possible states is denoted $M = |\mathcal{S}| = \prod_k m_k$. We also need to augment the state-space, $\mathcal{S}$, with a "death" state, such that the individual can transition between all $\mathbf{s} \in \mathcal{S}$ as well as death. We represent the death state with $\mathbf{s} = \mathbf{0}$, so the augmented state space is $\mathcal{S}^\dagger = \mathcal{S} \bigcup \{\mathbf{0}\}$. The observed data for an individual on a particular (re)capture occasion is denoted as $\mathbf{c} = (c_1, \ldots, c_K)$. The observed state can differ from the true state in two ways. First, an individual may not be observed (detected) on a particular capture occasion. Second, it is possible, and highly probable in many situations, that even if an individual is physically observed, state specification may not be completely determined by the researcher. Thus, the space of the observation vector is augmented with two additional levels corresponding to "not observed" and "unknown,", i.e., $c_k \in \{0, 1, \ldots, m_k, m_k + 1\}$, where 0 is for undetected individuals and $m_k + 1$ is associated with an unknown level for that state variable. In heuristic descriptions we will use $c_k = u$ to be clear we are referring to the unknown state. In addition, we take the convention that if $c_k = 0$ for any k, it is assumed that all $c_k = 0$. We use $\mathbf{c}_{ij}$ to denote the observed state of individual i on occasion j. Table A.1 of Supplement A (Johnson et al., 2016a) provides a glossary of the notation used in this paper.

2.1 Two state variable model

Specifying a model for a general multivariate discrete state is notationally cumbersome due to the complex dependence structure. Therefore, we begin with a description for a bivariate state model, and then follow up in the next section with a general model formulation for > 2 state variables with additional covariates. The general parameter types are the same as in any state-based CJS model: survival (ϕ), detection (p), and state transition (ψ) from one occasion to the next. In addition, we have parameters associated with the inability to fully assess the state of an individual (δ) and the probability that an individual is in a particular state at first capture (π). Typically, each of these parameter types depends on the state of the individual; e.g., survival of individual i from occasion j to occasion $j + 1$ depends on the state of the individual on occasion

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

TABLE 1

Example transition probabilities, $\psi(\mathbf{s};\mathbf{r})$, for a hypothetical bivariate state $\mathbf{s}$ composed of state variables $s_1 \in \{1,2,3\}$ and $s_2 \in \{1,2\}$. Here we use the reference cell $\mathbf{s}^* = (1,1)$. The ψ probabilities are presented in their unnormalized form as in equation (2.1). The necessary normalizing constant is the sum over all the table entries.

		s_2	
		1	2
s_1	1	1	$\exp\{\beta^{(2)}(2;\mathbf{r})\}$
	2	$\exp\{\beta^{(1)}(2;\mathbf{r})\}$	$\exp\{\beta^{(1)}(2;\mathbf{r}) + \beta^{(2)}(2;\mathbf{r}) + \beta^{(1,2)}(2,2;\mathbf{r})\}$
	3	$\exp\{\beta^{(1)}(3;\mathbf{r})\}$	$\exp\{\beta^{(1)}(3;\mathbf{r}) + \beta^{(2)}(2;\mathbf{r}) + \beta^{(1,2)}(3,2;\mathbf{r})\}$

j. Each multivariate state variable may contribute independently to survival, or may interact at various levels to further enhance or degrade survival. Thus, models for these parameters must accommodate this interaction (dependence) when modeling state transition. We take a log-linear approach (Christensen, 1997) for modeling multiway contingency tables which allows for all levels of dependence between state variables.

We begin with modeling state transitions from one occasion to the next. Suppose individual i is alive on occasion $j < J$, then the probability of a transition from state $\mathbf{r} = (r_1, r_2) \in \mathcal{S}$ on occasion j to state $\mathbf{s} = (s_1, s_2) \in \mathcal{S}$ on occasion $j+1$ is given by the conditional log-linear model,

$$\psi(\mathbf{s};\mathbf{r}) \propto \exp\left\{\beta^{(1)}(s_1;\mathbf{r}) + \beta^{(2)}(s_2;\mathbf{r}) + \beta^{(1,2)}(s_1,s_2,\mathbf{r})\right\}, \tag{2.1}$$

where, for each $\mathbf{r} \in \mathcal{S}$, $\beta^{(1)}(s_1;\mathbf{r})$, $\beta^{(2)}(s_2;\mathbf{r})$, and $\beta^{(1,2)}(s_1,s_2;\mathbf{r})$ are parameters which depend on the previous state, $\mathbf{r}$, and are indexed by the values of s_1, s_2, and the double index of (s_1, s_2), respectively. For example, the collection of $\beta^{(1)}$ parameters is $\{\beta^{(1)}(1;\mathbf{r}), \ldots, \beta^{(1)}(m_1;\mathbf{r})\}$. If we denote $\mathbf{\Psi}$ to be the transition matrix describing the probability of transitioning from any state on occasion j to any other state on occasion $j+1$, equation (2.1) evaluated for all $\mathbf{s} \in \mathcal{S}$ would form the rows associated with each possible current state $\mathbf{r}$. There are $m_1 + m_2 + m_1 m_2$ parameters for each $\mathbf{r}$. Thus, there are $m_1 m_2 \times (m_1 + m_2 + m_1 m_2)$ β parameters within the $m_1 m_2 \times m_1 m_2$ transition matrix $\mathbf{\Psi}$. Hence, the model is overparameterized. To alleviate this problem it is customary to pick a reference cell say, $\mathbf{s}^* \in \mathcal{S}$, and fix all β parameters to 0 when $s_1 = s_1^*$ *or* $s_2 = s_2^*$. For example, if $\mathbf{s}^* = (r_1, r_2)$ (the current state), $\beta^{(1)}(r_1;\mathbf{r}) \equiv 0$, $\beta^{(2)}(r_2;\mathbf{r}) \equiv 0$, $\beta^{(1,2)}(s_1, r_2;\mathbf{r}) \equiv 0$ for any s_1 value, and $\beta^{(1,2)}(r_1, s_2;\mathbf{r}) \equiv 0$ for any s_2 value. Of course it is not necessary to use the current state as the reference state, but it has some advantages, namely, the β parameters are interpreted as controlling movement away from the current state and the current state is usually an acceptable possibility for a realization at the next occasion (i.e., non-zero probability of occurring). This may not be the case if you define a fixed state as the reference (see Section 3 for example). See Table (1) for an example of $\psi(\mathbf{s};\mathbf{r})$ specification with the reference cell constraint.

Markovian dependence in time is explicitly included by allowing the β parameters to depend on the previous state. In addition, dependence between state variables is accommodated via the $\beta^{(1,2)}$ interaction parameters. If $\beta^{(1,2)}(s_1, s_2;\mathbf{r}) = 0$ for all values of s_1 and s_2 then this is a necessary and sufficient condition for s_1 and s_2 to be independent given the previous state $\mathbf{r}$. Sufficiency is easily observed

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

by the fact that under that assumption

$$\begin{aligned}\psi(\mathbf{s};\mathbf{r}) &\propto \exp\left\{\beta^{(1)}(s_1;\mathbf{r}) + \beta^{(2)}(s_2;\mathbf{r})\right\} \\ &\propto \psi^{(1)}(s_1;\mathbf{r}) \times \psi^{(2)}(s_2;\mathbf{r}).\end{aligned} \tag{2.2}$$

Thus, the probability of transition to state $\mathbf{s}$ is the product of independent transitions from $r_1 \to s_1$ and $r_2 \to s_2$. Proof of necessity is quite cumbersome and beyond the scope of this paper. See Lauritzen (1996) for more theoretical details of log-linear modeling. One caveat of these independence results is that they assume there are no structural zeros (impossible state transitions). If there are, one should check that independence still holds if it is a desirable property. Otherwise, zero valued interaction terms can just be used for parsimonious modeling.

The next set of parameters of scientific interest is survival. Given that individual i is in state $\mathbf{s}$ on occasion j the probability of survival from occasion j to $j+1$ is described by

$$\phi(\mathbf{s}) = \ell^{-1}\{\eta_{\emptyset} + \eta^{(1)}(s_1) + \eta^{(2)}(s_2) + \eta^{(1,2)}(s_1, s_2)\}, \tag{2.3}$$

where ℓ is a link function such as logit or probit, and the η parameters are functions of (indexed) by the current state, $\mathbf{s}$. The η parameters have the same structure and reference cell constraints as the β parameters in the ψ model. Readers should also observe that there is a $\eta_{\emptyset}$ vector. This term is not indexed by the state and represents baseline survival. Even though the notation may appear different for some readers, it should also be noted that this ϕ model is identical to a model with s_1 and s_2 being thought of as known categorical covariates with interaction effects included. So, as far as the survival model is concerned, s_1 and s_2 are equivalent to (individual $\times$ occasion) indexed factor covariates.

We now move to the observation and detection portion of the model. Given that individual i is in state $\mathbf{s}$ on occasion j, the probability that the individual is detected (i.e., recaptured or resighted) by the researcher is given by

$$p(\mathbf{s}) = \ell^{-1}\{\zeta_{\emptyset} + \zeta^{(1)}(s_1) + \zeta^{(2)}(s_2) + \zeta^{(1,2)}(s_1, s_2)\}, \tag{2.4}$$

as with the survival model, the state variables can be considered as factor variables for the purposes of conceptualizing the model. There is one more set of parameters working in concert with p, these are related to the ability of the researcher to jointly classify a detected individual according to all of the state variables. If an individual i is detected on occasion j, then the joint probability of the state observation, $\mathbf{c} = (c_1, c_2)'$, conditioned on the true state is modeled via,

$$\delta(\mathbf{c};\mathbf{s}) \propto \exp\{\alpha^{(1)}(c_1,\mathbf{s}) + \alpha^{(2)}(c_2,\mathbf{s}) + \alpha^{(1,2)}(c_1,c_2,\mathbf{s})\}, \mathbf{c} \neq \mathbf{0}. \tag{2.5}$$

Note that we do not consider misidentification of states; a state variable is either observed correctly ($c_k = s_k$) or it is not observed and the value is unknown ($c_k = m_k + 1$ or u). As with the multivariate transition model, ψ, a reference cell, say $\mathbf{c}^*$ is necessary for parameter identifiability. One could use $\mathbf{c}^* = (m_1 + 1, m_2 + 1)$, the completely unknown state, to be consistent with previous univariate state uncertainty models (e.g., Laake 2013; Kendall et al. 2012), but it is not necessary and maybe undesirable if the completely unknown state is

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

not observable; e.g., double-tag studies where no third permanent mark is available. We suggest using the current state, $\mathbf{c}^* = \mathbf{s}$, as the reference. As with the ψ models, one can incorporate dependence in recording the level of each state variable. That dependence implies, for example, that if the researcher is able to record the level for one state variable, the researcher is more (less) likely to record the level of the other. There may be some state variables which are always recorded with certainty, e.g., location. So, for any $\mathbf{c}$ that contains $c_k = m_k + 1$ for those state variables, we can also fix the $\delta \equiv 0$ and remove the corresponding α parameters from the overall parameter set. As with the state transition probabilities, ψ, if these structural zeros exist, then the dependence interpretation needs to be investigated because conditional dependence can be induced by the zeros.

Finally, the last set of parameters is associated with the initial capture of the individual. We are considering only CJS type models here, thus, we are still conditioning the model on the occasion when an individual is first observed. When an individual is first captured (marked), however, the researcher may not be able to observe all state variables, i.e., some $c_k = u$. Therefore, we need to model the probability that individual i, who is first marked on occasion j, is in state $\mathbf{s}$. So, we describe this probability via,

$$\pi(\mathbf{s}) \propto \exp\left\{\lambda^{(1)}(s_1) + \lambda^{(2)}(s_2) + \lambda^{(1,2)}(s_1, s_2)\right\}. \tag{2.6}$$

Here, there is no previous cell from which to form the reference cell. Therefore the researcher will have to decide on the most appropriate state to serve as the reference. The π model parameters, however, may not necessarily have to be estimated. For example, if it is certain that all state variables will be observed at first marking then it can be fixed to $\pi(\mathbf{s}) = 1$, where $\mathbf{s}$ is the state individual i was in when it was first marked, and $\pi(\mathbf{r}) = 0$ for all $\mathbf{r} \neq \mathbf{s}$ (see Section 3). Or, a researcher might specify, say, $\pi(\mathbf{s}) = 1/M$, where M is the number of cells in $\mathcal{S}$. The π probabilities can be interpreted as a prior distribution, in the Bayesian sense, on the state of a newly marked individual, thus, $1/M$ would be a uniform prior over the state-space.

2.2 General multivariate state model

In this section we generalize the bivariate state models presented in the previous section. Here, researchers can consider state specifications that index a cell in a general K dimensional hypercube and models can depend on covariates, as well. To accomplish this we introduce the notation $\mathcal{V} = \{1, \ldots, K\}$ to be the set which indexes the collection of K state variables and $v \subseteq \mathcal{V}$ is a subset. The partial state $\mathbf{s}^{(v)}$ represents only those elements of $\mathbf{s}$ whose index is contained in $v \subseteq \mathcal{V}$. The state subspace $\mathcal{S}^{(v)}$ denotes the set of all possible values of $\mathbf{s}^{(v)}$.

Now, the full specified model describing the transition of individual i in state $\mathbf{r} = (r_1, \ldots, r_K)$ on occasion j to state $\mathbf{s} = (s_1, \ldots, s_K)$ on occasion $j+1$ is given by,

$$\psi_{ij}(\mathbf{s}; \mathbf{r}) \propto \exp\left\{\sum_{v \subseteq \mathcal{V}} \mathbf{x}_{ij}^{(v)} \cdot \boldsymbol{\beta}_{ij}^{(v)}(\mathbf{s}; \mathbf{r})\right\}, \tag{2.7}$$

where $\mathbf{x}_{ij}^{(v)}$ is a vector of covariates, the notation "$\mathbf{x} \cdot \boldsymbol{\beta}$" represents the dot product $\mathbf{x}'\boldsymbol{\beta}$, and $\boldsymbol{\beta}_{ij}^{(v)}(\mathbf{s}; \mathbf{r})$ is a coefficient vector that depends on the state, $\mathbf{s} \in \mathcal{S}$, only

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

through the value of the sub-state $\mathbf{s}^{(v)} \in \mathcal{S}^{(v)}$ and $\mathbf{r} \in \mathcal{S}$. As with the bivariate model in the previous section, we must satisfy the reference cell constraint to have an identifiable set of parameters. Therefore, all $\boldsymbol{\beta}^{(v)}(\mathbf{s};\mathbf{r}) \equiv 0$ if any element of the sub-state $\mathbf{s}^{(v)}$ is equal to the corresponding elements of the reference cell, $\mathbf{s}^*$. Independence of state variables is more complex than described in the previous section due to a number of possible interactions between different sets of variables. To provide a generalization, let v_1, v_2, and v_3 be subsets of state variables that partitions $\mathcal{V}$. Then $\mathbf{s}^{(v_1)}$ and $\mathbf{s}^{(v_3)}$ are conditionally independent of one another given $\mathbf{s}^{(v_2)}$ (and $\mathbf{r}$) if and only if $\boldsymbol{\beta}^{(v)}(\mathbf{s};\mathbf{r}) = 0$ for all $\mathbf{s}$, where v contains elements of both v_1 and v_3 (i.e., $v \cap v_1 \neq \emptyset$ and $v \cap v_3 \neq \emptyset$) (Frydenberg, 1990). For example, suppose $\mathcal{V} = \{1, 2, 3\}$, then s_1 is conditionally independent of s_3 given s_2 (and $\mathbf{r}$) if $\beta^{(1,3)}(\mathbf{s};\mathbf{r}) = \beta^{(1,2,3)}(\mathbf{s};\mathbf{r}) = 0$ for all $\mathbf{s}$. Further, if we assume that $\beta^{(1,2)}(\mathbf{s};\mathbf{r}) = 0$, then $\mathbf{s}^{(1)} = s_1$ is independent (given $\mathbf{r}$) of $\mathbf{s}^{(2,3)} = (s_2, s_3)$. In that case, we can follow the example in (2.2) and write

$$\psi_{ij}(\mathbf{s};\mathbf{r}) = \psi_{ij}^{(1)}(\mathbf{s};\mathbf{r}^{(1)})\psi_{ij}^{(2,3)}(\mathbf{s};\mathbf{r}^{(2,3)}). \tag{2.8}$$

This independence factorization can extend to more than two groups using the Gibbs factorization theorem (see Frydenberg 1990). Again, independence theorems are based on the absence of structural zeros.

We can progress through the other parameter groups in a similar fashion to produce a general multivariate state model. The survival portion of the model can be generalized to

$$\phi_{ij}(\mathbf{s}) = \ell^{-1}\left\{\sum_{v \subseteq \mathcal{V}} \mathbf{x}_{ij}^{(v)} \cdot \boldsymbol{\eta}_{ij}^{(v)}(\mathbf{s})\right\}, \tag{2.9}$$

where, the covariate vector $\mathbf{x}_{ij}^{(v)}$ need not be the same as that shown in equation (2.7), we use the same notation simply to avoid extra clutter. Cycling through the remaining parameters, we have:

$$p_{ij}(\mathbf{s}) = \ell^{-1}\left\{\sum_{v \subseteq \mathcal{V}} \mathbf{x}_{ij}^{(v)} \cdot \boldsymbol{\zeta}_{ij}^{(v)}(\mathbf{s})\right\}, \mathbf{s} \neq \mathbf{0}, \tag{2.10}$$

$$\delta_{ij}(\mathbf{c};\mathbf{s}) \propto \exp\left\{\sum_{v \subseteq \mathcal{V}} \mathbf{x}_{ij}^{(v)} \cdot \boldsymbol{\alpha}_{ij}^{(v)}(\mathbf{c};\mathbf{s})\right\}, \mathbf{c} \neq \mathbf{0} \text{ and } \mathbf{s} \neq \mathbf{0}, \tag{2.11}$$

where, we suggest the reference cell $\mathbf{c}^* = \mathbf{s}$. In general, there will be 2^d possible values of $\mathbf{c}$ associated with each $\mathbf{s}$, where d is the number of state variables for which a u can be recorded. Finally, the probability distribution of states upon first capture is

$$\pi_{ij}(\mathbf{s}) \propto \exp\left\{\sum_{v \subseteq \mathcal{V}} \mathbf{x}_{ij}^{(v)} \cdot \boldsymbol{\lambda}_{ij}^{(v)}(\mathbf{s})\right\}. \tag{2.12}$$

This completes the general model for a K dimensional state vector. In the next section we focus on methods for efficient parameter estimation and inference.

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Similar to the ψ models, the δ and π parameters can be factored into independent products as well. This will facilitate easier overall model construction if certain states variables can be assumed to be independent, making it easier to parameterize dependence over a small subset of variables.

2.3 Statistical inference

Statistical inference for multistage mark-recapture models can be challenging (McCrea and Morgan, 2014). We take the approach described by Laake (2013) who uses a hidden Markov model (HMM) formulation. By framing a state-based CJS model as a specific HMM the efficient *forward algorithm* can be used to maximize the likelihood (Zucchini and MacDonald, 2009). Laake (2013) provides a detailed description of HMM formulated mark-recapture models and likelihood calculation, but we briefly revisit it here to place the forward algorithm in the context of multivariate state CJS models.

To begin the description of the forward algorithm in the multivariate state CJS model case, we must define a few matrices. The transition probability matrix for movement within $\mathcal{S}^\dagger$ is given by,

$$\mathbf{\Gamma}_{ij} = \begin{bmatrix} \phi_{ij}(\mathbf{r}_1)\psi_{ij}(\mathbf{s}_1;\mathbf{r}_1) & \cdots & \phi_{ij}(\mathbf{r}_1)\psi_{ij}(\mathbf{s}_M;\mathbf{r}_1) & 1-\phi_{ij}(\mathbf{r}_1) \\ \vdots & & \vdots & \vdots \\ \phi_{ij}(\mathbf{r}_M)\psi_{ij}(\mathbf{s}_1;\mathbf{r}_M) & \cdots & \phi_{ij}(\mathbf{r}_M)\psi_{ij}(\mathbf{s}_M;\mathbf{r}_M) & 1-\phi_{ij}(\mathbf{r}_M) \\ 0 & 0 & \cdots & 1 \end{bmatrix}, \tag{2.13}$$

where M is the size of the state-space, $\mathbf{r}_m$, $m = 1, \ldots, M$, refers to the current state on occasion j, $\mathbf{s}_m$ refers to the state transitioning to on occasion $j+1$, and the last column (row) is associated with transition to (from) the death state. In addition to the state-space augmentation, we define the observation probability matrix which describes the probability of observing state $\mathbf{c}_n$, $n = 0, \ldots, N$, for animal i on occasion j (rows) given it was actually in state $\mathbf{s} \in \mathcal{S}^\dagger$ (columns),

$$\mathbf{D}_{ij} = \begin{bmatrix} p_{ij}(\mathbf{s}_1)\delta_{ij}(\mathbf{c}_1;\mathbf{s}_1) & \cdots & p_{ij}(\mathbf{s}_M)\delta_{ij}(\mathbf{c}_1;\mathbf{s}_M) & 0 \\ \vdots & \vdots & & \vdots \\ p_{ij}(\mathbf{s}_1)\delta_{ij}(\mathbf{c}_N;\mathbf{s}_1) & \cdots & p_{ij}(\mathbf{s}_M)\delta_{ij}(\mathbf{c}_N;\mathbf{s}_M) & 0 \\ 1-p_{ij}(\mathbf{s}_1) & \cdots & 1-p_{ij}(\mathbf{s}_M) & 1 \end{bmatrix}, \tag{2.14}$$

where, observed state $\mathbf{c}_0 = (0, \ldots, 0)$, the undetected state, is represented by the last row. For use in the forward algorithm, we also define the $K \times K$ matrix, $\mathbf{P}(\mathbf{c}_{ij})$ to be a diagonal matrix with the row of $\mathbf{D}_{ij}$ associated with observed state for individual i on occasion j as the diagonal entries.

Using the $\mathbf{\Gamma}_{ij}$ and $\mathbf{P}(c_{ij})$ matrices we can describe the forward algorithm specific to maximum likelihood estimation in the multivariate state MR model. The algorithm proceeds as follows for individual $i = 1, \ldots, I$.

1. Initial conditions:
   - Set $w_{if_i} = \boldsymbol{\pi}'_{if_i}\mathbf{P}(\mathbf{c}_{if_i})\mathbf{1}$ where $f_i = 1, \ldots, J-1$, is the occasion on which individual i was first captured/marked, $\boldsymbol{\pi}_{if_i} = (\pi_{if_i}(\mathbf{s}_1), \ldots, \pi_{if_i}(\mathbf{s}_M))'$, and $\mathbf{1}$ is a vector of all 1s.
   - Set $\boldsymbol{\mu}'_{if_i} = \boldsymbol{\pi}_{if_i}\mathbf{P}(\mathbf{c}_{if_i})/w_{if_i}$
   - Set $L_i = \log w_{if_i}$

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2. For $j = f_i + 1, \ldots, J$:
   - Set $w_{ij} = \boldsymbol{\mu}'_{i,j-1}\boldsymbol{\Gamma}_{ij}\mathbf{P}(\mathbf{c}_{ij})\mathbf{1}$
   - Set $\boldsymbol{\mu}'_{ij} = \boldsymbol{\mu}_{i,j-1}\boldsymbol{\Gamma}_{ij}\mathbf{P}(\mathbf{c}_{ij})/w_{ij}$
   - Set $L_i = L_i + \log w_{ij}$

The forward probabilities, $\boldsymbol{\mu}_{ij}$ represent the conditional probability distribution of the state of individual i on occasion j, given the observed states up to and including occasion j and the parameters. After the algorithm has been run for every individual, the log-likelihood is given by,

$$L = \sum_{i=1}^{I} L_i. \tag{2.15}$$

The log-likelihood can be maximized to provide parameter estimates and biological inference. In the next section we will provide a real-world example where multivariate states are collected in mark-resight study of California sea lions.

3. EXAMPLE: MOVEMENT, SURVIVAL, AND MARK LOSS IN CALIFORNIA SEA LIONS

3.1 Data and Model Description

As an example of a multivariate multistate model we use 18 years of annual resighting data collected on a single cohort of California sea lion (*Zalophus californianus*) pups. In 1996, a total of 485 4-5 month old pups born on San Miguel Island (SMI), off the west coast near Santa Barbara, California (Melin et al., 2011) were herded into a pen over a total period of five days. The sex and weight of each pup was recorded and a unique permanent number was applied as hot brand to their left side. In addition, a uniquely numbered yellow roto-tag was applied to each of their foreflippers. Pups were resighted by their brand during a three month period from 15 May to 15 August each year at SMI and Año Nuevo Island (ANI). SMI is a breeding rookery, whereas ANI is a haul out where animals rest, and only a few pups are born and raised. When the animal was resighted the presence (+) or absence (−) of the tag was noted. In many cases, however, the status was not recorded for one or both of the tags. Although, the flipper tags were not directly relevant to the resighting effort, their presence allowed a rare opportunity to assess loss and dependence for other pinniped studies that rely solely on flipper tags for marking (e.g., Testa et al. 2013).

From the sea lion data, a capture history was constructed with annual occasions and each entry was composed of 3 characters, with the first for the area, the second for the left tag and the third for the right tag. There were 8 different possible states for a live animal ($a++$, $a-+$, $a+-$, $a--$, $s++$, $s-+$, $s+-$, and $s--$). Here we are using symbolic representation of the states for ease of discussion. In terms of the notation presented in the previous sections, s_1 is the location state ($s_1 = 1$ for "a" and 2 for "s"), s_2 and s_3 are the left and right tag status variables ($s_k = 1$ for "+" and 2 for "−"; $k = 2, 3$).

To begin the model specification used to analyze these data we will start with the observation portions of the model (p and δ). The possible capture history values were 0 if not seen, the 8 possible states if seen and each tag status was

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

recorded, and an additional 10 observations with unknown tag status ($a+u$, $au+$, $a-u$, $au-$, auu, $s+u$, $su+$, $s-u$, $su-$, suu). For these data, location is always known with certainty when the animal is detected but each tag status could be unknown. For the δ model we used the fully known tag status as the reference cell which implies that the parameters can be interpreted as controlling whether tag status was not obtained (i.e., unknown). The following formulation was used,

$$\begin{aligned}\delta_{ij}(\mathbf{c};\mathbf{s}) &= \delta^{(2,3)}(\mathbf{c}^{(2,3)},\mathbf{s});\ \text{for } c_1 = s_1. \\ &\propto \exp\{\alpha^{(2)}(\mathbf{c}) + \alpha^{(3)}(\mathbf{c}) + \alpha^{(2,3)}(\mathbf{c})\}.\end{aligned} \tag{3.1}$$

Table A.2 in Supplement A (Johnson et al., 2016a) illustrates the observation model parameters. For detection, we used the occasion $\times$ area model

$$\text{logit } p_{ij}(\mathbf{s}) = \zeta_0 + \zeta_j + \zeta^{(1)}(\mathbf{s}) + \zeta_j^{(1)}(\mathbf{s}), \tag{3.2}$$

where ζ_j is an occasion effect, $\zeta^{(1)}(\mathbf{s})$ is an area effect, and $\zeta_j^{(1)}(\mathbf{s})$ is an occasion $\times$ area interaction. Resighting of animals was based on the permanent brand mark, thus, tag status was not included in the model.

Now, we describe the more scientifically relevant, process portions of the model. First, all animals started on SMI with two tags (i.e., $\mathbf{s}_{i1} = (s++)$) on the first release occasion. Thus, for this example $\pi_{i1}(\mathbf{s}) = 1$ for $\mathbf{s} = (s++)$. The survival model was formulated with sex, age, and area specific effects,

$$\text{logit } \phi_{ij}(\mathbf{s}) = \mathbf{x}_{ij} \cdot \boldsymbol{\eta}^{(\emptyset)} + B_{sex(i),j} + \eta^{(1)}(\mathbf{s}), \tag{3.3}$$

where $\mathbf{x}_{ij} = (1, w_{ij})$, where w_{ij} is the mass anomaly (difference from sex specific mean) if $j = 1$, zero elsewise (i.e., pup survival is influenced by mass), $B_{sex(i),j}$ is a sex specific b-spline smooth (df = 3; Hastie 1992) over age, and $\eta^{(1)}(\mathbf{s})$ is an area effect. For the state transitions we used the previous state as the reference state for the following occasion, i.e., $\mathbf{s}^* = \mathbf{r}$ for transitions from $\mathbf{r} \to \mathbf{s}$. Thus, the parameters control movement away from the current area (s_1) and loss of flipper tags (s_2, s_3). The state transitions were modeled by

$$\psi_{ij}(\mathbf{r},\mathbf{s}) \propto \exp\{B^{(1)}_{sex(i),j}(\mathbf{s};\mathbf{r}) + \mathbf{x}_j \cdot \boldsymbol{\beta}^{(2)}(\mathbf{s};\mathbf{r}) + \mathbf{x}_j \cdot \boldsymbol{\beta}^{(3)}(\mathbf{s};\mathbf{r}) + \beta^{(2,3)}(\mathbf{s};\mathbf{r})\}, \tag{3.4}$$

where $B^{(1)}_{sex(i),j}(\mathbf{s};\mathbf{r})$ is a sex and area specific b-spline (df=3) over age, $\mathbf{x}_j = (1, age_j)$, and $\beta^{(2,3)}(\mathbf{s};\mathbf{r})$ is an interaction term between the two tags which models dependence in loss. Although, not immediately obvious, the b-spline effect, $B^{(1)}_{sex(i),j}(\mathbf{s};\mathbf{r})$, can be written in the form $\mathbf{b}_j \cdot \boldsymbol{\beta}^{(1)}_{sex(i)}(\mathbf{s};\mathbf{r})$, where $\mathbf{b}_j$ is a vector of the b-spline basis function values for occasion j. Therefore, it fits into the proposed log-linear framework. Also, because the same type of tag was placed on each flipper, we constrained the marginal loss rates to be identical for each side. This is accomplished by setting the tag loss parameters $\boldsymbol{\beta}^{(2)}(-;+) = \boldsymbol{\beta}^{(3)}(-;+)$. The probabilities for impossible transitions (i.e., going from "$-$" to "$+$" for either flipper tag variable (s_2, s_3) were fixed to zero. The tag loss portion of the state transition generally follows the tag loss model proposed by Laake et al. (2014). Note, that the structural zeros only depend on the $\mathbf{r}^{(2,3)} \to \mathbf{s}^{(2,3)}$ transitions and there is no interaction term for $\mathbf{s}^{(1)}$ and $\mathbf{s}^{(2,3)}$, therefore, we can write $\psi_{ij}(\mathbf{s};\mathbf{r}) = \psi^{(1)}_{ij}(\mathbf{s};\mathbf{r})\psi^{(2,3)}_{ij}(\mathbf{s};\mathbf{r})$. Table A.3 in Supplement A (Johnson et

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

al., 2016a) illustrates the transition probabilities for each current state, $\mathbf{r}$, to each possible state, $\mathbf{s}$.

The full model was fitted with maximum likelihood methods using the forward HMM algorithm in Section 2.3. The ability to fit multivariate state models was added to the R package `marked` (Laake, Johnson and Conn, 2013). R code used for this example is available in Supplement B (Johnson et al., 2016b). Confidence intervals for derived quantities were approximated using a multivariate normal parametric bootstrap with mean equal to the maximum likelihood estimate and covariance matrix equal to the negative Hessian of the log-likelihood function (see Supplement B; Johnson et al. 2016b).

3.2 Results

The 1996 cohort of sea lion pups on SMI experienced a very strong El Niño in 1997 which influenced both the movement of young animals to ANI and created a large amount of mortality before the sea lions reached age 2. With a single cohort and a large amount of early mortality, the sample sizes were small at older age classes and the confidence intervals became very wide and that should be considered in evaluating any patterns across age. This unusual event may have also influenced differences in survival across sex and age. However, as an example, it does illustrate the usefulness and flexibility of building multivariate state models.

We begin describing the results with the observation model, δ. This is a somewhat unusual example with regard to observation of each state's status. Observers use the brand for resighting and the tag status was often not observed. Observers were more likely to record status of left tag because the brand is on the sea lion's left side (Figure 1). The odds of missing second side when the first is missed increases 5-fold [95% CI: 3.5−8.2] (this CI notation used hereafter). This again is due to use of brand for resighting. If the observer didn't record the status of the left side they were unlikely to get the status of the right side. The fitted detection model, p, is illustrated in Figure A.1 of Supplement A (Johnson et al., 2016a). It is relatively standard with respect to multistate (and CJS) modeling, so, we do not elaborate further here.

Male pups had a slightly better overall pup survival (Figure 2). Although, this may be due to the fact that male pups are heavier on average. For both sexes, pup survival increased with increasing mass which has also been observed in northern fur seals (Baker and Fowler, 1992; Baker, Fowler and Antonelis, 1994), Steller sea lions (Hastings, Gelatt and King, 2009), and Hawaiian monk seals (Baker, Fowler and Antonelis, 1994; Craig and Ragen, 1999). For non-pups (age > 0) survival increased to adult age (about 5 years old) and then declined in older ages (Figure 3). The decline was much more dramatic for males whose cost of reproduction is high due to their polygynous breeding system in which breeding males have to defend a territory (Peterson and Bartholomew, 1967; Johnson, 1968). For both sexes, higher survival at ANI was likely due to two factors. First, in El Niño years prey resources are distributed farther north, closer to ANI, thus animals not constrained by reproductive commitments at SMI had better access to food. Second, animals at ANI during the breeding season were likely non-reproductive (at least for a given year) and, thus, did not have to invest in costly activities such as holding territories (males) or lactation (females), which allowed them to

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

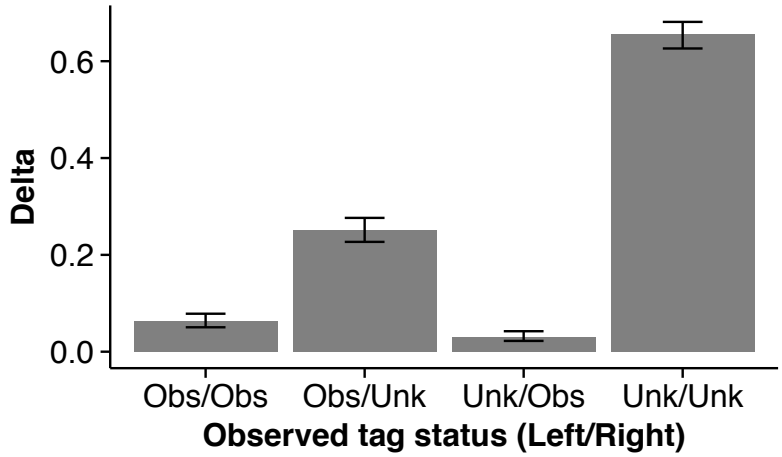


FIG 1. *Estimated δ parameters for the California sea lion model. If the tag status was observed it appears as "Obs" in the plot. Thus, for example, the column "Obs/Unk" indicates the probability that the left tag was observed and the right tag was not observed and its status is unknown.*

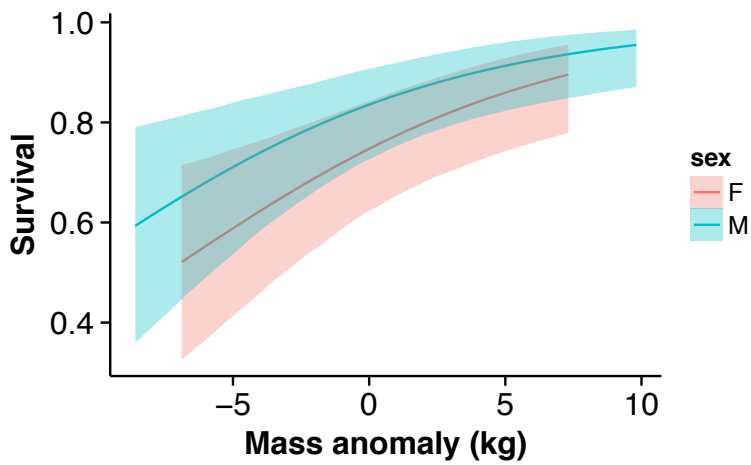


FIG 2. *Mass- and sex-specific survival probability for California sea lion pups on San Miguel Island, California. Mass anomaly is the deviation from the sex-specific mean mass.*

preserve or invest in better body condition.

With respect to movement between SMI and ANI, we could look at the fitted age and sex-specific annual area transition probabilities (Figure A.2; Supplement A, Johnson et al. 2016b), however, a more biologically interesting, derived quantity is the sex specific proportion of animals that were be located in each area for a given age, $\boldsymbol{\pi}_{ij}^{(1)}$. This is because animals at SMI during the breeding season were likely reproductive. This derived quantity can be calculated by

$$\widehat{\boldsymbol{\pi}}_{ij}^{(1)} = \boldsymbol{\pi}_{i1}^{(1)} \cdot \prod_{j=1}^{J-1} \widehat{\boldsymbol{\Psi}}_{ij}^{(1)}, \tag{3.5}$$

where, for the sea lion data $\boldsymbol{\pi}_{i1} = (1, 0)'$ because all pups began at SMI on occasion 1, and $\widehat{\boldsymbol{\Psi}}_{ij}^{(1)}$ is a 2×2 matrix with $\widehat{\psi}_{ij}^{(1)}(\mathbf{s}; \mathbf{r})$ entries. Figure 4 illustrates

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

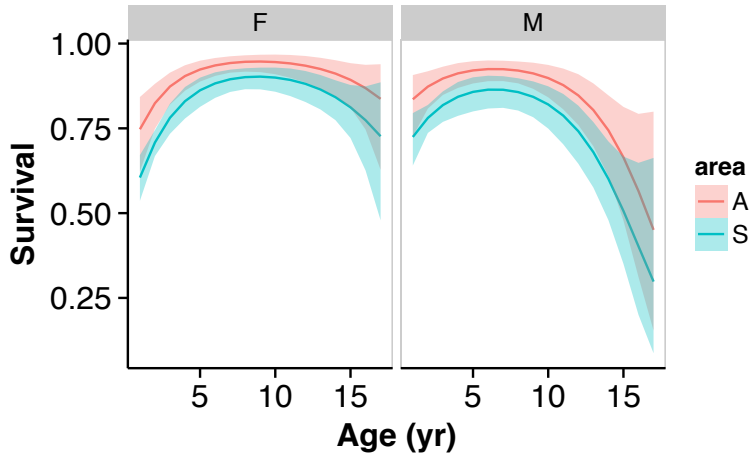


FIG 3. *Sex- and age-specific annual survival probabilities for California sea lion nonpups (age > 0). Separate curves are plotted for each area A = Año Nuevo I. and S = San Miguel I. Although, plotted as a solid curve, an animals transitioned between the curves depending on the area occupied.*

the occupancy rate of SMI by both sexes. In general, females began returning to SMI at around age 3 and occupancy stabilized to ≈ 0.7 by age 5. Thus, about 70% of females from the 1996 cohort were reproductively active each year based on location alone. This is similar to the 0.77 average natality estimated by Melin et al. (2011) for several cohorts. A large proportion of males stayed at ANI till the age of 10 or 11 when the began to return and to hold territories. Because a large proportion of males were located at ANI during the breeding season, they experienced better annual survival (for the reasons discussed previously) and hence, had higher survivorship than females (Figure 4). By age 12−13, however, the lower survival for breeding males on SMI resulted in equal survivorship.

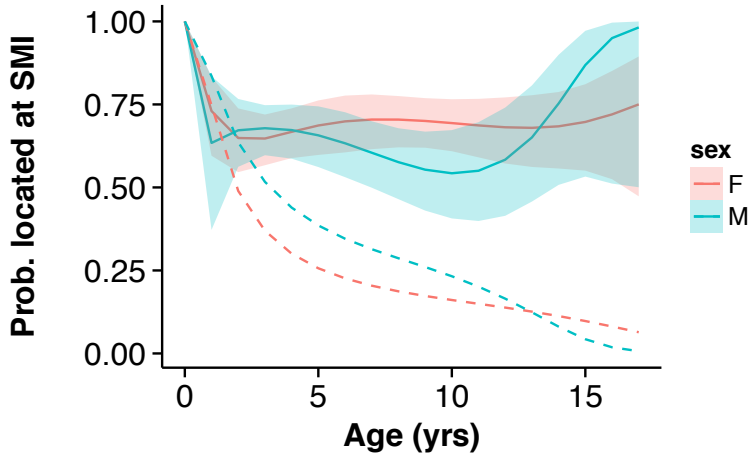


FIG 4. *Age- and sex-specific probability that an individual was located at San Miguel I. during the breeding season. Solid lines and envelopes represent probability that an animal was located at SMI and 95% confidence interval. The dashed lines represent the expected proportion of animals that were alive at a given age; i.e., survivorship.*

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Finally, although it did not affect the estimation of survival in this case due to the permanent brand, the flipper tag results were quite revealing for pinniped studies where a third mark is not possible. First, tag loss rate increased with age. Marginally, the odds of loosing a specific tag (left or right) increased 1.56-fold [1.38−1.75] every year. There was also substantial dependence between tags. The odds of loosing a tag increased 406-fold [76−2,204] when the other had been lost. Therefore, tags were almost always lost simultaneously. The probability of loosing both tags at or before the animal was 5 years old was 0.49 [0.38−0.61], by age 9, the probability increased to 0.91 [0.86−0.95] (Figure 5; CDF). The most probable age for an animal to become $(-,-)$ tag status was 5-6, when approximately 40% of animals entered double loss status (Figure 5; PDF). This was probably due to the fact that females and males are in periods of strong growth and development at these ages as they prepare for reproductive activities.

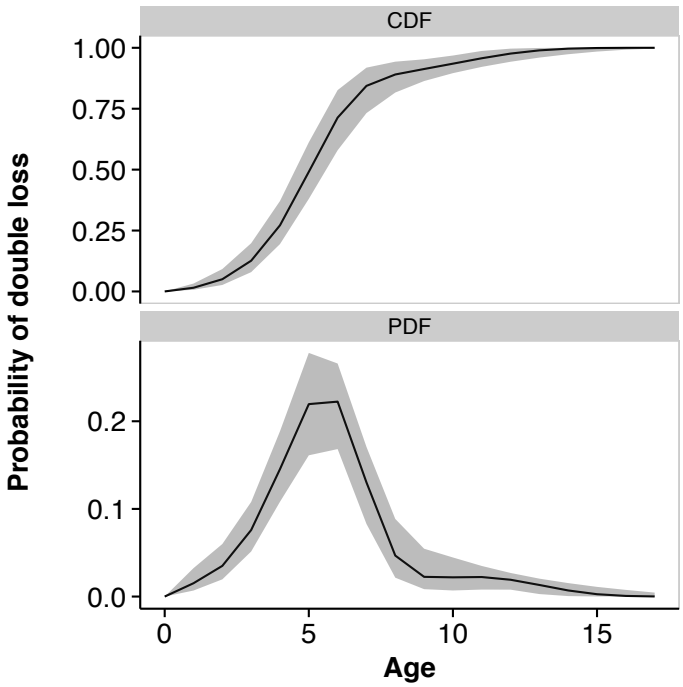


FIG 5. *Age-specific probabilities of double tag loss. The upper plot (CDF) is the cumulative probability distribution of entering the double tag loss state at or before each age. The lower plot (PDF) is the difference of the upper plot and provides the probability that an animal entered the double tag loss state at each age.*

4. DISCUSSION

In the on-going development of state-based CJS modeling we present a generalization of the multievent CJS models to the case where states are multivariate vectors of categorical variables. This allows the user to analyze complex state transitions and handle partial knowledge of states upon observation of the individual. Using a log-linear model parameterization for multiway contingency table data allows parsimonious modeling of multivariate state transitions by enforcing desired independence (or dependence) between state variables by exclusion

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(inclusion) of specified interaction terms. In addition, because of the necessary and sufficient conditions for determining independence from log-linear models, inference on dependence of state variables can also be tested by examining for significant deviation from zero for the appropriate interaction terms. We use the term "testing" in a loose sense, in that one might eliminate that effect through a model selection procedure as well (Burnham and Anderson, 2002; Hooten and Hobbs, 2015).

The parsimonious construction and hypothesis testing properties were both demonstrated in the example analysis of the sea lion data. We assumed that tag loss was primarily due to growth of the foreflippers from the time the tags were placed on the animals as pups. Therefore, it is reasonable to conclude that loss of tags will be independent of the location of the animal (i.e., tag loss variables s_2 and s_3 are independent of the area s_1). Therefore, we specifically omitted any interaction terms that would imply dependence of area on tag loss. Second, we examined the parameter measuring interaction between the loss of each tag, $\beta^{(2,3)}(\mathbf{s};\mathbf{r})$, for significant deviation from zero and found that there is a high degree of dependence in the loss of flipper tags in California sea lions. This conclusion has strong implications for future studies that do not use a permanent third mark such as a brand, as survival estimation may be negatively biased, depending on the assumed tag loss process (Laake et al., 2014). By using the multivariate state framework we were able to directly extend the double tag loss model of Laake et al. (2014) to account for movement between different areas as well. Another practical extension the the multivariate model provides over the single multievent framework is that the probability of state uncertainty can depend on the true state. For example, although we did not examine it here, it is possible that the probability of an observer missing the status of a flipper tag depends on whether it is present or not. Or, if there are different observers at each area, one set may ignore the tag status, while the other dutifully records it. Thus, probability of correct observation of one state variable may depend on another state variable.

Although the proposed multivariate framework is quite general, there are further extensions that are straightforward to develop given the HMM formulation. First, our model was based on the assumption that the state variable is either observed correctly or it is not observed at all and hence, unknown. There is no reason that this cannot be relaxed. All that needs to be done is to augment the observation space for each variable for which errors can be made. For example, $\mathcal{C}_k = \{0, 1, \ldots, m, u, e_1, \ldots\}$, where $e_1, \ldots$ refers to states that are not biological in origin, but observational errors. For example, in the tag loss example, an observer might see a tag is physically present, but fail to read it. In this case it is uncertain whether or not the tag should be considered lost (i.e., unreadable = lost), but there is more information than a simple "unknown." In addition, it is also possible that an observer might declare a tag to be lost when it fact it is present or *vice versa*. All that needs to be done in this case is to remove the constraint that $\delta_{ij}(\mathbf{c};\mathbf{s}) = 0$ when any $c_k \neq s_k$ or u. Both of these instances are directly handled by the HMM inference framework. Some of these extensions might produce a model of such complexity that some parameters become unidentifiable. Therefore, placement the multivariate model in a robust design framework similar to Kendall et al. (2012) might improve estimation of parameters. This can be accomplished in the multivariate setting (and in `marked` package in general)

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

by constraining parameters to be constant (most likely 0 or 1) within secondary sampling periods, but free between primary periods.

We would, again, like to again acknowledge the seminal work by the Cormack (1964), Jolly (1965), and Seber (1965) papers which initiated 50 years of development and analysis of capture-recapture data. Through the years, the same basic framework has been generalized to add additional inference capability as well as overcome assumptions of the original CJS models, but the same basic framework exists in each of these model, including the extension proposed herein.

ACKNOWLEDGMENTS

The findings and conclusions in the paper are those of the authors and do not necessarily represent the views of the National Marine Fisheries Service, NOAA. Reference to trade names does not imply endorsement by the National Marine Fisheries Service, NOAA. The authors would like to thank Profs. Byron Morgan and Steven Buckland for organizing this special issue on the anniversary of the influential Cormack, Jolly and Seber papers.

SUPPLEMENTARY MATERIAL

Supplement A: Notation summary and additional sea lion analysis details
(doi: 10.1214/00-AOASXXXXSUPP; supplement_A.pdf). Contains additional tables that summarize notation used throughout the paper and provide additional details and results for the analysis of sea lion data in Section 3.

Supplement B: R code used to analyze sea lion data
(doi: 10.1214/00-AOASXXXXSUPP; supplement_B.pdf). Contains the R code used to run the sea lion example analysis in Section 3. The most up-to-date version of the `marked` package can be installed directly using the R package `devtools` with the command:

```
devtools::install_github("jlaake/marked/marked")
```

.

However, in order to install the this version users need to ensure that their machines are equipped to compile R packages with source code (namely Fortran and C++). CRAN releases (version $\geq$ 1.1.10) of `marked` contain the multivariate state fitting capability and the sea lion data. Using the R console command:

```
install.packages("marked")
```

a precompiled version can be installed from CRAN.

REFERENCES

Arnason, A. N. (1973). The estimation of population size, migration rates and survival in a stratified population. *Researches on Population Ecology* **15** 1-8.

Arnason, A. N. and Schwarz, C. J. (1999). Using POPAN-5 to analyse banding data. *Bird Study* **46** 157–168.

Baker, J. and Fowler, C. (1992). Pup weight and survival of northern fur seals *Callorhinus ursinus*. *Journal of Zoology* **227** 231–238.

Baker, J. D., Fowler, C. W. and Antonelis, G. A. (1994). Body weight and growth of juvenile male northern fur seals, *Callorhinus ursinus*. *Marine Mammal Science* **10** 151–162.

Brownie, C., Hines, J. E., Nichols, J. D., Pollock, K. H. and Hestbeck, J. B. (1993). Capture-Recapture Studies for Multiple Strata Including Non-Markovian Transitions. *Biometrics* **49** 1173–1187.

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

BURNHAM, K. P. and ANDERSON, D. R. (2002). *Model Selection and Multimodel Inference: a Practical Information-theoretic Approach*. Springer-Verlag Inc.

CHOQUET, R., ROUAN, L. and PRADEL, R. (2009). Program E-Surge: A Software Application for Fitting Multievent Models. In *Modeling Demographic Processes In Marked Populations*, (D. Thomson, E. Cooch and M. Conroy, eds.). *Environmental and Ecological Statistics* **3** 845-865. Springer US.

CHOQUET, R., REBOULET, A. M., PRADEL, R., GIMENEZ, O. and LEBRETON, J. D. (2004). M-SURGE: new software specifically designed for multistate capture-recapture models. *Animal Biodiversity and Conservation* **27** 207–215.

CHRISTENSEN, R. (1997). *Log-Linear Models and Logistic regression*. Springer, New York.

CORMACK, R. M. (1964). Estimates of survival from the sighting of marked animals. *Biometrika* **51** pp. 429-438.

CORMACK, R. M. (1989). Log-linear models for capture-recapture. *Biometrics* **45** 395–413.

CRAIG, M. P. and RAGEN, T. J. (1999). Body size, survival, and decline of juvenile Hawaiian monk seals, *Monachus schauinslandi*. *Marine Mammal Science* **15** 786–809.

DARROCH, J. N. (1958). The multiple-recapture census: I. Estimation of a closed population. *Biometrika* 343–359.

DARROCH, J. and RATCLIFF, D. (1980). A note on capture-recapture estimation. *Biometrics* **36** 149–153.

FRYDENBERG, M. (1990). The chain graph Markov property. *Scandinavian Journal of Statistics* **17** 333-353.

GOURLAY-LAROUR, M.-L., PRADEL, R., GUILLEMAIN, M., GUITTON, J.-S., L'HOSTIS, M., SANTIN-JANIN, H. and CAIZERGUES, A. (2014). Movement patterns in a partial migrant: a multi-event capture-recapture approach. *PLoS ONE* **9** e96478.

HASTIE, T. J. (1992). Generalized additive models. In *Statistical Models in S* (J. M. Chambers and T. J. Hastie, eds.) Chapter 7 Wadsworth & Brooks.

HASTINGS, K. K., GELATT, T. S. and KING, J. C. (2009). Postbranding survival of Steller sea lion pups at Lowrie Island in Southeast Alaska. *The Journal of Wildlife Management* **73** 1040–1051.

HESTBECK, J. B., NICHOLS, J. D. and MALECKI, R. A. (1991). Estimates of movement and site fidelity using mark-resight data of wintering Canada geese. *Ecology* **72** 523–533.

HILBORN, R. (1990). Determination of fish movement patterns from tag recoveries using maximum likelihood estimators. *Canadian Journal of Fisheries and Aquatic Sciences* **47** 635-643.

HOOTEN, M. B. and HOBBS, N. T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs* **85** 3–28.

JOHNSON, A. M. (1968). Annual mortality of territorial male fur seals and its management significance. *The Journal of Wildlife Management* **32** 94–99.

JOHNSON, D. S., LAAKE, J. L., MELIN, S. R. and DELONG, R. L. (2016a). Supplement A: Notation summary and additional sea lion analysis details. *Supplement to "Multivariate State Hidden Markov Models for Mark-Recapture Data"*. DOI: 10.1214/00-AOASXXXXSUPP.

JOHNSON, D. S., LAAKE, J. L., MELIN, S. R. and DELONG, R. L. (2016b). Supplement B: R code to analyze sea lion data. *Supplement to "Multivariate State Hidden Markov Models for Mark-Recapture Data"*. DOI: 10.1214/00-AOASXXXXSUPP.

JOLLY, G. M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika* **52** pp. 225-247.

KENDALL, W. L., LANGTIMM, C. A., BECK, C. A. and RUNGE, M. C. (2004). Capture-recapture analysis for estimating manatee reproductive rates. *Marine Mammal Science* **20** 424–437.

KENDALL, W. L., WHITE, G. C., HINES, J. E., LANGTIMM, C. A. and YOSHIZAKI, J. (2012). Estimating parameters of hidden Markov models based on marked individuals: use of robust design data. *Ecology* **93** 913–920.

KING, R. and MCCREA, R. (2014). A generalised likelihood framework for partially observed capture–recapture–recovery models. *Statistical Methodology* **17** 30–45.

LAAKE, J. L. (2013). Capture-Recapture Analysis with Hidden Markov Models Technical Report No. AFSC Processed Rep. 2013-04, Alaska Fish. Sci. Cent., NOAA Natl. Mar. Fish Serv., 7600 Sand Point Way NE, Seattle, WA 98115.

LAAKE, J. L., JOHNSON, D. S. and CONN, P. B. (2013). marked: an R package for maximum likelihood and Markov Chain Monte Carlo analysis of capture–recapture data. *Methods In Ecology and Evolution* **4** 885–890.

LAAKE, J. L., JOHNSON, D. S., DIEFENBACH, D. R. and TERNENT, M. A. (2014). Hidden

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Markov model for dependent mark loss and survival estimation. *Journal of Agricultural, Biological, and Environmental Statistics* **19** 1–17.

LAURITZEN, S. L. (1996). *Graphical Models*. Claredon Press.

LEBRETON, J. D. and PRADEL, R. (2002). Multistate recapture models: modelling incomplete individual histories. *Journal of Applied Statistics* **29** 353–369.

LEBRETON, J. D., BURNHAM, K. P., CLOBERT, J. and ANDERSON, D. R. (1992). Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monographs* **62** 67–118.

MCCLINTOCK, B. T. (2015). multimark: an R package for analysis of capture-recapture data consisting of multiple "non-invasive" marks. *Ecology and Evolution* **In press**.

MCCREA, R. S. and MORGAN, B. J. (2014). *Analysis of capture-recapture data*. CRC Press.

MELIN, S. R., LAAKE, J. L., DELONG, R. L. and SINIFF, D. B. (2011). Age-specific recruitment and natality of California sea lions at San Miguel Island, California. *Marine Mammal Science* **28** 751-776.

NICHOLS, J. D. and KENDALL, W. L. (1995). The use of multi-state capture-recapture models to address questions in evolutionary ecology. *Journal of Applied Statistics* **22** 835-846.

NICHOLS, J. D., SAUER, J. R., POLLOCK, K. H. and HESTBECK, J. B. (1992). Estimating transition-probabilities for stage-based population projection matrices using capture recapture data. *Ecology* **73** 306–312.

PETERSON, R. and BARTHOLOMEW, G. (1967). The natural history and behavior of the California sea lion. Special publication No. 1, The American Society of Mammalogists.

PRADEL, R. (2005). Multievent: An extension of multistate capture-recapture models to uncertain states. *Biometrics* **61** 442–447.

SCHWARZ, C. J., SCHWEIGERT, J. F. and ARNASON, A. N. (1993). Estimating migration rates using tag-recovery data. *Biometrics* **49** 177–193.

SEBER, G. A. F. (1965). A Note on the Multiple-Recapture Census. *Biometrika* **52** pp. 249-259.

R DEVELOPMENT CORE TEAM (2015). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

TESTA, J. W., THOMASON, J. R., REAM, R. R. and GELATT, T. S. (2013). Demographic Studies Of Northern Fur Seals On The Pribilof Islands, Alaska, 2007-2012. In *Fur Seal Investigations, 2012* (J. W. Testa, ed.) U.S. Dept. of Commer., NOAA Tech. Memo. NMFS-AFSC-257, 90p.

WHITE, G. C. and BURNHAM, K. P. (1999). Program MARK: survival estimation from populations of marked animals. *Bird Study* **46** 120–139.

WHITE, G. C., KENDALL, W. L. and BARKER, R. J. (2006). Multistate survival models and their extensions in Program MARK. *Journal of Wildlife Management* **70** 1521–1529.

ZUCCHINI, W. and MACDONALD, I. L. (2009). *Hidden Markov models for time series: an introduction using R*. CRC Press.

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplement A: Notation summary and additional example results and details

Supplement to "Multivariate state hidden Markov models for mark-recapture data"

Devin S. Johnson, Jeff L. Laake, Sharon R. Melin, and Robert L. DeLong

National Marine Mammal Laboratory,
Alaska Fisheries Science Center,
National Marine Fisheries Service, NOAA,
Seattle, Washington, U.S.A.

Email: *devin.johnson@noaa.gov*

Acknowledgments

The findings and conclusions in the paper are those of the authors and do not necessarily represent the views of the National Marine Fisheries Service, NOAA. Reference to trade names does not imply endorsement by the National Marine Fisheries Service, NOAA.

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

TABLE A.1

Notation glossery. Here we present a list of notation used throughout the paper. The individual index i runs from 1 to n, the total number of marked individuals and j runs from 1 to T, the total number of capture/resighting occasions.

I	Number of individuals marked
J	Total number of capture/resight occasions
$\mathcal{V}$	Set of state variable index, i.e, $\mathcal{V} = \{1, \ldots, K\}$
K	Number of state variables, $K = \lvert\mathcal{V}\rvert$
$\mathcal{S}$	Space of possible multivariate state values
$\mathbf{s}$	Multivariate state in the $\mathbf{s} = (s_1, \ldots, s_K) \in \mathcal{S}$
s_k	Value of the kth state variable, $k = 1, \ldots, K$
m_k	The number of different values for the kth state variable, i.e., $s_k \in \{1, \ldots, m_k\}$
M	Total size of $\mathcal{S}$, i.e., $M = \lvert\mathcal{S}\rvert = \prod_{k=1}^{K} m_k$
$\mathcal{S}^\dagger$	State space augmented with the death state $\mathcal{S}^\dagger = \mathcal{S} \bigcup \{\mathbf{0}\}$
$\psi_{ij}(\mathbf{s};\mathbf{r})$	Probability of individual i transitioning from multivariate state $\mathbf{r} \in \mathcal{S}$ on occasion j to state $\mathbf{s} \in \mathcal{S}$ on occasion $j+1$
$\mathbf{\Psi}_{ij}$	Transition probability matrix for movement within $\mathcal{S}$, entries are given by $\psi_{ij}(\mathbf{s};\mathbf{r})$
$\mathbf{\Gamma}_{ij}$	Transition probability matrix for movement within $\mathcal{S}^\dagger$
$\phi_{ij}(\mathbf{s})$	Probability of individual i surviving from occasion j to occasion $j+1$ given they where in state $\mathbf{s}$ on occasion j
$p_{ij}(\mathbf{s})$	Probability individual i is detected on occasion j given it is in state $\mathbf{s}$ on the jth occasion
$\mathbf{c}$	Observed state vector (capture history), $\mathbf{c} = (c_1, \ldots, c_K)$, where $c_k \in \{0, 1, \ldots, m_k, m_k+1\}$, 0 represents unobserved, and m_k+1 represents unknown
$\mathcal{C}$	Space of possible observations
N	Total number of possible observations, i.e., $N = \lvert\mathcal{C}\rvert = 1 + \prod_{k=1}^{K}(m_k+1)$.
$\delta_{ij}(\mathbf{c};\mathbf{s})$	Conditional probability of observing $\mathbf{c}$ for individual i on occasion j given the true state of the animal is $\mathbf{s}$ and the animal is detected.
$\pi_{ij}(\mathbf{s})$	Probability individual i is in state $\mathbf{s}$ when it is first captured/marked on the jth occasion

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

TABLE A.2

Conditional probabilities, $\delta(\mathbf{c};\mathbf{s})$, for the observations, $\mathbf{c}$ given the true state, $\mathbf{s}$, of detected sea lion in the example analysis of Section 3. Because the location, s_1, is known with certainty when the animal is detected, we can write $\delta(\mathbf{c};\mathbf{s}) = \delta(\mathbf{c}^{(2,3)};\mathbf{s}^{(2,3)})$. Here we use the true state and the reference cell. Structural zeros occur for observations where the state is mis-observed rather that just unobserved. The δ probabilities are presented in their unnormalized form. The necessary normalizing constants are the sums over each column. The value in the last row holds for all columns.

		$\mathbf{s}^{(2,3)}$			
		$(+,+)$	$(+,-)$	$(-,+)$	$(-,-)$
$\mathbf{c}^{(2,3)}$	$(+,+)$	1	0	0	0
	$(+,-)$	0	1	0	0
	$(-,+)$	0	0	1	0
	$(-,-)$	0	0	0	1
	$(+,u)$	$\exp\{\alpha^{(3)}\}$	$\exp\{\alpha_3\}$	0	0
	$(-,u)$	0	0	$\exp\{\alpha^{(3)}\}$	$\exp\{\alpha^{(3)}\}$
	$(u,+)$	$\exp\{\alpha^{(2)}\}$	0	$\exp\{\alpha^{(2)}\}$	0
	$(u,-)$	0	$\exp\{\alpha^{(2)}\}$	0	$\exp\{\alpha^{(2)}\}$
	$(u,\ u)$		$\exp\{\alpha^{(2)}+\alpha^{(3)}+\alpha^{(2,3)}\}$		

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

TABLE A.3A

Transition probabilities, $\psi_{ij}^{(1)}(\mathbf{s};\mathbf{r}) = \psi_{ij}^{(1)}(\mathbf{s}^{(1)};\mathbf{r}^{(1)})$, for the movement between areas, $s_1 = s$ (San Miguel I.) and $s_1 = a$ (Año Nuevo I.). Readers should recall that $B_{sex(i),j}^{(1)}(\cdot,\cdot)$ is a sex-specific, 3 df b-spline smooth over age $= j-1$. Probabilities are formed by normalizing over rows of the table entries.

		$\mathbf{s}^{(1)}$	
		s	a
$\mathbf{r}^{(1)}$	s	1	$\exp\{\beta_{sex(i)}^{(1)}(a;s) + B_{sex(i),j}^{(1)}(a;s)\}$
	a	$\exp\{\beta_{sex(i)}^{(1)}(s;a) + B_{sex(i),j}^{(1)}(s;a)\}$	1

TABLE A.3B

Transition probabilities, $\psi_{ij}^{(2,3)}(\mathbf{s};\mathbf{r}) = \psi_{ij}^{(2,3)}(\mathbf{s}^{(2,3)};\mathbf{r}^{(2,3)})$, for the tag loss model presented in Section 3. The state of the $k-1$th flipper tag, s_k, $k = 2,3$ is represented by $s_k \in \{+,-\}$, where a "+" indicates the tag is present and "−" indicates the tag has been lost. Recall from the main portion of the paper that because the tags are identical, we assumed that the marginal loss rate would be the same for each side. Therefore, we constrained the non-zero parameter values $\boldsymbol{\beta}^{(2)}(-;+) = \boldsymbol{\beta}^{(3)}(-;+) = \boldsymbol{\beta}$. In addition, we modeled a linear effect of age, so the covariate vector $\mathbf{x}_j = (1, j-1)$. Values of zero, are due to the fact that tag loss is an absorbing state from which the animal cannot return. The ψ probabilities are presented in their unnormalized form, necessary normalizing constants are the sums over each row.

		$\mathbf{s}^{(2,3)}$			
		$(+,+)$	$(+,-)$	$(-,+)$	$(-,-)$
$\mathbf{r}^{(2,3)}$	$(+,+)$	1	$\exp\{\mathbf{x}_j \cdot \boldsymbol{\beta}\}$	$\exp\{\mathbf{x}_j \cdot \boldsymbol{\beta}\}$	$\exp\{2(\mathbf{x}_j \cdot \boldsymbol{\beta}) + \beta^{(2,3)}\}$
	$(+,-)$	0	1	0	$\exp\{\mathbf{x}_j \cdot \boldsymbol{\beta}\}$
	$(-,+)$	0	0	1	$\exp\{\mathbf{x}_j \cdot \boldsymbol{\beta}\}$
	$(-,-)$	0	0	0	1

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

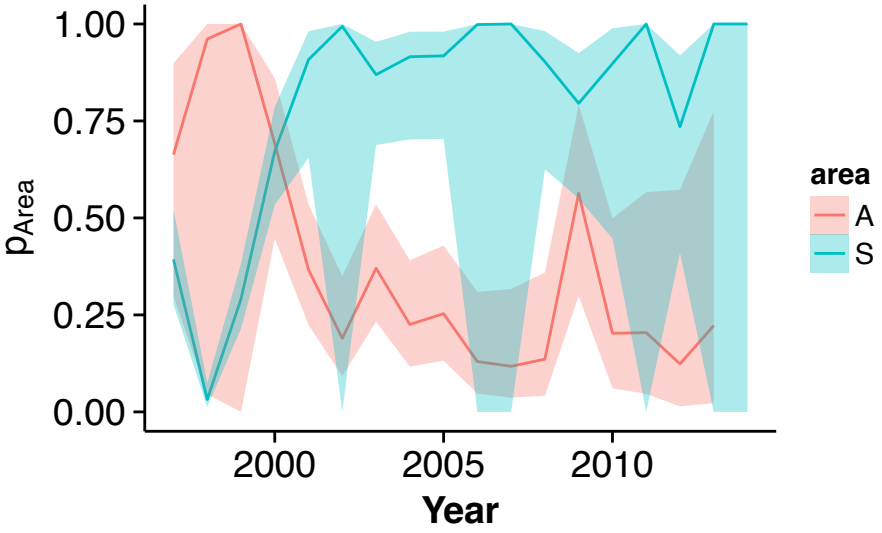


FIG. A.1. *Area- and occasion-specific detection probabilities for California sea lions.*

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

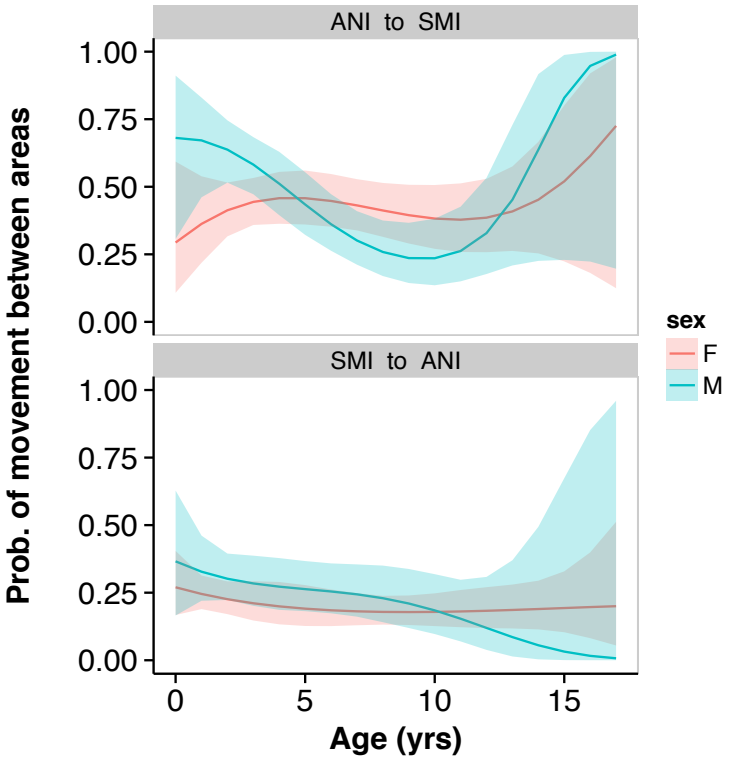


FIG. A.2. *Sex-specific probability of movement between San Miguel Island and Año Nuevo Island, California*

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplement B: R code used to analyze sea lion data

Supplement to "Multivariate state hidden Markov models for mark-recapture data"

Devin S. Johnson, Jeff L. Laake, Sharon R. Melin, and Robert L. DeLong

National Marine Mammal Laboratory,
Alaska Fisheries Science Center,
National Marine Fisheries Service, NOAA,
Seattle, Washington, U.S.A.

Email: *devin.johnson@noaa.gov*

Acknowledgments

The findings and conclusions in the paper are those of the authors and do not necessarily represent the views of the National Marine Fisheries Service, NOAA. Reference to trade names does not imply endorsement by the National Marine Fisheries Service, NOAA.

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data analysis with the multivariate state models can be accomplished with the marked package but a certain level of proficiency with R programming is required. In addition, a thorough knowledge of modeling mark-recapture data and use of multinomial logit link function is needed. Here we provide the code used to fit the model to the sea lion data as an example. It can be used as a template for other analysis. Laake et al (2013) provide an overview of the marked package which will suffice as an introduction but here we describe the extensions that enable multivariate state models to be fitted with the package.

First we attach the marked package and additional packages used for analysis, prediction and graphics.

```
### Load packages ###
# Must have 'marked' v >= 1.1.10
library(marked)
# The splines package is only necessary for fitting b-spline curves used in the paper
# It is not required for the multivate models in the marked package
library(splines)
library(mvtnorm)
library(dplyr)
library(ggplot2)
library(cowplot)
```

The example data are provided in the marked package and can be retrieved with `data(sealions)`. The data contain a capture history string, a numeric weight value which is the anomaly from the sex-specific mean in kg and a sex value (`F` or `M`) for 485 sea lions. The sea lions were initially marked and tagged as pups in 1996 on San Miguel Island, CA. The capture history string has a 3-character representation for each occasion and the occasion values are separated by a comma. The first character is either `A` (Año Nuevo Island) or `S` (San Miguel Island) for the animals location when it was seen. The second and third characters represent the status of the left and right tags respectively. A "`+`" means the tag was present, a "`-`" means the tag was absent and a "`u`" means the tag status was not observed (unknown).

```
data(sealions)
```

The first step in using the marked package is to process the data. Processing the data includes specifying mark-recapture model to be used (mvmscjs) and additional data attributes (e.g., strata.labels). For this model, the labels and values for the capture history string must be provided from left to right in the string. The following specifies that the first variable is called "area" and the possible characters are "`A`" and "`S`" as described above. The second variables are named `ltag` and `rtag` and have values "`+`","`-`", or "`u`". The tag status variables can be unknown but the area is always known because a "`u`" was not included in the vector for area.

```
# Process data for multivariate models in marked
dp=process.data(sealions,model="mvmscjs",
               strata.labels=list(area=c("A","S"),ltag=c("+","-","u"),rtag=c("+","-","u")))
```

```
## 485 capture histories collapsed into 334
```

In processing the data, if there are duplicate data records they are collapsed into unique records (referenced by `id`) and a field named `freq` contains the number of sea lions with the same data. In this case the 485 records are collapsed to $I = 334$ ids with frequencies ranging from 1 to 16. The most commonly collapsed records are those that are released and never seen again. Collapsing into the unique records reduces execution time and the size of the design data list.

The processed data list (`dp`) contains the data and the model and data attributes. The processed data list is passed to the function `make.design.data` which creates a list of design dataframes with a dataframe for each parameter. The parameters for the "`mvmscjs`" models are `Phi`, `p`, `Psi` and `delta` for this model. It does not include `pi` because this implementation assumes that the variables are all known at the time of release.

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

```
### Make design data
ddl=make.design.data(dp)
```

The dataframe for `Phi` contains a record for each id-interval-stratum where stratum is all M combinations of the values of the variables used to define the states. For this example there are 8 strata (`A++`, `A+-`, `A-+`, `A-`, `S++`, `S+-`, `S-+`, `S-`) and 19 occasions and 18 intervals. Thus, there are $I(J-1)M = 18 \times 8 = I \times 144$ records in the design data. Likewise, for `p` there is a record for id-occasion-stratum ($I(J-1)M = I \times 144$) where occasions are resight occasions from 2 to 19. A logit link function is used for `Phi` and `p`.

```
names(ddl$Phi)
```

```
##  [1] "id"      "occ"     "stratum" "area"    "ltag"    "rtag"    "time"
##  [8] "cohort"  "age"     "sex"     "weight"  "freq"    "Time"    "Cohort"
## [15] "Age"     "order"
```

```
nrow(ddl$Phi[ddl$Phi$id==1,])
```

```
## [1] 144
```

```
names(ddl$p)
```

```
##  [1] "id"      "occ"     "stratum" "area"    "ltag"    "rtag"    "time"
##  [8] "cohort"  "age"     "sex"     "weight"  "freq"    "Time"    "Cohort"
## [15] "Age"     "fix"     "order"
```

```
nrow(ddl$p[ddl$p$id==1,])
```

```
## [1] 144
```

For the `delta` design data, there are $I(J-1)Mn$ records where n is the possible number of observations for each state. In this example, $n = 4$ because you can either observe the state variable or unknown value for two of the three variables, `ltag` and `rtag`. For example, for state `A++` you can observe `A++`, `A+u`, `Au+`, or `Auu`. Each set of n records is a multinomial with the sum of the four probabilities equal to 1. The multinomial logit is implemented using a log link function and then normalizing by the sum over the set of n values. To be identifiable, one of the n values must be fixed to 1 (reference cell). We will explain fixed values later. As a convention we suggest specifying the fully known observation (e.g., `A++`) as the reference cell. Had we specified no `u` values for the variables then there would have only been $n = 1$ record which would be fixed to 1 (i.e., no unknown values). For each id and occasion, there are $Mn = 8 \times 4 = 32$ records.

```
head(ddl$delta)
```

```
##   id occ stratum area ltag rtag obs.area obs.ltag obs.rtag time cohort age
## 1  1   2     A--    A    -    -        A        -        -    2      1   1
## 2  1   2     A--    A    -    -        A        -        u    2      1   1
## 3  1   2     A--    A    -    -        A        u        -    2      1   1
## 4  1   2     A--    A    -    -        A        u        u    2      1   1
## 5  1   2     A-+    A    -    +        A        -        +    2      1   1
## 6  1   2     A-+    A    -    +        A        -        u    2      1   1
##   sex      weight freq Time Cohort Age order
## 1   F -0.08534202    4    0      0   1     1
```

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

```
## 2   F -0.08534202    4    0      0   1     2
## 3   F -0.08534202    4    0      0   1     3
## 4   F -0.08534202    4    0      0   1     4
## 5   F -0.08534202    4    0      0   1     5
## 6   F -0.08534202    4    0      0   1     6
```

```
nrow(ddl$delta[ddl$delta$id==1&ddl$delta$occ==2,])
```

```
## [1] 32
```

For the Psi design data, there are $I(J-1)M^2 = I \times 18 \times 64 = I \times 1152$ records. For each id-occasion-state, there are a set of M records that represent a multinomial for the probability of transitioning from that state to any of the M states. The reference cell is set up by default to be Ψ_{rr} of staying in the same state ($\mathbf{r} \rightarrow \mathbf{r}$). This is done when the design data are created by adding the field "`fix`" and assigning its value to `NA` except for the reference cell which get the value 1. Values with `NA` are estimated and any other value fixes the real parameter (inverse link) at that value. The value 1 is the real parameter value for a log link with a beta value of 0 (1=exp(0)). The multinomial parameters, `delta` and `Psi` are set up with a log link but then normalized by the sum over the multinomial set. For Psi that is a set of M records with $\Psi_{sr} = \exp(\beta_{sr}) / \sum_{s=1}^{M} \exp(\beta_{sr})$ and $\exp(\beta_{rr}) = 1$. A different reference cell could be used by changing the values of the field `fix`. The reference cell does not have to be the same for each state or it could vary by id or occasion as well but specifying the formula could become tricky in the latter cases.

```
head(ddl$Psi)
```

```
##   id occ stratum tostratum area ltag rtag toarea toltag tortag time cohort
## 1  1   1     A--       A--    A    -    -      A      -      -    1      1
## 2  1   1     A--       A-+    A    -    -      A      -      +    1      1
## 3  1   1     A--       A+-    A    -    -      A      +      -    1      1
## 4  1   1     A--       A++    A    -    -      A      +      +    1      1
## 5  1   1     A--       S--    A    -    -      S      -      -    1      1
## 6  1   1     A--       S-+    A    -    -      S      -      +    1      1
##   age sex      weight freq Time Cohort Age fix order
## 1   0   F -0.08534202    4    0      0   0   1     1
## 2   0   F -0.08534202    4    0      0   0  NA     2
## 3   0   F -0.08534202    4    0      0   0  NA     3
## 4   0   F -0.08534202    4    0      0   0  NA     4
## 5   0   F -0.08534202    4    0      0   0  NA     5
## 6   0   F -0.08534202    4    0      0   0  NA     6
```

```
nrow(ddl$Psi[ddl$Psi$id==1&ddl$Psi$occ==1,])
```

```
## [1] 64
```

Default design data are created by `make.design.data` but you are free to add your own fields which may create other fields derived from the default fields or possibly add entirely new data fields by merging in occasion data like environmental variables. If you do merge in other data with the merge function, make sure to re-order the design dataframe using the variable order. Below, a variable pup is defined for `Phi` and `sex` is changed from a character to a factor variable. Had the latter been done in the data, this would not have been needed.

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

```
# Create pup variable for Phi
ddl$Phi$pup=ifelse(ddl$Phi$Age==0,1,0)
ddl$Phi$sex=factor(ddl$Phi$sex)
```

Data from Año Nuevo Island in 2014 were not available at the time, so `p` is set to `0` for `area=="A"` on the last occasion:

```
# Detection model
# Set final year p=0 (no resight data) for ANI
ddl$p$fix = ifelse(ddl$p$Time==17 & ddl$p$area=="A", 0, NA)
```

For `delta`, a 0/1 variable is created for the left and right tag when it is unknown. These variables are used in the formula such that the implied reference cell is when the status of both tags is known.

```
# Delta model
# create indicator variables for 'unknown' tag observations
ddl$delta$obs.ltag.u = ifelse(ddl$delta$obs.ltag=="u", 1, 0)
ddl$delta$obs.rtag.u = ifelse(ddl$delta$obs.rtag=="u", 1, 0)
```

The changes to the `Psi` data are slightly more involved. We cannot have transitions from "`-`" (missing) to "`+`" present for a tag, so that transition is set to `0` for both the left and right tag. Dummy 0/1 variables are created for transitions from `A` to `S` and from `S` to `A` to be used in the formulas for `Psi`. In addition, dummy variable `lpm` and `rpm` are set to `1` for transitions from "`+`" to "`-`" (losing a tag) and `0` otherwise. Finally, `sex` is converted to a factor variable.

```
# Psi model
# Set Psi to 0 for cases which are not possible - missing tag to having tag
ddl$Psi$fix[as.character(ddl$Psi$ltag)=="-"&as.character(ddl$Psi$toltag)=="+"]=0
ddl$Psi$fix[as.character(ddl$Psi$rtag)=="-"&as.character(ddl$Psi$tortag)=="+"]=0
# Create indicator variables for transitioning between states
ddl$Psi$AtoS=ifelse(ddl$Psi$area=="A"&ddl$Psi$toarea=="S",1,0)  # ANI to SMI movement
ddl$Psi$StoA=ifelse(ddl$Psi$area=="S"&ddl$Psi$toarea=="A",1,0)  # SMI to ANI movement
ddl$Psi$lpm=ifelse(ddl$Psi$ltag=="+"&ddl$Psi$toltag=="-",1,0)   # Losing left tag
ddl$Psi$rpm=ifelse(ddl$Psi$rtag=="+"&ddl$Psi$tortag=="-",1,0)   # Losing right tag
ddl$Psi$sex=factor(ddl$Psi$sex)
```

Formulas are specified in a list and named by the variable with a dot and an extension (eg 1 in each case below). This form is suggested to use the crm.wrapper function for fitting multiple sets of models.

```
Psi.1=list(formula=~-1+ AtoS:sex + AtoS:sex:bs(Age) + StoA:sex + StoA:sex:bs(Age) +
                    I(lpm+rpm) +I(lpm+rpm):Age + lpm:rpm)
p.1=list(formula=~time*area)
delta.1=list(formula= ~ -1 + obs.ltag.u + obs.rtag.u + obs.ltag.u:obs.rtag.u)
Phi.1=list(formula=~sex*bs(Age)+pup:weight+area)
```

The formula for `p` is the simplest. It specifies time varying resighting probabities that differ across the two areas (`A` and `S`). The formula for `Phi` specifies a cubic spline across age that varies by `sex`, survival varying by weight for `pup==1`, and an additive `area` difference in survival. The formula for delta specifies a dependence model for the probability of not observing the status of a tag in which if one tag is not known may influence whether the other tag is known. By removing the intercept (`-1`), the intercept value is `0` which fixes the reference cell to be observed status for both tags (`obs.ltag.u`=0 and `obs.rtag.u`=0). The formula for `Psi` is the most complicated involving two additive portions with the first for movements between `A` and `S` and `S`

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

to `A` that are a function of `sex` and `Age` represented by a spline that is sex specific. The second portion is for tag loss (transitions from "+" to "-"). It specifies tag loss that is constant for left and right (I(lpm+rpm)), but changes linearly with age (`I(lpm+rpm):Age`) and dependence (`lpm:rpm`) whereby losing one tag influences loss of the other tag.

For this single model a call to `crm` fits the model and returns the results into `mod`. The arguments are the processed data list (`dp`), the design data (`ddl`), the `model.parameters` which is a list of the parameter lists and an argument hessian set to TRUE to get the variance-covariance matrix. Note that this model fitting with estimation of the hessian may take about an hour to run. We are working on finding more efficient ways of fitting these models. With the generality of fitting models in marked, comes the concomitant cost of slow execution.

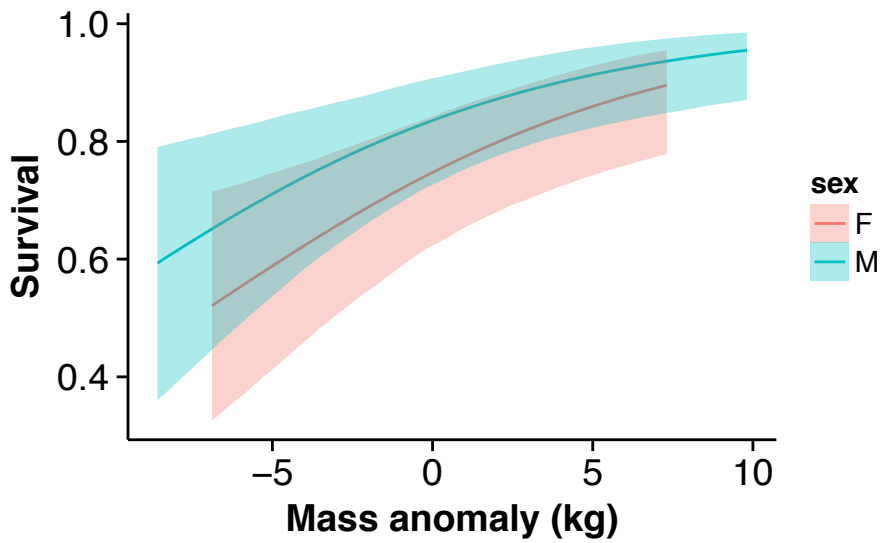
```
# Fit model
mod=crm(dp,ddl,model.parameters=list(Psi=Psi.1,p=p.1,delta=delta.1,Phi=Phi.1,hessian=TRUE)
```

Various plots were produced from the model results. Here we provide the code to produce plots for survival. The first plot shows the pup survival estimates as a function of their weight anomaly. The comments below describe the code used.

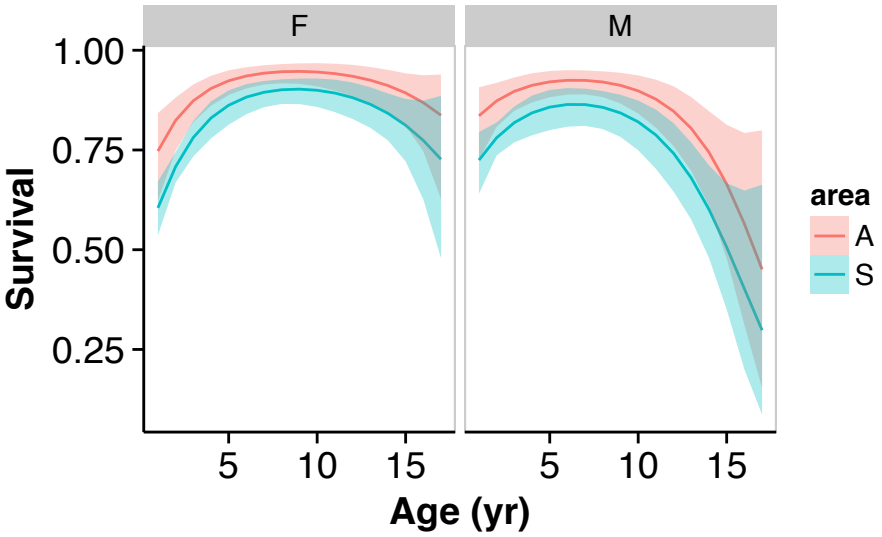
```
#pup - create a dataframe with a sequence of weights for females and males
Phi_pup = rbind(
  data.frame(
    sex="F",
    weight=seq(min(ddl$Psi$weight[ddl$Psi$sex=="F"]),
               max(ddl$Psi$weight[ddl$Psi$sex=="F"]),0.1)
  ),
  data.frame(
    sex="M",
    weight=seq(min(ddl$Psi$weight[ddl$Psi$sex=="M"]),
               max(ddl$Psi$weight[ddl$Psi$sex=="M"]),0.1)
  )
)
# create a design matrix for the pup survival with the data defined above
Phi_pup_dm = model.matrix(~sex  + weight, data=Phi_pup)
Phi_coef = mod$results$beta$Phi
# extract the variance-covariance matrix for the betas
vcv = mod$results$beta.vcv
vcv_names = rownames(mod$results$beta.vcv)
# extract the portion of the v-c matrix for the Phi parameters
Phi_vcv = vcv[grep("Phi", vcv_names),grep("Phi", vcv_names)]
# compute 10,000 multivariate normal variables with means Phi_coef
# and variance-covariance matrix Phi_vcv
Phi_rep = rmvnorm(10000, Phi_coef, Phi_vcv)
# using the random variable values for the intercept, sex and weight variables compute the
# survival estimates for the 10000 variates
pred_rep=apply(Phi_rep[,c(1,2,10)], 1, FUN=function(b,D){plogis(D%*%b)}, D=Phi_pup_dm)
# compute the 95% confidence intervals from the simulated values
tmp = t(apply(pred_rep, 1, quantile, prob=c(0.025, 0.975)))
colnames(tmp) = c("lower","upper")
# create data frame for plotting and call ggplot to create the plot
Phi_pup = cbind(Phi_pup,tmp)
Phi_pup$estimate = plogis(Phi_pup_dm%*%Phi_coef[c(1,2,10)])
p_pup_surv = ggplot(data=Phi_pup) + geom_line(aes(x=weight, y=estimate, color=sex)) +
  geom_ribbon(aes(ymin=lower, ymax=upper, x=weight, fill=sex), alpha=0.33) +
```

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

```
  ylab("Survival") + xlab("Mass anomaly (kg)") + geom_rangeframe()
p_pup_surv
```



The second plot shows the non-pup survival estimates as a function of age. The procedure used is very similar to the code used above.

```
#adult - Follow a similar
Phi_ad = data.frame(
  expand.grid(
    sex=c("F","M"),
    Age=c(1:17),
    area=c("A","S"),
    pup=0,
    weight=0
  )
)
Phi_ad_dm = model.matrix(mod$model.parameters$Phi$formula, Phi_ad)
Phi_ad$estimate = plogis(Phi_ad_dm%*%Phi_coef)
tmp = Phi_rep %>% apply(., 1, FUN=function(b,D){plogis(D%*%b)}, D=Phi_ad_dm) %>%
      apply(., 1, quantile, prob=c(0.025, 0.975)) %>% t(.)
colnames(tmp)=c("lower","upper")
Phi_ad = cbind(Phi_ad,tmp)
p_ad_surv = ggplot(data=Phi_ad) + geom_line(aes(x=Age, y=estimate, color=area)) +
  geom_ribbon(aes(ymin=lower, ymax=upper, x=Age, fill=area), alpha=0.33) +
  facet_wrap(~sex) +
  ylab("Survival") + xlab("Age (yr)")
p_ad_surv
```

bioRxiv preprint doi: https://doi.org/10.1101/025569; this version posted August 29, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.