Hierarchical modeling of abundance in closed population capture–recapture models under heterogeneity

Matthew R. Schofield · Richard J. Barker

Abstract Hierarchical modeling of abundance in space or time using closed-population mark-recapture under heterogeneity (model $M_h$) presents two challenges: (i) finding a flexible likelihood in which abundance appears as an explicit parameter and (ii) fitting the hierarchical model for abundance. The first challenge arises because abundance not only indexes the population size, it also determines the dimension of the capture probabilities in heterogeneity models. A common approach is to use data augmentation to include these capture probabilities directly into the likelihood and fit the model using Bayesian inference via Markov chain Monte Carlo (MCMC). Two such examples of this approach are (i) explicit trans-dimensional MCMC, and (ii) superpopulation data augmentation. The superpopulation approach has the advantage of simple specification that is easily implemented in BUGS and related software. However, it reparameterizes the model so that abundance is no longer included, except as a derived quantity. This is a drawback when hierarchical models for abundance, or related parameters, are desired. Here, we analytically compare the two approaches and...
show that they are more closely related than might appear superficially. We exploit this relationship to specify the model in a way that allows us to include abundance as a parameter and that facilitates hierarchical modeling using readily available software such as BUGS. We use this approach to model trends in grizzly bear abundance in Yellowstone National Park from 1986 to 1998.

Keywords  Bayesian · Capture recapture · Complete data likelihood · Data augmentation · Hierarchical · MCMC · Reversible jump · Superpopulation · Trans-dimensional

1 Introduction

Despite their long history and widespread use in ecology and epidemiology capture–recapture models continue to be the focus of research. Usually we treat the closed population problem as being one of inference about abundance, $N$, a quantity that indexes the size of the population. While often the case that a single estimate of $N$ may represent the endpoint of the analysis it is inevitable that analysts will think in terms of describing a collection of abundance estimates across space or time (indexed by $k$). We use $N$ generically for abundance, both for a single instance or when describing a collection across space or time. We also use the notation $N_k$ when we wish to be clear that we are referring to members of a larger collection.

One approach to modeling $N$ is the open population model in which a single population is marked and an integrated model used to describe the dynamics of the population in terms of recruitment and survival probabilities. Here we consider an alternative where closed-population capture–recapture surveys are replicated in space or time with marked cohorts that are distinct and hierarchical models are used to describe dynamics of $N$.

1.1 A motivating example

Boyce et al. (2001) estimated the abundance of the female grizzly bears, *Ursus arctos*, in the Yellowstone ecosystem each year in a 13 year study from 1986 to 1998 using daily counts of individually identified bears. To account for sighting heterogeneity (i.e., individual variation in detection probabilities), they modeled the capture frequencies as over-dispersed Poisson counts by fitting a negative binomial model conditioned on the fact that the counts are non-zero. An alternative approach, which we follow below, is to treat the capture frequencies as summary statistics from a closed population capture recapture study with $t$ the total number of days surveyed.

Boyce et al. (2001) used simulated rank-correlations to assess the strength of evidence in support of an increasing trend in the numbers of grizzly bear females. A more natural approach would be to fit a two-stage hierarchical model. In the first stage, we would model female grizzly abundance each year (indexed by time) using a closed population mark-recapture model parameterized in terms of $N_k$ and with heterogeneity. In the second stage we would model grizzly female abundance as following some sort of trend with error. Questions concerning long-terms changes in grizzly bear abundance could then be addressed in terms of parameters of the trend model.
Hierarchical capture–recapture modeling presents two challenges. The first is finding a flexible likelihood in which \( N \) appears as an explicit parameter when there is heterogeneity in the capture probabilities which we denote by \( p \). The problem is that the parameters \( N \) determine the dimension of \( p \). The second challenge is fitting the hierarchical model for \( N \). We focus our initial discussion on this first challenge but show that it can be resolved in a way that provides a flexible solution to the second.

The reference model for closed-population capture–recapture under heterogeneity is \( M_h \) (Otis et al. 1978). A fundamental problem with \( M_h \) is that it is over-parameterized. The standard solution (e.g. Burnham and Overton 1978) is to model the components of \( p \) as random effects drawn from a distribution indexed by \( N \) and to then integrate them from the model. A drawback however is that explicit integration of the random effects yields a closed-form likelihood only in a few restricted cases thereby limiting the number of models available. A useful alternative is to integrate over the random effects using Monte Carlo methods. One such approach is to model in terms of a complete data likelihood (CDL) where the likelihood is ‘completed’ by including the unobserved \( p \); inference can then proceed via the (Monte Carlo) EM algorithm (Dempster et al. 1977; Levine and Casella 2001) or the Gibbs sampler (Geman and Geman 1984).

An impediment to use of the CDL is that its dimension changes with \( N \) through \( p \). Three seemingly different approaches have been used to circumvent this problem. The first is to eliminate the dependence of \( N \) on \( p \) in a Gibbs sampler by integrating out \( p \) from the joint full conditional distribution of \( N \) and \( p \) (Fienberg et al. 1999). Here, we do not consider combining MCMC and other numerical integration techniques, although we briefly discuss the use of quadrature in evaluating the likelihood in Sect. 4.

The second approach is to use a trans-dimensional (TD) algorithm to update the parameter \( N \). Sisson (2005) defines a TD algorithm as one that admits transitions between states that have corresponding parameter vectors of differing dimension (in our case the states are indexed by values of \( N \)). One example of a TD algorithm is the reversible jump Markov chain Monte Carlo (RJ-MCMC) algorithm (Green 1995) that explicitly accounts for the differing dimension when moving between states. Another example is the product space approach of Chib (1995) that avoids the need to explicitly account for the differing dimension when moving between states. Such TD algorithms have been used by Fienberg et al. (1999), King and Brooks (2008), Durban and Elston (2005) (subsequently DE) and others in the context of \( M_h \) or related capture–recapture models.

The third approach is to use superpopulation data augmentation to specify the model. The term ‘superpopulation’ refers to an approach in which the likelihood is completed by augmenting \( p \) in a way that fixes its dimension. The phrase ‘data augmentation’ (DA) refers to a family of computational methods that add missing data/auxiliary variables into the likelihood (Gelman 2004) to facilitate modeling. DA is used, both for the CDL models mentioned above, as well as in the superpopulation models of Royle et al. (2007) (subsequently RDL), Royle and Dorazio (2008), Link and Barker (2010) and others in the context of \( M_h \) or related capture–recapture models.
The idea behind the approach of RDL came from identifying a relationship between capture–recapture and occupancy modeling (MacKenzie et al. 2002).

While elegantly simple, the approach suggested by RDL is unnatural for capture–recapture modeling in that it relegates \( N \), the parameter of interest, to the status of derived parameter. A practical consequence is that in Bayesian modeling it can make prior specification difficult, particularly for hierarchical models in which interest is in modeling \( N \) through space and time.

1.3 Outline of the rest of the paper

Our purpose is twofold. Firstly, to provide a critical examination of the methodological issues underlying various augmented data representations of closed population capture–recapture models when capture probabilities are heterogeneous (approaches two and three above). In Sect. 2 we review integrated likelihood, full-likelihood and super-population data augmentation and investigate their relationship and how parameters can be modeled within a Bayesian framework. Our second purpose is to show how the CDL can be explicitly expressed in terms of the abundance parameter \( N \) for the capture recapture model \( M_h \) in a manner that is straight-forward to implement using generic Bayesian hierarchical modeling. In Sect. 3 we introduce our hierarchical extensions to \( M_h \) and illustrate these methods using the grizzly bear example and show how to fit this model using the readily available software BUGS (Lunn et al. 2000) or JAGS (Plummer 2003) for carrying out Bayesian analysis of data. We offer concluding comment in Sect. 4.

2 \( M_h \) likelihoods and varying dimensions

Likelihoods for model \( M_h \) include an integrated likelihood (Burnham and Overton 1978), a “full” likelihood (King and Brooks 2008), and a super-population likelihood of RDL. The three most obvious differences between these approaches are (i) whether or not \( N \) is explicitly included in the likelihood, (ii) the presence or absence of combinatorial terms involving \( N \), and (iii) the presence or absence of a superpopulation (with corresponding value \( M \)) in the model. We do not consider conditional approaches, such as those discussed by Sanathanan (1972).

In the following we use \([x|p]\) to denote the model for the latent data \( x \) given the vector of capture probabilities \( p \). Unless we specify otherwise, \( x_{ij} \) is 1 if individual \( i (i = 1, \ldots, N) \) was captured in sample \( j (j = 1, \ldots, t) \) and 0 otherwise, and \( p_i \) is the capture probability of individual \( i \). We also consider the observed data \( x^{obs} \), where \( x^{obs}_{ij} \) is 1 if individual \( i = 1, \ldots, n \) was captured in sample \( j \) and \( n \) is the number of distinct individuals caught in the study. Other than the inclusion of unobserved individuals in \( x \), the only difference between \( x \) and \( x^{obs} \) is being unable to align the row indices of \( x^{obs} \) with the row indices of \( x \). There are \( \prod_{z=0}^{N} z_h^{N} \) possible \( x \) matrices that are compatible with \( x^{obs} \) for a given value of \( N \geq n \), where \( z_h \) is the number of individuals observed with capture history \( h \). Note that \( z_0 = N - n \) denotes the number of individuals with the null history 000 . . . 0.
2.1 Burnham’s integrated likelihood

Burnham’s integrated likelihood (Burnham and Overton 1978) can be considered as the standard form of the observed data likelihood (ODL) for model \( M_h \). The ODL is obtained by modeling the unknown capture probabilities \( p \) as exchangeable random effects sampled from a distribution with parameters \( \theta_p \), \( [p|\theta_p] \) with \( cdf \, F(p) \equiv F_{\theta_p}(p) \). Starting with

\[
[x|p] = \prod_{i=1}^{N} \prod_{j=1}^{t} p_i^{x_{ij}} (1 - p_i)^{1-x_{ij}} \tag{1}
\]

and summing over the permutations of \( x \) and integrating out \( p \) leads to:

\[
[x^{obs}|\theta_p, N] = \frac{N!}{\prod_{h} z_{hh}!} \prod_{j=0}^{t} \xi_{jF} \tag{2}
\]

where \( f_j \) is the number of individuals caught \( j \) times, and \( \xi_{jF} = \int_0^1 p^j (1 - p)^{t-j} dF(p) \). Our focus here is on parametric distributions \( F \) that do not allow \( \xi_{jF} \) to be evaluated directly. Tardella (2002) and Farcomeni and Tardella (2010) consider a case where \( F \) is a finite-dimensional non-parametric distribution and the resulting MCMC algorithm does not require a TD step.

2.2 Complete data likelihood (CDL): \( N \) as a parameter

Starting with (1) and including a hierarchical model for \( p \), we can write a complete data likelihood as

\[
[x^{obs}, p|\theta_p, N] = \frac{N!}{\prod_{h} z_{hh}!} \prod_{i=1}^{n} \prod_{j=1}^{t} p_i^{x^{obs}_{ij}} (1 - p_i)^{1-x^{obs}_{ij}} \prod_{i=n+1}^{N} (1 - p_i)^{\prime} \prod_{i=1}^{N} [p_i|\theta_p] \tag{3}
\]

If we integrate across \( p_i (i = 1, \ldots, N) \), we obtain Eq. (2). We refer to (3) as the CDL-\( N \) model. More details of CDL-\( N \) are available in the supplementary materials, Sect. 1.1.

2.3 Complete data likelihood (CDL): \( w \) as a parameter

RDL describe a complete data likelihood in terms of a super-population \( M \), with \( x \) now an \( M \) by \( t \) matrix. The idea is that we replace \( N \) in the likelihood by \( w \), a vector of binary labels of dimension \( M \) where \( w_i \) takes the value 1 if individual \( i \) is included in the population and 0 otherwise such that \( N = \sum_{i=1}^{M} w_i \). RDL model the elements
\[ w_i \text{ of } w \text{ using an exchangeable Bernoulli prior with parameter } \phi. \text{ Thus, their complete model can be written as} \]

\[
[x, w, p | \theta_p, \phi] = [w | \phi][p | w, \theta_p][x | w, p] \\
= [w | \phi][p | \theta_p][x | w, p] \\
= \prod_{i=1}^{M} \left\{ \left[ p_i | \theta_p \right] \phi^{w_i} (1 - \phi)^{1-w_i} \prod_{j=1}^{t} (p_i w_j)^{x_{ij}} (1 - p_i w_j)^{1-x_{ij}} \right\}, \tag{4}
\]

where \( \phi \) is the probability that an individual in the superpopulation (of size \( M \)) is also a member of the population (of size \( N \)).

Two observations on this approach are relevant for subsequent discussion. The first is that the RDL approach features two distinct data augmentation steps, one involving \( w \) and a second involving \( p \). The first DA step involving \( w \) is equivalent to replacing the scalar parameter \( N \) in the usual formulation of model \( M_h \) with the vector parameter \( w \). The second DA step specifies a value \( p_i (i = 1, \ldots, M) \) for all individuals in the superpopulation, whether they are in the population or not.

The second observation is that starting with (4), then integrating over \( p \), summing over \( w \), and finally summing over permutations of \( x \) leads to

\[
[x^{\text{obs}}, N | \theta_p, \phi, M] = \frac{N!}{\prod_{h \leq h} z_h!} \prod_{j=0}^{t} \xi_{jF} \frac{M}{N} \phi^N (1 - \phi)^{M-N}, \tag{5}
\]

which is equivalent to the ODL in (2) multiplied by the prior for \( N \) induced by modeling the set of values \( w_i \) as exchangeable Bernoulli random variables with parameter \( \phi \). Thus, replacing \( N \) with \( w \) leads to a Bayesian reparameterization of \( M_h \) in which the implied prior on \( N \) is specified by a binomial distribution. We refer to (4) as the CDL-\( w \) model. More details of CDL-\( w \), including the move from (4) to (5) are in the supplementary materials, Sects. 1.2 and 1.3.

Note that RDL do not use (5) to make inference. Instead, they only consider one partitioning of \( x \), where \( x^{\text{obs}} \) is augmented by an \((M - n) \times t\) matrix of zeros. This leads to an integrated likelihood proportional to that shown in (5), see supplementary materials, Sect. 1.3 for details.

2.4 Model fitting

Modeling via the ODL in (2) is straight-forward with the trans-dimensional nature of the model eliminated, provided we can explicitly integrate over the random effects distribution of \( p \). Unfortunately we can only do this for a restricted choice of random effects distribution. We can solve this problem through use of Bayesian CDL approaches and Markov chain Monte Carlo (MCMC) for model fitting.

The CDL-\( w \) model can be easily implemented using Gibbs sampling. An associated advantage is the simplicity of specifying model in BUGS. We refer the reader to Royle...
et al. (2007) for details of the Gibbs sampler for the CDL-$w$ model, including implementation in BUGS. A point we will return to later is that it is potentially difficult to include flexible priors for $N$ in the CDL-$w$ model. In particular, it is difficult to include any prior that is not a binomial prior that is induced through the specification outlined earlier. This is unlikely to be a problem when we have a single closed population data set. However, it becomes more important if we have replicated closed population data sets, across space or time, for which we wish to specify a hierarchical model for $N$.

Here we outline several approaches that can be used to fit model CDL-$N$. As already noted, the problem is that the random variable $N$ defines the dimension of the parameter vector $p$. To overcome this, King and Brooks (2008) implement a reversible-jump MCMC algorithm to update $N$ within a MCMC procedure. All other unknowns can be updated using standard Gibbs sampling steps.

A different approach was implemented by DE who use the algorithm of Carlin and Chib (1995). A necessary part of this algorithm is to specify an upper size limit $M$ and values $p_{N+1}, \ldots, p_M$ in the same way as in the CDL-$w$ model. To complete specification of the algorithm, DE specify ‘pseudo-priors’ for the values of $p_{N+1}, \ldots, p_M$. The pseudo-priors are not expressed in the likelihood and their sole purpose is to improve the efficiency of the algorithm. DE used pilot runs of the algorithm to determine appropriate pseudo-priors.

The fitting of CDL-$N$ has been criticized by RDL due to the need to move between ‘models’ using an explicit TD step—either (i) the reversible-jump step of King and Brooks (2008), or (ii) the need for pseudo-priors (and pilot runs) in DE (Royle et al. 2007; Royle and Dorazio 2012). To show this criticism is unjustified, we consider an additional approach to fitting the CDL-$N$ model. It amounts to a special case of the DE algorithm, where we overcome the need for a pilot run by choosing the pseudo-prior for $p_i, i > N$ to be the hierarchical distribution for $p_i, i = 1, \ldots, N$. Such an approach appears to share much in common with the approach used to fit CDL-$w$: both include an upper size limit $M$ and specify the same model for ‘ghost’ individuals (i.e. those individuals that are in $M$ but not $N$). Details of the algorithm as well as associated BUGS code is in the supplementary materials, Sects. 2.1 and 3.

A closer comparison between the algorithms used to fit CDL-$N$ and CDL-$w$ reveals that the problem of moving between models does exist in the CDL-$w$ formulation. The key to understanding this is to consider model fitting using CDL-$N$ as a problem of selecting between models, each indexed by a different value for the parameter $N$. Similarly, model fitting using CDL-$w$ is a problem of selecting between models, each indexed by a different value for $w$. Moreover, the MCMC algorithm used to fit CDL-$w$ is identical to a specific TD algorithm for a model where we (i) reparameterize model $M_h$ by replacing the scalar parameter $N$ by the vector parameter $w$, including the capture probabilities $p_i$ in the likelihood for individuals included in the population (i.e. with $w_i = 1$), and (ii) assume the choice of distribution for augmenting variables has been fixed. The distinction between the likelihood for (i) above and CDL-$w$ is subtle; in (i) the capture probabilities $p_i$ for the ghost individuals ($w_i = 0$) are not included in the likelihood as they are not defined. More details of the equivalence of the two algorithms, including showing that pseudo-priors are a latent feature in the model of RDL as well as the corresponding MCMC algorithms are in the supplementary materials, Sects. 2.2 and 4.
3 Hierarchical extensions

We are interested in situations where replicated closed population data sets are observed through time or across space. In general, we suppose that \( m \) distinct closed-population studies have been carried out and that we are unable to link individuals from one study to another. In addition to estimating abundance from each of these data sets, we also wish to model changes in abundance. We outline the changes required in the CDL-\( N \) and CDL-\( w \) models to include this extension below.

3.1 CDL-\( N \)

The latent capture history matrix for the \( k \)th data set is denoted by the \( N_k \times t_k \) matrix \( x_k \), \( k = 1, \ldots, m \), where \( N_k \) is the abundance for data set \( k \) and \( t_k \) is the number of samples conducted in data set \( k \). The value \( x_{ijk} = 1 \) if individual \( i \) in sample \( j \) of data set \( k \) was observed and \( x_{ijk} = 0 \) otherwise. Note that individual \( i = i^* \) associated with data set \( k = k^* \) is not linked to individual \( i = i^* \) associated with data set \( k \neq k^* \).

Likewise, the observed capture history matrix for data set \( k \) is denoted by the \( n_k \times t_k \) matrix \( x_{\text{obs}}^k \), \( k = 1, \ldots, m \). In general, we model the data from the \( m \) data sets as being exchangeable

\[
[x_{\text{obs}}^1, \ldots, x_{\text{obs}}^m, p_1, \ldots, p_m | \theta_p, N_1, \ldots, N_m] = \prod_{k=1}^{m} [x_{\text{obs}}^k, p_k | \theta_p, N_k]
\]

\[
= \prod_{k=1}^{m} \left\{ \frac{N_k!}{\prod_h z_{hk}} \prod_{i=1}^{N_k} \prod_{j=1}^{t_k} p_{ij}^{x_{ijk}} (1 - p_{ij})^{1 - x_{ijk}} \prod_{i=n_k+1}^{N_k} (1 - p_{ik})^{y_k} \prod_{i=1}^{N_k} [p_{ik} | \theta_p, N_k] \right\},
\]

where \( p_k \) are the vector of capture probabilities from data set \( k \) and \( z_{hk} \) is the number of individuals caught with capture history \( h \) in data set \( k \). That is, we model the \( m \) distinct data sets separately as described in Sect. 2, although we may choose to specify hierarchical models on the parameters of these models.

Instead of keeping the model general, we focus on allowing a hierarchical model for \( N_k \) that includes a trend for the abundance through time. One such possible model is a Poisson model with parameter \( \lambda_k \),

\[
[N_k | \lambda_k] = \prod_{k=1}^{m} \frac{\lambda_k^{N_k} \exp(-\lambda_k)}{N_k!}
\]

with a model for \( \log(\lambda_k) \) that is a linear regression through time,

\[
[\log(\lambda_k) | \beta, \sigma_\lambda] = \prod_{h=1}^{k} \frac{1}{\sqrt{2\pi \sigma_\lambda}} \exp \left( \frac{1}{2\sigma_\lambda} (\log(\lambda_h) - \beta_0 - \beta_1 h)^2 \right).
\]
This model includes $\sigma_\lambda$, which accounts for overdispersion in the Poisson model for $N_k$. Note that this model is useful to illustrate the inclusion of a hierarchical model for $N_k$ but we do not consider this the definitive model for modeling abundance through time. Many other approaches are reasonable including different choices of distribution and trend model.

To fit this hierarchical extension requires a standard extension of the previous MCMC algorithm for model CDL-$N$. We can use a Gibbs sampler and update those additional parameters with known full conditional distributions directly. If a parameter cannot be updated by sampling from a known distribution we use an algorithm that allows sampling from unnormalized density (e.g., the Metropolis-Hastings algorithm). Such extensions are easily included in programs such as BUGS.

### 3.2 CDL-$w$

As with CDL-$N$, we extend the capture history matrices to allow for replicated data sets. The only difference from CDL-$N$ is that the dimension of the latent capture history matrix is $M_k \times t_k$ for the $k$th data set, where $M_k$ is the superpopulation for data set $k$. We also extend the vector $w$ to allow for multiple data sets, with $w_k$ denoting the vector for the $k$th data set. The value $w_{ik} = 1$ if individual $i$ in data set $k$ is included in the population and $w_{ik} = 0$ otherwise.

As with CDL-$N$ we model $x_k$ and $p_k$ from the $m$ data sets as being exchangeable

$$\begin{align*}
[x_1, \ldots, x_m, p_1, \ldots, p_m | w_1, \ldots, w_m, \theta_p, \omega_1, \ldots, \omega_m] &= \prod_{k=1}^{m} [x_k | w_k, p_k] [p_k | w_k, \theta_p] \\
&= \prod_{k=1}^{m} \left\{ \prod_{i=1}^{M_k} \left[ p_{ik} | \theta_p \right] \prod_{j=1}^{t_k} (p_{ik} w_{ik})^{x_{ijk}} (1 - p_{ik} w_{ik})^{1-x_{ijk}} \right\}
\end{align*}$$

Including the required hierarchical model for $N_k$ is more difficult for CDL-$w$. The problem is that we need to express the model for $N_k$ in terms of the random variables $w_k$. That is, given a desired model (or prior) on $N_k$, with density $[N_k | \lambda]$, we seek a distribution $[w_k | \lambda]$, such that

$$[w_k | \lambda] = [N_k | \lambda],$$

where $w_k = \sum_i w_{ik} \equiv N_k$. For this specific case, we need to specify a distribution $[w_k | \lambda]$ that induces

$$[w_k | \lambda] = [N_k | \lambda] = \frac{\lambda_k^{N_k} \exp(-\lambda_k)}{N_k!}.$$ 

One such distribution that satisfies this is

$$[w_k | \lambda_k] \propto \lambda_k^{w_k} \exp(-\lambda_k) (M_k - w_k)!, \quad (6)$$
since we know that for any particular value of $N_k$ there are $\binom{M_k}{N_k}$ realizations of $w_k$ that satisfy $\sum_i w_{ik} = N_k$.

An alternative approach is to use parameter expansion (Liu and Wu 1999); that is, we add a (non-identifiable) parameter to the distribution to facilitate modeling (Gelman et al. 2004). In particular, we want to specify the distribution for the indicator variables $w_k$ such that when we integrate over the additional parameter $\phi_k$ we obtain

$$\int_{\phi_k} [w, k | \phi_k, \lambda] [\phi_k] d\phi_k = [w, k | \lambda] = [N_k | \lambda].$$

One example that satisfies this requirement is

$$[w_k | \phi_k, \lambda_k] \propto \prod_{i=1}^{M_k} \left\{ \phi_k^{w_{ik}} (1 - \phi_k)^{1-w_{ik}} \right\} \lambda_k^{w_k} \exp(-\lambda_k) \frac{\exp(-\lambda_k)}{(w_k)!} \quad (7)$$

with $\phi_k$ having a uniform distribution between 0 and 1.

Once we have specified the distribution for $w_k$ to obtain the required model, we include the distribution for $\log(\lambda_k)$ as in the CDL-N model,

$$[\log(\lambda_h) | \beta, \sigma] = \prod_{h=1}^k \frac{1}{\sqrt{2\pi} \sigma} \exp\left(\frac{-1}{2\sigma^2} (\log(\lambda_h) - \beta_0 - \beta_1 h)^2\right).$$

How we include this hierarchical extension in model fitting depends on the method of fitting. If we are using self-written code, we can fit the model using either (6) or (7) above. However, to the best of our knowledge, we are unable to specify the model using (6) in BUGS. We can, however, include (7) using a “ones/zeros trick” as described in the supplementary materials Section 5.

3.3 Example: grizzly bear abundance

Boyce et al. (2001) estimated the abundance of the female grizzly bears, *Ursus arctos*, in the Yellowstone ecosystem each year in a $m = 13$ years study from 1986 to 1998 using negative binomial models on the counts of individual bears. Instead of modeling the counts we treat the counts as arising from $t$ capture–recapture samples in each year (across which we assume closure). Unfortunately, the number of sampling occasions is not specified (and in fact may be ill-defined), but the methods section of the manuscript suggests that we could treat the sampling as undertaken each day during summer resulting in $t \approx 150$. To test the sensitivity of the model to the value of $t$ we also fit the model with $t = 20$.

We use the hierarchical model specified above with one additional extension. We model the capture probabilities within each data set as being exchangeable random
variables from a logit-normal distribution,

$$\text{logit}(p_{ik}) \overset{\text{exch.}}{\sim} \mathcal{N}(\mu_k, \sigma^2_p), \quad i = 1, \ldots, N_k, \ k = 1, \ldots, m,$$

and model the mean values by an exchangeable normal prior,

$$\mu_k \overset{\text{exch.}}{\sim} \mathcal{N}(\mu_\mu, \sigma^2_\mu), \quad k = 1, \ldots, m.$$

To complete the specification we include prior distributions. We place normal prior distributions on $\beta_0$ and $\beta_1$ with mean of 0 and variance 100,000, a logistic prior distribution on $\mu_\mu$ with location 0 and scale 1 and Student’s $t$ prior distributions (truncated above zero) on $\sigma_p$, $\sigma_\mu$ and $\sigma_\lambda$ with location 0, scale 25 and 3 degrees of freedom. The prior distributions for the standard deviations were specified using the redundant parameterization of Gelman (2006). We fitted the model in JAGS, with full details of the models in the JAGS code at www.maramatanga.com.

For inference we used three chains of length 10,000 each with a tuning/burn-in phase of length 5,000. The value of $t$ (20 or 150) has little impact on the prediction of $N$ and the resulting trend (Fig. 1), but does, as expected, have a large impact on the parameter estimates related to the capture probability random effect (Fig. 2). A substantial amount of mass in the posterior distribution near 0 for $\sigma_\lambda$ suggests that there is little evidence of overdispersion in the Poisson model.

The yearly estimates obtained largely agree with those obtained by Boyce et al. (2001). The major difference between the estimates (and corresponding uncertainties) of Boyce et al. (2001) and ours is due to the partial pooling towards the Poisson regression model evident in our results. Such pooling modestly shrinks the uncertainty of most estimates, with large shrinkage evident for estimates with high uncertainty in Boyce et al. (2001). Shrinkage in the estimates is also evident in 1992 when the estimate from Boyce et al. (2001) is high relative to the estimates 2 years either side.

Evidence for an increasing population of grizzly bear females over time is compelling: a 95% credible interval for the mean increase in total female grizzlies from 1986 to 1998 is $(7.4, 45.6)$ for $t = 20$ or $(6.7, 41.0)$ for $t = 150$. This corresponds to a 95% credible interval for the relative population increase of $(22.5, 206.4\%)$ for $t = 20$ and $(21.5, 201.5\%)$ for $t = 150$ over the duration of the study (Fig. 1).

### 3.3.1 Computational comparison

Above we have compared the different parameterizations and their implementation theoretically. Here we explore the practical differences in the MCMC algorithms for CDL-$N$ and CDL-$w$. In particular, we fit the model for the grizzly bears mentioned above in JAGS both for CDL-$N$ (as described above) and CDL-$w$ (as described in Sect. 3.2 and the supplementary materials Section 5). To evaluate performance of the two approaches we look at (i) the Gelman diagnostic, $\hat{R}$ (Brooks and Gelman 1998), (ii) the time taken to run the algorithm and (iii) the effective sample size for the parameters. Both (i) and (iii) were computed using the coda package in R. These
Fig. 1 Posterior distribution summaries for abundance through time for female grizzly bears in the Yellowstone ecosystem between 1986 and 1998 from fitting model $M_h$ to the data of Boyce et al. (2001). For the models with $t = 150$ and $t = 20$, the thin vertical lines span the central 95% credible interval, the thick vertical line span the quartiles of the posterior distribution and the central symbol corresponds to the median. For the Boyce model, the thin vertical lines represent the 95% confidence interval and the central symbol corresponds to the point estimate found in Boyce et al. (2001). The plotted trend for $N$ is the median, 2.5% and 97.5% quantiles of the posterior distribution for $\exp(\beta_0 + \beta_1 h)$ for the models with $t = 150$ (black) and $t = 20$ (gray).

Fig. 2 Posterior distribution summaries of parameters relating to abundance of female grizzly bears in the Yellowstone ecosystem between 1986 and 1998 from fitting model $M_h$ to the data of Boyce et al. (2001). The thin horizontal lines span the central 95% credible interval, the thick horizontal line span the quartiles of the posterior distribution and the vertical line corresponds to the median. The model fitted to $t = 150$ is denoted in black, and $t = 20$ denoted in gray. A solid gray area is an inadmissible region in the parameter space. The value $\beta_0^* = \beta_0 - 4$ and $\mu_\mu^* = \mu_\mu + 4$

measures are not a definitive measure of performance, as how we express the models, as well as the specific algorithms and programming language used can have a dramatic effect on both the convergence and running time of the algorithms.
For this example, the CDL-\(w\) model has larger effective sample sizes in general (Table 1). However, once we incorporate the time required to run the algorithm (on the lead authors desktop machine the CDL-\(N\) model took \(\sim 30\) vs \(\sim 50\) min for the CDL-\(w\) model), the CDL-\(N\) model has better efficiency than the CDL-\(w\) model. The effective sample size per minute is higher for every parameter we considered when using CDL-\(N\).

4 Discussion

Here we have shown how the CDL can be explicitly expressed in terms of the abundance parameter \(N\) for the capture recapture model \(M_h\) in a manner that is straightforward to implement using generic Bayesian hierarchical modeling software such as BUGS. In particular, the trans-dimensional nature of the model is no major impediment, and offers an advantage over the RDL approach in that it facilitates hierarchical modeling of \(N\). Although we have shown how the RDL model can be modified to allow hierarchical modeling of \(N\) this is not straightforward owing to the fact that \(N\) is a derived quantity in the CDL-\(w\) model.

Another way to fit these models that we have not discussed in detail is to integrate any latent random variables (including those whose dimension is defined by \(N_k\)) out of the likelihood. In general, we are unable to explicitly evaluate the necessary integrals. Numerical integration in the form of quadrature is an attractive alternative for models where the integral is of low dimension. In particular, finding the MLE using quadrature is often faster than obtaining an appropriately sized sample from the posterior distribution using MCMC and has the advantage of well defined convergence conditions (Gimenez and Choquet 2010). The problem is that quadrature gets extremely difficult (and resource hungry) as we begin to explore hierarchical models that require high dimensional integrals. This is the case for the hierarchical models we consider here that extend model \(M_h\) to describe data from replicated studies with distinct cohorts.

In discussing use of DA for the CDL-\(w\) model, RDL emphasize the distinction between their data augmentation scheme and use of TD methods. Similar commentary is subsequently provided by Royle and Dorazio (2012). In particular, RDL argue that standard approaches for fitting CDL-\(N\) require specialized TD algorithms and treat the problem as one of model-selection. They then state that the model selection aspect is not a feature of the DA approach used to fit the CDL-\(w\) model and go on to characterize TD approaches as being difficult and inaccessible (Royle and Dorazio 2012).

As we have described above (and shown in the supplementary materials), the DA model fitting scheme of RDL corresponds to a specific TD implementation. This means that the problem of moving between models does exist in the CDL-\(w\) formulation. In presenting this argument we do not claim that their method is useful only because it is a special case of a TD algorithm. We are arguing that their characterization of TD algorithms and model selection for fitting \(M_h\) as being difficult, inaccessible and unnecessary is without foundation. In spite of a different underlying motivation, RDL have used an equivalent method in their solution to the problem.

The close relationship between the RDL approach and more usual TD formulations is hardly surprising in that Godsill (2001) has shown that algorithms such as RJMCMC
Table 1  The Gelman diagnostic $\hat{R}$, effective sample size (ESS) and ESS per minute (ESS/min) for selected parameters in the hierarchical model for the CDL-$N$ model and the CDL-$w$ model

| Parameter | $\hat{R}$ | ESS | ESS/min |
|-----------|-----------|-----|---------|
|           |           | Total | CDL-$N$ | CDL-$w$ | Total | CDL-$N$ | CDL-$w$ |
| $\beta_0$ | 1.01      | 916   | 1.00    | 32.3    | 1.73  |
|           | 1.02      | 918   |         | 18.7    |       |
| $\beta_1$ | 1.00      | 3,295 | 0.74    | 116.3   | 1.29  |
|           | 1.00      | 4,446 |         | 90.4    |       |
| $N_1$     | 1.01      | 1,604 | 0.75    | 56.6    | 1.31  |
|           | 1.02      | 2,126 |         | 43.2    |       |
| $N_2$     | 1.00      | 2,449 | 0.89    | 86.4    | 1.55  |
|           | 1.00      | 2,742 |         | 55.7    |       |
| $N_3$     | 1.01      | 2,041 | 0.87    | 72.0    | 1.51  |
|           | 1.01      | 2,352 |         | 47.8    |       |
| $N_4$     | 1.01      | 1,794 | 0.83    | 63.3    | 1.44  |
|           | 1.01      | 2,165 |         | 44.0    |       |
| $N_5$     | 1.00      | 2,065 | 0.79    | 72.9    | 1.36  |
|           | 1.01      | 2,629 |         | 53.4    |       |
| $N_6$     | 1.01      | 1,428 | 0.72    | 50.4    | 1.25  |
|           | 1.01      | 1,989 |         | 40.4    |       |
| $N_7$     | 1.01      | 2,285 | 0.90    | 80.6    | 1.56  |
|           | 1.00      | 2,548 |         | 51.8    |       |
| $N_8$     | 1.00      | 2,269 | 1.01    | 80.1    | 1.76  |
|           | 1.01      | 2,243 |         | 45.6    |       |
| $N_9$     | 1.01      | 1,900 | 1.12    | 67.0    | 1.95  |
|           | 1.01      | 1,695 |         | 34.5    |       |
| $N_{10}$  | 1.01      | 2,173 | 1.00    | 76.7    | 1.74  |
|           | 1.00      | 2,172 |         | 44.1    |       |
| $N_{11}$  | 1.01      | 1,618 | 0.89    | 57.1    | 1.54  |
|           | 1.00      | 1,822 |         | 37.0    |       |
| $N_{12}$  | 1.00      | 1,552 | 0.91    | 54.8    | 1.59  |
|           | 1.01      | 1,699 |         | 34.5    |       |
| $N_{13}$  | 1.01      | 1,255 | 0.80    | 44.3    | 1.39  |
|           | 1.01      | 1,568 |         | 31.9    |       |
| $\sigma_N$ | 1.00    | 3,734 | 0.86    | 131.8   | 1.49  |
|           | 1.00      | 4,360 |         | 88.6    |       |
| $\mu\mu$ | 1.02      | 526   | 0.97    | 18.6    | 1.68  |
|           | 1.04      | 545   |         | 11.1    |       |
| $\sigma_\mu$ | 1.02    | 1,515 | 0.90    | 53.5    | 1.57  |
|           | 1.00      | 1,677 |         | 34.1    |       |
| $\sigma_p$ | 1.02    | 668   | 0.88    | 23.6    | 1.53  |
|           | 1.03      | 760   |         | 15.4    |       |

For each parameter the top line refers to the CDL-$N$ model and the bottom line the CDL-$w$ model. Columns three and five take the ratio of ESS and ESS/min for the CDL-$N$ and CDL-$w$ models.
and the algorithm of Carlin and Chib (1995) are all special cases of a general class of methods based on a product-space representation of the TD problem. In this product-space representation, the TD problem is solved by exploring an expanded parameter space that is of fixed dimension and that includes parameters in the likelihood for each model as well as any supplemental variables. The RDL model fits into this class of problems as we have illustrated in the supplementary materials.

A key difference between the RDL approach and more typical TD algorithms is the reparameterization from $N$ to the nonidentified parameter $w$. The practical consequence of this in terms of fitting $M_h$ is the introduction of a binomial prior on $N$. Importantly, the observed data likelihood is unchanged; the reparameterization of the model does not change its description of the data other than through the introduction of this prior (Gelman 2004).

The product Bernoulli prior on $w$ corresponds to an induced binomial prior on $N$. But $N$ is the parameter of fundamental interest in closed-population capture recapture problems. While computationally convenient, reparameterization in terms of $w$ has the unfortunate effect of relegating $N$ to the status of a derived parameter. As we have showed here, while still possible to model in terms of $N$, the specification of such a model requires additional effort and care to ensure it is appropriate. The natural focus of closed-population mark-recapture modeling is on $N$ and it would seem desirable to allow hierarchical modeling explicitly in terms of $N$.

While we have focused on model $M_h$, we are able to extend these ideas to more complex closed population models. The CDL-$N$ representation can be used for any capture–recapture model that can be fit using CDL-$w$. For example, we could use a the CDL-$N$ representation when including individual covariates (for example Royle 2008; Schofield and Barker 2011).

For one such extension, consider the data of Hook et al. (1980), who estimated the number of people born in upstate New York between 1969 and 1974 with spina bifida by comparing recorded cases found on three “lists”: birth certificates, death certificates and medical rehabilitation records. Two features of the study are relevant here. Firstly, a quantity of interest is the trend in spina bifida prevalence. In particular, Hook et al. (1980) discuss the apparent “temporal decrease in [spina bifida prevalence] rates with birth year” after considering the yearly prevalence estimates. Secondly, there appear to be differences in dependencies among the lists (Hook et al. 1980; Madigan and York 1997) that limit a simple Rasch-type implementation of model $M_h$ or $M_h$ (Agresti 1994; Bartolucci and Forcina 2001; Stanghellini and van der Heijden 2004; Bartolucci and Forcina 2006) for multiple list data. In the supplementary materials, Section 6, we show how we can use a latent class model (Goodman 1974; Bartolucci et al. 2003) for this multilist capture–recapture dataset. The model we consider includes a hierarchical model for spina bifida prevalence that allows us to formally explore potential trends through time.

The ability to model abundance in multiple closed population studies as illustrated here provides an alternative to modeling using traditional robust design models (Pollock 1982). While robust design models make full use of the data, there are many situations where complicated models are required in order to adequately describe components of abundance change such as survival and recruitment, e.g. Bailey et al. (2010). Provided we are (i) willing to accept a small loss in efficiency, and (ii) only
interested in modeling the overall changes in abundance, the approach outlined here may provide a simpler alternative for modeling such data.

Acknowledgments M.R.S. was partially funded by NSF Grants Nos. 0934516 and 0814194.

References

Agresti A (1994) Simple capture-recapture models permitting unequal catchability and variable sampling effort. Biometrics 50:494–500
Bailey LL, Converse SJ, Kendall WL (2010) Bias, precision, and parameter redundancy in complex multi-state models with unobservable states. Ecology 91(6):1598–1604
Bartolucci F, Forcina A (2001) Analysis of capture–recapture data with a Rasch-type model allowing for conditional dependence and multidimensionality. Biometrics 57(3):714–719
Bartolucci F, Forcina A (2006) A class of latent marginal models for capture–recapture data with continuous covariates. J Am Stat Assoc 101(474):786–794
Bartolucci F, Mira A, Scaccia L (2003) Answering two biological questions with a latent class model via MCMC applied to capture-recapture data. Appl Bayesian Stat Stud Biol Med 7
Boyce MS, MacKenzie DI, Manly BFJ, Haroldson MA, Moody D (2001) Negative binomial models for abundance estimation of multiple closed populations. J Wildl Manag 65:498–509
Brooks SP, Gelman A (1998) General methods for monitoring convergence of iterative simulations. J Comput Graph Stat 7(4):434–455
Burnham KP, Overton WS (1978) Estimation of the size of a closed population when capture probabilities vary among animals. Biometrika 65:625–633
Carlin B, Chib S (1995) Bayesian model choice via Markov chain Monte Carlo methods. J Royal Stat Soc Series B 57(3):473–484
Chib S (1995) Marginal likelihood from the Gibbs output. J Am Stat Assoc 90(432):1313–1321
Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Royal Stat Soc B 39:1–38
Durban JW, Elston DA (2005) Mark-recapture with occasion and individual effects: abundance estimation through Bayesian model selection in a fixed dimensional parameter spaces. J Agric Biol Environ Stat 10:291–305
Farcomeni A, Tardella L (2010) Reference Bayesian methods for recapture models with heterogeneity. Test 19(1):187–208
Fienberg S, Johnson M, Junker B (1999) Classical multilevel and Bayesian approaches to population size estimation using multiple lists. J Royal Stat Soc Series A (Stat Soc) 162(3):383–405
Gelman A (2004) Parameterization and Bayesian modeling. J Am Stat Assoc 99(466):537–545
Gelman A (2006) Prior distributions for variance parameters in hierarchical models. Bayesian Anal 1:515–533
Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis, 2nd edn. Chapman & Hall/CRC, London
Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans Patt Anal Mach Intell 6:721–741
Gimenez O, Choquet R (2010) Individual heterogeneity in studies on marked animals using numerical integration: capture-recapture mixed models. Ecology 91(4):951–957
Godsill SJ (2001) On the relationship between Markov chain Monte Carlo methods for model uncertainty. J Comput Graph Stat 10(2):230–248
Goodman LA (1974) Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika 61(2):215–231
Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82(4):711–732
Hook EB, Albright SG, Cross PK (1980) Use of Bernoulli census and log-linear methods for estimating the prevalence of spina bifida in livebirths and the completeness of vital record reports in New York State. Am J Epidemiol 112(6):750–758
King R, Brooks SP (2008) On the Bayesian estimation of a closed population size in the presence of heterogeneity and model uncertainty. Biometrics 64(3):816–824
Levine RA, Casella G (2001) Implementations of the Monte Carlo EM algorithm. J Comput Graph Stat 10(3):422–439
Link WA, Barker RJ (2010) Bayesian inference with ecological applications. Academic Press, London
Liu JS, Wu YN (1999) Parameter expansion for data augmentation. J Am Stat Assoc 94:1264–1274
Lunn D, Thomas A, Best N, Spiegelhalter D (2000) WinBUGS: a Bayesian modelling framework—
concepts, structure, and extensibility. Stat Comput 10:325–337
MacKenzie DI, Nichols JD, Lachman GB, Droege S, Royle JA, Langtimm C (2002) Estimating site occupancy rates when detection probabilities are less than one. Ecology 83:2248–2255
Madigan D, York J (1997) Bayesian methods for estimation of the size of a closed population. Biometrika 84(1):19–31
Otis DL, Burnham KP, White GC, Anderson DR (1978) Statistical inference from capture data on closed animal populations. Wildl Monogr 62:1–135
Plummer M (2003) JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In:
Proceedings of the 3rd international workshop on distributed statistical computing, Vienna, Austria.
Pollock KH (1982) A capture–recapture design robust to unequal probability of capture. J Wildl Manag 46(3):752–757
Royle JA (2008) Modeling individual effects in the Cormack-Jolly-Seber model: a state-space formulation.
Biometrics 64(2):364–370
Royle JA, Dorazio RM (2012) Parameter-expanded data augmentation for Bayesian analysis of capture–
recapture models. J Ornithol 152:521–537
Royle JA, Dorazio RM, Link WA (2007) Analysis of multinomial models with unknown index using data
augmentation. J Comput Graph Stat 16:67–85
Royle J, Dorazio R (2008) Hierarchical modeling and inference in ecology: the analysis of data from
populations, metapopulations and communities. Academic Press, London
Sanathanan L (1972) Estimating the size of a multinomial population. Ann Math Stat 43(1):142–152
Schofield MR, Barker RJ (2011) Full open population capture–recapture models with individual covariates.
J Agric Biol Environ Stat 16:253–268
Sisson SA (2005) Transdimensional Markov chains. J Am Stat Assoc 100(471):1077–1089
Stanghellini E, van der Heijden PGM (2004) A multiple-record systems estimation method that takes
observed and unobserved heterogeneity into account. Biometrics 60(2):510–516
Tardella L (2002) A new Bayesian method for nonparametric capture-recapture models in presence of
heterogeneity. Biometrika 89(4):807–817

Matthew R. Schofield is an Assistant Professor in the Department of Statistics at the University of Ken-
tucky. He received his Ph.D. from the University of Otago in 2007, under the direction of Prof. Richard Barker. His research interests include ecological statistics and problems involving missing data.

Richard J. Barker is Professor of Statistics at the University of Otago. He received his Ph.D. in Statistics from Massey University in 1996, under the direction of Prof. Jeffrey J. Hunter. His research interests include models for estimating animal population parameters and ecological statistics.