Stat Methods Appl (2013) 22:45–66
DOI 10.1007/s10260-012-0221-4

Improved inference on capture recapture models with behavioural effects

Danilo Alunni Fegatelli · Luca Tardella

Accepted: 24 October 2012 / Published online: 13 November 2012

Abstract In the context of capture-recapture modeling for estimating the unknown size of a finite population it is often required a flexible framework for dealing with a behavioural response to trapping. Many alternative settings have been proposed in the literature to account for the variation of capture probability at each occasion depending on the previous capture history. Inference is typically carried out relying on the so-called conditional likelihood approach. We highlight that such approach may, with positive probability, lead to inferential pathologies such as unbounded estimates for the finite size of the population. The occurrence of such likelihood failures is characterized within a very general class of behavioural effect models. It is also pointed out that a fully Bayesian analysis overcomes the likelihood failure phenomenon. The overall improved performance of alternative Bayesian estimators is investigated under different non-informative prior distributions verifying their comparative merits with both simulated and real data.

Keywords Capture history · Capture–recapture · Conditional likelihood · Behavioural effect · Likelihood failure · Bayesian inference

1 Introduction

Capture-recapture models have been developed in biological sciences to estimate the size of a finite population. Ecology has been one of the original fields of development although many other fields such as epidemiology, genetics and software reliability

D. A. Fegatelli · L. Tardella (✉)
Department of Statistical Sciences, Sapienza University of Rome, Rome, Italy
e-mail: luca.tardella@uniroma1.it

D. A. Fegatelli
e-mail: danilo.alunnifegatelli@uniroma1.it

make nowadays extensive use of capture recapture models and promote further developments. In modeling capture recapture experiments with a finite number of trapping occasions there is a classical tripartition of the sources of variability which can affect the probability that a particular unit is caught in one of the trapping occasions: i) behavioural variability due to the change of behaviour of each unit after trapping experience; ii) individual heterogeneity due to observable or unobservable specific characteristics of each unit; iii) temporal, due to the external conditions (weather, season, trapping effort, etc.) which can influence the success of the specific trapping occasion.

Especially in sampling wild animal populations there is empirical evidence that units of the population often exhibit a change of behaviour after trapping experience and this can affect either permanently, or temporarily, their capture probabilities.

In this paper we will be concerned with inferential tools available to model and understand behavioural patterns in capture recapture experiments. This will lead us to restrict the attention from the most general embedding model framework denoted with M_{TBH} in Otis et al. (1978) to the more restrictive M_{TB} framework so that we can highlight some pathologies which typically occur with behavioural models without overlapping them with those due to another layer of difficulty related to individual heterogeneity (Link 2003). Another critical point that is not addressed here is related to possible relaxation of the independence hypotheses among units (Fattorini et al. 2007).

One of the aims of our paper is to point out that this sort of "estimation failure" does hold for the conditional MLE in some classical behavioural models (as well as in some alternative ones) whereas it does not necessarily hold for other inferential approaches. This point is relevant from a theoretical as well as practical perspective. There is a lot of very recent and less recent papers which are concerned with modeling and inferring behavioural patterns. Different approaches have been used ranging from most frequent and classical conditional-likelihood-based inference of Huggins (1989, 1991) to more recent Markov-chain (Yang and Chao 2005) and extensions thereof (Farcomeni 2011), extended latent class models based on Hidden Markov Models (Bartolucci and Pennoni 2007), semiparametric covariate dependent approach (Hwang and Huggins 2011) and others (Ramsey and Severns 2010). Previous papers based on a latent class approach are Bartolucci and Forcina (2001), Stanghellini and van der Heijden (2004) and Bartolucci and Forcina (2006).

Fewer authors have adopted a Bayesian approach for the simplest permanent behavioural settings (Lee and Chen 1998; Lee et al. 2003; Ghosh and Norris 2005) while alternative estimating approaches have been more recently proposed to cope with behavioural modeling in continuous-time recapture settings (Chaiyapong and Lloyd 1997; Yip et al. 2000; Chao et al. 2000; Hwang et al. 2002).

In Sect. 2 we will adopt as model set up a rather general modeling framework recently proposed by Farcomeni (2011) where the contingency table probabilities of all possible capture histories are reparameterized in terms of conditional probabilities. In that setting many classical and new alternative models are derived by imposing restrictions on such conditional probabilities. One of the main motivation for considering the classes of models derived by the approach in Farcomeni (2011) is that the conditional probability reparameterization of the whole joint probability of the entire capture history is particularly suitable for modeling behavioural changes. In fact we believe that one can easily understand the fact that some pattern present in the (initial) partial capture history may result in a change of the subsequent conditional probabilities. In Sect. 3 we point out that a common form of pathology arises when the conditional-likelihood approach is adopted to draw inference on the main parameter of interest within a general class of behavioural models in that modeling framework. Such annoying phenomenon called *likelihood failure* consists of unbounded estimates for the unknown population size which also implies some degree of non robustness of the estimator even when it is guaranteed to yield a finite estimate. This form of degeneracy is not true in general. It is true only sometimes, and especially if you estimate the unknown population size N by means of the maximization of the conditional likelihood. We connect this phenomenon to a problem pointed out similarly by Seber and Whale (1970) in modeling removal studies and later on faced by Carle and Strub (1978) who suggested a weighted likelihood approach as a possible overcome. We also highlight the generality of this likelihood failure phenomenon providing general conditions for its occurrence. In Sect. 4 we show that a fully Bayesian approach theoretically overcomes the possible unboundedness of estimates and we propose alternative Bayesian estimators for further investigation. In Sect. 5 we compare the conditional maximum likelihood estimator (CMLE) and the proposed alternative Bayesian estimators via simulation studies providing empirical evidence of overall improved performance of Bayesian alternatives. In Sect. 6 we also compare the two approaches with a real dataset concerning Great Copper butterflies.

2 Capture–recapture behavioural effect modeling

Let us consider a discrete-time closed capture-recapture experiment in which the unknown population size N is assumed to be constant (no birth–death or immigration-emigration during the whole sampling stages) and individual trappings are recorded in t consecutive times. Suppose that all units act independently and there is no misclassification i.e. all individuals are always recorded correctly and do not lose their marks. Ideally underlying data can be represented as an $N \times t$ binary matrix $\mathbf{X} = \left[x_{ij}\right]$ where $x_{ij} = 1$ if the i-th unit is captured in the j-th occasion and $x_{ij} = 0$ otherwise. Assume that units captured during the study are conveniently labelled from 1 to M and those not captured from $M + 1$ to N. It is clear that we can observe only the firsts M rows of the matrix $\mathbf{X}$. Denoting with $\mathcal{X} = \{0, 1\}$, the space of all possible capture histories for each unit is $\mathcal{X}^t = \{0, 1\}^t$ while the set of all observable capture histories is $\mathcal{X}^t_* = \mathcal{X}^t \setminus (0, \ldots, 0)$ since the unobserved units are not sampled.

In this work we review the main aspects of modeling the behavioural effect to capture revisiting some general model frameworks proposed in literature. Indeed, mice, voles and small mammals often modify their behaviour after being trapped and this change can reduce or increase the probability of later recaptures. Originally Otis et al. (1978) introduced the basic behavioural model M_b, where individual capture probabilities vary only once when first capture occurs. Model M_b is the simplest way to consider behavioural effects. In particular it considers an *enduring* effect to capture since the behaviour, and, consequently, the recapture probability change permanently until the end of the experiment. In model M_b the initial capture probability is denoted with p. It is the same for each unit and remains constant from occasion to occasion until the first capture. Once the unit is captured for the first time the (re)capture probability p changes in r and it remains the same until the end of trapping stages. Formally, in order to distinguish the first capture probability from the recapture probability we will make use of the conditioning with respect to the quantity $\sum_{l=1}^{j-1} x_{il}$ corresponding to the number of recaptures prior to the current time j

$$M_b : \begin{cases} Pr(x_{ij} = 1 \mid \sum_{l=1}^{j-1} x_{il} = 0) = p & \forall i = 1, \dots N \;\; \forall j = 1, \dots t \\ Pr(x_{ij} = 1 \mid \sum_{l=1}^{j-1} x_{il} > 0) = r & \forall i = 1, \dots N \;\; \forall j = 2, \dots t \end{cases}$$

where if the upperbound of the summation index is such that $j - 1 \leq 0$ then the conditioning event $\sum_{l=1}^{j-1} x_{il} = 0$ is dropped. When $r < p$ the capture probability decreases for all subsequent recaptures and this corresponds to modeling the so called *trap shyness*. This behavioural pattern could be due to the traumatic event associated to the capture experience. On the other hand, when $r > p$ there is the so called *trap happiness* effect.

Alternative model frameworks have been recently proposed to model more flexibly behavioural patterns during trapping stages. Yang and Chao (2005) propose to model the capture history sequence by a bivariate Markov chain in which the states incorporate the information on both capture status (captured/non-captured) and marking status (marked/non-marked). Notice that, obviously, if a unit is captured in the previous occasions it is also marked. Yang-Chao's model allows to handle both enduring effects where individuals exhibit a long lasting behavioural response to capture and the so called *ephemeral* effect where individuals have a short term memory and the capture probabilities depend only on the capture occurrence in the previous occasion. When the marking status is not considered we have the simple first-order Markov chain model allowing for ephemeral effect only. A generalized k-th order Markov chain model is considered in Farcomeni (2011) and it is denoted by M_{c_k}. In model M_{c_k}, for each unit, capture probability at some stage j depends only on the capture status of the unit in the previous k occasions. More formally for $k = 1$ in model M_{c_1} we have

$$M_{c_1} : \begin{cases} p(x_{ij} = 1 | x_{ij-1} = 0) = p_{(0)}, & \forall i = 1, \dots N \;\; \forall j = 1, \dots t \\ p(x_{ij} = 1 | x_{ij-1} = 1) = p_{(1)}, & \forall i = 1, \dots N \;\; \forall j = 2, \dots t \end{cases}$$

while for $k = 2$ in model M_{c_2} we have

$$M_{c_2} : \begin{cases} Pr(x_{ij} = 1 | x_{ij-2} = 0, x_{ij-1} = 0) = p_{(00)}, & \forall i = 1, \dots N \;\; \forall j = 1, \dots t \\ Pr(x_{ij} = 1 | x_{ij-2} = 0, x_{ij-1} = 1) = p_{(01)}, & \forall i = 1, \dots N \;\; \forall j = 2, \dots t \\ Pr(x_{ij} = 1 | x_{ij-2} = 1, x_{ij-1} = 0) = p_{(10)}, & \forall i = 1, \dots N \;\; \forall j = 3, \dots t \\ Pr(x_{ij} = 1 | x_{ij-2} = 1, x_{ij-1} = 1) = p_{(11)}, & \forall i = 1, \dots N \;\; \forall j = 3, \dots t \end{cases}$$

For $k = 1, 2$ if $j - k \leq 0$ the conditioning events related to x_{ij-k} are dropped. We remark that in all the models considered so far the probability of never being observed during all t occasions, denoted by P_0, depends only on one parameter. More precisely we have for the previous models

$$\begin{aligned} M_b &: \quad P_0 = (1-p)^t \\ M_{c_1} &: \quad P_0 = (1-p_{(0)})^t \\ M_{c_2} &: \quad P_0 = (1-p_{(00)})^t \end{aligned}$$

As we will see the probability P_0 plays a crucial role in determining the estimate of the population size. Indeed it is also possible to consider an encompassing model which allows for both ephemeral and enduring effects together and it will be denoted with $M_{c_k b}$. It basically consists of a generalized k-th order Markov chain model where, in correspondence of the same conditioning k-th order event $x_{j-k} = 0, \ldots, x_{j-1} = 0$, one distinguishes those histories where a previous first capture has occurred. Only for the partial capture histories formed by k zeroes in the last k occasions we need to specify if a unit is marked or not. In conceiving an appropriate notation for the different capture probabilities the fact that a unit has been captured previously (and hence marked) can be denoted by the digit 0 or 1 before the comma. For example, in model $M_{c_3 b}$, $p_{0,(000)}$ is the probability that a unit is captured at a generic stage j given it is not captured previously and hence it is unmarked; while, $p_{1,(000)}$ is the probability that a unit is captured at time j given it is not captured in the previous $k = 3$ stages but it is captured at least once previously and hence it is marked. Indeed, Yang-Chao's model framework corresponds to $M_{c_1 b}$.

Farcomeni (2011) provides a much more flexible framework based on the capture probabilities conditioned on each possible partial capture history as follows

$$\begin{cases} p_1() = Pr(x_{i1} = 1) \\ p_j(x_{i1}, \ldots, x_{ij-1}) = Pr(x_{ij} = 1 | x_{i1}, \ldots, x_{ij-1}) \quad \forall j > 1 \ , \ \forall (x_{i1}, \ldots, x_{ij-1}) \in \mathcal{X}^{j-1} \end{cases}$$

All these conditional probabilities can be arranged with a natural order in a $2^t - 1$ dimensional vector as follows

$$\mathbf{p} = (p_1(), p_2(0), p_2(1), p_3(0,0), p_3(0,1), p_3(1,0), \ldots, p_t(0, \ldots, 0), \ldots, p_t(1, \ldots, 1))$$

where, for example, the element $p_3(0, 1)$ represents the probability of being captured at time 3 given that the unit is not captured in the first occasion while it is captured in the second occasion. The initial empty brackets () is understood as the absence of previous capture history at time 1. The vector $\mathbf{p}$ can be seen as a convenient reparameterization of the joint probabilities corresponding to all $2^t - 1$ complete capture history configurations in $\mathcal{X}_*^t$. The conditional probabilities, rather than the joint probabilities, are more easily interpreted in the process of modeling the consequences determined by the change of behaviour due to a particular previous trapping history.

Notice that under the saturated reparameterization the probability of never being observed during trapping stages is

$$P_0 = \left[(1 - p_1()) \prod_{j=2}^{t} (1 - p_j(0, \ldots, 0)) \right] \tag{1}$$

From the saturated parametrization one can specify a parsimonious nested model based on a suitable partition of the conditional probabilities in $\mathbf{p}$ in terms of equivalence classes. Let H be the set of all partial capture histories: $H = \{$ (), (0), (1), (00), (10), (01), (11), $\ldots \} = \cup_{j=0}^{t-1} \mathcal{X}^j$ where $\mathcal{X}^0 = \{()\}$. Denote by $\mathcal{H}_B$ one of the possible partitions of H in B disjoint subsets

$$\mathcal{H}_B = \{H_1, \ldots, H_b, \ldots, H_B\}$$

where each $H_b \subset H$. The role of the index set H is to list all the partial capture histories which may yield possible changes in the conditional capture probability depending on the past.

There is a corresponding parameter vector of probabilities denoted with $\mathbf{p}_{\mathcal{H}_B} = (p_{H_1}, \ldots, p_{H_B})$. The partition of capture histories in equivalence classes is such that

$$\forall \ \ \mathbf{h}, \mathbf{h}^{'} \in H_b \quad \Rightarrow \quad p_{(l_h+1)}(\mathbf{h}) = p_{(l_{h'}+1)}(\mathbf{h}^{'}) = p_{H_b} \qquad \forall b = 1, \ldots, B$$

where l_h is the length of the binary vector of a generic partial capture history $\mathbf{h} = (h_1, \ldots, h_{l_h})$. Notice that when there is absence of previous capture history ($\mathbf{h} = ()$) we have $l_h = 0$.

With the partition $\mathcal{H}_B$ of subsets of H representing equivalence classes we make more explicit the fact that the set of very specific constraints formalized in Farcomeni (2011) as $\mathbf{Cp} = \mathbf{0}$ are nothing but a way to identify blocks of conditional probabilities corresponding to the same common value hence reducing the number of free parameters with respect to the saturated model. Indeed in the $\mathbf{Cp} = \mathbf{0}$ formalization the entries of the constraint matrix C must obey further restrictions (only entries -1,0 or 1 and no more that one 1 entry in each column) and this, we believe, is not very natural. No other specific use of those linear constraints are suggested in that paper.

In the following we will denote by $\mathcal{M}$ the class of models based on conditional probabilities parameterization and specified in terms of a suitable partition $\mathcal{H}_B$.

As an example of such formalization based on partitions of subsets of H one can consider a model which assumes that only after being captured for more than 2 times in a row the behaviour of an animal/unit can be affected so that the probability of being trapped again could be lower (or greater). This simple model denoted with $M_{\bullet\bullet}$ can be formalized using the following (bi)partition of the partial capture histories $\mathcal{H}_2(M_{\bullet\bullet}) = \{H_1, H_2\}$ where

$$\begin{cases} H_1 = \{\mathbf{h} \in H : l_h \ < 2\} \cup \left\{\mathbf{h} \in H : l_h \geq 2 \ , \ h_{l_h - 1} + h_{l_h} < 2\right\} \\ H_2 = H \setminus H_1 \end{cases}$$

As another example, we can build up a model, denoted with $M_{\#}$ where the number of captures occurred may influence the capture probability. The corresponding partition denoted with $\mathcal{H}_t(M_{\#})$ splits the set H in t equivalence classes each corresponding to a specific total number of captures as follows

$$\begin{cases} H_1 = \mathcal{X}^0 \cup \left\{h \in \cup_{s=1}^{t-1}\mathcal{X}^s : \sum_{s=1}^{l_h} h_s = 0\right\} \\ H_2 = \left\{h \in \cup_{s=1}^{t-1}\mathcal{X}^s : \sum_{s=1}^{l_h} h_s = 1\right\} \\ \dots \\ H_r = \left\{h \in \cup_{s=1}^{t-1}\mathcal{X}^s : \sum_{s=1}^{l_h} h_s = r-1\right\} \\ \dots \\ H_{t-1} = \left\{h \in \cup_{s=1}^{t-1}\mathcal{X}^s : \sum_{s=1}^{l_h} h_s = t-2\right\} \\ H_t = \left\{h \in \cup_{s=1}^{t-1}\mathcal{X}^s : \sum_{s=1}^{l_h} h_s = t-1\right\} \end{cases}$$

Indeed Farcomeni (2011) provides this general framework where the generic partition is rather specified equivalently in terms of linear constraints on $\mathbf{p}$. This constraints are specified by a $2^t - 1 \times 2^t - 1$ matrix $\mathbf{C}$ as follows

$$\mathbf{Cp} = \mathbf{0}$$

where the generic element of the matrix $\mathbf{C}$ denoted by c_{ij} is such that $c_{ij} \in \{0, 1, -1\}$ with the restriction that each column of $\mathbf{C}$ can not have positive and negative values at the same time. The number of free parameters in the constrained model is the number of columns without negative values which are in one-to-one correspondence with the representative elements of each equivalence class. For example, model M_b can be obtained by using two blocks of equality constraints

$$\begin{cases} p_1() = p_2(0) = p_3(0,0) = \cdots = p_t(0,\dots,0) = p \\ p_2(1) = p_3(10) = p_3(01) = \cdots = p_t(1,\dots,1) = r \end{cases}$$

Equivalently model M_b corresponds to a bipartition $\mathcal{H}_2(M_b) = \{H_1, H_2\}$ such that

$$\begin{cases} H_1 = \{(), (0), (00), \dots, (0\dots0)\} = \mathcal{X}^0 \cup \left\{\mathbf{h} \in \cup_{j=1}^{t-1}\mathcal{X}^j : \sum_{j=1}^{l_h} h_j = 0\right\} \\ H_2 = H \setminus H_1 \end{cases}$$

In the original paper it is also shown that many models proposed in the literature such as model $M_0, M_b, M_{c_k}, M_{c_k b}, M_t$ can be recovered as special cases of model with saturated parameterization $\mathbf{p}$ subject to specific linear constraints corresponding to $\mathbf{C}$.

In the following we prefer to index parameters with the partition notation and we refer to the reduced parametrization $\mathbf{p}_{\mathcal{H}_B} = (p_{H_1}, \dots, p_{H_B})$ corresponding to the uniquely identified conditional probabilities associated to the partition $\mathcal{H}_B$.

3 Conditional likelihood approach and likelihood failure

Under individual independence assumption, the likelihood function can be written in terms of saturated parameterization as follows

$$L(N,\mathbf{p}) = \prod_{i=1}^{N} p_1()^{x_{i1}}(1-p_1())^{1-x_{i1}} \prod_{j=2}^{t} p_j(x_{i1},...,x_{ij-1})^{x_{ij}}$$
$$(1-p_j(x_{i1},...,x_{ij-1}))^{1-x_{ij}}$$

In order to highlight the generality of some pathological likelihood features of behavioural models we focus on the subclass of all models associated to a generic partition $\mathcal{H}_B$ where all the conditioning partial capture histories corresponding to no capture belong to the same partition set, say H_1. This means that all the conditional probabilities $(p_1(), p_2(0), \ldots, p_t(0,\ldots,0))$ determining P_0 as in (1) correspond to the same parameter value. Notice that in the class of models we are considering the first partition set denoted as H_1 can contain also other partial capture histories beside those corresponding to no capture. In the following we will denote by $\tilde{\mathcal{M}}$ this special class of models where, by convention, the first set H_1 listed in the partition $\mathcal{H}_B$ contains (at least) all the aforementioned capture histories defining P_0. Of course $\tilde{\mathcal{M}} \subset \mathcal{M}$.

It is easy to verify that model M_0, M_b, M_{c_k}, $M_{c_k b}$ belong to $\tilde{\mathcal{M}}$.

As an example, model M_{c_1} corresponds to the partition $\mathcal{H}_2(M_{c_1}) = \{H_1, H_2\}$ such that

$$\begin{cases} H_1 = \{(), (0), (00), (10), \ldots\} = \mathcal{X}^0 \cup \left\{\mathbf{h} \in \cup_{j=1}^{t-1}\mathcal{X}^j : h_{l_h} = 0\right\} \\ H_2 = H \setminus H_1 \end{cases}$$

where H_1 contains the void capture history () and all partial capture histories $h = (h_1, \ldots, h_{l_h})$ such that the terminal digit $h_{l_h} = 0$ for $l_h = 1, 2, ..., t-1$. Of course the conditioning capture histories corresponding to no capture are contained in H_1.

On the other hand, model M_t does not belong to $\tilde{\mathcal{M}}$. It can be expressed as $\mathcal{H}_t(M_t) = \{H_1, \ldots, H_t\}$ where $H_j = \mathcal{X}^{j-1}$ for $j = 1, \ldots, t$. Notice also that within the class $\tilde{\mathcal{M}}$ all the models such as M_b and $M_{c_k b}$ do have the first set of the partition H_1 containing all and solely the partial capture histories with no capture i.e. with no 1 digit, while models such as M_{c_k} do have H_1 containing also other partial capture histories (see the particular partition $\mathcal{H}_2(M_{c_1})$ described above).

Hence, for all models belonging to $\tilde{\mathcal{M}}$ it will be

$$P_0 = (1-p_{H_1})^t$$

The likelihood function corresponding to the generic model $M_{\mathcal{H}_B} \in \tilde{\mathcal{M}}$ parametrized with the vector of conditional probabilities $\mathbf{p}_{\mathcal{H}_B}$ will have the following form

$$L(N,\mathbf{p}_{\mathcal{H}_B}) \propto \left[\binom{N}{M} p_{H_1}^{n(H_1 1)}(1-p_{H_1})^{n(H_1 0)+t(N-M)}\right] \prod_{b=2}^{B} p_{H_b}^{n(H_b 1)}(1-p_{H_b})^{n(H_b 0)} \quad (2)$$

where $n_{(H_b0)}$ is the number of times that all the observed units which experience partial capture history $\mathbf{h}$ belonging to H_b are not captured at time l_h+1; similarly $n_{(H_b1)}$ is the number of times that the observed units which experience partial capture history $\mathbf{h}$ belonging to H_b are captured at time l_h+1. Formally $\forall\, b=1,\ldots,B$

$$n_{(H_b0)} = \sum_{i=1}^{M}\sum_{\mathbf{h}\in H_b} I\left[(x_{i1},\ldots,x_{il_h})=\mathbf{h}\ ,\ x_{i(l_h+1)}=0\right]$$

$$n_{(H_b1)} = \sum_{i=1}^{M}\sum_{\mathbf{h}\in H_b} I\left[(x_{i1},\ldots,x_{il_h})=\mathbf{h}\ ,\ x_{i(l_h+1)}=1\right]$$

These are easily recognized as the sufficient statistics in this model framework. The classical estimation procedure considered in Farcomeni (2011) is based on the factorization of the likelihood function in (2) as in Sanathanan (1972) as follows

$$\begin{aligned} L(N,\mathbf{p}_{\mathcal{H}_B}) &\propto \binom{N}{M}(1-P_0)^M P_0^{(N-M)} \times \frac{1}{(1-P_0)^M}\prod_{b=1}^{B} p_{H_b}^{n_{(H_b1)}}(1-p_{H_b})^{n_{(H_b0)}} \\ &= L^r(N,p_{H_1})\times L^c(\mathbf{p}_{\mathcal{H}_B}) \end{aligned}$$

where L^c is the conditional likelihood while L^r is the residual (binomial) likelihood. The conditional maximum likelihood estimator $\hat{N}_{CMLE}$ of N is obtained in 2 steps: first we compute $\hat{\mathbf{p}}_{\mathcal{H}_B}$ maximizing $L^c(\mathbf{p}_{\mathcal{H}_B})$ and then using $\hat{p}_{H_1}\in\hat{\mathbf{p}}_{\mathcal{H}_B}$ maximize $L^r(N,\hat{p}_{H_1})$ with respect to N. Let $q_{H_1}=1-p_{H_1}$; the CMLE of N is given by

$$\hat{N}_{CMLE} = \frac{M}{1-\hat{q}_{H_1}^{\,t}} = \frac{M}{1-\hat{P}_0} \tag{3}$$

where $\hat{q}_{H_1}=1-\hat{p}_{H_1}$ must satisfy the conditional likelihood equation

$$\frac{q_{H_1}}{1-q_{H_1}}\frac{n_{(H_11)}}{M} - \frac{tq_{H_1}^t}{1-q_{H_1}^t} = \frac{n_{(H_10)}}{M}(\equiv R_{H_1}) \tag{4}$$

Equation (4) can be numerically solved and then the estimate $\hat{q}_{H_1}=1-\hat{p}_{H_1}$ is plugged into (3). This corresponds to the Horvitz-Thompson estimator which can be also derived as the classical maximum likelihood estimator of the number of trials in a binomial experiment when the probability of success is known and it is equal to $1-\hat{P}_0$.

However in Seber and Whale (1970) it is pointed out for the first time that in a related removal model the conditional likelihood approach may end up with an unbounded estimate $\hat{N}_{CMLE}$ yielding an annoying inferential pathology called likelihood failure. In the removal model of Seber and Whale (1970), similarly to our models in the class $\tilde{\mathcal{M}}$, units act independently and at each trapping time the capture probability is p and it is the same for each unit. When a unit is captured for the first time it is removed from the population. The likelihood function for the removal model is

$$L_R(N, p) = \binom{N}{M} p^M (1-p)^{n_{0p}+t(N-M)}$$

where n_{0p} is the number of times that observed units are not captured i.e. $n_{0p} = \sum_{i=1}^{M} \sum_{j=1}^{t} I(\sum_{l=1}^{j} x_{il} = 0)$. Notice that the likelihood for a removal model has the same functional form of the factor within brackets in (2) on the right hand side. Since the CML estimation of N and p_{H_1} from (2) depend only on the expression within brackets it could end up with the same pathological unbounded estimates as the removal model.

Notice that the argument which shows that the estimates of N depend only on the expression within brackets makes all models $M_{\mathcal{H}_B} \in \tilde{\mathcal{M}}$ sharing the same element $H_1 \in \mathcal{H}_B$ equivalent in terms of the resulting estimates of N. For instance the partitions corresponding to models M_b and $M_{c_k b}$ do share the same H_1.

Hence we claim that it is important to be aware of the possible occurrence of likelihood failures within general frameworks for behavioural modeling like the one proposed in Farcomeni (2011) once the conditional maximum likelihood is pursued. In particular we show that it is possible to characterize the likelihood failure occurrence for the generic subclass of models $\tilde{\mathcal{M}}$. Adapting from Seber and Whale (1970) we provide the conditions which guarantee the finiteness and the uniqueness of the CML solution in that class of models.

In order to understand the behaviour of the solving Eqs. (3) and (4) consider the left-hand side of (4) as a function f of q_{H_1}

$$f(q_{H_1}) = \frac{q_{H_1}}{1-q_{H_1}} \frac{n_{(H_1 1)}}{M} - \frac{t q_{H_1}^t}{1-q_{H_1}^t}$$

Notice that we always have $n_{(H_1 1)} \geq M$. In fact, the number of times that observed units with partial capture history $h \in H_1$ are not captured at time $l_h + 1$ is at least M. For models such as M_b, $M_{c_k b}$ and $M_{\#}$ the statistic $n_{(H_1 1)}$ is always equal to M. For $0 \leq q_{H_1} < 1$ we have that

$$\frac{df(q_{H_1})}{dq_{H_1}} = \frac{1}{(1-q_{H_1})^2} \left[1 - \frac{t^2 q_{H_1}^{t-1} (1-q_{H_1})^2}{(1-q_{H_1}^t)^2} \right] > 0$$

hence, $f(q_{H_1})$ is an increasing function in $[0, 1)$. Consider the limit of $f(q_{H_1})$ for $q_{H_1} \to 1^-$; we have to distinguish 2 cases

$$\begin{cases} \lim_{q_{H_1} \to 1^-} f(q_{H_1}) = \frac{1}{2}(t-1) & n_{(H_1 1)} = M \\ \\ \lim_{q_{H_1} \to 1^-} f(q_{H_1}) = \infty & n_{(H_1 1)} > M \end{cases}$$

When $n_{(H_1 1)} = M$ there exists a unique solution $0 < q_{H_1} < 1$ if and only if R_{H_1} defined in (4) is such that

$$0 < R_{H_1} < \frac{1}{2}(t-1) \tag{5}$$

In fact, $R_{H_1} > (t-1)/2$ implies that q_{H_1} maximizing the conditional likelihood will be a boundary estimate $\hat{q}_{H_1} = 1$ which implies $\hat{P}_0 = 1$ and hence an infinite estimate of the population size $\hat{N}_{CMLE} = M/(1-\hat{P}_0) = \infty$ (likelihood failure!). Of course restricting q_{H_1} in (0, 1) does not overcome this issue. On the other hand, when $n_{(H_1 1)} > M$ the fact that $\lim_{q_{H_1} \to 1^-} f(q_{H_1}) = \infty$ leads to a unique solution $0 < q_{H_1} < 1$ and hence a finite estimate $\hat{N}_{CMLE}$.

The likelihood failure problem is not overcome by using the unconditional likelihood. The unconditional MLE (UMLE) can be easily derived maximizing $L(N, \mathbf{p}_{\mathcal{H}_B})$ as a function of $\mathbf{p}_{\mathcal{H}_B}$ for N fixed so that once obtained $\hat{\mathbf{p}}_{\mathcal{H}_B}(N)$ one gets the profile likelihood $L_p(N) = L(N, \hat{\mathbf{p}}_{\mathcal{H}_B}(N))$ which can be in turn maximized as a function of N.

In Carle and Strub (1978) within the context of removal model it is pointed out the existence of the likelihood failure also for the unconditional likelihood approach providing the following conditions under which failure occurs

$$M(t-1) - n_{0p} \leq \frac{(M-1)(t-1)}{2} - 1 \Leftrightarrow \frac{n_{0p}}{M} \geq \frac{1}{2}(t-1) + \frac{t-1}{2M} \tag{6}$$

Notice that (6) is less restrictive than condition (5) in the removal model case and hence if failure occurs for the unconditional likelihood approach it occurs also for the conditional likelihood approach while vice-versa is not necessary true.

In order to completely overcome the likelihood failure problem for a removal study, Carle and Strub (1978) proposed to weight the likelihood function with a 2-parameter Beta distribution and then integrate out the nuisance parameter p. It is easy to understand that this procedure is equivalent to locating the posterior mode in a Bayesian approach with an improper non-informative uniform prior on N and a prior Beta distribution for p. In the original paper they show only by simulation how the integrated likelihood approach does not come across the problem of the likelihood failure.

4 Bayesian approach

Motivated by the solution proposed by Carle and Strub (1978) for the removal model we propose to extend the weighted likelihood approach as a fully Bayesian approach for the general class of behavioural models (2) and in particular for models in the class $\tilde{\mathcal{M}}$.

We will make use of Beta densities as convenient conjugate priors for each conditional probability $p_{H_b} \in \mathbf{p}_{\mathcal{H}_B}$. On the other hand we will consider a prior distribution on N as well. As reference recipes we have evaluated 4 non-informative prior distributions on N: Uniform, $1/N$ (Jeffreys'prior), $1/N^2$ and Rissanen's prior which represents a universal non-informative prior for discrete parameters (Rissanen 1983). Since we would like to pursue a fully Bayesian approach we need to verify whether the first two improper priors lead to proper posterior distributions. If this is the case we can fully exploit alternative summaries of the posterior distribution on N. In particular we will consider as alternative summaries the mean, the median, the mode and a minimizer of a specific loss function $\mathcal{L}$ connected with the Relative Mean Square Error (RMSE) as in Tardella (2002)

$$\mathcal{L}(a, N) = \left(\frac{a}{N} - 1\right)^2$$

Let $\pi(N, \mathbf{p}_{\mathcal{H}_B})$ be the joint prior distribution on the whole parameter vector $(N, \mathbf{p}_{\mathcal{H}_B})$ such that

$$\pi(N, \mathbf{p}_{\mathcal{H}_B}) = \pi(N) \times \prod_{b=1}^{B} \pi(p_{H_b}) \propto \pi(N) \times \prod_{b=1}^{B} p_{H_b}^{\alpha_b - 1}(1 - p_{H_b})^{\beta_b - 1} \qquad (7)$$

Hence, given (2) and (7) the joint posterior distribution for model $M_{\mathcal{H}_B}$ is

$$\begin{aligned}
&\pi(N, \mathbf{p}_{\mathcal{H}_B}|\mathbf{X}) \propto L(N, \mathbf{p}_{\mathcal{H}_B})\pi(N, \mathbf{p}_{\mathcal{H}_B}) \propto \\
&\pi(N)\binom{N}{M} p_{H_1}^{n_{(H_1 1)}+\alpha_1 - 1}(1 - p_{H_1})^{n_{(H_1 0)}+t(N-M)+\beta_1 - 1} \prod_{b=2}^{B} p_{H_b}^{n_{(H_b 1)}+\alpha_b - 1} \\
&\quad (1 - p_{H_b})^{n_{(H_b 0)}+\beta_b - 1}
\end{aligned}$$

The choice of Beta densities as prior distributions for the conditional probability parameters makes the marginal posterior distribution of N available in closed form up to a normalizing constant as follows

$$\begin{aligned}
\pi(N|\mathbf{X}) &= \int_0^1 \ldots \int_0^1 \pi(N, \mathbf{p}_{\mathcal{H}_B}|\mathbf{X}) d\mathbf{p}_{\mathcal{H}_B} \\
&\propto \pi(N) \frac{N!}{(N-M)!} B(n_{(H_1 1)} + \alpha_1, t(N-M) + n_{(H_1 0)} + \beta_1) \qquad (8)
\end{aligned}$$

where $B(,)$ is the Beta function. The posterior marginal distribution of N in closed form as in (6) makes it easy to compute quickly all the posterior summaries. We will show how the choice of the prior distribution on N has a relevant impact on the posterior summaries while the sensitivity with respect to the choice of the parameters of the Beta distribution is less relevant. In the following we consider a uniform density on p_{H_b} corresponding to Beta parameters $\alpha_b = \beta_b = 1$, for $b = 1, \ldots, B$. As preliminary step we formally verify whether the choice of improper prior distributions on N such as $\pi(N) \propto 1$ and $\pi(N) \propto 1/N$ leads to a proper marginal distribution on N.

Lemma 1 *Consider a generic model within the class $\tilde{\mathcal{M}}$ parametrized in terms of $p_{\mathcal{H}_B}$. If one chooses independent uniform priors for all its components and a noninformative prior on N with probabilities $\pi(N) \propto 1/N^r$ the Bayes rule always yields a proper posterior distribution for any $r > 0$ while for $r = 0$ the condition $n_{H_1 1} > M$ suffices.*

Proof Considering $\pi(N) \propto 1/N^r$ the posterior marginal distribution of N is proportional to

$$\pi(N|\mathbf{X}) \propto \frac{1}{N^r} \frac{\Gamma(N+1)}{\Gamma(N-M+1)} \frac{\Gamma(t(N-M) + n_{(H_1 0)} + 1)}{\Gamma(t(N-M) + n_{(H_1 0)} + n_{(H_1 1)} + 2)}$$

Using the inequalities

$$(2\pi)^{\frac{1}{2}}\, x^{x-\frac{1}{2}} \exp(-x) \leq \Gamma(x) \leq (2\pi)^{\frac{1}{2}}\, x^{x-\frac{1}{2}} \exp(-x + 1/12x)$$

one gets

$$\frac{\Gamma(N+1)}{\Gamma(N-M+1)} < \mathcal{O}(N^M)\ ; \quad \frac{\Gamma(t(N-M)+n_{(H_10)}+1)}{\Gamma(t(N-M)+n_{(H_10)}+n_{(H_11)}+2)} < \mathcal{O}(N^{-(n_{(H_11)}+1)})$$

Hence, $\pi(N|\mathbf{X}) < \mathcal{O}(N^{M-(r+n_{(H_11)}+1)})$ which corresponds to a proper pmf if and only if $M - (n_{(H_11)} + r + 1) < -1$. We have to distinguish 2 cases: when $n_{(H_11)} > M$ the marginal posterior distribution on N is always proper for any $r \geq 0$ while when $n_{(H_11)} = M$ this is true only for $r > 0$ ◇

From the proof of the previous lemma one can easily argue formally that the Bayes rule always provides an eventually vanishing function in the numerator of (8) for $N \to \infty$. This important result shows that if one makes inference on N maximizing

$$W(N|\mathbf{X}) = \pi(N) \frac{N!}{(N-M)!}\, B(n_{(H_11)} + 1, t(N-M) + n_{(H_10)} + 1) \qquad (9)$$

one will never get unbounded estimates for the finite population size no matter what improper prior is chosen within the class of $1/N^r$ for $r \geq 0$. Hence the Bayesian approach can never provide unbounded estimates in the following sense.

Corollary 1 *Consider a generic model within the class $\tilde{\mathcal{M}}$ parametrized in terms of $\mathbf{p}_{\mathcal{H}_B}$. If one chooses independent uniform priors for all p_{H_b} components and a noninformative prior on N with probabilities $\pi(N) = 1/N^r$ then there exists $\hat{N}_{mode} < \infty$ such that $W(\hat{N}_{mode}|\boldsymbol{X}) \geq W(N|\boldsymbol{X})$ for any N.*

Notice that $\hat{N}_{mode}$ is the mode of the posterior distribution only in those cases where it is well defined otherwise it can be considered only as a weighted likelihood. Hence we claim that from a theoretical inferential point of view the Bayesian approach should be regarded in this context as a favorite inferential tool since it always yields valid inference. Unfortunately it is not easy to get explicit formulas to determine how likely the occurrence of the likelihood failure is. Of course that will depend on the true model and parameter configurations. In the following section we investigate the issue with a little simulation study with replicated data from the same model.

Moreover we will show that even when we remove from the analysis those data which yield likelihood failure the comparative performance of Bayesian output versus conditional maximum likelihood is still always in favor of the former.

5 Simulation study

In order to evaluate the comparative performance of the Bayesian approach with respect to the classical approach based on conditional likelihood we propose a small

Table 1 Parameter configurations for simulation experiments

Trial	Model	Probability parameters	$E[M]/N$
Tr.1	M_b	$p = 0.2; r = 0.4$	0.67
Tr.2	M_b	$p = 0.1; r = 0.3$	0.41
Tr.3	M_{c_1}	$p_{(0)} = 0.2;\ p_{(1)} = 0.4$	0.67
Tr.4	M_{c_1}	$p_{(0)} = 0.1;\ p_{(1)} = 0.3$	0.41
Tr.5	M_{c_2}	$p_{(00)} = 0.2;\ p_{(10)} = 0.3;\ p_{(01)} = 0.35;\ p_{(11)} = 0.4$	0.67
Tr.6	M_{c_2}	$p_{(00)} = 0.1;\ p_{(10)} = 0.2;\ p_{(01)} = 0.3;\ p_{(11)} = 0.4$	0.41

For each parameter configuration $K = 1,000$ datasets have been simulated

simulation study. We consider the set of simulation trials described in Table 1. The true population size is $N = 100$ and the number of trapping occasions is $t = 5$. We evaluate three different kinds of behavioural models within the extended Markovian structure $M_{c_k b}$ following the ideas in Yang and Chao (2005) to account for both enduring and ephemeral effects. Indeed, Markov order is restricted to 2 and we have also excluded $M_{c_1 b}$ and $M_{c_2 b}$ from consideration since they yield inference on N which is identical to model M_b for the reasons we have explained in the previous section. The true (conditional) capture probability parameters for the different simulation trials are chosen so that they correspond to different degrees, from medium-high to medium-low, of expected capture sample coverage defined as the fraction of distinct individuals observed during the t trapping stages, in symbols

$$\frac{E[M]}{N} = 1 - P_0$$

Notice that we have used for each simulated trial the same sequence of pseudo-random numbers so that the observed number of distinct units in each trial is the same when the probabilities p, $p_{(0)}$, and $p_{(00)}$ are the same. To summarize the posterior distribution of the main parameter of interest N we consider the usual mean, median and mode together with the posterior loss minimizer for the loss function described in Sect. 4

$$m_R = \arg\min_a E_{\pi(N|\mathbf{X})}(\mathcal{L}(a, N)).$$

In Table 2 we report the root of the relative mean square error (RMSE) of the estimates of N based on simulations from the correct model. RMSE is evaluated empirically on the basis of $K = 1,000$ replicated datasets for each trial. As we can see the Bayesian approach outperforms the CMLE and UMLE in terms of RMSE. Indeed the occurrence of likelihood failure is reported in the last lines of Table 2 as a percentage of the K datasets. In reporting the estimated RMSE the * sign denotes the presence of likelihood failure so that the RMSE is indeed computed as restricted RMSE conditioning on the absence of failure. This means that RMSE is computed conditioning only on datasets which lead to a finite value of $\hat{N}_{CMLE}$ and $\hat{N}_{UMLE}$. Table 2 allows to assess the comparative performance of alternative choices as far as $\pi(N)$ is concerned. We remark that the choice has some impact on the frequentist performance. Similarly, the choice of posterior summary has a remarkable effect on the precision

Table 2 Simulated data: estimated $\sqrt{RMSE}$ based on 1,000 replicated datasets for each trial. For each simulation setting (column) bold values highlight the best performing estimation method and the corresponding $\sqrt{RMSE}$

Prior	Estimator	Tr. 1	Tr. 2	Tr. 3	Tr. 4	Tr. 5	Tr. 6
$1/N$	Mean	0.999	1.400	0.188	1.265	0.535	2.787
	Median	0.378	0.435	0.163	0.589	0.286	0.594
	Mode	0.173	0.391	0.137	0.288	0.163	0.330
	m_R	0.220	**0.306**	0.145	0.289	0.194	0.288
$1/N^2$	Mean	0.374	0.356	0.163	0.463	0.271	0.454
	Median	0.216	0.323	0.146	0.313	0.195	0.295
	Mode	**0.167**	0.421	**0.132**	0.286	**0.150**	0.345
	m_R	**0.167**	0.350	0.134	**0.262**	0.159	0.291
Rissanen	Mean	0.688	0.807	0.177	0.895	0.410	1.109
	Median	0.293	0.342	0.155	0.445	0.241	0.407
	Mode	0.170	0.407	0.135	0.285	0.156	0.330
	m_R	0.194	0.327	0.140	0.273	0.178	**0.280**
	CMLE	1.149*	1.284*	0.176	0.642*	0.337*	0.652*
	% of $\hat{N}_{CMLE} < \infty$	(99.3%)	(80.5%)	(100.0%)	(98.2%)	(99.8%)	(83.8%)
	% of $\hat{N}_{CMLE} = \infty$	(0.7%)	(19.5%)	(0.0%)	(1.8%)	(0.2%)	(16.2%)
	UMLE	**1**.341*	1.835*	0.166	0.592*	0.280*	0.757*
	% of $\hat{N}_{UMLE} < \infty$	(99.7%)	(86.3%)	(100.0%)	(98.2%)	(99.8%)	(85.2%)
	% of $\hat{N}_{UMLE} = \infty$	(0.3%)	(13.7%)	(0.0%)	(1.8%)	(0.2%)	(14.8%)

of the resulting estimator. Our simulations show that the combination of $\pi(N)$ and summary which produces a better performance corresponds to either one of posterior mode and m_R combined with $1/N^2$. Overall the option m_R with Rissanen shows a more robust behaviour even when they are not the best combination since its RMSE is always close to the best one.

Notice also that the $\hat{N}_{CMLE}$ seems to be more accurate than $\hat{N}_{UMLE}$ in trial 1,2,6 but this is due to the fact that the RMSE are restricted RMSE computed considering different subsets of the K datasets.

We have also considered the performance of alternative approaches with respect to interval estimators. In Table 3 we report the actual percentage of trials in which the 95% interval estimates covered the true value of N and also the average length of the intervals. For the classical approach $1-\alpha$ confidence intervals for the population size are obtained through the profile log-likelihood as (N^-, N^+) where N^- and N^+ are the two roots of the following equation

$$2(\log(L_p(\hat{N})) - \log(L_p(N))) = z^2_{\alpha/2}$$

where $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal and L_p is the profile likelihood. As in Table 2 the * sign denotes the presence of likelihood failure while the $ sign warns that the actual average length is greater than the reported value since we

Table 3 Simulated data: empirical coverage and average length in simulated data of alternative interval estimates with nominal confidence level 0.95 and posterior probability 0.95 respectively

	Interval estimate	Coverage	Average length
Tr.1	Bayes $(1/N^2)$	95.3%	120.87
	Bayes (Rissanen)	96.0%	194.46
	Classical PLI	95.1%*$	$\geq 2,388.94$*$
Tr.2	Bayes $(1/N^2)$	88.0%	163.94
	Bayes (Rissanen)	92.8%	342.04
	Classical PLI	94.5%*$	$\geq 7,201.63$*$
Tr.3	Bayes $(1/N^2)$	94.7%	57.07
	Bayes (Rissanen)	95.1%	59.94
	Classical PLI	94.5%	69.43
Tr.4	Bayes $(1/N^2)$	90.6%	144.4
	Bayes (Rissanen)	97.3%	204.13
	Classical PLI	94.5%*	488.80*
Tr.5	Bayes $(1/N^2)$	94.8%	81.92
	Bayes (Rissanen)	95.7%	97.31
	Classical PLI	94.9%*	175.10*
Tr.6	Bayes $(1/N^2)$	89.8%	172.67
	Bayes (Rissanen)	93.0%	319.40
	Classical PLI	97.2%*$	≥ 855.24*$

have arrested the root finding to an upper-bound $N^+_{upper} = 10,000$. In those cases we have set $N^+ = N^+_{upper}$. In fact in some dataset, although the failure condition is not met the flatness of the profile likelihood prevent us from locating the root N^+ before N^+_{upper}. For the Bayesian approach we have computed the HPD credible set with the same nominal $1 - \alpha$ posterior probability value. The prior $\pi(N) = 1/N^2$ leads to the smallest interval estimates, but the actual coverage is not always sufficiently close to the level $1 - \alpha$ desired for a frequentist match. For trial 2, 4 and 6, characterized by a moderately low sample coverage $E[M]/N$, the coverage of the Bayesian interval estimator corresponding to $\pi(N) = N^{-2}$ is significantly lower than 95 % while this is not true for the Rissanen prior. Even for interval estimate purposes Rissanen's prior represents a good compromise: the average length is reasonably small and the coverage is appropriately close to the nominal frequentist match.

6 Real data

We reanalyze the Great-Copper butterfly dataset originally studied in Ramsey and Severns (2010) to support the use of more flexible behavioural models to account for possibly decreasing/increasing recapture probability patterns likely to occur after the first capture of each unit. It is supposed that butterflies are subject to a change of behaviour which persists with different intensity until the end of trapping stages. In Ramsey and Severns (2010) three alternative models denoted with M_p, M_{pt}, M_{pb}

are proposed and referred to as *persistence models* (see the original paper for a more detailed description of these models). Indeed they do not belong to the class $\mathcal{M}$ of conditional probability models within the framework proposed in Farcomeni (2011). This persistence phenomenon can be considered as a *trap-happiness* response and it can be justified from the fact that butterflies are used to return to the same place where the food is in great quantity. Analogously, researchers are used to return to the same place where they find butterflies. The same dataset is also reviewed in Farcomeni (2011) to show that the class $\mathcal{M}$ is flexible enough to accommodate behavioural models which fit the same data better. The experiment is made of $t = 8$ trapping occasions and the number of distinct butterflies captured during all trapping stages is $M = 45$. In Table 4 we report only the observed complete capture histories associated with the respective frequencies.

We fit several models based on different partitions of the set H some of which correspond to alternative versions of $M_{c_k b}$. Model M_L originally proposed in Farcomeni (2011) considers a 3-rd order Markov-chain-like structure where capture probabilities depend only on the previous three occasions but, differently from the full model M_{c_3} which contains $2^3 = 8$ probability parameters, it considers only 2 parameters corresponding to the following (bi)partition $\mathcal{H}_2(M_L) = \{H_1, H_2\}$ such that

$$\begin{cases} H_1 = \big\{(), (0), (10), (x_1, \ldots, x_{j-4}, 0, 0, 0), \\ \qquad\quad (x_1, \ldots, x_{j-4}, 1, 0, 0), (x_1, \ldots, x_{j-4}, 0, 1, 0), \\ \qquad\quad (x_1, \ldots, x_{j-4}, 1, 1, 0), (x_1, \ldots, x_{j-4}, 0, 0, 1)\big\} \\ \qquad\qquad\qquad\qquad\qquad \forall (x_1, \ldots, x_{j-4}) \in \mathcal{X}^{j-4}; \quad \forall j \geq 4 \\ H_2 = H \setminus H_1 \end{cases}$$

The parameter p_{H_1} corresponding to the first partition identifies a vanishing behavioural effect which occurs if the unit is not captured in the most recent occasion, or captured only once in the last three occasions.

In Table 5 we display point and interval estimates at level 95 % of population size N derived with both classical and Bayesian approach. As described in Sect. 3 the confidence intervals are built considering the normal approximation of the profile log-likelihood while for the Bayesian approach we have proposed the HPD interval. Furthermore in Table 5 in order to drive model selection we report both the AIC index and the log-marginal likelihood associated to each model.

In order to get insights on the pattern of behavioural effects we look at the posterior distribution of $p_{H_2} - p_{H_1}$ for models which involve $\mathbf{p}_{\mathcal{H}_2} = (p_{H_1}, p_{H_2})$ as nuisance parameter.

In Fig. 1 we display the posterior densities of $p_{H_2} - p_{H_1}$ for models M_b, M_{c_1} and M_L. Model M_b which considers only the classical enduring effect to capture provides evidence of trap-shyness. In fact the distribution $p_{H_2} - p_{H_1} = r - p$ is well concentrated almost entirely below the value zero. On the other hand both models M_{c_1} and M_L present trap-happiness effect ($p_{H_2} - p_{H_1} > 0$) more consistent with the underlying biological assumptions.

Following the recommendation suggested by our simulation study we have used Rissanen's prior as prior distribution of N since it yields more convincing results than

Table 4 Great Copper Butterfly data: frequencies of observed capture histories

History	Butterflies
0 0 0 0 0 0 0 1	3
0 0 0 0 0 0 1 0	3
0 0 0 0 0 0 1 1	1
0 0 0 0 0 1 0 0	4
0 0 0 0 1 0 0 0	4
0 0 0 0 1 0 0 1	1
0 0 0 0 1 1 0 0	1
0 0 0 1 0 0 0 0	3
0 0 0 1 0 1 0 0	2
0 0 0 1 1 0 0 0	1
0 0 1 0 0 0 0 0	4
0 1 0 0 0 0 0 0	5
0 1 0 0 0 0 1 0	1
0 1 0 1 0 1 1 0	1
0 1 1 0 0 0 0 0	1
0 1 1 0 1 0 0 0	1
0 1 1 1 1 0 1 1	1
1 0 0 0 0 0 0 0	5
1 1 0 0 0 0 0 0	1
1 1 1 0 0 0 0 0	1
1 1 1 1 1 1 0 0	1

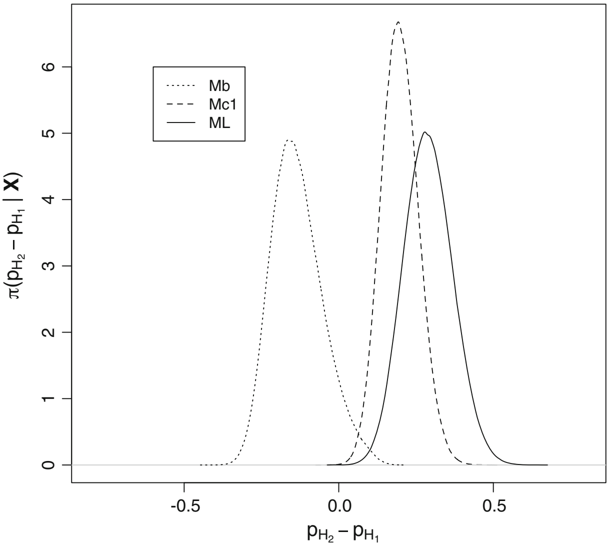

Fig. 1 Great Copper Butterfly data: posterior distribution of $p_{H_2} - p_{H_1}$ for models M_b, M_{c_1} and M_L.

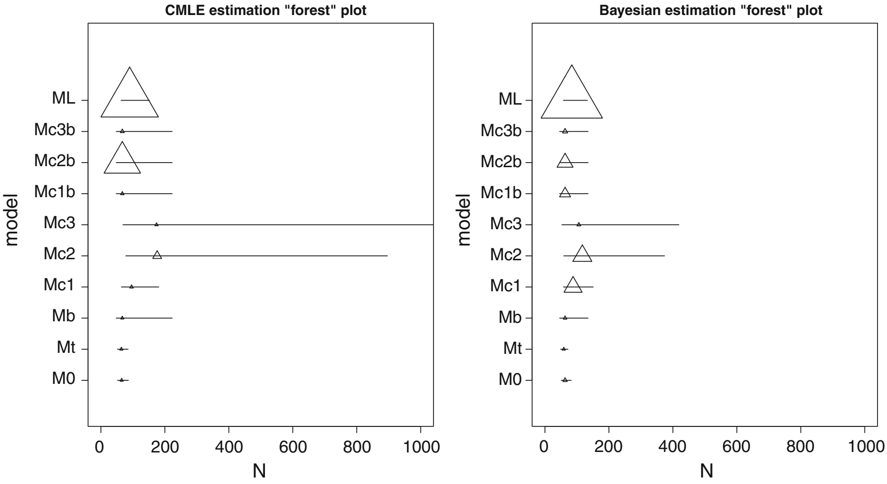

Fig. 2 Great Copper Butterfly data: forest plots for interval estimates of N. Triangles locate point estimates while their sizes are proportional to the amount of comparative evidence evaluated either by AIC or by log-ML

those provided by $\pi(N) = 1/N^2$. From Table 5 one can observe how the Bayesian approach always yields estimates of the population size N which are smaller than CMLE. This is indeed expected from the fact that the Bayesian approach makes full use of the (integrated) unconditional likelihood and the well known monotonicity properties with respect to the estimation based on the conditional likelihood (Sanathanan 1972). From the kind of forest plot in Fig. 2 it is also easy to appreciate that Bayesian approach provides narrower and more stable interval estimates than those provided by a frequentist approach based on the profile likelihood corresponding to comparable $1 - \alpha$ levels. In particular model M_{c_2} and M_{c_3} yield very wide classical confidence intervals which reflect the relative flatness of the profile likelihood. In Table 5 we report for completeness the results of Ramsey and Severns (2010) for their proposed models M_p, M_{pt} e M_{pb} to highlight how instability of classical estimators based on CMLE together with wide confidence intervals may be present also in behavioural models which are outside the $\tilde{\mathcal{M}}$ class of models unraveling that likelihood flatness problems lurks behind.

Notice also that the AIC index and the log marginal likelihood (log-ML) agree on the choice of L_2 as the best model. However, the log-ML gives stronger support than AIC to more parsimonious Markovian models such as M_{c1} and M_{c1b} while it rather penalizes M_{c3} and M_{c3b} which include a higher number of parameters.

7 Concluding remarks

In order to understand behavioural patterns in capture recapture experiments we have focussed on a general class of models following the approach of Farcomeni (2011). Instead of adopting more conventional tools for categorical (binary) data it relies on the reparameterization of the joint probability of the multivariate binary outcome corresponding to the entire individual capture history in terms of subsequent

Table 5 Great Copper Butterfly data: AIC, log marginal likelihood, point and interval estimates

Model	# parameters	Approach	$\hat{N}$	(N^-, N^+)	AIC	Log-ML
M_0	1+1	CMLE	65	(52,86)	336.80	
		Bayesian	63	(51,82)		−174.68
M_t	1+8	CMLE	64	(52,85)	350.84	
		Bayesian	59	(49,72)		−187.14
M_b	1+2	CMLE	67	(48,223)	342.77	
		Bayesian	63	(46,135)		−176.57
M_{c_1}	1+2	CMLE	96	(64,181)	328.92	
		Bayesian	88	(58,151)		−169.63
M_{c_2}	1+4	CMLE	176	(78,896)	326.26	
		Bayesian	117	(59,374)		−169.51
M_{c_3}	1+8	CMLE	174	(69,2315)	330.16	
		Bayesian	106	(53,419)		−175.83
$M_{c_1 b}$	1+3	CMLE	67	(48,223)	329.24	
		Bayesian	63	(46,135)		−170.91
$M_{c_2 b}$	1+5	CMLE	67	(48,223)	324.50	
		Bayesian	63	(46,135)		−169.93
$M_{c_3 b}$	1+9	CMLE	67	(48,223)	328.19	
		Bayesian	63	(46,135)		−173.62
M_L	1+2	CMLE	90	(63,152)	324.01	
		Bayesian	84	(58,133)		−166.91
M_p	1+2	CMLE	97	(70,215)	328.92	
M_{pt}	1+9	CMLE	64	(54,103)	339.46	
M_{pb}	1+2	CMLE	69	(60,1006)	330.16	

conditional probabilities. This is in the same spirit of the so-called transitional model reviewed in Zeng and Cook (2007). Our choice is appropriate since we believe that a behavioural pattern is more easily understood and formalized in terms of conditional probabilities.

We have then pointed out that with the conditional likelihood approach a possible unbounded estimate of the parameter of interest can occur and such pathological inferential feature is indeed shared by a large class of behavioural models both with enduring and ephemeral effects. This phenomenon is rather neglected in the literature since most of the analyses are based on conditional likelihood.

In the literature there are other classes of capture-recapture models where likelihood failure may occur. In particular some parametric and nonparametric heterogeneity models labelled as M_h have been considered in Mao and You (2009) following some critical remarks raised by Link (2003) on model identifiability. They showed with simulated examples similar likelihood pathologies (see Table 5 therein and related comments). However, as said in the introduction, we opted for distinguishing the pathologies derived by the heterogeneity from those due to behavioural effect modeling.

Hence focussing on classes of behavioural models with no heterogeneity such as those derived from the approach of Farcomeni (2011) we have characterized with the subclass $\tilde{\mathcal{M}}$ some models and conditions under which likelihood failure occurs and we have shown that even when there is no likelihood failure the inferential output can be very large and unstable. On the other hand in a very flexible model framework for behavioural patterns we have shown that a fully Bayesian approach is a viable solution which brings a two-fold beneficial effect on inference: i) a simple conjugate structure with closed form expressions for the marginal posterior probabilities $\pi(N|X)$ up to a normalizing constant ii) the complete overcome of unbounded inference under any observed dataset. Since Bayesian inference requires the specification of prior distributions on the unknown parameters we have investigated the sensitivity of the analysis with respect to few alternative default priors using their frequentist properties as performance criterion. Our analysis strongly supports the use of a fully Bayesian approach within the class of models $\tilde{\mathcal{M}}$ based on grouping of the conditional probabilities in equivalence classes. As default choice we advocate the use of uniform priors on the conditional probability parameters and a Rissanen prior on the integer parameter representing the unknown population size N. In our simulations this choice provided improved inference in terms of reduced relative mean square error and shorter interval estimates in the presence of equivalent frequentist coverage. This remains true, although at a lesser extent, when the comparison with unconditional MLE is considered.

An anonymous referee suggested the possibility of using a generalized log-linear parameterization as in Lang (1996) to get Farcomeni's model framework as a particular instance. However we found the implementation of such idea not straightforward and we will look forward to further investigation on that. Indeed we point out the possibility of using a logistic reparameterization of the probability of each binary outcome of the capture history to derive unconditional MLE. In fact one can consider the logit of the conditional probability of each binary outcome regressed as a suitable function of the previous partial capture history. When such function corresponds to a categorical covariate assuming levels corresponding to each equivalence class the derivation of the unconditional MLE can be easily carried out by maximizing the profile likelihood of N. In fact for each value of N one can augment the observed capture histories with $N - M$ histories corresponding to units which were not observed and obtain the profile likelihood corresponding to N from the standard output of GLM routines of any statistical software. A similar logistic model structure has been previously sketched in Huggins (1989) and Alho (1990) although the focus there was in developing conditional likelihood estimators in the presence of individual covariates different from partial capture histories.

We believe that the generality of the pathological features of classical likelihood analysis (CMLE and UMLE) of behavioural capture recapture models suggests a wider use of Bayesian alternative analysis even in those more realistic and complex frameworks such as, for instance, those developed in Bartolucci and Pennoni (2007) where latent Markov structure is embedded to model more flexibly ephemeral effects and heterogeneity of individual capture probability.

Acknowledgements We would like to thank an anonymous referee for his/her stimulating comments.

References

Alho JM (1990) Logistic regression in capture-recapture models. Biometrics 46:623–635

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, Pennoni F (2007) A class of latent Markov models for capture–recapture data allowing for time, heterogeneity, and behavior effects. Biometrics 63(2):568–578

Carle FL, Strub MR (1978) A new method for estimating population size from removal data. Biometrics 34(4):621–630

Chaiyapong Y, Lloyd CJ (1997) Accurate inference for recapture experiments with behavioural response. J Stat Comput Simul 56(2):97–115

Chao A, Chu W, Hsu CH (2000) Capture–recapture when time and behavioral response affect capture probabilities. Biometrics 56(2): 427–433

Farcomeni A (2011) Recapture Models under Equality Constraints for the Conditional Capture Probabilities. Biometrika, 98(1): 237–242

Fattorini L, Marcheselli M, Monaco A, Pisani C (2007) A critical look at some widely used estimators in mark-resighting experiments. J Anim Ecol 76(5):957–965

Ghosh SK, Norris JL (2005) Bayesian capture–recapture analysis and model selection allowing for heterogeneity and behavioral effects. J Agric Biol Environ Stat 10(1):35–49

Huggins RM (1989) On the statistical analysis of capture experiments. Biometrika 76(1):133–140

Huggins RM (1991) Some practical aspects of a conditional likelihood approach to capture experiments. Biometrics 47(2):725–732

Hwang WH, Huggins R (2011) A semiparametric model for a functional behavioural response to capture in capture-recapture experiments. Aust NZ J Stat 53(4):403–421

Hwang WH, Chao A, Yip PSF (2002) Continuous-time capture-recapture models with time variation and behavioural response. Aust NZ J Stat 44(1):41–54

Lang JB (1996) Maximum likelihood methods for a generalized class of log-linear models. Ann Stat 24(2):726–752

Lee SM, Chen CWS (1998) Bayesian inference of population size for behavioral response models. Statistica Sinica 8(4):1233–1248

Lee SM, Hwang WH, Huang LH (2003) Bayes estimation of population size from capture-recapture models with time variation and behavior response. Statistica Sinica 13(2):477–494

Link WA (2003) Nonidentifiability of population size from capture–recapture data with heterogeneous detection probabilities. Biometrics 59(4):1123–1130

Mao CX, You N (2009) On comparison of mixture models for closed population capture–recapture studies. Biometrics 65(2):547–553

Otis DL, Burnham KP, White GC, Anderson DR (1978) Statistical inference from capture data on closed animal populations. Wildlife Monographs,

Ramsey F, Severns P (2010) Persistence models for mark-recapture. Environ Ecol Stat 17(1):97–109

Rissanen J (1983) A universal prior for integers and estimation by minimum description length. Ann Stat 11(2):416–431

Sanathanan L (1972) Estimating the size of a multinomial population. Ann Math Stat 43(1):142–152

Seber GAF, Whale JF (1970) The removal method for two and three samples (Corr: V27 p1104). Biometrics 26(3):393–400

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

Yang HC, Chao A (2005) Modeling animals' behavioral response by Markov chain models for capture-recapture experiments. Biometrics 61(4):1010–1017

Yip PSF, Xi L, Chao A, Hwang WH (2000) Estimating the population size with a behavioral response in capture-recapture experiment. Environ Ecol Stat 7(4):405–414

Zeng L, Cook RJ (2007) Transition models for multivariate longitudinal binary data. J Am Stat Assoc 102(477):211–223