Research Repository

A Pólya Tree modelling framework for batch-mark data

A PÓLYA TREE MODELLING FRAMEWORK FOR BATCH-MARK DATA

BY IOANNIS ROTOUS[1,a], ALEX DIANA[2,b], AND ELENI MATECHOU[1,c]

[1]*Department of Statistical Science, University College London,* [a]*i.rotous@ucl.ac.uk*

[2]*Department of Mathematics, Statistics and Actuarial Science, University of Essex,* [b]*ad23269@essex.ac.uk*

[3]*Statistical Ecology @ Kent, School of Mathematics, Statistics and Actuarial Science, University of Kent,* [c]*e.matechou@kent.ac.uk*

Wildlife population surveys typically consist of multiple sampling occasions, where individuals are followed over time, enabling estimation of population size and, in open populations, of entry and exit patterns. Batch-mark (BM) surveys, where newly sampled individuals are given the same marking, often unique for each sampling occasion or each sampling period but not for each individual, provide the only viable monitoring tool for many species of amphibians, birds and fish. Modelling BM data for open populations has proven more challenging than modelling data where individuals are uniquely marked, and approaches proposed in the literature thus far rely on approximate inference or do not scale well with the number of individuals, and do not readily extend to the joint modelling of different observation processes often employed in practice. In this paper, we propose a novel approach for modelling BM data, by defining a bivariate grid for modelling the latent entry and exit patterns, as well as population size. We employ the Bayesian nonparametric Pólya Tree (PT) prior for defining a model on the grid cells, which enables exact and highly efficient Bayesian inference on the number of individuals in each cell, and hence of the population size and the entry/exit pattern. Our approach scales with the number of sampling occasions, instead of the number of individuals, and allows us to easily write the likelihood function for BM data under different observation processes. We demonstrate our new PT Batch Mark (PTBM) approach using extensive simulations and two case studies, comparing its performance with two recently proposed approaches.

1. Introduction.

Ecological monitoring is of paramount importance in safeguarding our natural ecosystems. By studying wildlife populations and accurately estimating essential demographic parameters, such as population size, birth/death rates or phenological patterns, we can gain invaluable insights into their dynamics. One approach that can be employed for population monitoring is batch-marking (BM), which involves repeated sampling occasions when individuals are physically caught, with newly caught individuals marked using a sampling occasion-specific mark, which is a mark that is typically different among sampling occasions but not among individuals. BM surveys are particularly useful in cases where individual marking cannot be employed, such as with species that are highly abundant or small in size, for example fish or insects, and have been extensively utilized providing valuable data for population monitoring (Cowen et al., 2017; Vavassori, Saddler and Müller, 2019; Davidson et al., 2019; Doll et al., 2021; Rosser, Willden and Loeb, 2022).

BM data on each sampling occasion consist of the number of individuals caught that are unmarked and the number of individuals caught that were first marked on previous sampling occasions. Therefore, as opposed to standard capture-recapture studies (McCrea and Morgan,

Keywords and phrases: Bayesian nonparametrics, Stopover model, Population size, Removal models, Survival probability.

2014; Seber and Schofield, 2019; King and McCrea, 2019), where individuals are uniquely marked, it is not known how many times and on which sampling occasions each individual is re-caught. This aggregated nature of BM data means that likelihood-based inference is more challenging, and the literature on appropriate models for BM data is limited. To discuss existing approaches we need to introduce the concept of individual presence histories and capture histories. Both are vectors of length equal to the number of sampling occasions. The former indicates when the corresponding individual is present at the surveyed site, while the latter indicates when the individual has been caught. Individual presence histories are always latent in ecological data, while individual capture histories are observed in capture-recapture data but latent in BM data. Current approaches for modelling BM data include the work by Huggins, Wang and Kearns (2010), who derived estimating equations and closed-form solutions for survival and capture probabilities, along with abundance estimates using a Horvitz-Thompson-type estimator. Another important contribution came from Cowen et al. (2014), who developed a tractable likelihood function for marked individuals only, but with an associated computational cost that increases substantially with more sampling occasions, since the calculation of the likelihood involves nested summations of all possible individual latent presence histories. The Cowen et al. (2014) work was further extended by Cowen et al. (2017) who employed Hidden Markov Models (HMM) to model BM data for both marked and unmarked individuals, but this HMM approach does not scale well with the number of individuals, since it involves high dimensional state-transition and state-dependent probability matrices. More recently, Zhang, Bonner and McCrea (2023) introduced an innovative approach to approximating the likelihood function for BM data under the robust design (RD, Kendall and Pollock, 1992). When the RD is employed, it is assumed that the population is open between primary periods, e.g. months, but closed within secondary periods, e.g. days within a month, so that individuals can enter/exit between primary periods but not between secondary periods. Zhang, Bonner and McCrea (2023) proposed an approach that relies on the saddlepoint approximation (SPA Zhang, Bravington and Fewster, 2019) to the likelihood function, which requires reconstructing the latent capture history as well as the latent presence history for each individual, so the computational burden increases considerably with the number of sampling occasions.

In this paper, we present a novel Bayesian nonparametric approach utilizing the Pólya Tree (PT) prior for BM data analysis. This approach, which builds on the work by Diana et al. (2023) for count and ring-recovery data, offers numerous benefits. The fundamental idea is that we build a lower-right triangular grid with $\frac{(K+1)(K+2)}{2}$ latent cells, as shown in Figure 1, where K is the number of sampling occasions. The grid cells represent the latent number of individuals with a specific combination of entry and exit intervals, where the intervals are defined by pairs of consecutive sampling occasions, so that $n_{i,j}$ is the number of individuals with entry between sampling occasions i and $i+1$ and exit between sampling occasions j and $j+1$. We note that entry and exit can correspond to birth/death, arrival/departure, or any other process that introduces/removes individuals from the population. By using our proposed grid approach and corresponding PT prior, as we discuss below, we naturally account for the aggregated nature of BM data because we do not need to infer, impute or marginalize over latent individual presence histories or capture histories, and instead only need to infer the latent cells of the grid. Consequently, within this framework, we overcome previous challenges related to BM model inference and establish a tractable likelihood function and a corresponding efficient model-fitting algorithm for standard BM data, as well as BM data collected under different sampling designs, as we discuss in the two case studies of the paper.

Inference of the cells $n_{i,j}$ is efficient and flexible within a Bayesian framework based on the PT prior, which allows us to define and infer the grid probabilities. By relying on the PT

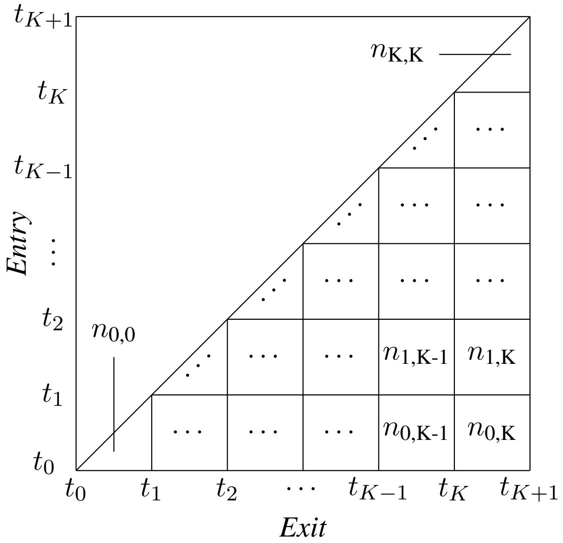

Fig 1: Entry and exit sample space for K sampling occasions, taking place at times $t_1, t_2, ..., t_K$ with convention that $t_0 = -\infty, t_{K+1} = \infty$. The latent number of individuals in cell (i, j), that is with entry between the ith and $(i+1)$th and exit between the jth and $(j+1)$th sampling occasions, is denoted by $n_{i,j}$, with $i = 0, 1, ..., K$ and $j = i, ..., K$.

prior, we can build a model directly on the distributions of entry and exit patterns, with minimal parametric assumptions on the shape of these distributions. Additionally, the replicate PT framework, first introduced in Diana et al. (2023), allows us to impose constraints on the entry and exit processes, leading to more parsimonious models, as we discuss in Section 3.

We present results for two case studies previously analysed in the literature. Both consider open populations and fit models of Jolly-Seber (JS Jolly, 1965; Seber, 1965) type, which are models that allow direct estimation of population size while accounting for entry and exit from the population. The data from the two case studies result from different observation processes. The first case study considers data on weather loaches (*Misgurnus anguillicaudatus*), which are freshwater fish, monitored in South-Eastern Australia. In this case, during the BM survey, individuals caught as unmarked are either marked and returned to the population in the standard BM approach, or are removed from the population, akin to removal sampling (Matechou et al., 2016). We show how our framework can easily be adapted to model the two processes jointly, which was not considered by Cowen et al. (2017). A corresponding simulation study demonstrates the resulting improvement in inference when the removal process is accounted for. The second case study considers data on golden mantella frogs (*Mantella aurantiaca*), collected in Central Madagascar under the RD across six primary periods. In this case, we demonstrate how our framework can easily account for BM data collected under a RD sampling, using the exact likelihood, and compare our results to those obtained by Zhang, Bonner and McCrea (2023). A corresponding simulation study explores issues around allocation of effort within a RD framework for BM surveys.

The article is organized as follows: In Section 2 we introduce our modelling approach of the latent entry/exit pattern and population size from BM data, and in Section 3 we describe the PT prior, which provides the foundation of our modelling framework. We present simulation and real data results for each of the two case studies in Sections 4 and 5. The article concludes with a discussion of the findings and outlines potential avenues for future research in Section 6.

2. Model for batch-mark data. First, we introduce the notation and modelling of standard BM data, with K sampling occasions, taking place at times $t_1, ..., t_K$, with $t_0 = -\infty, t_{K+1} = \infty$, for convenience. Additional notation for the two different observation processes employed in the case studies is introduced in the corresponding sections. In what follows, we refer to individuals in (grid) cell (i, j) as those individuals that entered between sampling occasions i and $(i+1)$ and exited between sampling occasions j and $(j+1)$.

Data

u_t : observed number of unmarked individuals caught on sampling occasion t.
$m_{k,t}$: observed number of individuals that were marked on sampling occasion k and were recaptured on sampling occasion t, with $m_{k,t} = 0$ for $t \leq k$.

Parameters

N : population size, with $\sum_{i=0}^{K} \sum_{j=i}^{K} n_{i,j} = N$ (see Figure 1). We model N as a Poisson random variable with mean ω, $N \sim \text{Poisson}(\omega)$.
$w_{i,j}$: probability of an individual belonging to cell (i, j).
p_t : probability of capturing an individual that is present on sampling occasion t.

Latent

$\mathbf{n}$: $(K+1) \times (K+1)$ matrix where $n_{i,j}$ corresponds to the latent number of individuals in cell (i, j). The latent cell counts $n_{i,j}$, conditional on N are modelled as

$$\{n_{i,j}\}_{i=0,\ldots,K, j=i,\ldots,K} \, |N \sim \text{Multinomial}(N, \{w_{i,j}\}_{i=0,\ldots,K, j=i,\ldots,K})$$

so that, since $N \sim \text{Poisson}(\omega)$, the marginal distribution of the counts in the cells of matrix $\mathbf{n}$ is given by

$$n_{i,j} \sim \text{Poisson}(\omega \times w_{i,j}), \; i = 0, \ldots, K, \; j = i, \ldots, K \tag{1}$$

$\mathbf{U^k}$: $(K+1) \times (K+1)$ matrix where $U_{i,j}^k$ corresponds to the latent number of individuals in cell (i, j) that are present and unmarked on sampling occasion k.
$\mathbf{M^k}$: $(K+1) \times (K+1)$ matrix where $M_{i,j}^k$ corresponds to the number of individuals in cell (i, j) that were caught as unmarked (and subsequently marked) on sampling occasion k. Each of the latent $U_{i,j}^k$ individuals can be caught with probability p_k, independent of other individuals, so that

$$M_{i,j}^k \sim \text{Binomial}(U_{i,j}^k, p_k)$$

We note that for $t \leq i$ or $t > j$, $U_{i,j}^t = 0$ (the latent number of individuals in cell (i, j) that are unmarked on sampling occasion t) by definition, since individuals from cell i, j are not present on sampling occasion t in this case. Otherwise, for $i < t \leq j$, $U_{i,j}^t$ is equal to the difference between the total latent number of individuals in that cell and the latent number of individuals in that cell that were marked before sampling occasion t, so that

$$U_{i,j}^t = n_{i,j} - \sum_{k=1}^{t-1} M_{i,j}^k, \text{ with } i < t \leq j, \; t > 1.$$

Note that when $t = 1$ there are no marked individuals yet and hence $U_{i,j}^1 = n_{i,j}$ for $i < 1 \leq j$, and 0 otherwise.

Therefore, conditional on the latent matrices $\mathbf{M^k}$, the observed number of unmarked individuals caught on sampling occasion t, u_t, is

$$u_t = \sum_{i=0}^{t-1} \sum_{j=t}^{K} M_{i,j}^t .$$

The number of individuals marked on sampling occasion k that are still present on sampling occasion t is equal to $\sum_{i=0}^{k-1} \sum_{j=t}^{K} M_{i,j}^k$ for $i < k < t \leq j$, that is the sum of $M_{i,j}^k$ terms for individuals that entered before k and have not yet exited at t and each of these individuals is recaptured on sampling occasion t with probability p_t, independent of other individuals, so that the observed number of individuals marked on sampling occasion k and recaptured on sampling occasion t is distributed as

$$m_{k,t} \sim \text{Binomial}\left(\sum_{i=0}^{k-1} \sum_{j=t}^{K} M_{i,j}^k, p_t \right), \; i < k < t \leq j$$

The complete model for the standard BM data is given in Equation 2.

(2)

latent number of individuals in cell (i,j)	$n_{i,j} \sim \text{Poisson}(\omega \times w_{i,j}), \; i = 0, ..., K, \; j = i, ..., K$
latent number of individuals in cell (i,j) that are unmarked on sampling occasion t for $i < t \leq j$	$U_{i,j}^t = \begin{cases} n_{i,j}, & t = 1, \\ n_{i,j} - \sum_{k=1}^{t-1} M_{i,j}^k, & t > 1, \end{cases}$
latent number of individuals in cell (i,j) caught as unmarked and subsequently marked on sampling occasion t	$M_{i,j}^t \sim \text{Binomial}(U_{i,j}^t, p_t),$
observed number of unmarked individuals caught on occasion t	$u_t = \sum_{i=0}^{t-1} \sum_{j=t}^{K} M_{i,j}^t, \; t \geq 1$
observed number of individuals that were marked on sampling occasion k and were recaptured on sampling occasion t, for $i < k < t \leq j$	$m_{k,t} \sim \text{Binomial}\left(\sum_{i=0}^{k-1} \sum_{j=t}^{K} M_{i,j}^k, p_t \right).$

We note that, in our model, N is not the equivalent of the super-population of individuals, N^S, that became available for capture at least once (Schwarz and Arnason, 1996). Instead, N also accounts for individuals that entered and exited without ever becoming available for capture (Matechou and Caron, 2017), that is individuals in cells with $i = j$. Inference on these individuals is possible within this framework thanks to the PT prior, as also discussed in Diana, Griffin and Matechou (2019). However, inference on the super-population size is also readily available from the $\mathbf{n}$ matrix, since $N^S = N - \sum_{i=0}^{K} n_{i,i}$, i.e. by excluding individuals that never became available for capture, and in the case studies we report the super-population size.

A simulation study for this standard BM model is presented in Section B of the Supplementary Material (Rotous, Diana and Matechou, 2025).

3. Pólya Tree Prior. Inferring the grid matrix $\mathbf{n}$ relies on inferring the grid probabilities $w_{i,j}$, through the application of the PT prior. The PT prior requires partitioning the sample space, which in our case is the entry and exit space, the latter conditional upon entry, and subsequently, specifying a sequence of parameters to assign probabilities to each set in these partitions. Specifically, this partitioning is achieved by first splitting the sample space into the sets $B_i = (t_i, t_{i+1}] \times (t_i, t_{K+1})$ for $i = 0, 1, 2, ..., K$, corresponding to the individuals entering in the interval between the ith and $(i+1)$th sampling occasions. The probability of an individual belonging to B_i is V_i, with $V_0 + V_1 + ... + V_K = 1$. Next, we split each entry set, B_i into the sets $B_{i,j} = (t_i, t_{i+1}] \times (t_j, t_{j+1}],\ j = i, ..., K$, where $(t_i, t_{i+1}] \times (t_j, t_{j+1}]$ corresponds to the individuals entering between the ith and $(i+1)$th sampling occasion and exiting between the jth and $(j+1)$th sampling occasion. The probability of an individual belonging to $B_{i,j}$ conditional on being in B_i is $V_{i,j}$, with $V_{i,i} + V_{i,i+1} + \ldots + V_{i,K} = 1$. We note that our framework can consider individuals that never became available for capture, i.e. individuals that belong to sets $(t_i, t_{i+1}] \times (t_i, t_{i+1}]$.

The PT prior provides an efficient approach for calculating the probabilities $\{V_i\}_{i=0}^{K}$ and $\{V_{i,j}\}_{i=0,j=i}^{K,K}$ through conditional probabilities. We can express the entry probabilities as $V_0 = \tilde{V}_0, V_1 = \tilde{V}_1(1-\tilde{V}_0), ..., V_K = (1-\tilde{V}_{K-1})\ldots(1-\tilde{V}_0)$, with $\tilde{V}_i$ corresponding to the conditional probability of an individual entering between the ith and $(i+1)$th sampling occasions, given that the individual has not entered before the ith sampling occasion. The conditional probabilities are assumed i.i.d. and distributed as $\tilde{V}_i \sim \text{Beta}(a_{i,0}, a_{i,1})$. The exit probabilities, conditional on entry between the ith and $(i+1)$th sampling occasions, can be expressed as $V_{i,i} = \tilde{V}_{i,i}, V_{i,i+1} = \tilde{V}_{i,i+1}(1-\tilde{V}_{i,i}), ..., V_{i,K-1} = \tilde{V}_{i,K-1}(1-\tilde{V}_{i,K-2})...(1-\tilde{V}_{i,i}), V_{i,K} = (1-\tilde{V}_{i,K-1})...(1-\tilde{V}_{i,i})$, with $\tilde{V}_{i,j}$ corresponding to the conditional probability of an individual exiting between the jth and $(j+1)$th sampling occasion, given that the individual is present on the jth sampling occasion. The conditional probabilities are assumed i.i.d. and distributed as $\tilde{V}_{i,j} \sim \text{Beta}(a_{i,j,0}, a_{i,j,1})$.

Finally, the PT prior distribution of the probabilities $w_{i,j}$, in terms of the conditional entry/exit probabilities is defined as

$$\{w_{i,j}\} = \{V_i V_{i,j}\} \sim PT(\{a_{i,0}, a_{i,1}\}, \{a_{i,j,0}, a_{i,j,1}\}),\ i = 0, 1, 2, ..., K \quad j = i, ..., K \tag{3}$$

In this PT prior framework we can impose constraints on population dynamic parameters, that is the entry and exit probabilities, which is standard practice in JS-type models (Schwarz and Arnason, 1996). For example, to impose a constraint that the probability of entry is constant, i.e. time-invariant, we set $\tilde{V}_0 = \tilde{V}_1, = ... = \tilde{V}_{K-1}$. Similarly, to place constraints on the conditional exit probabilities, we use the idea of the replicated PT prior described in Diana et al. (2023). This corresponds to assuming that the conditional exit probabilities satisfy the constraint $\tilde{V}_{0,j} = \tilde{V}_{1,j} = ... = \tilde{V}_{j,j},\ j = 0, ..., K-1$, for each j i.e. we replicate the conditional exit probabilities across the entry intervals, so that the conditional probability of exit in a given interval is the same for all individuals, regardless of entry. As a result of using these constraints we can accommodate different JS-type models and derive parsimonious models, since we reduce the number of inferred parameters. For instance, we can accommodate models where individuals have conditional entry/exit probabilities that either change over time or remain constant across the sampling occasions. We could further impose age dependent constraints, or simply no constraints at all, as described in Diana et al. (2023).

The models are presented throughout in their most general, unconstrained, form. However, in the case studies considered in this paper in Sections 4.5 and 5.5, we impose constraints that agree with previous analyses of the data to make the corresponding results comparable to those obtained by Cowen et al. (2017) and Zhang, Bonner and McCrea (2023). First, we exclude individuals never exposed to capture by only considering individuals in cells i, j of matrix $\mathbf{n}$ for $i = 0, \ldots, K-1$ and $j = i+1, \ldots, K$. Then, we note that the entry probabilities of our model, V_i, and the conditional entry probabilities, $\tilde{V}_i$, are denoted in Zhang, Bonner and McCrea (2023) by β_i and β_i^*, respectively.

This parametrization implies that the conditional entry probabilities are time-dependent. On the other hand, Cowen et al. (2017) assume a time-invariant entry rate (η) between sampling occasions, which in our model can be enforced by assuming that $\tilde{V}_i = \tilde{V}$. We discuss how η, as defined by Cowen et al. (2017), can be inferred from our model output in Section D of the Supplementary Material (Rotous, Diana and Matechou, 2025).

Finally, we introduce the survival probabilities, $\phi_{i,j}$ corresponding to the probability that an individual with entry between the ith and $(i+1)$th sampling occasions, remains in the population between the jth and $(j+1)$th sampling occasions conditional on being present on the jth sampling occasion (Cowen et al., 2017; Zhang, Bonner and McCrea, 2023), so that $\tilde{V}_{i,j} = 1 - \phi_{i,j}$. Zhang, Bonner and McCrea (2023) assume a time-dependent constraint on survival probabilities so that $\phi_{i,j} = \phi_j,\ i = 0, 1, \ldots, j$ which we can impose on $\tilde{V}$'s with the replicated PT, described in Diana et al. (2023), by assuming $\tilde{V}_{0,j} = \tilde{V}_{1,j} = \ldots = \tilde{V}_{j-1,j} = 1 - \phi_j$. Finally, Cowen et al. (2017) assume a time-invariant survival probability so that $\tilde{V}_{i,j} = 1 - \phi$.

3.1. *Inference.* We describe our employed Markov Chain Monte Carlo (MCMC), with corresponding conditional distributions, where appropriate, in Section A of the Supplementary Material (Rotous, Diana and Matechou, 2025). Briefly, the latent matrices $\{\mathbf{M^k}\}_{k=1}^{K}$ and $\mathbf{n}$ are updated using a standard Metropolis-Hastings (MH) random walk sampler (Robert and Casella, 2004) whereas the $\{w_{i,j}\}_{i=0,j=i}^{K,K}$ parameters are updated using a Gibbs algorithm (Gelfand and Smith, 1990) since we exploit the conjugacy properties of the PT prior (Lavine, 1992). Parameters $\{p_t\}_{t=1}^{K}$ and ω are also updated with the use of a Gibbs sampler.

4. Case Study 1.

BM data on weather-loch were collected across 11 sampling occasions, but in contrast to standard BM sampling, a proportion of unmarked individuals caught on each sampling occasion were removed from the population (Cowen et al., 2017). Here we demonstrate how our modelling framework introduced in Section 2 can naturally account for removals on captures in BM surveys. From Section 2, we make use of data $\{u_t\}_{t=1}^{K}, \{m_{k,t}\}_{k=1,t=k+1}^{K-1,K}$, parameters $N, \{w_{i,j}\}_{i=0,j=i}^{K,K}, \{p_t\}_{t=1}^{K}$, and latent terms $\mathbf{n}, \{\mathbf{M^k}\}_{k=1}^{K}$ and we introduce additional data, parameters and latent terms corresponding to removals per sampling occasion.

4.1. *Data and Notation.* First we introduce some case-study-specific notation.

Data

r_k : observed number of individuals that were removed from the population on sampling occasion k.

Parameters

l_k : probability of removing an unmarked individual caught on sampling occasion k.

Latent

$\mathbf{Y^k}$: $(K+1) \times (K+1)$ matrix where $Y_{i,j}^k$ corresponds to the latent number of unmarked individuals in cell (i, j) caught on sampling occasion k. We note that this is not the same as $M_{i,j}^k$ introduced in Section 2 since not all newly caught individuals are marked in this case, as we discuss below. Similarly to the standard BM model described in Section 2,

$$Y_{i,j}^k \sim \text{Binomial}(U_{i,j}^k, p_k),\ i < k \leq j$$

$\mathbf{R^k}$: $(K+1)\times(K+1)$ matrix where $R^k_{i,j}$ corresponds to the latent number of individuals in cell (i,j) that were removed on sampling occasion k. Each of the $Y^k_{i,j}$ individuals caught as unmarked has the same probability, l_k, independent of other individuals, of being removed instead of being marked and returned to the population, so that

$$R^k_{i,j} \sim \text{Binomial}(Y^k_{i,j}, l_k)$$

We note that now the number of individuals in cell (i,j) that are present and marked on sampling occasion t, for $i < t \leq j$, is $M^t_{i,j} = Y^t_{i,j} - R^t_{i,j}$, and 0 otherwise, and the latent number of individuals in cell (i,j) that are available as unmarked on sampling occasion t is $U^t_{i,j} = n_{i,j} - \sum_{k=1}^{t-1} M^k_{i,j} - \sum_{k=1}^{t-1} R^k_{i,j},\ \ i < t \leq j$ since individuals in cell (i,j) are no longer available as unmarked on sampling occasion t if they were previously marked or removed from the population.

Therefore it follows that, conditional on $R^k_{i,j}$, the observed number of individuals removed from the population on sampling occasion k are

$$r_k = \sum_{i=0}^{k-1}\sum_{j=k}^{K} R^k_{i,j}$$

4.2. *Model.* The model of Equation (2) is extended to account for the removal process, in addition to the standard BM process, as shown in Equation (4).

(4)

latent number of individuals in cell (i,j)	$n_{i,j} \sim \text{Poisson}(\omega \times w_{i,j}),\ i = 0,...,K,\ j = i,...,K$
latent number of individuals in cell (i,j) that are unmarked on sampling occasion t for $i < t \leq j$	$U^t_{i,j} = \begin{cases} n_{i,j},\ t = 1, \\ n_{i,j} - \sum_{k=1}^{t-1} M^k_{i,j} - \sum_{k=1}^{t-1} R^k_{i,j},\ t > 1, \end{cases}$
latent number of individuals in cell (i,j) caught as unmarked on sampling occasion t	$Y^t_{i,j} \sim \text{Binomial}(U^t_{i,j}, p_t)$
observed number of unmarked individuals caught on sampling occasion t	$u_t = \sum_{i=0}^{t-1}\sum_{j=t}^{K} Y^t_{i,j},\ t \geq 1$
latent number of individuals in cell (i,j) removed on sampling occasion t	$R^t_{i,j} \sim \text{Binomial}(Y^t_{i,j}, l_t),$
observed number of individuals removed on sampling occasion t	$r_t = \sum_{i=0}^{t-1}\sum_{j=t}^{K} R^t_{i,j}\ ,\ t \geq 1$
latent number of individuals in cell (i,j) marked on sampling occasion t	$M^t_{i,j} = Y^t_{i,j} - R^t_{i,j}$
observed number of individuals that were marked on sampling occasion k and were recaptured on sampling occasion t for $i < k < t \leq j$	$m_{k,t} \sim \text{Binomial}\left(\sum_{i=0}^{k-1}\sum_{j=t}^{K} M^k_{i,j}, p_t\right)$

4.3. *Case study 1: Inference.* Inference for $\left\{\mathbf{M^k}\right\}_{k=1}^{K}$, $\mathbf{n}$, $\{w_{i,j}\}_{i=0,j=i}^{K,K}$, $\{p_t\}_{t=1}^{K}$ and ω is described in Section 2. Parameters l_k and latent matrix $\left\{\mathbf{R^k}\right\}_{k=1}^{K}$, are updated using a MH random walk sampler, as described in Section A of the Supplementary Material (Rotous, Diana and Matechou, 2025).

4.4. *Simulation study.* We conducted a simulation study with the aim of exploring how ignoring removed individuals affects inference quality. For 100 replications, we simulated BM data for 11 sampling occasions, involving 6,000 individuals and with removal probabilities of 0.05, 0.1, or 0.2. We used time varying conditional entry/exit probabilities as well as capture probabilities. The chosen values are given in Section C of the Supplementary Material (Rotous, Diana and Matechou, 2025), together with the prior distribution choices for all parameters. In each case, we ran an MCMC algorithm for 200,000 iterations, with a burn-in of 50,000 iterations.

We calculate the posterior root mean squared error (RMSE) for each parameter as $\frac{1}{R}\sum_{r=1}^{R}\frac{1}{T}\sum_{i=1}^{T}(x_r^{(i)}-x^\star)^2$, where $x_r^{(i)}$ is the ith MCMC sample of the rth replication, and $x^\star$ is the true value of that parameter. Inference for N when removed individuals are accounted for (denoted by PTBM-R) and when they are not (denoted by PTBM) is summarized in Figure 2 and for all other parameters in Section C of the Supplementary Material (Rotous, Diana and Matechou, 2025).

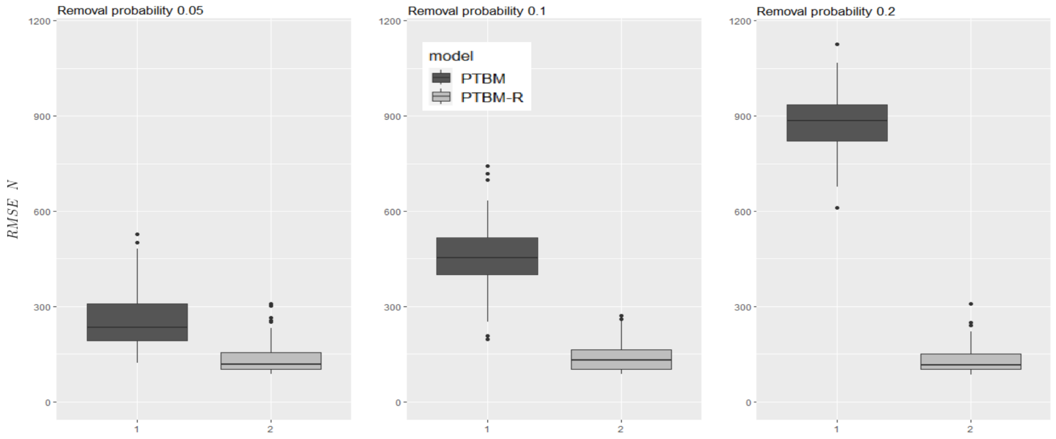

Fig 2: RMSE of population size, N. Each column corresponds to a different removal probability. We compare RMSE across the models PTBM-R and PTBM, i.e. when we account for removals and when we do not.

As expected, when the removed individuals are not accounted for, RMSE increases, especially for parameters concerning the number of individuals in the survey such as N, U_t and N_t, which is defined as the number of individuals available on sampling occasion t (see Section C of the Supplementary Material (Rotous, Diana and Matechou, 2025)).

RMSE inflation becomes more severe as the proportion of unmarked individuals removed increases (for more details see Section 1 in Section C of the Supplementary Material (Rotous, Diana and Matechou, 2025)). The rest of the parameters, such as conditional entry/exit and capture probabilities exhibit less sensitivity, especially when removal probability is low, since they are mainly informed by the marked individuals that are followed over time, and hence are not affected as much by the removal of unmarked individuals. Finally, in Table 1 of Section C of the Supplementary Material (Rotous, Diana and Matechou, 2025), we display computational time in minutes varying the population size, N, and number of sampling occasions, K. We have used the same number of MCMC iterations for all scenarios to make comparisons meaningful. As expected, computational times scale with the number of sampling occasions, K, and are invariant to the number of individuals, N. This is a consequence of the grid approach, where the number of latent cells to be inferred only changes as a function of K and not of N.

4.5. *Weather-loch data.* We present results for this case study in Figures 3 and Table 1 and compare them to those obtained by Cowen et al. (2017). As stated in Section 3, we assume a time-invariant conditional entry probability $\tilde{V}$ and time-invariant conditional exit probability $1 - \phi$. The prior distribution choices for our parameters, number of MCMC iterations and computational times are given in Section D of the Supplementary Material (Rotous, Diana and Matechou, 2025).

The overall patterns in our findings agree with those in the simulation study. Namely, in Table 1, we observe that estimation of recruitment rate, η, to the unmarked population and conditional exit probabilities for the PTBM-R, PTBM, and HMM models are, on average, aligned. Additionally, in Figure 3, we observe that, on average, our estimates for the number of unmarked individuals present on each sampling occasion, and for the population size are larger when accounting for removals compared to the other two models PTBM and HMM. Finally, the posterior mean population and super-population for the PTBM-R model are 3280 with a 95% posterior credible interval of $(2892, 3771)$ and 2805 with a 95% posterior credible interval of $(2554, 3047)$, respectively, whereas for the PTBM model are 2198 with a 95% posterior credible interval of $(1958, 2419)$ and 2182 with a 95% posterior credible interval of $(2095, 2276)$ respectively. The super-population for the Cowen et al. (2017) model is estimated as 2242 (no interval was reported) which is in alignment with out PTBM model, which is expected since neither model accounts for removals.

Figures 3 in this section and 8 in Section D of the Supplementary Material (Rotous, Diana and Matechou, 2025) demonstrate that the number of available individuals increases and peaks on the 5th sampling occasion before gradually decreasing. This could potentially be interpreted as a result of seasonality. Notably, Huggins, Wang and Kearns (2010) mentioned that they estimated the smallest number of available individuals on the 7th sampling occasion, which they attributed to the winter season, with the population then increasing during the spring. In contrast, Cowen et al. (2017) stated that they estimated a decrease in the number of available individuals over time when using time-varying capture probabilities, as opposed to Huggins, Wang and Kearns (2010). Our PTBM-R model manages to identify some seasonality, with population size peaking before the winter time and subsequently estimating a decrease in the population.

Finally, we conducted a goodness of fit assessment, by comparing the observed number of unmarked individuals $\{u_t\}_{t=1}^{K}$ caught on each sampling occasion with summaries of the corresponding data simulated from each model (see Section D of the Supplementary Material (Rotous, Diana and Matechou, 2025)). It is evident that the PTBM-R model, which better describes the data-generating process, outperforms the other two models in generalizing observed patterns in the data.

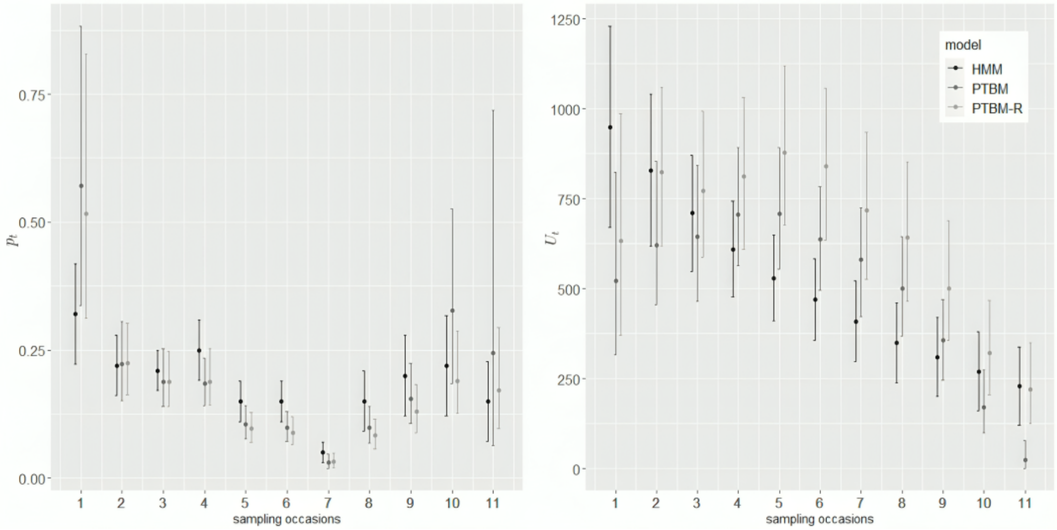

Fig 3: Summaries of capture probabilities p_t, (a), and of the inferred number of unmarked individuals present on each sampling occasion U_t, $t = 1, 2, ..., K$, (b). The HMM model of Cowen et al. (2017) was fitted classically, so the summaries correspond to the maximum likelihood estimate and corresponding 95% confidence interval in each case, whereas our PTBM and PTBM-R models are fitted in a Bayesian framework, so the summaries correspond to the posterior mean and 95% posterior credible interval.

TABLE 1

Summaries of recruitment rate to the unmarked population η and survival probability ϕ. In the HMM case, these correspond to the maximum likelihood estimate and corresponding 95% confidence interval, while in the PTBM and PTBM-R models, these correspond to the posterior mean and corresponding 95% posterior credible interval.

	Parameter	
Model	η	ϕ
HMM	0.24 (0.18, 0.30)	0.63 (0.58,0.69)
PTBM	0.22 (0.18,0.28)	0.61 (0.27,0.81)
PTBM-R	0.24 (0.16,0.32)	0.63 (0.40,0.87)

5. Case Study 2.

Next, we consider the case study first published in Zhang, Bonner and McCrea (2023) where sampling follows Pollock's robust design (RD Pollock, 1982). The sampling of golden mantella took place over 6 primary periods with $(3, 3, 3, 4, 4, 4)$ secondary sampling occasions for each primary period. During each primary period, captured individuals receive a distinct mark and are then released so can be recaptured on later secondary sampling occasions within the same or different primary periods.

5.1. *Data and Notation.*

First we introduce some case-study-specific notation. In the case of the RD, there are K primary periods with T_k secondary sampling occasions within primary period k. The underlying latent process of entry and exit refers to intervals between primary periods, as described in Section 2, while the observation process has the same structure as in Equation (2), but in this case with multiple secondary sampling occasions per primary period, as we describe below.

Data

- $u_{k,t}$: observed number of unmarked individuals caught on secondary sampling occasion t of primary period k.
- $m_{k,\nu,t}$: observed number of individuals that were marked in primary period k and were recaptured on secondary sampling occasion t of primary period ν for $\nu > k$.

Parameters

$p_{k,t}$: probability of capturing an individual present on secondary sampling occasion t of primary period k.

Latent

- $\mathbf{U^{k,t}}$: $(K+1)\times(K+1)$ matrix where $U_{i,j}^{k,t}$ corresponds to the latent number of unmarked individuals in cell (i,j) present on secondary sampling occasion t of primary period k for $i<k\leq j$.
- $\mathbf{M^{k,t}}$: $(K+1)\times(K+1)$ matrix, where $M_{i,j}^{k,t}$ corresponds to the latent number of individuals in cell (i,j) caught as unmarked on secondary sampling occasion t of primary period k.

$$M_{i,j}^{k,t} \sim \text{Binomial}(U_{i,j}^{k,t}, p_{k,t})$$

We note that now the latent number of individuals in cell (i,j) that are available as unmarked on secondary sampling occasion t of primary period k is $U_{i,j}^{k,t} = n_{i,j} - \left(\sum_{\nu=1}^{k-1}\sum_{\ell=1}^{T_\nu} M_{i,j}^{\nu,l} + \sum_{\ell=1}^{t-1} M_{i,j}^{k,l}\right)$ for $k \neq 1, t \geq 1$, $U_{i,j}^{k,t} = n_{i,j} - \sum_{l=1}^{t-1} M_{i,j}^{k,l}$, for $k = 1,\ t > 1$ and $U_{i,j}^{k,t} = n_{i,j}$ for $k = 1, t = 1$, where $\sum_{\nu=1}^{k-1}\sum_{\ell=1}^{T_\nu} M_{i,j}^{\nu,l}$ is the total number of individuals from cell (i,j) marked in any primary period before period k and $\sum_{\ell=1}^{t-1} M_{i,j}^{k,l}$ is the total number of individuals from cell (i,j) marked on any secondary sampling occasion before t within primary period k.

Therefore, similarly to Sections 2 and 4, it follows that the observed number of unmarked individuals caught on secondary sampling occasion t of primary period k is

$$u_{k,t} = \sum_{i=0}^{k-1}\sum_{j=k}^{K} M_{i,j}^{k,t}$$

while the total number of individuals that were marked in primary period k and recaptured on secondary sampling occasion t of primary period ν are modelled as

$$m_{k,\nu,t} \sim \begin{cases} \text{Binomial}\left(\sum_{i=0}^{k-1}\sum_{j=\nu}^{K}\sum_{\ell=1}^{t-1} M_{i,j}^{k,l}, p_{k,t}\right), & \nu = k,\ t = 2,3,...,T_\nu \\ \text{Binomial}\left(\sum_{i=0}^{k-1}\sum_{j=\nu}^{K}\sum_{\ell=1}^{T_k} M_{i,j}^{k,l}, p_{k,t}\right), & \nu > k,\ t = 1,2,...,T_\nu \end{cases}$$

In the above, we distinguish between two cases: recaptures taking place in the same primary period as the first capture ($\nu = k$), and recaptures taking place in subsequent primary periods ($\nu > k$), where the sum over ℓ in each case is used to calculate the total number of individuals from cell (i,j) that were marked in primary period k.

5.2. *Model.* The complete model for BM data collected under a RD is given in Equation (5).

(5)

latent number individuals in cell (i,j)

$$n_{i,j} \sim \text{Poisson}(\omega \times w_{i,j}),\ i=0,...,K,\ j=i,...,K$$

latent number of individuals in cell (i,j) that are unmarked on secondary sampling occasion t of primary period k, with $i<k\leq j$

$$U_{i,j}^{k,t} = \begin{cases} n_{i,j},\ \ k=1,\ t=1 \\ n_{i,j} - \sum_{l=1}^{t-1} M_{i,j}^{k,l},\ \ k=1,\ t>1 \\ n_{i,j} - \sum_{\nu=1}^{k-1} \sum_{l=1}^{T_\nu} M_{i,j}^{\nu,l} - \sum_{l=1}^{t-1} M_{i,j}^{k,l},\ k\neq 1,\ t\geq 1 \end{cases}$$

latent number of individuals in cell (i,j) marked on secondary sampling occasion t of primary period k

$$M_{i,j}^{k,t} \sim \text{Binomial}(U_{i,j}^{k,t}, p_{k,t})$$

observed number of unmarked individuals caught on secondary sampling occasion t of primary period k

$$u_{k,t} = \sum_{i=0}^{k-1} \sum_{j=k}^{K} M_{i,j}^{k,t},\ k=1,...,K\ \ t=1,...,T_k$$

observed number of individuals that were marked in primary period k and were recaptured on secondary sampling occasion t of primary period ν

$$m_{k,\nu,t} \sim \begin{cases} \text{Binomial}\left(\sum_{i=0}^{k-1} \sum_{j=\nu}^{K} \sum_{l=1}^{t-1} M_{i,j}^{k,l}, p_{k,t}\right), \\ \quad \nu=k,\ t=2,3,...,T_\nu \\ \text{Binomial}\left(\sum_{i=0}^{k-1} \sum_{j=\nu}^{K} \sum_{l=1}^{T_k} M_{i,j}^{k,l}, p_{k,t}\right), \\ \quad \nu>k,\ t=1,2,...,T_\nu \end{cases}$$

5.3. *Case Study 2: Inference.* We employ an MCMC with a MH random walk sampler for inferring the latent matrices $\{\mathbf{M^k}\}_{k=1}^{K}$ and $\mathbf{n}$, and a Gibbs sampler for the probabilities $\{w_{i,j}\}_{i=0,j=i}^{K,K}, \{p_{k,t}\}_{k=1,t=1}^{K,T_k}$ and for ω. Details, about the conditional distributions are given in Section A of the Supplementary Material (Rotous, Diana and Matechou, 2025).

5.4. *Simulation study.* We simulated a population size of $5,000$ individuals across 5 primary periods, considering time varying conditional entry/exit probabilities and capture probabilities as described in Section 3. Their values are given in Section E of the Supplementary Material (Rotous, Diana and Matechou, 2025), together with the prior distribution choices for the parameters.

To assess the impact of the number of secondary sampling occasions on the quality of inference, we simulated data using 1, 2, 4, 8, and 16 secondary sampling occasions.

We ran the MCMC for 200,000 iterations in each case, with the first 50,000 iterations as burn-in. The results are presented in Table 2 and in Section E of the Supplementary Material (Rotous, Diana and Matechou, 2025). Here, we summarize the effect, in terms of percentage decrease in RMSE, in conditional entry and exit probabilities and in population size when the number of secondary occasions increases (Table 2). Our findings indicate that, as expected, incorporating more sampling occasions within each primary period leads to smaller RMSE for all parameters. However, the benefit in terms of decrease in RMSE is not proportionate to the increase in effort for larger numbers of secondary periods, and the effects of the additional effort practically diminish for more than eight secondary periods, especially in the case of population size. These results highlight the benefit of the RD within BM studies in comparison to the standard BM without any periods of closure, and can support decisions around study design and allocation of effort in BM surveys.

Finally, we present the computational time in minutes for different numbers of secondary sampling occasions T_k in Section E of the Supplementary Material (Rotous, Diana and Matechou, 2025), Table 2. We considered the same number of MCMC iterations, making comparisons meaningful and observe that computational time does not change, with average run time 13.2 minutes. Demonstrating that computational time is not a function of the number of secondary sampling occasions but only of the number of primary periods K as shown in Section 4.4.

TABLE 2

The mean decrease change in RMSE as we increase the number of secondary sampling occasions from 1 to 2, 2 to 4, 4 to 8 and lastly from 8 to 16. The displayed parameters are the population size, N, and the conditional entry and exit probabilities $\tilde{V}_i, (1-\phi_j)$ for $i = 0, 1, ..., K-1$ and $j = 1, 2, ..., K-1$, for K primary periods.

	Secondary sampling occasions			
Parameters	1→2	2→4	4→8	8→16
N	8%	5%	3%	0.3%
$\tilde{V}_i$	20.7%	19.2%	11.5%	4.2%
$(1-\phi_j)$	22.6%	21%	18.4%	14.9%

5.5. *Golden mantella data.* We apply our model to the data considered in Zhang, Bonner and McCrea (2023), with prior distributions and MCMC iterations same with the ones used in Section F of the Supplementary Material (Rotous, Diana and Matechou, 2025) and computational time displayed in Section E of the Supplementary Material (Rotous, Diana and Matechou, 2025).

In Figure 4, the results demonstrate that the Zhang, Bonner and McCrea (2023) and our PTBM model provide similar inference on entry and survival probabilities with PTBM yielding consistently narrower posterior credible intervals than the corresponding confidence intervals obtained by Zhang, Bonner and McCrea (2023) using SPA. Additionally, Figure 4 demonstrates that the survival probabilities are larger for the midpoints of the breeding seasons, which is expected, as also reported by Zhang, Bonner and McCrea (2023). Additionally, the entry probabilities, on average, are higher in the early primary periods within each breeding season.

We present posterior summaries of the capture probabilities and the number of available individuals in Figure 13 in Section F of the Supplementary Material (Rotous, Diana and Matechou, 2025). The posterior mean of the number of individuals present on each primary period from PTBM is similar to the point estimates from SPA for the early primary periods however, in our case we infer a smaller number of individual present for the later primary periods, with the PTBM having narrower intervals as before. Furthermore, the posterior mean of the super-population size is smaller for our PTBM model with 95% posterior credible interval $(4808, (4371, 5310)$ by PTBM) compared to the point estimate and corresponding 95% confidence interval $(5567\ (5145, 6063)$ by SPA). Finally, Figure 14 in Section F of the Supplementary Material (Rotous, Diana and Matechou, 2025) displays the goodness of fit for our PTBM model, which does not raise any concerns.

6. Discussion.

BM surveys are an important monitoring approach for several species and our paper contributes to the, fairly limited thus far, statistical literature for modelling the corresponding data. Our introduced latent grid for modelling the entry and exit pattern,

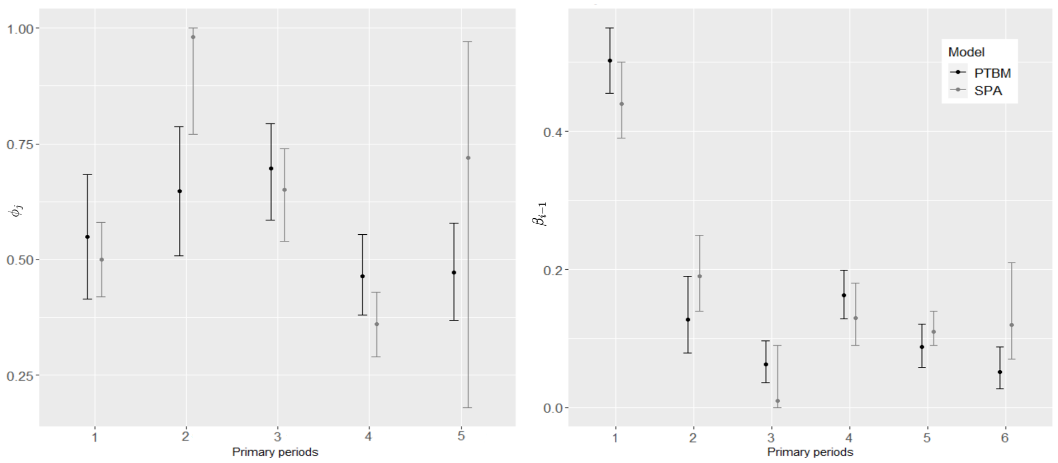

Fig 4: Posterior summaries of survival probabilities ϕ_j, $j = 1, 2, \ldots, K-1$, (a), and of entry probabilities β_i, $i = 0, 1, ..., K-1$, (b). The SPA model of Zhang, Bonner and McCrea (2023) was fitted classically, so the summaries correspond to the maximum likelihood estimate and corresponding 95% confidence interval in each case, whereas our PTBM model is fitted in a Bayesian framework, so the summaries correspond to the posterior mean and 95% posterior credible interval.

together with the associated PT prior for the grid probabilities, provides a new, general, flexible and computationally efficient modelling framework for BM data. Our model scales with the number of sampling occasions, which is typically much smaller than the number of individuals, as it infers the number of individuals with a specific entry/exit interval, rather than inferring latent individual presence or capture histories, and gives us access to exact inference under different observation processes. The PT model described in this article may appear similar to a Dirichlet-Multinomial model (Royle, Converse and Link, 2012), which is, in fact, a special case of the PT framework when considering the partition described in Section 3. However, the two methods are not identical since we utilize the PT machinery, such as the replicate PT, to impose restrictions on the dynamic parameters of the model. The replicate PT structure allows us to place constraints on the model parameters to build parsimonious models that are also ecologically meaningfully.

In our case, we have assumed a replicate PT structure that leads to the assumption that the probability of exit depends on the sampling occasion, and not on the time of entry (time-varying exit). The PT framework gives us access to efficient approaches for model comparisons of different constraints, such as constant, time, or age-varying probabilities (Holmes et al., 2015), and hence future work could explore establishing the required methodology for building and comparing models with different constraints for the model parameters.

Additionally, in the case studies and corresponding simulations, we have employed flat prior distributions for all parameters, to make our results as comparable as possible to the classical models considered thus far in the literature. However, the PT can be centered on parametric distributions, such as the normal, that express our prior expectation of the entry/exit pattern, and in that case inference benefits from both the smoothness of the parametric curve and the flexibility of the nonparametric PT prior.

We considered two case studies with different observation processes to demonstrate the flexibility of our proposed PT prior approach to accommodate different data-generating processes. Our framework can easily be extended to jointly model BM data with capture-recapture, count or other types of ecological data that are often collected in practice. However, an open challenge that remains is the incorporation of covariates in the model parameters, and in particular when these are measured at the individual level, as in that case the grid-approach does not apply, at least not in its current form.

Acknowledgements. We are grateful to Professor Rachel S. McCrea, Dr Panagiotis (Takis) Besbeas, Professor Byron J. T. Morgan, Dr Wei Zhang, Dr Simon J. Bonner and Professor Richard Griffiths for providing data, insights and feedback on our analysis and to two anonymous referees for their comments that greatly improved this manuscript.

Supporting Information. Section A of the Supplementary Material (Rotous, Diana and Matechou, 2025), referenced in Subsections 3.1, 4.3 and 5.3, Section B of the Supplementary Material (Rotous, Diana and Matechou, 2025), referenced in Section 2,Section C of the Supplementary Material (Rotous, Diana and Matechou, 2025), referenced in Subsection 4.4, Section D of the Supplementary Material (Rotous, Diana and Matechou, 2025), referenced in Subsection 4.5, Section E of the Supplementary Material (Rotous, Diana and Matechou, 2025), references in Subsections 5.4 and 5.5, Section F of the Supplementary Material (Rotous, Diana and Matechou, 2025), in Subsection 5.5, are available with this paper at the Annals of Applied Statistics Online Library. The code for fitting all of the models presented in the paper is available on https://github.com/IoannisRs/BatchMark_scripts.git.

REFERENCES

Cowen, L. L. E., Besbeas, P., Morgan, B. J. T. and Schwarz, C. J. (2014). A comparison of abundance estimates from extended batch-marking and Jolly–Seber-type experiments. *Ecology and Evolution* **4** 210–218.

Cowen, L. L. E., Besbeas, P., Morgan, B. J. T. and Schwarz, C. J. (2017). Hidden Markov models for extended batch data. *Biometrics* **73** 1321–1331.

Davidson, J. R., Sudirman, R., Wahid, I., Baskin, R. N., Hasan, H., Arfah, A. M., Nur, N., Hidayat, M. Y., Syafruddin, D. and Lobo, N. F. (2019). Mark-release-recapture studies reveal preferred spatial and temporal behaviors of Anopheles barbirostris in West Sulawesi, Indonesia. *Parasites & Vectors* **12** 1–11.

Diana, A., Griffin, J. and Matechou, E. (2019). A Pólya tree based model for unmarked individuals in an open wildlife population. In *Bayesian Statistics and New Generations: BAYSM 2018, Warwick, UK, July 2-3 Selected Contributions* 3–11. Springer.

Diana, A., Matechou, E., Griffin, J., Arnold, T., Tenan, S. and Volponi, S. (2023). A general modeling framework for open wildlife populations based on the Pólya tree prior. *Biometrics* **79** 2171–2183.

Doll, J. C., Wood, C. J., Goodfred, D. W. and Rash, J. M. (2021). Incorporating batch mark–recapture data into an integrated population model of brown trout. *North American Journal of Fisheries Management* **41** 1390–1407.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85** 398–409.

Holmes, C. C., Caron, F., Griffin, J. E. and Stephens, D. A. (2015). Two-sample Bayesian Nonparametric Hypothesis Testing. *Bayesian Analysis* **10** 297 – 320.

Huggins, R., Wang, Y. and Kearns, J. (2010). Analysis of an extended batch marking experiment using estimating equations. *Journal of Agricultural, Biological, and Environmental Statistics* **15** 279–289.

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

Kendall, W. L. and Pollock, K. H. (1992). The robust design in capture-recapture studies: a review and evaluation by Monte Carlo simulation. *Wildlife 2001: Populations* 31–43.

King, R. and McCrea, R. (2019). Capture–Recapture methods and models: estimating population size. In *Handbook of Statistics*, **40** 33–83. Elsevier.

Lavine, M. (1992). Some Aspects of Pólya Tree Distributions for Statistical Modelling. *The Annals of Statistics* **20** 1222 – 1235.

Matechou, E. and Caron, F. (2017). Modelling individual migration patterns using a Bayesian nonparametric approach for capture–recapture data. *The Annals of Applied Statistics* **11** 21 – 40.

Matechou, E., McCrea, R. S., Morgan, B. J., Nash, D. J. and Griffiths, R. A. (2016). Open models for removal data. *The Annals of Applied Statistics* **10** 1572-1589.

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

POLLOCK, K. H. (1982). A capture-recapture design robust to unequal probability of capture. *The Journal of Wildlife Management* **46** 752–757.

ROBERT, C. P. and CASELLA, G. (2004). *The Metropolis—Hastings Algorithm* In *Monte Carlo Statistical Methods* 267–320. Springer New York, New York, NY.

ROSSER, E., WILLDEN, S. A. and LOEB, G. M. (2022). Effects of SmartWater, a fluorescent mark, on the dispersal, behavior, and biocontrol efficacy of Phytoseiulus persimilis. *Experimental and Applied Acarology* **87** 163–174.

ROTOUS, I., DIANA, A. and MATECHOU, E. (2025). Supporting Information for
"A Pólya Tree modelling framework for batch-mark data".

ROYLE, J. A., CONVERSE, S. J. and LINK, W. A. (2012). Data augmentation for hierarchical capture-recapture models. *arXiv preprint arXiv:1211.5706*.

SCHWARZ, C. J. and ARNASON, A. N. (1996). A General Methodology for the Analysis of Capture-Recapture Experiments in Open Populations. *Biometrics* **52** 860–873.

SEBER, G. A. (1965). A note on the multiple-recapture census. *Biometrika* **52** 249–259.

SEBER, G. A. F. and SCHOFIELD, M. R. (2019). *Capture-recapture: Parameter estimation for open animal populations*. Springer.

VAVASSORI, L., SADDLER, A. and MÜLLER, P. (2019). Active dispersal of Aedes albopictus: a mark-release-recapture study using self-marking units. *Parasites & Vectors* **12** 1–14.

ZHANG, W., BONNER, S. J. and MCCREA, R. S. (2023). Latent multinomial models for extended batch-mark data. *Biometrics* **79** 2732–2742.

ZHANG, W., BRAVINGTON, M. V. and FEWSTER, R. M. (2019). Fast likelihood-based inference for latent count models using the saddlepoint approximation. *Biometrics* **75** 723–733.