Citation for published version

Matechou, Eleni and Argiento, Raffaele (2022) Capture-recapture models with heterogeneous temporary emigration. Journal of the American Statistical Association . ISSN 0162-1459.

DOI

https://doi.org/10.1080/01621459.2022.2123332

Link to record in KAR

https://kar.kent.ac.uk/96926/

Document Version

Publisher pdf

Journal of the American Statistical Association

September 2022 ■ Volume 117 ■ Number 539

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Capture-recapture models with heterogeneous temporary emigration

Eleni Matechou & Raffaele Argiento

To cite this article: Eleni Matechou & Raffaele Argiento (2022): Capture-recapture models with heterogeneous temporary emigration, Journal of the American Statistical Association, DOI: 10.1080/01621459.2022.2123332

To link to this article: https://doi.org/10.1080/01621459.2022.2123332

Capture-recapture models with heterogeneous temporary emigration

Eleni Matechou[a] and Raffaele Argiento[b]

[a]School of Mathematics, Statistics and Actuarial Science, University of Kent, UK

[b]Department of Economics, University of Bergamo, Italy

E.Matechou@kent.ac.uk

Abstract

We propose a novel approach for modelling capture-recapture (CR) data on open populations that exhibit temporary emigration, whilst also accounting for individual heterogeneity to allow for differences in visit patterns and capture probabilities between individuals. Our modelling approach combines changepoint processes – fitted using an adaptive approach – for inferring individual visits, with Bayesian mixture modelling – fitted using a nonparametric approach – for identifying clusters of individuals with similar visit patterns or capture probabilities. The proposed method is extremely flexible as it can be applied to any CR data set and is not reliant upon specialised sampling schemes, such as Pollock's robust design. We fit the new model to motivating data on salmon anglers collected annually at the Gaula river in Norway. Our results when analysing data from the 2017, 2018 and 2019 seasons reveal two clusters of anglers – consistent across years – with substantially different visit patterns. Most anglers are allocated to the "occasional visitors" cluster, making infrequent and shorter visits with mean total length of stay at the river of around seven days, whereas there also exists a small cluster of "super visitors", with regular and longer visits, with mean total length of stay of around 30 days in a season. Our estimate of the probability of catching salmon whilst at the river is more than three times higher than that obtained when using a model that does not account for temporary emigration, giving us a better understanding of the impact of fishing at the river. Finally, we discuss the effect of the COVID-19 pandemic on the angling population by modelling data from the 2020 season. Supplementary materials for this article are available online.

Keywords: Angling, Clustering, Chinese Restaurant process, Population size, Stopover model

1 Introduction

Capture-recapture (CR) data provide a method for monitoring wildlife populations, or more generally populations for which a census is infeasible because the probability of detecting individuals is lower than one. Appropriate means for capturing individuals are employed in a series of capture occasions. During these capture occasions, newly caught individuals are uniquely marked and all caught individuals are released back into the population. Repeating the process T times gives rise to a capture history, of length T, for each caught individual with entries equal to one (zero) indicating the capture occasions at which the particular individual was caught (not caught). At the end of the study, there remains an unknown number of individuals that were never caught and hence have capture histories with all entries equal to zero.

There exists a rich literature dealing with models for CR data. This covers closed-population models, which assume that the same individuals are present and available for capture at all T occasions (Darroch, 1958; Otis et al., 1978; Pledger, 2000), open-population models that allow for arrivals and departures during the study (Jolly, 1965; Seber, 1965; Schwarz and Arnason, 1996; Pledger et al., 2009; Lyons et al., 2016) and a recent body of work exploring Bayesian nonparametric methods for CR data (Manrique-Vallier, 2016; Matechou and Caron, 2017, closed and open populations, respectively).

Typically, one of the assumptions of open-population CR models is that emigration is permanent; once individuals depart they are assumed to never arrive again. However, the assumption of permanent emigration is not always satisfied. Existing approaches for modelling temporary emigration in population ecology (see for example Zhou et al., 2019) so far mostly use Pollock's robust design (Pollock, 1982), which assumes that there are two types of sampling periods: primary periods, for example seasons, between which the population is open, and secondary periods, for example sampling days within a season, between which the population is closed. Clearly, there exist cases where methods relying on Pollock's robust design cannot be employed.

In this paper we propose a novel solution for modelling temporary emigration for CR data by developing a flexible, general and mathematically tractable Bayesian mixture model that also accounts for individual heterogeneity in the visit pattern and capture probability. We treat the visit history of each individual as a changepoint process with visits, when an individual transitions from non-present to present, signifying a changepoint. We model the corresponding departure times as dependent on these changepoints and we extend recently developed adaptive MCMC methods (Benson and Friel, 2018) for updating the number and position of changepoints of each individual. These adaptive methods provide us with an elegant and computationally efficient way to infer individual visit histories, without requiring any tuning or complicated proposal distributions.

Additionally, we model individual heterogeneity using a random finite measure as the intensity of a Poisson process that jointly models the number of individuals and their visit histories. Conditionally on the number of individuals, the visit histories are shown to be a sample from a mixture model, which, although technically is finite dimensional, we fit using Bayesian nonparametric techniques. Our approach builds upon recent work by Argiento and De Iorio (2022) that exploits the relationship between finite and infinite mixtures in a Bayesian framework. Argiento and De Iorio (2022) show that a mixture model with a random number of components is indeed a nonparametric model since its complexity is a-priori unbounded and inferred by the data. Moreover, while an infinite mixture model can lead to an a posteriori inconsistency on the estimation of the number of clusters (Miller and Harrison, 2018), this is not the case for finite mixture models with a random number of components. The use of a finite mixture model allows us to devise a *blocked conditional Gibbs sampler*, which is faster than the marginal algorithms that are usually adopted, as in Matechou and Caron (2017). Finally, we show that when the number of elements being clustered is itself random, the clustering can be expressed via a function, which we call $N-$ exchangeable random probability function, that can be used to derive the marginal distribution of *N*, and gives rise to an interpretation of the clustering in terms of the Chinese restaurant process.

Our proposed modelling approach has several distinctive advantages. In contrast to methods relying on Pollock's robust design, we do not need to assume population closure. From a computational point of view, we design an adaptive scheme for the changepoint process part of the model that is free from complicated tuning strategies and we tackle the mixture model from a Bayesian point of view via a conditional algorithm that is computationally efficient and allows for full Bayesian inference.

We fit our model to CR data on salmon anglers in Norway. The anglers only report a visit to the river on a particular day if they catch at least one salmon. This gives rise to a standard CR data set indicating the days when each angler reported a catch. However, the number of anglers who visited the river, the number of days on which they were present each season and their probability of catching salmon whilst at the river are unknown. Our results in terms of types of visit patterns and fishing abilities are consistent across the 2017, 2018 and 2019 seasons. Specifically, we identify two clusters, with the largest cluster, consisting of around 95% of the angler population, making two to three visits on average each year, which last for around two days on average. The second cluster of anglers make around eight visits each year, each lasting on average two to three days. This leads to substantial differences in the total number of days spent at the river on average between the two clusters. Additionally, we estimate that the total number of anglers who visited the river each year is considerably higher than the number of anglers who caught at least one salmon, and, for the first time in the Gaula river, we estimate the total number of anglers who are present at the river on each day of the season. Finally, we quantify the effect of the COVID-19 pandemic on the angling population by analysing the 2020 season data. Our findings suggest that there was a considerable increase in the number of anglers who visited that year, with anglers spending longer on average per visit compared to previous seasons but having an overall lower ability to catch fish. Such information is invaluable for managing the river, and correctly and precisely setting fishing quotas and introducing further regulations, as required.

We present the model in Section 3 and discuss its clustering behaviour in Section 4. Our algorithm used for inference, including the adaptive MCMC changepoint sampler and the Bayesian nonparametric algorithm used for clustering, is presented in Section 5. The results from fitting the model to simulated data and to the angler data are presented in Section 6 while the paper concludes in Section 7. Additional technical details about our model and algorithm, as well as results of an extensive simulation study, are provided in the online Supplementary Material.

2 Salmon angler data

The Gaula is an unregulated river in Norway that is a very popular destination for salmon anglers. Any catches of salmon have to be reported by the anglers to the river management board. However, individual anglers only report that they were fishing on a particular day of the season, which lasts for $T = 92$ (1st of June to 31st of August) days, if they catch at least one salmon that day. This creates a capture history, which is a vector of length T, for each angler, with entry t equal to one if the angler caught salmon on day t, and zero otherwise, with $t = 1, \dots, T$. There also exists an unknown number of anglers who visited the river but did not catch salmon, and hence share the capture history with all T entries equal to 0. The population is open, with anglers entering and exiting the river throughout the season. Anglers need to purchase fishing licenses from land owners along the river in advance, but the information on the number of licenses that have been sold each year is considered sensitive and not shared. Licenses are often sold-out months in advance, and the only restriction placed on anglers is that they can only remove a maximum of four salmon a year for the period we are considering in this paper. However, they can catch as many salmon as they want and release them back in the river.

Individual anglers can return to the river to fish an arbitrary number of times in the season, meaning that their emigration from the site needs to be treated as temporary. Clearly, we cannot employ methods relying on Pollock’s robust design in this case as the population cannot be assumed closed for any period in the season. Additionally, our inference needs to account for individual heterogeneity in the visit behaviour and fishing ability, the latter expressed through the probability of an angler catching salmon whilst at the river, and hence being themselves “caught”, henceforth referred to as observed, while present at the river. For example, some anglers may visit often for shorter periods of time while others may visit less regularly for longer periods of time, with capture probability also linked to the visit behaviour, with, for example, frequent visitors being more skilled in fishing and vice versa.

We are interested in inferring the size of the population of anglers, *N*, and the visit history of each angler, which consists of the number, timing and duration of their visits, as well as the probability of an angler catching salmon on any given day. We note here that if an open population model that does not allow for temporary emigration is fitted to the data, such as the "POPAN" model in the R package RMark (Laake, 2013), then the resulting estimates of capture probability, which are between 0.06 and 0.07 for all years, are certainly under-estimates; since the model assumes that individuals are present between their first and last observation, it inevitably underestimates the probability of capture.

The total number of anglers observed in the 2017, 2018, and 2019 seasons is equal to 2057, 1901, and 1687, respectively, but we cannot infer the total number of visitors to the river each season using the raw data alone. The median number of times that each angler has been observed is equal to one in 2017 and 2019 and to two in 2018 but it is not possible to decipher from the data alone the total number of days each angler has spent at the river. Similarly, the number of visits made by individual anglers each year is unknown. Clearly, we can safely assume that consecutive days on which an individual angler was observed belong to a single visit. However, observations of individual anglers are also separated by days on which the particular angler was not observed. If we assume that visits are separated by at least one day of non-observation, then we find that the proportion of anglers making a single visit would be set to around 60% with an average visit length of around 1.2 days for all seasons, while if we assume that visits are separated by at least two days of non-observation, then these figures change to around 70% and 1.4-1.5 days on average. Finally, Figure 1 shows that the pattern in the number of anglers observed each day of the season varies greatly between years, but it is of interest to compare the arrival and departure patterns between years, which cannot be achieved using the raw capture histories. These questions motivate the development of the model presented in Section 3.

Inferring the number of anglers present at any one time within the season enables the river management to assess the fishing pressure and consequently decide on fishing quotas or additional regulations. Information on the probability of catching salmon is valuable to the river management board for the purposes of fisheries management, for ensuring the conservation of salmon in the river and the availability of resources, such as accommodation and food, for anglers. Knowledge about the number of anglers who visit the river each season and the number of visits they each make can help demonstrate to local politicians and stakeholders the importance and attractiveness of the river Gaula as a tourist destination and increase the appreciation of the value of salmon fishing to the local community, resulting in policies that ensure protection and conservation of the river and the salmon stock. Finally, by modelling the 2020 season and quantifying the differences compared to previous seasons in terms of numbers of visitors, visit patterns and capture probabilities can shed light on the effect of the COVID-19 pandemic and the corresponding global travel restrictions on the angling and hence salmon population at the Gaula river.

3 Model

We denote the number of individuals caught at least once by *n* and the number of sampling occasions by *T*. The population size is denoted by *N*. Let $i = 1, \cdots, N$ index individuals and $t = 1, \ldots, T$, which we refer to as time, index equally-spaced sampling occasions.

The data consist of *n* capture histories, each of length *T*. We denote the capture history of individual *i* by $\mathbf{CH}_i$ with entry $\mathrm{CH}_{it} = 1$ if individual *i* was caught at time *t* and 0 otherwise. Individuals with index $n + 1, \ldots, N$ have capture histories with all entries equal to 0.

Individual *i*, $i = 1, \ldots, N$, makes an unknown number of visits, k_i, and has an unknown visit history summarised by the arrival times $\boldsymbol{a}_i = (a_{i1}, a_{i2}, \ldots, a_{ik_i})$ and the corresponding departure times $\boldsymbol{d}_i = (d_{i1}, d_{i2}, \ldots, d_{ik_i})$. The arrival times are an increasing sequence of length k_i with $k_i > 0 \ \forall i$, in $\{1, \ldots, T\}$ and the corresponding departure times are an increasing sequence in $\{1, \ldots, T\}$, such that $a_{iv} \leq d_{iv} < a_{iv+1}$ for each $v = 1,\dots,k_i$, with $a_{k_i+1} = T+1$ by definition. The visit history uniquely defines a corresponding presence history, with the presence history of individual *i* denoted by $\mathbf{PH}_i$ and $\mathrm{PH}_{it} = 1$ if individual *i* was present at time *t* and 0 otherwise. For example in Figure 2, the presence history $\mathbf{PH}_i$ is coded as a sequence of 0s (absent) and 1s (present) and represented by a piecewise constant function. In this case we have *k_i* = 2 visits, $a_{i1} = 27$ and $a_{i2} = 57$, corresponding to the points where the piecewise constant function jumps from 0 to 1, represented by the two filled circles on the x-axis, and k_i = 2 departure times, $d_{i1} = 56$ and $d_{i2} = 86$, corresponding to the points where the function jumps from 1 to 0, represented by the two filled triangles on the x-axis. We observe that, as mentioned above, $\boldsymbol{a}_i$ and $\boldsymbol{d}_i$ uniquely define $\mathbf{PH}_i$, but not vice versa. Nevertheless, to simplify notation, in the following we use $\mathbf{PH}_i \equiv (\boldsymbol{a}_i, \boldsymbol{d}_i)$, since it is the latter that we infer.

We refer to the set $\{(\mathbf{PH},\mathbf{CH})\} := \{(\mathbf{PH}_i,\mathbf{CH}_i), i = 1,\dots,N\}$ as the augmented data because it includes both the $1,\dots,n$ caught individuals and the $n+1,\dots,N$ uncaught individuals. We note that each observation $(\mathbf{PH}_i,\mathbf{CH}_i)$ is a pair of sequences such that its *t*th entry is (CH_{it},PH_{it}) $\in \{0,1\} \times \{0,1\}$. We denote by $(\mathcal{P},\mathcal{C})$ the space of all these possible sequences, i.e. the sample space, and we model the augmented data as a marked Poisson process over ($\mathcal{P},\mathcal{C}$), i.e.

$$\{(\mathbf{PH},\mathbf{CH})\} = \{(\mathbf{PH}_1,\mathbf{CH}_1),(\mathbf{PH}_2,\mathbf{CH}_2),\dots,(\mathbf{PH}_N,\mathbf{CH}_N) \mid \nu(\cdot,\cdot))\} \sim \mathrm{PP}(\nu(\cdot,\cdot))$$

where ν is a random intensity function. The approach originally developed by Kottas and Sansó (2007), extended by Taddy et al. (2012) and adapted for CR models by Matechou and Caron (2017) employs an infinite mixture model for ν. Here, we employ a mixture model with $M \in \{1,2,\dots\}$ components, where $M - 1 \sim \mathrm{Poisson}(\Lambda)$, i.e., the number of components, *M*, is distributed according to a *shifted Poisson*. Conditionally on *M*,

$$\nu(\mathbf{PH}_i,\mathbf{CH}_i) = \sum_{c=1}^{M} S_c\, f(\mathbf{PH}_i,\mathbf{CH}_i \mid \tau_c)$$

where $S_1,\dots,S_M$ are (unnormalized) weights of the mixture and $f(\mathbf{PH}_i,\mathbf{CH}_i \mid \tau_c)$ is the joint probability mass function of the presence history $\mathbf{PH}_i$ and capture history $\mathbf{CH}_i$, conditional on individual *i* belonging to component *c*, and is a function of the component-specific parameter τ_c for $c = 1, \ldots, M$.

The overall intensity of the Poisson process is

$$\Omega := \int_{\mathcal{P},\mathcal{C}} \nu(\mathbf{PH}_i, \mathbf{CH}_i) d\mathbf{PH}_i d\mathbf{CH}_i = \int_{\mathcal{P},\mathcal{C}} \sum_{c=1}^{M} S_c f(\mathbf{PH}_i, \mathbf{CH}_i \mid \tau_c) d\mathbf{PH}_i d\mathbf{CH}_i = \sum_{c=1}^{M} S_c .$$

The model can be represented in the following hierarchical form

$$\begin{aligned}
&N \mid \nu \sim \text{Poisson}(\Omega) \\
&(\mathbf{PH}_1, \mathbf{CH}_1), (\mathbf{PH}_2, \mathbf{CH}_2), \ldots, (\mathbf{PH}_N, \mathbf{CH}_N) \mid P \overset{\text{iid}}{\sim} \sum_{c=1}^{M} \frac{S_c}{\Omega} f(\mathbf{PH}_i, \mathbf{CH}_i \mid \tau_c) \\
&P = \sum_{c=1}^{M} \frac{S_c}{\Omega} \delta_{\tau_c} \qquad (1) \\
&M - 1 \sim \text{Poisson}(\Lambda), \quad S_1, \ldots, S_M \mid M \overset{\text{iid}}{\sim} \text{Gamma}(\eta, \zeta) \quad \text{and} \quad \tau_1, \ldots, \tau_M \mid M \overset{\text{iid}}{\sim} P_0, \\
&\Lambda \sim \text{Gamma}(a_\Lambda, b_\Lambda), \quad \eta \sim \text{Gamma}(a_\eta, b_\eta) \quad \text{and} \quad \zeta \sim \text{Gamma}(a_\zeta, b_\zeta)
\end{aligned}$$

where P_0 is the prior distribution of the component specific parameters, described in Section 5.2.

We let $\tau_c = (q_{1c}, q_{0c}, p_c)$, where

- q_{1c} is the probability that an individual from component *c* arrives at a particular time,
- q_{0c} is the probability that an individual from component *c* departs at a particular time, given presence at that time and, finally,
- p_c is the probability that an individual from component *c* is caught at a particular time given presence at that time.

Given q_{1c} , the time between consecutive arrivals is modelled as a $\text{Geometric}(q_{1c})$ random variable with support $\{1, 2 \ldots ,$ }, probability mass function g_1 and cumulative distribution function G_1. In other words, the process of arrival times $\boldsymbol{a}_i = (a_{i1}, \ldots , a_{ik_i})$ is a realization of a homogeneous Bernoulli process with parameter q_{1c} observed at times $\{1,\ldots,T\}$; consequently, the joint probability mass function of the vector $\boldsymbol{a}_i$ and its length *k_i*, with $k_i > 0$, is given by

$$f(\boldsymbol{a}_i, k_i \mid q_{1c}) = \frac{g_1(a_{i1} \mid q_{1c})\left\{\prod_{v=2}^{k_i} g_1(a_{iv} - a_{iv-1} \mid q_{1c})\right\}\left\{1 - G_1(T - a_{k_i} \mid q_{1c})\right\}}{G_1(T)} = \frac{q_{1c}^{k_i}(1 - q_{1c})^{T-k_i}}{1 - (1 - q_{1c})^{T}}.$$

Conditionally on $\boldsymbol{a}_i$, k_i and q_{0c}, the probability mass function of departure times $\boldsymbol{d}_i$ is given by

$$f(\boldsymbol{d}_i \mid \boldsymbol{a}_i, k_i, q_{0c}) = \prod_{v=1}^{k_i} \frac{g_0(d_{iv} - a_{iv} \mid q_{0c})}{G_0(a_{iv+1} - a_{iv} - 1 \mid q_{0c})}\mathbb{I}(a_{iv} \le d_{iv} < a_{iv+1}),$$

with g_0 and G_0, respectively, the probability mass function and cumulative distribution function of a geometric distribution with support on $\{0,1,\ldots\}$ and component-specific parameter q_{0c}.

The different support for g_1 and g_0 ensures that a departure can occur at the same time as an arrival, while no more than one arrival can occur at any one time.

Following the CR stopover literature (Pledger et al., 2009; Matechou et al., 2016) individuals are assumed to be available for capture on both their arrival and departure times. Hence, the probability mass function of $\mathbf{CH}_i$ conditionally on $\mathbf{PH}_i$ and p_c is

$$f(\mathbf{CH}_i \mid \mathbf{PH}_i, p_c) = \prod_{t=1}^{T}\left[p_c^{\mathrm{CH}_{it}}(1 - p_c)^{1-\mathrm{CH}_{it}}\right]^{\mathrm{PH}_{it}} = p_c^{\sum_{t=1}^{T}\mathrm{CH}_{it}\mathrm{PH}_{it}}(1 - p_c)^{\sum_{t=1}^{T}(1-\mathrm{CH}_{it})\mathrm{PH}_{it}}. \tag{2}$$

Finally, the joint sampling model in component *c* of augmented data $(\mathbf{PH}_i, \mathbf{CH}_i) \equiv (\boldsymbol{a}_i, \boldsymbol{d}_i, \mathbf{CH}_i)$ is

$$f\left(\mathbf{PH}_i, \mathbf{CH}_i \mid \tau_c\right) = f\left(\boldsymbol{a}_i, \boldsymbol{d}_i, \mathbf{CH}_i \mid q_{1c}, q_{0c}, p_c\right) = f(\mathbf{CH}_i \mid \mathbf{PH}_i, p_c)\, f(\boldsymbol{a}_i, k_i \mid q_{1c})\, f(\boldsymbol{d}_i \mid \boldsymbol{a}_i, k_i, q_{0c})$$

(3)

4 Clustering behaviour of the model

In this section we examine the clustering structure that our mixture model induces on the data.

In particular, we observe that the mixing measure *P* in model (1) belongs to the wide class of species sampling models, investigated in detail in Pitman (1996) and largely adopted in Bayesian nonparametric clustering (see Ishwaran and James, 2003; Argiento and De Iorio, 2022; Miller and Harrison, 2018). Building upon the latter literature, we derive quantities of interest for our model.

Given the population size *N*, each of the *N* individuals belongs to one of the *M* components of the mixture in (1). We denote the component to which individual *i* belongs by c_i. To study the clustering behaviour of the model, it is useful to represent the first two lines of model (1), conditionally on the parameters Λ, *η* and *ζ*, in terms of the cluster allocations

$$
\begin{aligned}
&(\mathbf{PH}_i, \mathbf{CH}_i) \mid c_i \overset{\text{ind}}{\sim} f(\mathbf{PH}_i, \mathbf{CH}_i \mid \tau_{c_i}), \quad i = 1, \dots, N \\
&c_1, \dots, c_N \mid w_1, \dots, w_M \overset{\text{iid}}{\sim} \text{Multi}(1, c_i \mid w_1, \dots, w_M) \qquad (4) \\
&w_c = S_c / \Omega \quad c = 1, \dots, M .
\end{aligned}
$$

With probability greater than zero we will obtain ties among the component allocations, $c_1, \dots, c_N$ and we denote by $c_1^\star, \dots, c_C^\star$ the unique values among these allocations. The vector of $c^\star$'s induces a random partition (i.e. clustering) among the augmented data that we denote by $\rho = \{J_1, \dots, J_C\}$, where the clusters are given by $J_1 = \left\{ i : c_i = c_1^\star \right\}, \dots, J_C = \left\{ i : c_i = c_C^\star \right\}$. Each cluster J_j, $j = 1, \dots C$ corresponds to a component of the mixture, namely the component *c* such that $c = c_j^\star$. Note that there is a random number *C* of occupied components, namely the clusters, and $M^{(\text{no})} = M - C$ empty components. For ease of notation we refer to $\theta_j^\star = \tau_{c_j^\star}$ as the parameters of the clusters.

When P_0 is continuous, we can explicitly write (see Section 1 of the Supplementary Material for details) the prior distribution that model (4) induces on N, ρ and $\theta_1^\star, \dots, \theta_C^\star$
:

$$\mathcal{L}(N,\rho,d\theta_1^*,\cdots,d\theta_C^*) = \mathcal{L}(\rho,N)\mathcal{L}(d\theta_1^*,\cdots,d\theta_C^* \mid C) = \mathrm{eppf}_N(N_1,\dots,N_C,C)\prod_{j=1}^{C} P_0(d\theta_j^*),$$

where N_j is the number of individuals in cluster j and $\sum_{j=1}^{C} N_j = N$. We call $\mathrm{eppf}_N(N_1,\dots,N_C,C)$ the *N-exchangeable partition probability function* given by

$$\mathrm{eppf}_N(N_1,\dots,N_C \mid N,C) = \frac{1}{N!}\frac{1}{(\zeta+1)^N}\prod_{j=1}^{C}\left\{\frac{\Gamma(N_j+\eta)}{\Gamma(\eta)}\Lambda\psi_1\right\}e^{-\Lambda(1-\psi_1)}\frac{\psi_1\Lambda+C}{\Lambda}, \qquad (5)$$

with $\psi_1 = \left(\frac{\zeta}{1+\zeta}\right)^{\eta}$. We observe that N ranges in $\{0,1,\dots\}$ and if $N = 0$ then ρ is not defined, otherwise ρ is a partition of the indices $\{1,\dots,N\}$. Moreover, $N_j = 0$ for each j if $N = 0$, while $N_j > 0$ and $\sum_{j=1}^{C} N_j = N$ if $N > 0$. Equation (5) is the prior probability of having N individuals (augmented observations) clustered in C exchangeable groups with frequencies $N_1,\dots,N_C$. We would like to highlight that our model induces a joint distribution on the number of individuals N and the number of clusters C. This can be obtained by marginalizing out the cluster frequencies $N_1,\dots,N_C$ from (5), as described in Section 1 of the Supplementary Material. The joint probability mass function of (N, C) at points $x = 0,1,\dots$ and $k \le x$ is

$$\mathbb{P}_{N,C}(x,k) = \frac{1}{x!}\frac{1}{(\zeta+1)^x}e^{-\Lambda(1-\psi_1)}\frac{\psi_1\Lambda+k}{\Lambda}(\Lambda\psi_1)^k\mathcal{C}(x,k;\eta)$$

where, for any non-negative integers $x \ge 0,\, 0 \le k \le n$ and real numbers η, $\mathcal{C}(x,k;\eta)$ denotes the central generalized factorial coefficient (see Charalambides (2005) formula 2.67 for details). Here we mention that these indices can be easily computed using the recursive formula $\mathcal{C}(x,k;\eta) = \eta\mathcal{C}(x-1,k-1;\eta) + (k\eta - x + 1)\mathcal{C}(x-1,k;\eta)$ with $\mathcal{C}(1,1,\eta) = \eta$. Furthermore, the marginal distribution of N can be computed by the hierarchical representation

$$N \mid \Omega \sim \mathrm{Poisson}(\Omega), \quad \Omega \mid M \sim \mathrm{Gamma}(M\eta,\zeta), \text{ and } \quad M-1 \sim \mathrm{Poisson}(\Lambda),$$

so that for $x = 0,1,\dots,$ the probability mass function of N at x is

$$\mathbb{P}_N(x) = \sum_{M=1}^{\infty} \mathrm{NegBin}(x \mid M\eta, \zeta)\,\mathrm{Poisson}(M-1 \mid \Lambda)$$

$$= \sum_{M=1}^{\infty} \binom{x+M\eta-1}{x} \left(\frac{\zeta}{\zeta+1}\right)^{M\eta} \left(\frac{1}{\zeta+1}\right)^{x} \mathrm{e}^{-\Lambda} \frac{\Lambda^{M-1}}{(M-1)!}.$$

Hence, we can easily calculate $\mathbb{E}(N) = (\Lambda+1)\eta/\zeta$ and $\mathrm{Var}(N) = \frac{\eta^2\Lambda + (\Lambda+1)\eta(\zeta+1)}{\zeta^2}$.

The $N-$ eppf in Equation (5) jointly regulates the size of the augmented data and the clustering behaviour of the process. We use a slightly modified version of the Chinese restaurant metaphor (Aldous, 1985) to describe this behaviour. This modification is required because the number of customers, *N*, arriving in this case is random.

The probability that the first customer arrives is equal to $1-\mathbb{P}_N(0)$. The first customer sits with probability one at the first table and the second customer arrives with conditional probability $1-\mathbb{P}_N(1)/(1-\mathbb{P}_N(0))$ and selects a table according to Equations (6) and (7) where in this case *C* = 1 and $N_1 = 1$. In general, given that *x* customers have arrived and are sitting on *C* tables with frequencies $N_1, \dots, N_C$, the probability that customer *x* + 1 arrives is $1-\mathbb{P}_N(x)/\left(1-\sum_{y=0}^{x-1}\mathbb{P}_N(y)\right)$ and selects a table again according to Equations (6) and (7). After *x* customers have arrived, the process stops with probability $\mathbb{P}_N(x)/\left(1-\sum_{y=0}^{x-1}\mathbb{P}_N(y)\right)$ and no more customers arrive.

We refer to the *C* groups $J_1, \dots, J_C$ as the occupied tables with $N_1, \dots, N_C$ customers each, and to J_{C+1} as the new empty table:

$$\mathbb{P}(\text{sits at table } l) \propto N_l + \eta, \quad l = 1, \dots, C \qquad (6)$$

$$\mathbb{P}(\text{sits at table } J_{C+1}) \propto \eta\Lambda\psi_1\left(1+\frac{1}{\psi_1\Lambda + C}\right). \quad (7)$$

We observe that, if $\eta := \alpha/\Lambda$ for some $\alpha > 0$ and let Λ go to ∞, so that the mixing measure *P* in (4) approaches in law the Dirichlet process with mass parameter *α*, then (6) reduces to *N_l* and (7) reduces to *α*, so that we obtain, a-priori, the clustering behaviour of the Chinese restaurant process (i.e. the clustering induced by the Dirichlet process with mass parameter *α*).

As a final note, we mention that the Chinese restaurant metaphor is quite popular in the machine learning and Bayesian nonparametric literature. It is widely adopted and it is customary to describe any related extensions, such as Favaro and Teh (2013) for normalised completely random measures and Teh et al. (2006) for Hierarchical Dirichlet processes, using this metaphor, and this is the approach we employ here.

5 Inference

Our MCMC algorithm iterates between the following steps: (a) conditionally on *N*, $c_1, \dots, c_N$ and $\tau_1, \dots, \tau_M$, we update the individual visit histories using an adaptive changepoint sampler, as described in Section 5.1, (b) conditionally on *N* and the individual visit histories, we update $c_1, \dots, c_N, S_1, \dots, S_M$ and $\tau_1, \dots, \tau_M$, using a conditional algorithm for mixture models, as described in Section 5.2. We note that as a result of this update, we also achieve an update for Ω, and, (c), conditionally on Ω and $\tau_1, \dots, \tau_M$, we update *N* using a rejection algorithm, as described in Section 5.3.

5.1 Adaptive MCMC changepoint sampler

The adaptive MCMC algorithm proposed by Benson and Friel (2018) gives rise to the posterior distribution of the latent vector indicating the presence (or not) of a changepoint, which in our case indicates an arrival time for an individual at each time point. We denote the iteration number, say *j*, for objects that are iteration specific using a superscript (*j*). The design of the algorithm ensures that time points that are unlikely to be changepoints are not proposed as frequently as time points that have been accepted as changepoints in earlier iterations of the algorithm. This is ensured by introducing individual- and time-specific weights where for individual *i*, each time point *t* has a specific weight, $\gamma_{it}^{(j)}$, of being proposed as a changepoint i.e. an arrival time and a specific weight, $\delta_{it}^{(j)}$, of being proposed to be deleted from a changepoint if it already is one. These weights are unnormalised selection weights and are updated if time *t* is accepted as or deleted from a changepoint, respectively as explained below.

For the current visit history described by $(\boldsymbol{a}_i^{(j)}, \boldsymbol{d}_i^{(j)})$ and corresponding presence history $\mathbf{PH}_i^{(j)}$, we define the set of all changepoints (arrivals) of individual *i* by $\mathcal{A}_i^{(j)} = \left\{ t : t \in \boldsymbol{a}_i^{(j)} \right\}$, the set of all departures by $\mathcal{D}_i^{(j)} = \left\{ t : t \in \boldsymbol{d}_i^{(j)} \right\}$, the set of points exterior to the visits, that is all *t* at which individual *i* is absent, by $\mathcal{E}_i^{(j)} = \left\{ t : \mathrm{PH}_{it}^{(j)} = 0 \right\}$, and the remaining *t*, which are interior to the visits, by $\mathcal{I}_i^{(j)}$; see Figure 2 for an illustration. Additionally, we define $\gamma_{+i}^{(j)} = \sum_{\left\{ t \notin \mathcal{A}_i^{(j)} \right\}} \gamma_{it}^{(j)}$ and $\delta_{+i}^{(j)} = \sum_{\left\{ t \in \mathcal{A}_i^{(j)} \right\}} \delta_{it}^{(j)}$. At iteration *j* of the algorithm we can propose to increase *k_i* by one, as long as $k_i < T$, or to decrease k_i by one, as long as $k_i > 1$. That is, we can propose to add a changepoint, with probability $\kappa_i^{(j)}$, at any $t \notin \mathcal{A}_i^{(j)}$ and a departure time while we can propose to delete any changepoint $t \in \mathcal{A}_i^{(j)}$ (and a departure time), with probability $1 - \kappa_i^{(j)}$. This leads to a proposed visit history described by $(\boldsymbol{a}_i^*, \boldsymbol{d}_i^*)$.

We describe the two cases in detail below.

Increasing k_i: We select time $a^* \notin \mathcal{A}_i^{(j)}$ as a proposed new changepoint with probability $\gamma_{ia^*}^{(j)} / \gamma_{i+}^{(j)}$.

We note that there is a non-zero probability of selecting any interior point $a^* \in \mathcal{I}_i^{(j)}$. The next step of the update, which is the proposal of the corresponding departure time, is conditional on whether $a^* \in \mathcal{I}_i^{(j)}$, which we refer to as a *split* move, or $a^* \in \mathcal{E}_i^{(j)}$, which we refer to as an *add* move.

Split move If $a^* \in \mathcal{I}_i^{(j)}$, then we denote the arrival time just before a^* by a_{im}, with $a_{im} = \max(t \in \mathcal{A}_i^{(j)} \backslash\& \, t < a^*)$ with corresponding departure time d_{im}, and we propose a departure time, d^* from $g_0(d^* - a_{im} \mid q_{0c})$, that is d^* is generated conditional on a_{im}, which forms a visit with this newly proposed departure time, while the newly proposed arrival time a^* forms a visit with the existing departure time that is the first to occur after a^* (see Figure 3 (a) for an illustration).

We accept this update with probability

$$\chi_1 = \min\left(1, \frac{f\left(\boldsymbol{a}_i^*, \boldsymbol{d}_i^*, q_{1c}^{(j)}, q_{0c}^{(j)}, p_c^{(j)} \mid \mathbf{CH}_i\right)}{f\left(\boldsymbol{a}_i^{(j)}, \boldsymbol{d}_i^{(j)}, q_{1c}^{(j)}, q_{0c}^{(j)}, p_c^{(j)} \mid \mathbf{CH}_i\right)} \frac{1-\kappa_i^{(j)}}{\kappa_i^{(j)}} \frac{\delta_{ia^\star}^{(j)} / (\delta_{ia^\star}^{(j)} + \delta_{i+}^{(j)})}{\gamma_{ia^\star}^{(j)} / (\gamma_{i+}^{(j)})} \frac{1/2}{g_0(d^\star - a_{im} \mid q_{0c})}\right)$$

Note that the 1/2 in the numerator of the last fraction is only needed if the visit to be deleted is not the first one as if that were the case then we cannot delete it by merging it with the previous visit.

The first fraction in χ_1 is the ratio of the joint posterior probabilities of the parameters and visit histories while the product of the three following fractions gives the ratio of the proposal probabilities when considering a delete move (numerator and also described below) and when considering an add move (denominator).

Add move If $a^\star \in \mathcal{E}_i^{(j)}$, we propose a departure time, $d^\star$, for this proposed visit from $g_0(d^\star - a^\star \mid q_{0c})$ (see Figure 3 (b) for an illustration).

We accept this update with probability

$$\chi_1 = \min\left(1, \frac{f\left(\boldsymbol{a}_i^*, \boldsymbol{d}_i^*, q_{1c}^{(j)}, q_{0c}^{(j)}, p_c^{(j)} \mid \mathbf{CH}_i\right)}{f\left(\boldsymbol{a}_i^{(j)}, \boldsymbol{d}_i^{(j)}, q_{1c}^{(j)}, q_{0c}^{(j)}, p_c^{(j)} \mid \mathbf{CH}_i\right)} \frac{1-\kappa_i^{(j)}}{\kappa_i^{(j)}} \frac{\delta_{ia^\star}^{(j)} / (\delta_{ia^\star}^{(j)} + \delta_{+i}^{(j)})}{\gamma_{ia^\star}^{(j)} / (\gamma_{+i}^{(j)})} \frac{1/2}{g_0(d^\star - a^\star \mid q_{0c})}\right)$$

We note here that the 1/2 in the numerator of the last fraction in χ_1 is needed only if the visit to be added is not going to be the first one as if that were the case then we could not delete it in the reverse move by merging it with the previous visit.

Increasing k_i: We select time $a^\star \in \mathcal{A}_i^{(j)}$ as a proposed changepoint to be deleted with probability $\delta_{ia^\star}^{(j)} / \delta_{i+}^{(j)}$. This arrival time to be considered for deletion forms a visit with departure time $d^\star$. Again we denote the arrival time just before $a^\star$ by a_{im}, with $a_{im} = \max(t \in \mathcal{A}_i^{(j)} \; \& \; t < a^\star)$ with corresponding departure time d_{im}.

At this stage, we propose, with probability 1/2, to delete the visit by merging it with the previous visit, which we refer to as a *merge move*, or not, which we refer to as a *delete move*. We describe the two cases below.

Merge move We note here that this is the inverse of the *split move* defined above (see Figure 3 (c) for an illustration). We accept this update with probability

$$\chi_0 = \min\left(1, \frac{f\left(\boldsymbol{a}_i^*, \boldsymbol{d}_i^*, q_{1c}^{(j)}, q_{0c}^{(j)}, p_c^{(j)} \mid \mathbf{CH}_i\right)}{f\left(\boldsymbol{a}_i^{(j)}, \boldsymbol{d}_i^{(j)}, q_{1c}^{(j)}, q_{0c}^{(j)}, p_c^{(j)} \mid \mathbf{CH}_i\right)} \frac{\kappa_i^{(j)}}{1-\kappa_i^{(j)}} \frac{\gamma_{ia^\star}^{(j)} / (\gamma_{ia^\star}^{(j)} + \gamma_{i+}^{(j)})}{\delta_{ia^\star}^{(j)} / \delta_{i+}^{(j)}} \frac{g_0(d_{im} - a_{im} \mid q_{0c})}{1/2}\right)$$

Delete move Alternatively, if we do not propose the merge move, then we propose to delete this visit and we accept this update with probability

$$\chi_0 = \min\left(1, \frac{f\left(\boldsymbol{a}_i^*, \boldsymbol{d}_i^*, q_{1c}^{(j)}, q_{0c}^{(j)}, p_c^{(j)} \mid \mathbf{CH}_i\right)}{f\left(\boldsymbol{a}_i^{(j)}, \boldsymbol{d}_i^{(j)}, q_{1c}^{(j)}, q_{0c}^{(j)}, p_c^{(j)} \mid \mathbf{CH}_i\right)} \frac{\kappa_i^{(j)}}{1-\kappa_i^{(j)}} \frac{\gamma_{ia^\star}^{(j)} / (\gamma_{ia^\star}^{(j)} + \gamma_{+i}^{(j)})}{\delta_{ia^\star}^{(j)} / \delta_{+i}^{(j)}} \frac{g_0(d^\star - a^\star \mid q_{0c})}{1/2}\right)$$

We note here that if $a^\star$ is the first change point then a_{im} does not exist and the visit cannot be merged but only be deleted and the probability that $a^\star$ is an interior point is zero.

Finally, at each iteration we propose to shift all arrival and departure times of all individuals using a Metropolis-Hastings move. We propose to shift each $a \in \mathcal{A}_i^{(j)}$ by adding to it a discrete Unif$\{-e_1, e_1\}$, while we propose to shift $d \in \mathcal{D}_i^{(j)}$ by adding to it a discrete Unif$\{-e_2, e_2\}$. We accept this update with probability

$$\min\left(1, \frac{f\left(\boldsymbol{a}_i^*, \boldsymbol{d}_i^*, q_{1c}^{(j)}, q_{0c}^{(j)}, p_c^{(j)} \mid \mathbf{CH}_i\right)}{f\left(\boldsymbol{a}_i^{(j)}, \boldsymbol{d}_i^{(j)}, q_{1c}^{(j)}, q_{0c}^{(j)}, p_c^{(j)} \mid \mathbf{CH}_i\right)}\right)$$

If the proposal to increase *k_i*, either via an add or a split move, is accepted, and hence a new visit at time *t* is introduced, then we update the corresponding weight according to the adaptation scheme proposed by Benson and Friel (2018),

$\log(\gamma_{it}^{(j+1)}) = \log(\gamma_{it}^{(j)}) + \frac{h}{j/T}(\chi_1 - \gamma_{\text{target}})$, where *h* > 0 is the initial adaptation parameter, *j* is the iteration number and γ_{target} is the target acceptance rate of the move.

Consequently, if the proposal to decrease *k_i*, either via a delete or a merge move, is accepted, and hence the visit that started at time *t* no longer exists, then we update the corresponding weight using $\log(\delta_{it}^{(j+1)}) = \log(\delta_{it}^{(j)}) + \frac{h}{j/T}(\chi_0 - \delta_{\text{target}})$.

5.2 Sampling the posterior distribution of *v* and the partition

Borrowing notation from the Bayesian nonparametric literature (Ishwaran and James, 2003), we develop a blocked Gibbs sampler for finite mixture models with a random number of components, as suggested by Argiento and De Iorio (2022), based on the posterior characterisation given in Section 4. As opposed to the marginal samplers considered in Taddy et al. (2012) and Matechou and Caron (2017), our approach does not require updating of the cluster-specific parameters every time an individual is removed from or added to a cluster. As a result, it is computationally more efficient.

Let $\mathcal{S} := \{s_1, ..., s_M\}$ be the unnormalized weights and $\mathcal{T} = \{\tau_1, \dots, \tau_M\}$ be the random locations of the model in (4). Drawing inference on $(\mathcal{S}, \mathcal{T}, M)$ is equivalent to drawing inference on *v*. Consequently, conditionally on *ρ* and $\{c_1^\star, \dots, c_C^\star\}$, our algorithm to update *v* is based on the following characterization of $(\mathcal{S}, \mathcal{T}, M)$ given the augmented data.

We refer to $c_1^\star, \dots, c_C^\star$ as the occupied components of the mixture and to the remaining *M* − *C* components as the unoccupied. An important observation in order to implement our algorithm is that we can always assume that the vector of unique values $c_1^\star, \dots, c_C^\star$ corresponds to the first *C* components of the mixture, i.e. $c_1^\star = 1, \dots, c_C^\star = C$. This statement as well as everything that follows in this Section are proven in Sections 1 and 2 of the Supplementary Material.

Under model (4), the intensity *v* given the augmented data $\{(\mathbf{PH}_i, \mathbf{CH}_i), i = 1, \dots, N\}$ and a sample $c_1, \dots, c_M$ of cluster allocation indices, such that $c_1^\star = 1, \dots, c_C^\star = C$, satisfies the following distributional equation

$$\nu(\mathbf{PH}_i, \mathbf{CH}_i) \overset{d}{=} \sum_{j=1}^{C} S_j f(\mathbf{PH}_i, \mathbf{CH}_i \mid \tau_j) + \sum_{c=C+1}^{M} S_c f(\mathbf{PH}_i, \mathbf{CH}_i \mid \tau_c)$$

In particular:

1. The number of unoccupied components, $M^{(\mathrm{no})} = M - C$, is distributed according to the mixture model
$$\frac{\psi_1 \Lambda}{\psi_1 \Lambda + C} P_1(M^{(\mathrm{no})} \mid \psi_1 \Lambda) + \frac{C}{\psi_1 \Lambda + C} \mathrm{Poisson}(M^{(\mathrm{no})} \mid \psi_1 \Lambda).$$
2. Conditionally on $M^{(\mathrm{no})}$, the weights of unoccupied components, S_j for $C < j \le M$, are iid with distribution Gamma$(\eta, \zeta + 1)$.
3. The locations of the unoccupied components τ_j, for $C < j \le M$ are iid from the prior distribution P_0.
4. The weights of occupied components, S_j for $1 \le j \le C$, are independent each with distribution Gamma$(\eta + N_l, \zeta + 1)$.
5. Conditionally on everything else,
$$\begin{aligned}\pi(\Lambda \mid \mathrm{rest}) &\propto \phi_1(C + a_\Lambda - 1)\mathrm{Gamma}(\Lambda \mid C + a_\Lambda, 1 - \phi_1 + b) \\ &\quad + C(1 - \phi_1 + b_\Lambda)\mathrm{Gamma}(\Lambda \mid C + a_\Lambda - 1, 1 - \phi_1 + b) \\ \pi(\eta \mid \mathrm{rest}) &\propto \mathrm{eppf}_N(N_1, \dots, N_C, C)\mathrm{Gamma}(\eta \mid a_\eta, b_\eta) \\ \pi(\zeta \mid \mathrm{rest}) &\propto \mathrm{eppf}_N(N_1, \dots, N_C, C)\mathrm{Gamma}(\zeta \mid a_\zeta, b_\zeta)\end{aligned}$$
6. The locations of occupied components $\theta_j^\star = \tau_j$, for $1 \le j \le C$ are independent and distributed as
$$\pi(d\theta_j^\star) \propto \prod_{i \in J_j} f\left(\mathbf{PH}_i, \mathbf{CH}_i \mid \theta_j^\star\right) P_0(d\theta_j^\star) \qquad (8)$$

The latter is the posterior density of a parametric Bayesian model where the sampling model is $f(\mathbf{PH}_i, \mathbf{CH}_i \mid \theta^\star)$, the prior distribution is P_0 and the data are the augmented observations in cluster J_j, namely $\{\mathbf{CH}, \mathbf{PH}\}_j := \{\mathbf{CH}_i, \mathbf{PH}_i, i \in J_j\}$. The sampling model of the posterior distribution in (8) is given in Equation (3) with $\theta^\star = (q_{1c}, q_{0c}, p_c)$, and a prior distribution, P_0, with independent components is specified on $\theta^\star$. In particular, assuming an independent Beta(α_{p_c}, β_{p_c}) on p_c gives as full conditional a

Beta($\alpha_{p_c} + \sum_{\{i \in c\}} \sum_{t=1}^{T} \mathrm{CH}_{it}\mathrm{PH}_{it}, \beta_{p_c} + \sum_{\{i \in c\}} \sum_{t=1}^{T} (1 - \mathrm{CH}_{it})\mathrm{PH}_{it}$). On the other hand, a conjugate prior cannot be identified for q_{0c} or q_{1c}. Hence, we specify a Beta($\alpha_{q_0}, \beta_{q_0}$) and a Beta($\alpha_{q_1}, \beta_{q_1}$), respectively, and resort to an MH update where the target is derived by Equation (8).

Conditionally on *v* and *N*, we update *ρ* and $\{c_1^\star, \dots, c_C^\star\}$ by, first, reparameterising *ρ* and $\{c_1^\star, \dots, c_C^\star\}$ in terms of the individual cluster allocation indices $c_1, \dots, c_N$ and, then, by sampling the new allocation *c_i* for individual *i* from $P(c_i = c) \propto S_c f(CH_i, PH_i \mid \tau_c), c = 1, \dots, M, \ \ i = 1, \dots, N$.

5.3 Sampling N

We update *N* using a rejection algorithm, as introduced by Matechou and Caron (2017) and also employed by Diana et al. (2020). We outline the algorithm below and provide more details in Section 2.3 of the Supplementary Material.

First, we propose a value for the number of individuals in *N* that were never caught from a Poisson(Ω) distribution. Consecutively, we thin this number as follows: we first allocate each of the proposed individuals to a cluster according to our process outlined in Section 4, then we simulate a presence history for each individual using the cluster-specific q_1 and q_0 parameters and finally, given the presence history, we simulate a capture history for individuals using the cluster-specific capture probability, p_c. Individuals with simulated capture histories with at least one non-zero entry are discarded, as any caught individuals are already part of the sample. Finally, the obtained sample from the posterior distribution of *N* at the particular iteration consists of the remainder of simulated individuals with capture histories with all entries equal to zero, together with the *n* individuals caught at least once in the original sample.

5.4 Summarising the clustering

Inference on clustering and cluster-specific parameters is made marginally with respect to the (latent) uncaught individuals, so that all of the summary statistics presented in this section are obtained using the *n* caught individuals.

Our iterative algorithm results in G draws from the posterior distribution of the random partition, ρ. In Bayesian model-based clustering, the choice of a single partition based on this posterior sample is a critical point. Here we employ the approach of Wade and Ghahramani (2018) because it also explores partitions that are not visited by the MCMC chain. This approach is essentially a greedy search algorithm informed by the posterior similarity matrix, which is obtained by calculating the average number of times each pair of individuals has been allocated to the same cluster.

Estimation of cluster-specific parameters is a well-known and open problem in Bayesian nonparametric model-based clustering, mainly due to the label-switching problem. Here we employ the simple and intuitive approach proposed by Molitor et al. (2010). Once we obtain an estimation of the data clustering, i.e. $\hat{c}_1, \ldots, \hat{c}_n$ that define the partition of $\{1, \ldots, n\}$ called $\hat{J}_1, \ldots, \hat{J}_{\hat{C}}$, as explained above, we average each of the cluster specific parameters across all G iterations to obtain

$$\hat{\theta}_j^{\star} = \frac{1}{G} \sum_{g=1}^{G} \mathrm{mean}_{i \in \hat{J}_j} \tau_{\hat{c}_i}^{(g)}, \qquad j = 1, \ldots, \hat{C}.$$

6 Results

6.1 Simulation

We simulated data on populations exhibiting temporary emigration and heterogeneity to assess the performance of our model and algorithm in identifying the number of clusters and in estimating the associated parameters as well as the size of the population.

We set T = 100, N = 500, C = 2, $q_{11} = 0.01, q_{12} = 0.1$, $q_{01} = 0.1, q_{02} = 0.3, p_1 = 0.2$ and $p_2 = 0.5$. We fit the model of Equation (1) using different scenarios for parameters Λ, η and ζ and to five different simulated data for each scenario. In all cases, we let P_0 be the product of three independent uniform distributions for p, q_0 and q_1 to represent absence of information for these parameters. Then, we investigate three different scenarios regarding parameters Λ, η and ζ. In the first two scenarios, the three parameters are fixed: in the first case $\mathbb{E}(N) = 500$, and in the second $\mathbb{E}(N) = 1000$. Finally, in the third scenario, the three parameters are assumed random with independent Gamma prior distributions (see Model 1). Details on the hyperparameter choices are given in Tables 1, 2 and 19 of the Supplementary Material.

For each simulation, we run our algorithm using initial adaptive weights equal to 1, *h* = 0.001 and $\gamma_{\text{target}} = \delta_{\text{target}} = 0.2$. The choice of *h* was made after running several chains for the same data set using the same starting values for the parameters but different values for *h* ($h = 0, 0.001, 0.005, 0.01$) and comparing effective sample sizes and mean squared errors for all model parameters. Our choice for the target acceptance rates was based on the general recommendation of a 0.234 acceptance rate in MCMC algorithms (Roberts et al., 1997) and the comment by Benson and Friel (2018) that values in the $[0.01, 0.2]$ range work well in practice. We note here that Benson and Friel (2018) suggest that the algorithm is insensitive to the choice of *h* and of the acceptance rates, provided that the first is set to a small value, such as the reciprocal of the length of the time-series.

Our results in all scenarios suggest that clustering estimation is quite robust with respect to choices on parameters Λ, *η* and *ζ*, thanks to the method of Wade and Ghahramani (2018). In fact, in most cases we obtain an a posteriori estimate of two clusters with high rand index (i.e. >0.76) between the true partition and the estimated partition (see Tables 5, 6 and 21 of the Supplementary Material). Estimation of *N* and of cluster specific parameters is also good in terms of coverage of the posterior credible intervals (PCIs).

We note that some issues in posterior inference can arise in the extreme case where the prior induced on *N*, *C* by the parameter choice is such that these two quantities are very strongly correlated (see for instance case a of Table 2 of the Supplementary material). We also mention that if a cluster consists of individuals with very low capture probabilities, the model can slightly overestimate this parameter. Finally, we highlight that the best performance in terms of estimation is obtained when all three parameters are assumed random with vague priors, such as independent Gamma(0.1,0.1) priors (this is denoted as case g in Table 19 of the Supplementary material), and this is indeed the approach we take for the angler data set.

Under this third hyperparameter scenario, clustering estimation is satisfying, since for all five replications we obtain a posterior mean of two clusters, with average width of the 95% PCI equal to 1.80, and Rand index of 0.76.

We present the average, over the five data sets, of the posterior mean $\hat{A}$ and the average width $\hat{L}$ of the 95% PCI as well as the coverage for N and all cluster-specific parameters in Table 1. Posterior summaries for the visit, arrival and departure patterns obtained for one of the simulated data sets are presented in Figure 4. In all cases, $\hat{A}$ is close to the true value used to simulate the data, while the coverage is satisfactory, and always higher than 0.8.

Hence, our results demonstrate that we can estimate parameters of interest, such as N, and retrieve the latent clustering of the population as well as estimate the corresponding cluster-specific parameters. More details about the simulation study are presented in Section 2.2 of the Supplementary Material.

6.2 Angler data

We fit model of Equation (1) to three data sets collected in 2017, 2018 and 2019. We set $p = 0$ on days when the river was closed for fishing in each season. We adopt an informative approach in the choice of the hyperparmameters of P_0, that is the choice of $\alpha_{q_1}, \beta_{q_1}$, $\alpha_{q_0}, \beta_{q_0}$ and α_p, β_p. Our information can be summarized as follows:

(a) an a priori 95% credible interval for the number of visits equal to $[1, 12]$, with the lower bound representing individuals who only visit the river once and the upper bound representing individuals who visit on a weekly basis during the season. Since $T = 92$ here, this gives a (1/92, 12/92) corresponding interval for q_1. Using a set of non-linear equations, we translated this into a Beta(2.8, 48.7) prior distribution for q_1.

(b) an a-priori mean length of stay for an angler equal to 6 days, with a 50% credible interval of roughly [3,7] days. This gives a, roughly, (0.12, 0.26) corresponding interval for q_0, which translates into a Beta(3, 12) prior distribution.

(c) a 95% prior credible interval for p equal to (0.1,0.5);

Finally, we assigned vague prior distributions on the parameters Λ, η and ζ, as described in Section 6.1.

Similarly to the simulation study, we run our algorithm for each data set using initial adaptive weights equal to 1, $h = 0.001$ and target acceptance rates equal to 0.2. Convergence was assessed by visual inspection of trace plots and using the Geweke diagnostic in package coda (Plummer et al., 2006), shown in Section 3 of the Supplementary Material.

Posterior summaries for N for each year are presented in Table 2. The posterior medians for N are all greater than 3000 and considerably greater than the number of anglers observed each year. There is substantial overlap between all 95% PCIs, suggesting that the population of anglers that visited the river each year between 2017 and 2019 is fairly stable, which is not surprising, as at least 70% of anglers are expected to be returning year after year.

As explained in Section 4, our model induces a clustering amongst anglers, described by our extension of the Chinese restaurant process. We clarify that in this case, the clients are the anglers and the tables are groups of anglers sharing the same characteristics (dishes), with characteristics being the cluster-specific parameters (q_0, q_1, p), which describe the visit behaviour and fishing ability of anglers. We summarise the clustering as described in Section 5.4 and obtain two clusters. Contour plots of the posterior distributions for q_0 and q_1 for each cluster and posterior summaries of the cluster-specific probabilities of capture are shown in Figure 5. The largest cluster each year (labelled cluster 1, consisting of 95% of the individuals) is found to consist of the “occasional visitors”, with the lowest q_1 and highest q_0. This group of anglers tend to perform between two and three visits per year, with posterior mean visit duration equal to around two days. The smallest cluster corresponds to the “super-visitors” who make on average around eight visits per year, with posterior mean visit duration between two and three days. We highlight that the estimated capture probability for both clusters, which, although higher for the second cluster is not found to differ significantly between clusters within each season, is considerably greater than what is obtained when a model that does not account for temporary emigration is fitted to the data, as mentioned in Section 2.

The estimated arrival, departure and presence patterns, presented in Figure 6, vary substantially between years, although a common feature is a larger proportion of individuals estimated to arrive on the first day of the season and similarly to depart on the last day of the season. A peak in the estimated number of individuals present (third column in Figure 6) is observed towards the end of June in all years. This is possibly due to anglers tending to avoid planning a trip to Gaula early in the season as some parts of the river may still be inaccessible due to snow, which is much less likely after the end of June. In 2018 and 2019, there is another peak in the number of anglers present, following river closure at the end of July in both years. We note that, as mentioned above, anglers book their trips typically months in advance, so they tend to remain at the river even when it is closed, waiting for it to reopen once there is sufficient rainfall.

The cluster-specific visit pattern can also be summarised by looking at the posterior probability of presence at any given time point for each individual. In Figure 7 we present these for the most representative individual in each of the two clusters for each year. Individuals representative for each cluster are identified by calculating the distance between each individual and the centroid of each cluster in terms of the number and length of visits and capture probability. The posterior probability of presence is equal to 1 at times when the individual has been observed. This probability decreases as we move away from times of observation. The rate of decrease depends on the cluster and on the number of days between observations. Specifically, the posterior probability of presence reduces slightly more sharply as we move away from the time of observation for individuals in cluster 1 compared to cluster 2. However, even for cluster 1, individuals observed only a few days apart have a high posterior probability of presence between these two observations. As expected, the number of peaks, i.e. visits, for each individual corresponds to what we have already identified as typical for each cluster, with individuals in cluster 1 having one or two peaks and individuals in cluster 2 multiple peaks. We focused our interpretation of the clustering results on the individuals observed at least once, since no information, other than the fact that they never caught salmon, is available on the remainder of the individuals. Since our inference on *N* is performed using a rejection algorithm, the number of individuals that were never observed as well as their corresponding presence histories and hence cluster membership change at every iteration of the algorithm. Therefore, no meaningful posterior summaries exist for particular individuals with index greater than *n*.

6.3 The effect of the COVID-19 pandemic

In this section, we discuss the results of our model when fitted to the 2020 data set, presented in Section 3 of the Supplementary Material, and compare them to those obtained for the 2017-19 seasons. In Norway, there were no national travel restrictions in the summer of 2020, so local anglers could still travel to Gaula, but the international travel restrictions meant that the number of anglers travelling from abroad to Norway was dramatically lower. As a result, both the visiting patterns of anglers and their probability of catching salmon whilst at the river were considerably different to pre-pandemic seasons, since the foreign visiting anglers are among the most dedicated and skilled.

We find that we again identify two clusters of anglers, with substantially different visit patterns. The largest cluster, consisting of almost 90% of anglers in the sample, visit one to two times on average in the season. However, the smallest cluster of what we termed “super-visitors” only visit on average around three times in the season, which is considerably lower than in the pre-pandemic seasons. On the other hand, both clusters spent longer at the river per visit on average compared to previous seasons (around six days), which results in average total lengths of stay at the river that are greater than in other seasons (around 15 and 37 days, respectively). Additionally, we estimate that the total number of anglers that visited the river in 2020 was greater than in pre-pandemic seasons (posterior median = 4034, 95% PCI = (3550, 4583)). We note here that once travel restrictions were imposed, the pre-sold fishing licenses where made available to locals in Norway and, given the global travel restrictions, also preventing Norwegians from going on their normal summer vacations, more locals took up the new opportunity to try fishing in Gaula. Finally, our estimated capture probabilities for both clusters are slightly lower than in pre-pandemic seasons (posterior means equal to 0.18 and 0.25, respectively), which is again not surprising as less experienced anglers took the opportunity to visit the river in 2020 compared to the regular and more experienced anglers who usually buy the fishing licenses months in advance.

7 Conclusions

We proposed a novel modelling approach that enables us to estimate parameters of interest, such as the size of a population, from CR data whilst allowing for temporary emigration and accounting for individual heterogeneity. To our knowledge, this is the only such model for open-populations that does not require the use of methods that rely on Pollock's robust design, which cannot always be employed, as is the case for the data set considered in this paper.

Our approach brings together CR models, changepoint process and Bayesian nonparametric methodology for the first time and gives rise to a flexible modelling technique for time-series that could be extended to other types of ecological data, such as those collected in occupancy studies (MacKenzie et al., 2002) when the assumption of population closure is not satisfied.

The model is fitted by combining and extending recently-developed algorithms in the area of changepoint processes (Benson and Friel, 2018) and a blocked Gibbs sampler for finite mixture models built upon a Bayesian nonparametric approach (Argiento and De Iorio, 2022; Argiento et al., 2016; Ishwaran and James, 2003). The first learn from past states of the MCMC and hence yield proposal distributions that quickly discover the position of changepoints while the latter updates the cluster-specific parameters conditioning on the mixing process P, which yields an independent update (for these parameters) that is more efficient than competing approaches.

We presented a simulation study that demonstrated the reliability of the results obtained by our model and the ability of our method to uncover the underlying clusters in the population as well as to estimate the corresponding cluster-specific parameters. Our analysis of CR data sets on anglers in Norway identified two clusters, each with fairly consistent visit patterns each year, except in 2020, when the effect of the COVID-19 pandemic led to a different angling population. Our results can be used to inform the decisions made by the river management board in terms of number and pricing of fishing licenses issued, based on the quality of fishing in terms of probability to catch a fish on a given day. Reliable estimates of the probability of anglers catching salmon whilst at the river are crucial for effective management of the river and can also be used as a marketing tool by the river management board, as success rates are used as indicators for the quality of fishing experience at rivers. On the other hand, estimates that are biased low, obtained from existing models that do not account for temporary emigration, risk underestimating the effect of angling on the salmon population and cannot guide actions for sustainable fishing at the river.

A natural extension of our model is the inclusion of covariates, for example the effect of rainfall on capture probability throughout the season, to inform estimation about clustering and cluster specific parameters. The use of covariates in mixture models is part of a lively stream of research in the Bayesian nonparametric community (see for instance Foti and Williamson, 2013), under the name of (non-exchangeable) dependent processes for mixtures. One of the main advantages of the framework of finite mixture models introduced in Argiento and De Iorio (2022), which we employ in this paper, is exactly the possibility of incorporating covariate dependence in the model, and this is a matter of current research. Another interesting extension could consider following the same individuals over different seasons to study the level of consistency in their visiting pattern and the potential improvement of their fishing ability over the years. In 2018, the river closed for the first time in its history because of low levels of rainfall, and since then river closure towards the middle of the season is a recurring phenomenon. As a result, it is expected that anglers will change their behaviour about when to book their fishing trip in the future, and our model, or extensions of it described above, could be used to model this behavioural shift.

SUPPLEMENTARY MATERIAL

Title: Supplementary Material Document with three sections: 1. Bayesian mixture model, 2: Posterior characterisation, 3: Angler data. The first two give details on the mathematical results of the Bayesian mixture model presented in the paper and the algorithm used for inference, with an additional extensive simulation study on prior sensitivity in section 2.2 and the third includes trace plots and diagnostics for all model parameters for the analysis of the data presented in the paper.

R code and data: The angler data analysed in Section 6.2 and the R code used to perform the simulation study and the angler data analysis and to elicit the prior distributions for a number of model parameters in the simulation study presented in the supplementary material.

Acknowledgements

We are grateful to Reinhard Fellner for bringing this case study to our attention, and for all the discussions on the prior knowledge about the system as well as his insights that highlight the importance of our findings and the potential impact of the work presented in this paper.

The authors report there are no competing interests to declare.

References

Aldous, D. J. 1985. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII–1983*, pages 1–198. Springer.

Argiento, R., Bianchini, I., Guglielmi, A., et al. 2016. Posterior sampling from *varepsilon*-approximation of normalized completely random measure mixtures. *Electronic Journal of Statistics*, 10(2):3516–3547.

Argiento, R. and De Iorio, M. 2022. Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. *Annals of Statistics (to appear)*.

Benson, A. and Friel, N. 2018. Adaptive MCMC for multiple changepoint analysis with applications to large datasets. *Electronic Journal of Statistics*, 12(2):3365–3396.

Charalambides, C. A. 2005. *Combinatorial methods in discrete distributions*, volume 600. John Wiley & Sons.

Darroch, J. N. 1958. The multiple-recapture census: I. Estimation of a closed population. *Biometrika*, 45(3/4):343–359.

Diana, A., Matechou, E., Griffin, J., Johnston, A., et al. 2020. A hierarchical dependent Dirichlet process prior for modelling bird migration patterns in the UK. *Annals of Applied Statistics*, 14(1):473–493.

Favaro, S. and Teh, Y. W. 2013. MCMC for normalized random measure mixture models. *Statistical Science*, 28(3):335–359.

Foti, N. J. and Williamson, S. A. 2013. A survey of non-exchangeable priors for Bayesian nonparametric models. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):359–371.

Ishwaran, H. and James, L. F. 2003. Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, pages 1211–1235.

Jolly, G. M. 1965. Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika*, 52(1/2):225–247.

Kottas, A. and Sansó, B. 2007. Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis. *Journal of Statistical Planning and Inference*, 137(10):3151–3163.

Laake, J. 2013. RMark: An R interface for analysis of capture-recapture data with MARK. AFSC Processed Rep. 2013-01, Alaska Fish. Sci. Cent., NOAA, Natl. Mar. Fish. Serv., Seattle, WA.

Lyons, J. E., Kendall, W. L., Royle, J. A., Converse, S. J., Andres, B. A., and Buchanan, J. B. 2016. Population size and stopover duration estimation using mark–resight data and Bayesian analysis of a superpopulation model. *Biometrics*, 72(1):262–271.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., and Langtimm, C. A. 2002. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255.

Manrique-Vallier, D. 2016. Bayesian population size estimation using Dirichlet process mixtures. *Biometrics*, 72(4):1246–1254.

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(1):21–40.

Matechou, E., Nicholls, G. K., Morgan, B. J., Collazo, J. A., and Lyons, J. E. 2016. Bayesian analysis of Jolly-Seber type models. *Environmental and Ecological Statistics*, 23(4):531–547.

Miller, J. W. and Harrison, M. T. 2018. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356.

Molitor, J., Papathomas, M., Jerrett, M., and Richardson, S. 2010. Bayesian profile regression with an application to the National Survey of Children's Health. *Biostatistics*, 11(3):484–498.

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. 1978. Statistical inference from capture data on closed animal populations. *Wildlife monographs*, (62):3–135.

Pitman, J. 1996. Some developments of the Blackwell-Macqueen urn scheme. In Ferguson, T. S., Shapley, L. S., and B., M. J., editors, *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell*, volume 30 of *IMS Lecture Notes-Monograph Series*, pages 245–267. Institute of Mathematical Statistics, Hayward (USA).

Pledger, S. 2000. Unified maximum likelihood estimates for closed capture–recapture models using mixtures. *Biometrics*, 56(2):434–442.

Pledger, S., Efford, M., Pollock, K., Collazo, J., and Lyons, J. 2009. Stopover duration analysis with departure probability dependent on unknown time since arrival. In *Modeling demographic processes in marked populations*, pages 349–363. Springer.

Plummer, M., Best, N., Cowles, K., and Vines, K. 2006. CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11.

Pollock, K. H. 1982. A capture-recapture design robust to unequal probability of capture. *The Journal of Wildlife Management*, 46(3):752–757.

Roberts, G. O., Gelman, A., and Gilks, W. R. 1997. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 7(1):110–120.

Schwarz, C. J. and Arnason, A. N. 1996. A general methodology for the analysis of capture-recapture experiments in open populations. *Biometrics*, 52(3):860–873.

Seber, G. A. F. 1965. A note on the multiple-recapture census. *Biometrika*, 52(1/2):249–259.

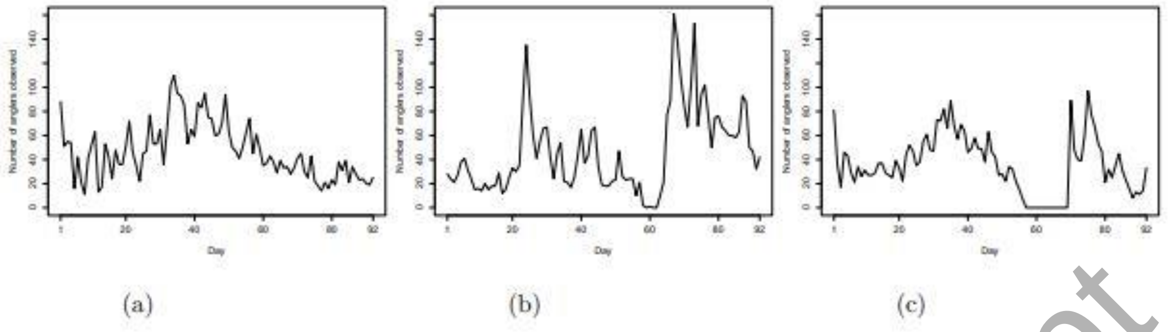
Taddy, M. A., Kottas, A., et al. 2012. Mixture modeling for marked Poisson processes. *Bayesian Analysis*, 7(2):335–362.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Wade, S. and Ghahramani, Z. 2018. Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, 13(2):559–626.

Zhou, M., McCrea, R. S., Matechou, E., Cole, D. J., and Griffiths, R. A. 2019. Removal models accounting for temporary emigration. *Biometrics*, 75(1):24–35.

Fig. 1 Number of anglers observed each day of the season in 2017, (a), 2018, (b), and 2019, (c). The river may be closed occasionally for fishing, which results in days when no fish are caught, such as the days around day 60 in 2018 and in 2019.



(a)

(b)

(c)

Fig. 2 Example presence history for individual *i*. Gray ticks indicate external points, $\mathcal{E}_i$, black ticks indicate internal points, $\mathcal{I}_i$, solid circles indicate changepoints, i.e. arrivals times, $\mathcal{A}_i$, and solid triangles indicate their corresponding departure times, $\mathcal{D}_i$. The arrival and departure times are indicated on the x-axis.

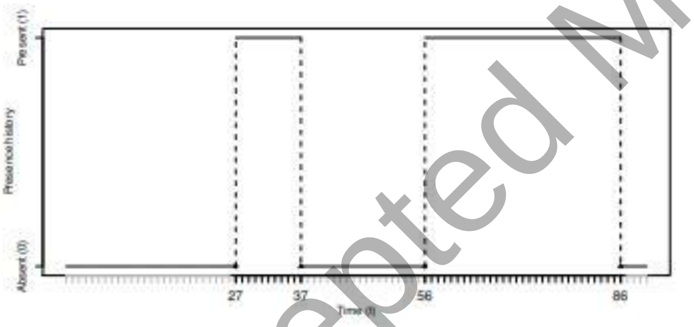
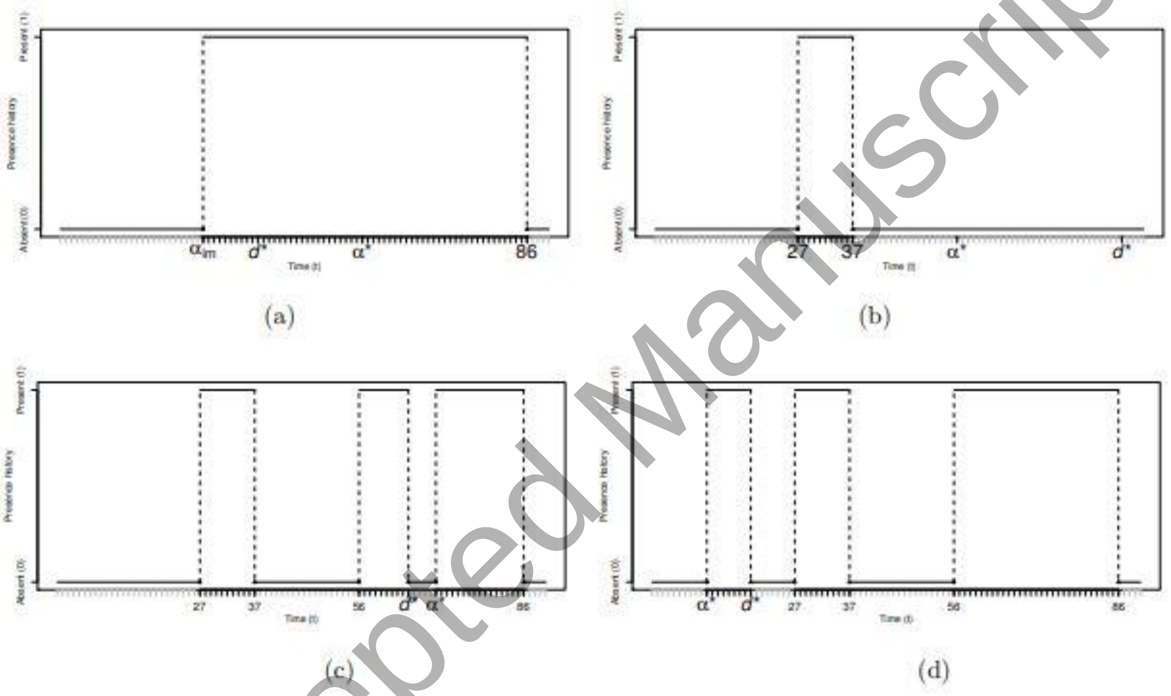


Fig. 3 Four examples of presence histories that can lead to the presence history shown in Figure 2: (a) Split move; we propose an internal point as a new changepoint ($a^* = 56$), which forms a visit with the next available departure point (d_{im} = 86), while the existing previous arrival time a_{im} = 27 forms a visit with the newly proposed departure time ($d^* = 37$), (b) Add move; we propose an external point as a new changepoint ($a^* = 56$), which forms a visit with the newly proposed departure time ($d^* = 86$), (c) Merge move; we propose to delete an existing changepoint ($a^* = 70$) and the departure time of the previous visit ($d^* = 56$), so the two visits are merged, and (d): Delete move: we propose to delete an existing changepoint ($a^* = 11$) and its corresponding departure time ($d^* = 19$).



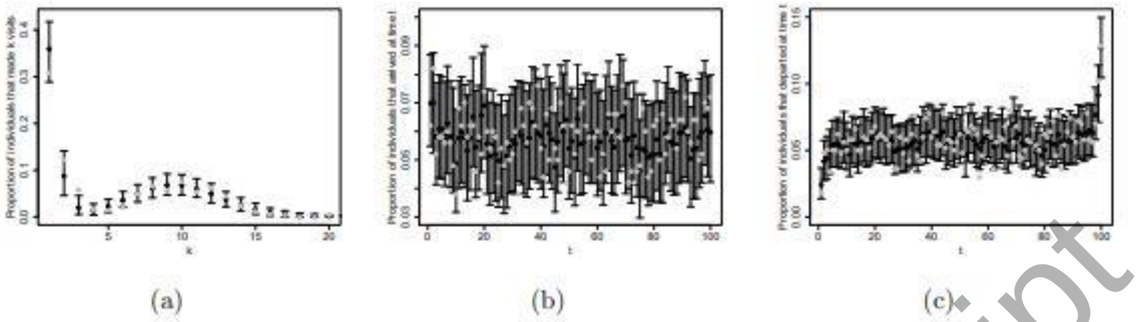
(a)

(b)

(c)

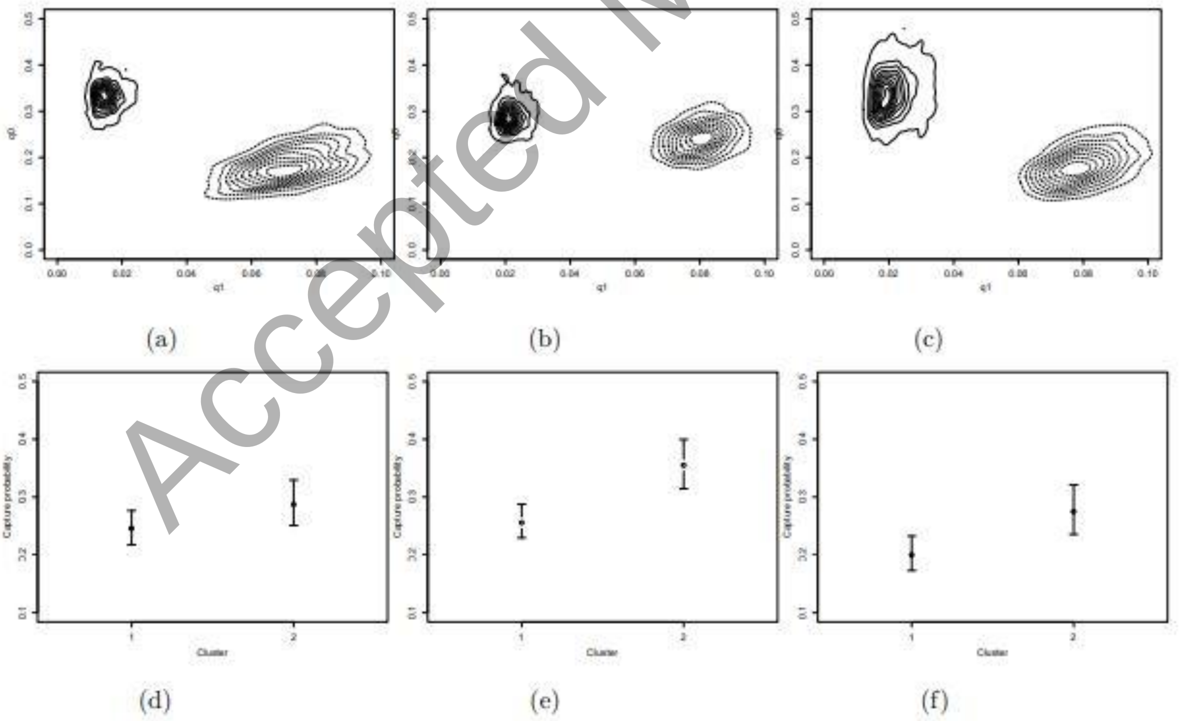
(d)

Fig. 4 Posterior medians, indicated by the empty circles, and 95% PCI, indicated by the vertical bars, of the visit, (a), arrival, (b), and departure, (c), patterns. The true values are indicated by the gray dots. The plots refer to one (among five) of the simulated data, results for other data sets are quite similar.



(a) (b) (c)

Fig. 5 Contour plots of cluster-specific q_0 and q_1 for the 2017, (a), 2018, (b) and 2019, (c), season for cluster 1, solid lines, and cluster 2, dashed lines. Summaries of posterior samples of cluster-specific capture probabilities for the 2017, (d), 2018, (e) and 2019, (f), season.



(a) (b) (c)

(d) (e) (f)

Fig. 6 Posterior medians, indicated by the empty circles, and 95% PCI, indicated by the vertical bars, of the number of individuals arriving (first column), departing (second column), and being present (third column), each day of the season in 2017 (first row), 2018 (second row) and 2019 (third row). The gray dots in the third column indicate the number of anglers who caught at least one salmon that day.

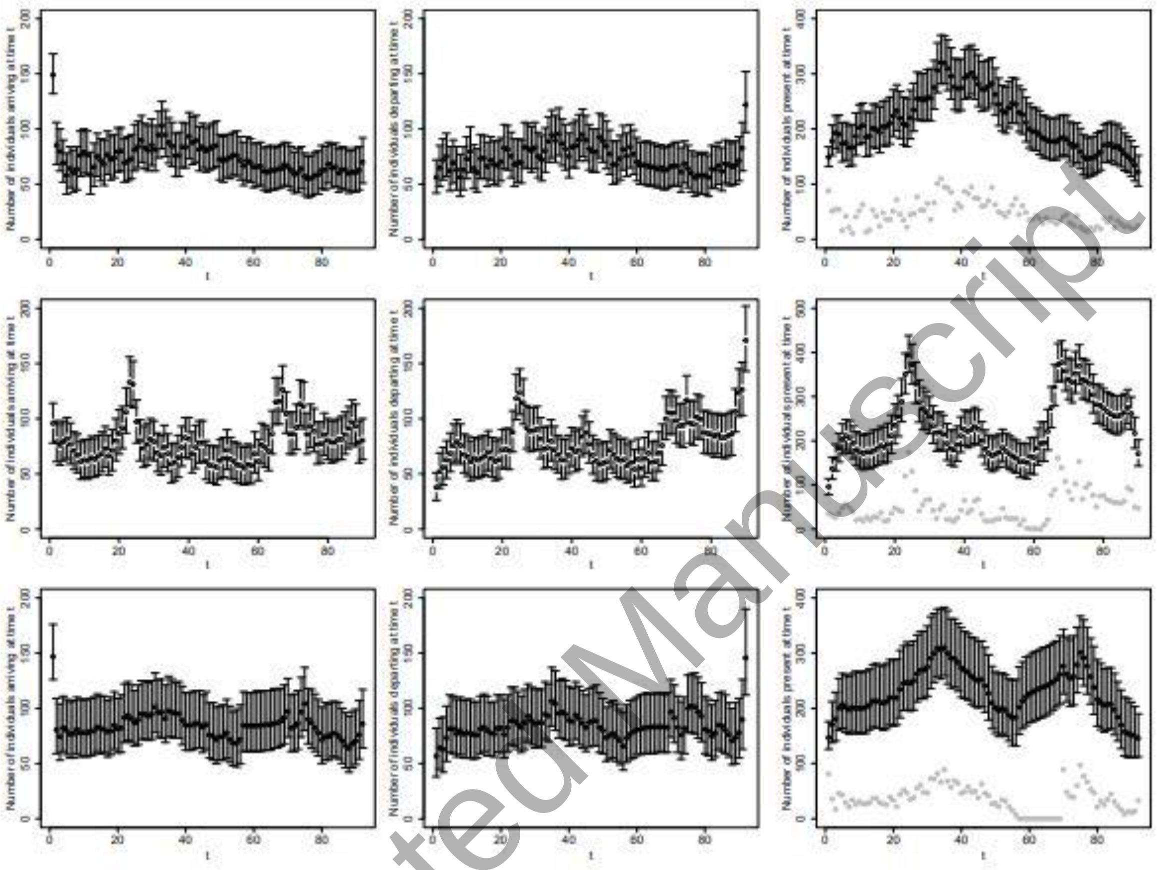


Fig. 7 Posterior probabilities of presence for the most representative individuals in each of the two identified clusters (rows 1 and 2) for 2017 (first column), 2018 (second column) and 2019 (third column).

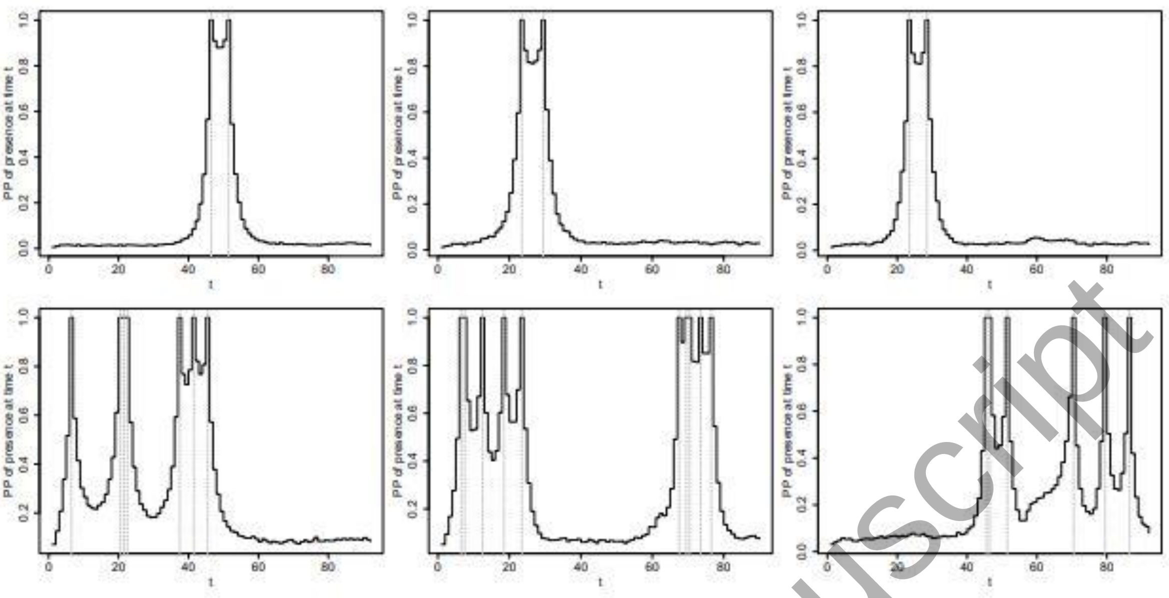


Table 1 Simulation results. Average over five replications of the posterior mean ($\hat{A}$), average width ($\hat{L}$) of the 95% PCI and coverage of population size and all cluster-specific parameters.

Parameter	N	q_{11}	q_{12}	q_{01}	q_{02}	p_1	p_2
$\hat{A}$	505.28	0.02	0.09	0.11	0.29	0.22	0.50
$\hat{L}$	45.20	0.01	0.01	0.05	0.08	0.06	0.08
cover.	1.00	1.00	1.00	0.80	0.80	0.80	0.80

Table 2 Posterior summaries of the population size, *N*, of anglers for each season. The number of anglers observed each season, *n*, is given in the second column of the table.

Year	*n*	Posterior Median	95% PCI
2017	2057	3499	(3208, 3871)
2018	1901	3142	(2857, 3472)
2019	1687	3495	(3036, 3988)