Efficient sequential Monte Carlo algorithms for integrated population models

Efficient sequential Monte Carlo algorithms for integrated population models

Axel Finke[1] a.finke@ucl.ac.uk
Ruth King[2,3] ruth.king@ed.ac.uk
Alexandros Beskos[1,3] a.beskos@ucl.ac.uk
Petros Dellaportas[1,3,4] p.dellaportas@ucl.ac.uk

[1]Department of Statistical Science, University College London, U.K.
[2]School of Mathematics, University of Edinburgh, U.K.
[3]The Alan Turing Institute, U.K.
[4]Department of Statistics, Athens University of Economics and Business, Greece

Abstract

In statistical ecology, state-space models are commonly used to represent the biological mechanisms by which population counts – often subdivided according to characteristics such as age group, gender or breeding status – evolve over time. As the population counts are typically only noisily or partially observed, the information from the count data alone is not sufficient for sensibly estimating demographic parameters of interest. Thus, the count data are combined with additional ecological observations to form an integrated data analysis. Unfortunately, fitting integrated models can be challenging, especially if the constituent state-space model is non-linear/non-Gaussian. We first propose an efficient particle Markov chain Monte Carlo algorithm to estimate demographic parameters without the need for resorting to linear or Gaussian approximations. We then incorporate this algorithm into a sequential Monte Carlo sampler in order to perform model comparison with regards to the dependence structure of demographic parameters. In particular, we exploit the integrated model structure to enhance the efficiency of both algorithms. We demonstrate the methods on two real data sets: little owls and grey herons. For the owls, we find that the data do not support an ecological hypothesis found in the literature. For the herons, our methodology highlights the limitations of existing models which we address through a novel regime-switching model.

Key words: Bayesian inference; Capture-recapture; Integrated population models; Model comparison; Sequential Monte Carlo; State-space models.

1 Introduction

State-space models are becoming an increasingly common and useful representation of many ecological systems (**???**). For example, they are used to describe population count data (**???**); telemetry data (**???**); longitudinal growth data (**?**); fisheries data (**?**); capture-recapture data (**???**).

Unfortunately, fitting state-space models leads to computational challenges as the likelihood – expressible only as integral or sum over the latent (unobserved) states – is typically intractable unless: the states take values in a small, finite set; or the model is linear Gaussian in which case the likelihood is evaluated via the Kalman filter (**??**). Two approaches are typically applied to circumvent this problem. The first is to approximate the state-space model with a model that is linear and Gaussian, e.g. as in **?**. Unfortunately, such approximations introduce a bias which is difficult to quantify. The second is to impute the unobserved states alongside the model parameters within a Markov chain Monte Carlo (MCMC) approach. Unfortunately, such data-augmentation schemes – in particular as normally implemented in BUGS (**?**) or JAGS (**?**) (see e.g. **?**) – can be slow and poorly mixing if the system states and parameters are highly correlated because only (small) subsets of them are updated individually (**?**).

To avoid the problems with these approaches, (**??**) proposed particle Markov chain Monte Carlo (PMCMC) algorithms (see **?** for a recent application in ecology). Such algorithms replace the intractable likelihood in the Metropolis–Hastings algorithm with an unbiased estimate obtained through a sequential Monte Carlo (SMC) algorithm (or "particle filter"). PMCMC methods do not require a modification of the model and, despite replacing the likelihood with an approximation, do not introduce bias.

Integrated population models. In this work, count data are available on some species of interest, i.e. estimates of population sizes over discrete times (**???**). These count data are modelled as a state-space model to account for measurement errors. In

addition to the count data, other types of data are available on the species, e.g. capture-recapture, ring-recovery or nest-record data. To utilise all available information for estimating demographic parameters of interest, it is necessary to combine these different data sets within a single *integrated population model*. Unfortunately, fitting such models is challenging, since they inherit all the above-mentioned difficulties with fitting the constituent state-space model.

Contributions. In this work, we devise efficient methodology for performing full Bayesian parameter estimation and model comparison in integrated population models without the need for linear or Gaussian approximations to the state-space model.

- In Section 3, we first review standard PMCMC methods for Bayesian parameter estimation in models with intractable likelihoods. Then, in Subsection 3.2, we exploit the integrated-model structure to reduce the computational cost of the PMCMC algorithm through a delayed-acceptance (**?**) technique.

- In Section 4, we first incorporate our PMCMC methodology into SMC samplers (**????**) so that we can estimate posterior model probabilities (Bayes factors) across a set of different integrated population models. This permits Bayesian model comparison without the need for reversible-jump MCMC (**?**) which often mixes poorly and can be difficult to implement and tune. Then, in Subsection 4.3, we again exploit the integrated-model structure to reduce the computational burden of the SMC sampler by separately tempering the different likelihood terms.

- In Sections 5 and 6, we apply the proposed methodology to two real data sets relating to little owls and grey herons. In both applications, our methodology yields reliable estimates of the model evidence, even in moderately high dimensions. In the case of owls, we find that some models proposed in the literature may be unnecessarily over-parametrised. For instance, we find no evidence for the hypothesis in **?** that the immigration rate of owls depends on the abundance of voles – their main prey. We also demonstrate the utility of the delayed-acceptance approach. In the case of herons, we show that state-of-the-art models used in the literature, such as

the threshold model from **?**, fit poorly; to remedy this, we propose a novel regime-switching state-space model which significantly outperforms all existing models in terms of model fit and model evidence.

2 Integrated model

2.1 Data

We combine multiple data sets, one of which being count data, obtained from a single population, within a single integrated model. Let $\mathbf{y} = \{y_1, \ldots, y_T\}$ denote *count data* collected at times $t = 1, \ldots, T$. Here, y_t is the observation (subject to measurement error) of the true population size at time t. The observed counts may be multivariate, e.g. counts for males and females or juveniles and adults, though in all the examples we consider later the count data are univariate. Let $\mathbf{w}$ denote all *additional data* available such as capture-recapture data, ring-recovery or nest-record data. The aim of this work is then to perform inference based on *all data* $\mathbf{z} = \{\mathbf{y}, \mathbf{w}\}$. To illustrate the methodology, we consider two data sets relating to little owls and grey herons.

Little owls. In Section 5, we consider little-owl data described by **?** and subsequently analysed in **?**. The count data represent the number of breeding females at nest boxes near Göppingen, South Germany, observed anually from 1978 to 2003 (i.e. $T = 26$). The nest boxes were checked multiple times annually and data were recorded relating to overall population size (number of occupied nest boxes and number of breeding females), capture-recapture histories of individuals observed at nest boxes and reproductive success of the nests. In addition, time-varying covariate information about the abundance of voles – the primary prey for little owls – is available. For further details see **?**.

Grey herons. In Section 6, we consider grey-heron data previously presented and analysed by **???**. The count data correspond to the estimated number of female herons (or breeding pairs) in the UK, from 1928 to 1998, i.e. for $T = 71$ time periods. Within our application we also have ring-recovery data for individuals released between 1955 and 1997.

2.2 Model structure

Given unknown model parameters $\theta \in \Theta$, the likelihood of the count data $\mathbf{y}$ and additional data $\mathbf{w}$ is $p(\mathbf{z}|\theta) = p(\mathbf{y}|\theta, \mathbf{w})p(\mathbf{w}|\theta)$. To simplify the notation, and consistent with ecological practice (see e.g. **?**), we assume that the count data are independent of the additional data given the parameters, i.e. $p(\mathbf{y}|\theta, \mathbf{w}) = p(\mathbf{y}|\theta)$ (see **?** for a justification). However, we stress that this conditional independence is purely a modelling choice and our methodology remains applicable without it. Our methodology only requires that the additional data are modelled in such way that $p(\mathbf{w}|\theta)$ can be evaluated pointwise.

The count data are modelled as a state-space model as follows. Let $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\} \in \mathsf{X}^T$ (for some space X) denote the true (unobserved) population counts with initial density $\mu_\theta(\mathbf{x}_1)$ and transitions $f_\theta(\mathbf{x}_t|\mathbf{x}_{t-1})$. Furthermore, let $g_\theta(y_t|\mathbf{x}_t)$ be the density of the tth observed count given $\mathbf{x}_t$. Then, conditionally on θ, the joint distribution of $\mathbf{y}$ and $\mathbf{x}$ is:

$$p(\mathbf{y}, \mathbf{x}|\theta) = \mu_\theta(\mathbf{x}_1)g_\theta(y_1|\mathbf{x}_1)\prod_{t=2}^{T} f_\theta(\mathbf{x}_t|\mathbf{x}_{t-1})g_\theta(y_t|\mathbf{x}_t).$$

The (marginal) count-data likelihood is thus given by the integral (or sum, if X is discrete)

$$p(\mathbf{y}|\theta) = \int_{\mathsf{X}^T} p(\mathbf{y}, \mathbf{x}|\theta)\,\mathrm{d}\mathbf{x}. \tag{1}$$

Throughout this work, we assume that this integral (sum) is intractable as is usually the case unless X is finite and sufficiently small or the state-space model is linear and Gaussian in which case (1) can be evaluated using the Kalman filter.

Let $p(\theta)$ denote the prior distribution of the parameters then the (marginal) *posterior* distribution of the parameters θ (given the full data $\mathbf{z}$) is given by

$$\pi(\theta) \coloneqq p(\theta|\mathbf{z}) = \frac{p(\mathbf{z}|\theta)p(\theta)}{p(\mathbf{z})}; \quad p(\mathbf{z}) \coloneqq \int_\Theta p(\mathbf{z}|\theta)p(\theta)\mathrm{d}\theta, \tag{2}$$

where $p(\mathbf{z})$ in the denominator is the *evidence* for the model. This quantity plays a key rôle in Bayesian model comparison as outlined in Section 4. The posterior distribution is typically intractable as the integrals in (1), (2) are not of closed form. Instead, we approximate it via Monte Carlo methods as described in the next section.

3 Parameter estimation

3.1 Particle MCMC

In this section, we describe MCMC methods for approximating the posterior distribution of the model parameters. We also propose modifications which exploit the structure of integrated models to improve efficiency of the algorithm. For now, we assume that the model is known – model uncertainty is dealt with in Section 4.

As the count-data likelihood $p(\mathbf{y}|\theta)$ (thus, also the overall likelihood $p(\mathbf{z}|\theta)$) is intractable, we cannot implement the *idealised* Metropolis–Hastings algorithm which, at each iteration, proposes new parameters $\vartheta \sim q(\vartheta|\theta)$ and accepts them with probability (w.p.) $\min\{1, \frac{q(\theta|\vartheta)}{q(\vartheta|\theta)} \frac{p(\vartheta)}{p(\theta)} \frac{p(\mathbf{z}|\vartheta)}{p(\mathbf{z}|\theta)}\}$. A common solution is to use data-augmentation approaches that impute the latent variables $\mathbf{x}$ (alongside the parameters). However, the number of states is typically large so that single-site updates are required. This approach, commonly used in 'black-box' samplers such as BUGS or JAGS, can lead to poor mixing if highly correlated variables or parameters are updated separately.

To avoid such problems, we employ particle Markov chain Monte Carlo (PMCMC) algorithms (**?**). These replace $p(\mathbf{y}|\theta)$ in the acceptance ratio of the idealised Metropolis–Hastings algorithm with an unbiased estimate $\hat{p}(\mathbf{y}|\theta)$ obtained through a sequential Monte Carlo (SMC) methods. Crucially, the resulting algorithm still targets the correct posterior distribution.

Before stating the PMCMC algorithm, we review SMC algorithms. Comprehensive discussions of the application of SMC to state-space models – usually termed particle filters (PFs) in this setting – can be found in **??**. A simple PF is outlined in Algorithm 1, where we use the convention that actions prescribed for the nth particle are to be performed independently for all $n \in \{1, \ldots, N\}$.

---

1 Algorithm (particle filter).

(1) Sample $\mathbf{x}_1^n \sim \mu_\theta(\mathbf{x}_1)$ and set $w_1^n := g_\theta(y_1|\mathbf{x}_1^n)$.

(2) At Steps $t = 2, \ldots, T$,

(a) sample $a_{t-1}^n = l$ w.p. $W_{t-1}^l \coloneqq w_{t-1}^l / \sum_{k=1}^N w_{t-1}^k$,

(b) sample $\mathbf{x}_t^n \sim f_\theta(\mathbf{x}_t | \mathbf{x}_{t-1}^{a_{t-1}^n})$ and set $w_t^n \coloneqq g_\theta(y_t | \mathbf{x}_t^n)$.

At the end of Algorithm 1, an unbiased (**?**) estimate of $p(\mathbf{y}|\theta)$ is

$$\hat{p}(\mathbf{y}|\theta) \coloneqq \prod_{t=1}^{T} \frac{1}{N} \sum_{n=1}^{N} w_t^n.$$

Numerous extensions exist for making Algorithm 1 more efficient. The particular version of PF we use in our applications is outlined in Web Appendix C.

We now describe the PMCMC algorithm. A single PMCMC update is outlined in Algorithm 2, where $\alpha \in [0, 1]$ is a parameter which will be used by the evidence-approximation algorithms in Section 4. For the moment simply consider $\alpha = 1$.

---

2 Algorithm (particle MCMC). At each iteration, given $(\theta, \hat{p}(\mathbf{y}|\theta))$,

(1) propose $\vartheta \sim q(\vartheta|\theta)$ and generate $\hat{p}(\mathbf{y}|\vartheta)$ using Alg. 1 (wherein $\theta = \vartheta$),

(2) return $(\vartheta, \hat{p}(\mathbf{y}|\vartheta))$ w.p. $\min\{1, r\}$, where $r \coloneqq \dfrac{q(\theta|\vartheta)}{q(\vartheta|\theta)} \dfrac{p(\vartheta)}{p(\theta)} \left[\dfrac{\hat{p}(\mathbf{y}|\vartheta) p(\mathbf{w}|\vartheta)}{\hat{p}(\mathbf{y}|\theta) p(\mathbf{w}|\theta)} \right]^\alpha$;
otherwise, return $(\theta, \hat{p}(\mathbf{y}|\theta))$.

3.2 Improving PMCMC efficiency for integrated models

The computational cost of the PMCMC update in Algorithm 2 is dominated by the PF used to evaluate the estimate of $p(\mathbf{y}|\vartheta)$ for each proposed parameter value ϑ. To improve the efficiency of algorithm, we utilise a delayed-acceptance (DA) approach (**??**) based on the integrated-model structure. The idea is to avoid invoking the PF for proposed values ϑ which are not compatible with the additional data $\mathbf{w}$ and which are therefore likely to be rejected in Algorithm 2. This can improve efficiency if $\mathbf{w}$ is highly informative about a large proportion of the model parameters. DA was previously combined with PMCMC updates in **?** (though in a slightly different way). Algorithm 3 summarises the approach whose validity may be established using the arguments of **??**. Again, assume for the moment that $\alpha = 1$.

3 Algorithm (delayed acceptance PMCMC). At each iteration, given $(\theta, \hat{p}(\mathbf{y}|\theta))$:

(1) Propose $\vartheta \sim q(\vartheta|\theta)$,

(2) Go to Step 3 w.p. $\min\{1, r\}$, where $r \coloneqq \frac{q(\theta|\vartheta)}{q(\vartheta|\theta)} \frac{p(\vartheta)}{p(\theta)} \left[\frac{p(\mathbf{w}|\vartheta)}{p(\mathbf{w}|\theta)}\right]^{\alpha}$; otherwise, return $(\theta, \hat{p}(\mathbf{y}|\theta))$.

(3) Generate $\hat{p}(\mathbf{y}|\vartheta)$ using Alg. 1 (with $\theta = \vartheta$).

(4) Return $(\vartheta, \hat{p}(\mathbf{y}|\vartheta))$ w.p. $\min\{1, r\}$, where $r \coloneqq \left[\frac{\hat{p}(\mathbf{y}|\vartheta)}{\hat{p}(\mathbf{y}|\theta)}\right]^{\alpha}$; otherwise, return $(\theta, \hat{p}(\mathbf{y}|\theta))$.

4 Model comparison

4.1 Posterior model probabilities

Let $\{\mathcal{M}_i \colon i \in \mathcal{I}\}$ denote a finite collection of plausible biological models. To indicate the ith model, we add the model indicator $\mathcal{M}_i$ to densities in Section 2. So, the prior of parameters $\theta \in \Theta_i$ is written as $p(\theta|\mathcal{M}_i)$, the likelihood as $p(\mathbf{z}|\theta, \mathcal{M}_i) = p(\mathbf{y}|\theta, \mathcal{M}_i)p(\mathbf{w}|\theta, \mathcal{M}_i)$ and the evidence as $p(\mathbf{z}|\mathcal{M}_i) = \int_{\Theta_i} p(\mathbf{z}|\theta, \mathcal{M}_i)p(\theta|\mathcal{M}_i)\,\mathrm{d}\theta$. Let $p(\mathcal{M}_i)$ denote the prior probability of the ith model. Bayesian model comparison is based on the *posterior model probabilities* (**?**, Chapter 6)

$$p(\mathcal{M}_i|\mathbf{z}) \coloneqq \frac{p(\mathcal{M}_i)p(\mathbf{z}|\mathcal{M}_i)}{\sum_{j\in\mathcal{I}} p(\mathcal{M}_j)p(\mathbf{z}|\mathcal{M}_j)}. \tag{3}$$

Unfortunately, the model evidence $p(\mathbf{z}|\mathcal{M}_i)$ – hence also the posterior model probabilities – in (3) is intractable. To perform model comparison, we replace the model evidence $p(\mathbf{z}|\mathcal{M}_i)$ with an estimate $\hat{p}(\mathbf{z}|\mathcal{M}_i)$ obtained via an SMC sampler. As a by-product, the SMC sampler also yields an approximation of the posterior of θ under the ith model.

4.2 SMC sampler for evidence approximation

For the moment, assume that $p(\mathbf{y}|\theta, \mathcal{M}_i)$ can be evaluated. A simple importance-sampling approximation of $p(\mathbf{z}|\mathcal{M}_i)$ is then given by $\frac{1}{M}\sum_{m=1}^{M} p(\mathbf{z}|\theta^m, \mathcal{M}_i)$, where $\theta^1, \ldots, \theta^M$ are sampled independently from $p(\theta|\mathcal{M}_i)$. However, this approach typically performs poorly

if there is a strong mismatch between the prior and the posterior (which is common, especially if θ is high-dimensional or if the data are highly informative). To circumvent this problem, we employ an SMC sampler (**??**) which uses successive importance-sampling steps to approximate a *sequence* of distributions to smoothly bridge the gap between the prior and the posterior,

$$p(\theta|\mathcal{M}_i) = \pi_0(\theta), \pi_1(\theta), \ldots, \pi_S(\theta) = p(\theta|\mathbf{z}, \mathcal{M}_i). \tag{4}$$

The idea behind SMC samplers is that each individual importance-sampling step (i.e. proposing samples from $\pi_{s-1}(\theta)$ to approximate $\pi_s(\theta)$) may be feasible even if the gap between prior $\pi_0(\theta)$ and posterior $\pi_S(\theta)$ is wide. We use a likelihood-tempering approach,

$$\pi_s(\theta) \propto p(\theta|\mathcal{M}_i)p(\mathbf{z}|\theta, \mathcal{M}_i)^{\alpha_s}, \tag{5}$$

where the *temperatures* $0 = \alpha_0 < \alpha_1 < \ldots < \alpha_S = 1$ (and the number of bridging distributions, S) can then be tuned to ensure that the interpolation between the prior and posterior in (4) is sufficiently smooth. Of course, in the models considered in this work, $p(\mathbf{y}|\theta)$ is intractable and is therefore again approximated using a PF (for any $0 < \alpha_s < 1$, the distributions targeted by the algorithm are then actually slightly different from (5) but we stress that this does not affect the validity of the algorithm). This idea was first employed by **?** and it shares some similarities with the SMC^2 approach from **?** which we discuss at the end of this section.

Algorithm 4 outlines the SMC sampler; we use the convention that any action specified for the mth particle is to be performed independently for *all* $m \in \{1, \ldots, M\}$.

---

4 Algorithm (SMC sampler).

(1) Sample $\theta_0^m \sim p(\theta|\mathcal{M}_i)$ and generate $\hat{p}_0^m(\mathbf{y}|\theta_0^m, \mathcal{M}_i)$ using Alg. 1 (with $\theta = \theta_0^m$),

(2) At Step $s = 1, \ldots, S$,

  (a) set $v_s^m \coloneqq (u_{s-1}^m)^{\alpha_s - \alpha_{s-1}}$, where $u_{s-1}^m \coloneqq \hat{p}_{s-1}^m(\mathbf{y}|\theta_{s-1}^m, \mathcal{M}_i)p(\mathbf{w}|\theta_{s-1}^m, \mathcal{M}_i)$,

  (b) sample $b_{s-1}^m = l$ w.p. $V_s^l \coloneqq v_s^l / \sum_{k=1}^M v_s^k$,

  (c) sample $(\theta_s^m, \hat{p}_s(\mathbf{y}|\theta_s^m, \mathcal{M}_i))$ using Alg. 2 (with $\alpha = \alpha_s$; $\theta = \theta_{s-1}^{b_{s-1}^m}$; $\hat{p}(\mathbf{y}|\theta) = \hat{p}_{s-1}^{b_{s-1}^m}(\mathbf{y}|\theta_{s-1}^{b_{s-1}^m}, \mathcal{M}_i)$).

We then approximate the evidence $p(\mathbf{z}|\mathcal{M}_i)$ by

$$\hat{p}(\mathbf{z}|\mathcal{M}_i) \coloneqq \prod_{s=1}^{S} \frac{1}{M} \sum_{m=1}^{M} v_s^m.$$

The algorithm can also be used to infer parameters in the ith model. That is, a posterior expectation $\mathbb{E}[\varphi(\theta)]$, for $\theta \sim p(\theta|\mathbf{z}, \mathcal{M}_i)$ and test function φ, is approximated by $\sum_{m=1}^{M} V_S^m \varphi(\theta_S^m)$. Numerous extensions can make Algorithm 4 more efficient. The particular version of SMC sampler we use in our applications is given in Web Appendix C.

Other methods for performing model comparison using SMC samplers can be found in **?** (see also **?** for extensions). In addition, **?** proposed another special case of the SMC-sampler framework from **?**, called *SMC²*. This algorithm can be useful when one wishes to perform inference *sequentially* because it can incorporate new data points as they arrive. However, as observed in **?**, SMC^2 can become unstable when the newly-arrived observation contains information about the parameters which contradicts the existing information. In such cases, the likelihood-tempering approach adopted here can lead to a smoother sequence of target distributions **?** and hence more accurate estimates.

4.3 Improving SMC efficiency for integrated models

We are able to exploit the structure of integrated population models to enhance the efficiency of the SMC sampler for evidence approximation. Firstly, we employ the DA approach from Subsection 3.2 to reduce the computational cost of the MCMC updates in the SMC sampler. Secondly, we propose to employ a likelihood-tempering approach which tempers the different parts of the likelihood separately. That is, for some $1 \leq S' < S$, the SMC sampler targets the distributions

$$\pi_s(\theta) \propto \begin{cases} p(\theta|\mathcal{M}_i)p(\mathbf{w}|\theta, \mathcal{M}_i)^{\alpha_s}, & \text{if } 0 \leq s \leq S', \\ p(\theta|\mathcal{M}_i)p(\mathbf{w}|\theta, \mathcal{M}_i)p(\mathbf{y}|\theta, \mathcal{M}_i)^{\beta_s}, & \text{if } S' < s \leq S, \end{cases}$$

where $0 = \alpha_0 < \alpha_1 < \ldots < \alpha_{S'} = 1$ and $0 < \beta_{S'+1} < \ldots < \beta_S = 1$. Of course, the intractable count-data likelihood is again replaced by an unbiased estimate. The advantage of this refined tempering scheme is that the approximation of the count-data

likelihood (obtained via the costly PF) is not needed in the first S' steps, so that S' can be taken to be large. Introducing the additional data likelihood first can be especially beneficial if these data are highly informative about the parameters relative to the count data. This refined tempering strategy was crucial for obtaining reliable estimates in the herons example from Section 6; its efficiency gains are also illustrated in Subsection 5.4.

5 Example 1: Little owls

5.1 Parameters

The main model parameters – potentially specific to age group $a \in \{1, \mathrm{A}\}$ (1: juvenile, i.e. first-year, A: adult) and gender $g \in \{\mathrm{m}, \mathrm{f}\}$ (f: female, m: male) of the owls, and to time index $t \in \{1, \ldots, T\}$ – are

$\phi_{a,g,t}$**:** probability of an owl of gender g surviving until time $t+1$ if alive and aged a at time t;

$p_{g,t+1}$**:** probability of observing an owl of gender g at time $t+1$ if alive at time $t+1$;

ρ_t**:** productivity rate governing the number of chicks produced per female at time t that survive to fledgling;

η_t**:** immigration rate governing the number of female immigrants at time $t+1$ per female of the population at time t.

5.2 Model specification

We consider the model defined by **?** and subsequently fitted in BUGS by **?** – for further information and biological rationale see these papers.

5.2.1 Count-data model

The system process, in terms of the true population sizes for the juvenile and adult females, $\mathbf{x}_t = \{x_{1,t}, x_{\mathrm{A},t}\}$, is described by

$$x_{1,t}|\mathbf{x}_{t-1}, \theta \sim \text{Poisson}\big([x_{1,t-1} + x_{\mathrm{A},t-1}]\rho_{t-1}\phi_{1,\mathrm{f},t-1}/2\big), \quad x_{\mathrm{A},t} = sur_t + imm_t,$$

where $sur_t|\mathbf{x}_{t-1}, \theta \sim \text{Binomial}(x_{1,t-1} + x_{\mathrm{A},t-1}, \phi_{\mathrm{A},\mathrm{f},t-1})$ is the number of female adults which survive from time $t-1$ to time t, and $imm_t|\mathbf{x}_{t-1}, \theta \sim \text{Poisson}\big([x_{1,t-1} + x_{\mathrm{A},t-1}]\eta_{t-1}\big)$ is the number of female adults which immigrate in this period. We take the initial population sizes $x_{1,1}$, $x_{\mathrm{A},1}$ to be a-priori independently distributed according to a discrete uniform law on $\{0, 1, \ldots, 50\}$. The observation process is specified by $y_t|\mathbf{x}_t, \theta \sim \text{Poisson}(x_{1,t}+x_{\mathrm{A},t})$.

5.2.2 Capture-recapture model

Capture-recapture data are available in the form of age-group and gender specific matrices $\mathbf{m} \coloneqq \{\mathbf{m}_{a,g} \colon a \in \{1, \mathrm{A}\}, g \in \{\mathrm{m}, \mathrm{f}\}\}$. The tth row, $\mathbf{m}_{a,g,t} \coloneqq \{m_{a,g,t,s} \colon 1 < s \leq T+1\}$, corresponds to the tth year of release ($t \in \{1, \ldots, T-1\}$). That is, $m_{a,g,t,s}$ is the number of individuals of gender g, last observed at age a at time t, that are recaptured at time s (if $t+1 \leq s \leq T$) or never recaptured again (if $s = T+1$). Note that $m_{a,g,t,s} = 0$ if $s \leq t$. For each year of release, we assume a multinomial distribution for the subsequent recaptures. The capture-recapture model specified as

$$\mathbf{m}_{a,g,t}|R_{a,g,t}, \theta \sim \text{Multinomial}(R_{a,g,t}, \mathbf{q}_{a,g,t}).$$

Here, $R_{a,g,t}$ denotes the number of owls in age group a and of gender g that are recorded as being observed (either an initial capture or, if $a = \mathrm{A}$, as a recapture) at time t and subsequently released. The multinomial cell probabilities $\mathbf{q}_{a,g,t} \coloneqq \{q_{a,g,t,s} \colon 1 < s \leq T+1\}$ are given by

$$q_{a,g,t,s} \coloneqq \begin{cases} 0, & \text{if } 1 < s \leq t, \\ \phi_{a,g,t} p_{g,s} \prod_{r=t+1}^{s-1} \phi_{\mathrm{A},g,r}(1 - p_{g,r}), & \text{if } t < s \leq T, \\ 1 - \sum_{r=1}^{T} q_{a,g,t,r}, & \text{if } s = T+1. \end{cases}$$

5.2.3 Fecundity model

Nest record data $\mathbf{n} \coloneqq \{N_t, n_t \colon 1 \leq t \leq T\}$ are also available to provide information relating to the fecundity rate of little owls. Specifically, N_t denotes the number of breeding females recorded at time t and n_t the number of chicks produced that survive to leave the nest. Following **?** we specify $n_t | N_t, \theta \sim \text{Poisson}(N_t \rho_t)$. With this notation, the set of all additional data is $\mathbf{w} = \{\mathbf{m}, \mathbf{n}\}$.

5.3 Parametrisation and priors

There is additional covariate information about the abundance of voles – the primary source of prey for little owls – classified as low ($vole_t = 0$) or high ($vole_t = 1$), for each year of the study. Following **??**, we parametrise

$$\begin{aligned}
\operatorname{logit} \phi_{a,g,t} &= \alpha_0 + \alpha_1 \mathbb{I}\{g = \mathrm{m}\} + \alpha_2 \mathbb{I}\{a = \mathrm{A}\} + \alpha_3 year_t, \\
\log \eta_t &= \delta_0 + \delta_1 vole_t, \\
\operatorname{logit} p_{g,t+1} &= \beta_1 \mathbb{I}\{g = \mathrm{m}\} + \beta_{t+1},
\end{aligned}$$

for $t = 1, \ldots, T-1$, where the additional covariate $year_t$ denotes the normalised year. Also, we specify $\log \rho_t = \gamma_t$, for $1 \leq t \leq T$.

For simplicity, we assume that all components of θ have independent Normal$(0, 2)$ priors, except δ_0 for which we use a Normal$(-2, 2)$ prior because preliminary runs of the algorithm indicated that this parameter is typically very small. We avoided diffuse priors (a) to improve efficiency of the first few steps of the SMC sampler and (b) to reduce the impact of the phenomenon known as Jeffreys–Lindley paradox (**?**) on the model comparison. We stress that other prior specifications could have been employed but investigating the choice of priors in integrated population models is beyond the scope of the work.

5.4 Results

We end this section by demonstrating the performance gains achievable for the PMCMC and SMC algorithms through the modifications proposed in Subsections 3.2 and 4.3. We also perform a model comparison to demonstrate the scientific utility of our methodology.

Delayed acceptance. We first illustrate the performance gains obtained through the delayed-acceptance (DA) approach proposed in Subsection 3.2. For simplicity, we only report results for the case that the productivity rate is constant over time and with immigration independent of the abundance of voles, i.e. $\gamma_1 = \ldots = \gamma_T$ and $\delta_1 = 0$, as this was one of the specifications which performed best in terms of model evidence. Figure 1 illustrates the utility of DA. It shows that even though DA decreases the acceptance rate (**?**), the computational savings due to only invoking the PF for "promising" parameters more than compensate for this.

Refined tempering. In Table 1, we illustrate efficiency gains attainable through the refined likelihood tempering scheme (Section 4.3) over standard likelihood tempering (Section 4.2). For eight different models (specified in Web Appendix A), Table 1 displays

$$(\textit{efficiency gain}) = \frac{MSE \times (\textit{computation time})}{MSE \times (\textit{computation time})} \Big\}^{\text{standard tempering}}_{\text{refined tempering}}. \tag{6}$$

Here, *MSE* denotes the average mean-square error (MSE) of the estimate of the posterior mean based on 20 independent repeats of the SMC samplers (the average is taken over all components of the vector of model parameters); (*computation time*) represents the average computation time over the independent repeats. Since the true posterior means are intractable, we ran an MCMC algorithm using a large number $(10,000,000)$ of iterations for each model and treated the resulting posterior mean estimates as the true values.

Model comparison. Finally, we perform a model comparison to investigate the hypothesis from **?** that little-owl immigration depends on the abundance of voles – their main prey. Figure 2 shows estimates of the evidence for the eight models specified in Web Appendix A in the case that the immigration rate may depend on the abundance of voles (i.e. $\delta_1 \neq 0$) and in the case that it is independent of vole abundance (i.e. $\delta_1 = 0$). The results indicate that the hypothesis is not supported by the data.

6 Example 2: Grey herons

6.1 Parameters

Following **?** we specify up to four age categories for the herons in order to account for different survival probabilities, with younger herons typically having a lower survival probability than older adults. We indicate the age group by the subscript $a \in \{1, \ldots, \mathrm{A}\}$, where $a = 1$ represents first-years, $a = 2$ represents second-years, etc. while $a = \mathrm{A}$ represents all the remaining adults. The main model parameters are then

$\phi_{a,t}$**:** probability of a heron surviving until time $t+1$ if alive and aged a at time t;

ρ_t**:** productivity rate governing the number of females produced per female at time t;

λ_t**:** probability of recovering a dead heron in $[t, t+1)$ if it died in that interval.

6.2 Model specification

We follow **?** with regard to the model specification, allowing for some judicial changes in the state-space model specification.

6.2.1 Count-data model

We once again specify state-space model for the count data $\mathbf{y} = \{y_1, \ldots, y_T\}$. We let $x_{a,t}$, denote the true population sizes of herons in age group a at time t. The system process is then described by

$$
\begin{aligned}
x_{t,1}|\mathbf{x}_{t-1}, \theta &\sim \text{Poisson}(\rho_{t-1}\phi_{1,t-1}\textstyle\sum_{a=2}^{\mathrm{A}} x_{a,t-1}),\\
x_{a,t}|\mathbf{x}_{t-1}, \theta &\sim \text{Binomial}(x_{a-1,t-1}, \phi_{a,t-1}), \quad \text{for } 1 < a < \mathrm{A},\\
x_{\mathrm{A},t}|\mathbf{x}_{t-1}, \theta &\sim \text{Binomial}(x_{\mathrm{A}-1,t-1} + x_{\mathrm{A},t-1}, \phi_{\mathrm{A},t-1}).
\end{aligned}
$$

For simplicity, we assume that the distribution of each component of the initial state is a negative-binomial distribution with probability $p = 1/100$ and size $n_0 = \mu_0 p/(1-p)$ for age groups $1 \leq a < \mathrm{A}$ and $n_1 = \mu_1 p/(1-p)$ for adults, respectively. We specify the means $\mu_0 = 5000/5$ and $\mu_1 = 5000 - (\mathrm{A}-1)\mu_0$ in such a way that a-priori, $\mathbb{E}[\sum_{a=1}^{A} x_{a,1}|\theta] = 5000$.

Such state-space model are typically approximated by a linear-Gaussian model in order to permit inference via the Kalman filter (**?**). However, the assumption that $g_\theta(y_t|\mathbf{x}_t)$ is Gaussian is typically unrealistic, since it implies that the observation error is independent of scale and continuous. Alternatively, to incorporate the effect of scale a lognormal distribution has been applied (**?**), but this too assumes a continuous distribution for the discrete observations. Instead, we consider a negative-binomial observation process (with probability/size parametrisation), such that

$$y_t|\mathbf{x}_t, \theta \sim \text{Negative-Binomial}\left(\frac{\kappa}{1-\kappa}\sum_{a=2}^{A} x_{a,t}, \kappa\right),$$

for some $\kappa \in (0,1)$. Notice that $\mathbb{E}[y_t|\mathbf{x}_t,\theta] = \sum_{a=2}^{A} x_{a,t} < \sum_{a=2}^{A} x_{a,t}/\kappa = \text{var}[y_t|\mathbf{x}_t,\theta]$, so this specification permits overdispersed observations.

6.2.2 Ring-recovery data model

Recall that count data are available from 1928 to 1998, i.e. for $T = 71$ time periods. In contrast, ring-recovery data are only available for individuals released between 1955 and 1997, i.e. released in time period $t \in \{t_1, \ldots, t_2\}$, where $t_1 = 28$ and $t_2 = 70$. These data are stored in a matrix $\mathbf{w}$ whose tth row is denoted $\mathbf{w}_t = \{w_{t,s} \colon t_1 + 1 \leq s \leq t_2 + 2\}$. Here, $w_{t,s}$ indicates the number of individuals released at time t which are subsequently recovered dead in the interval $(s-1, s]$; w_{t,t_2+2} corresponds to the number of individuals that are released at time t that are not seen again within the study.

For each year of release, we assume a multinomial distribution for the subsequent recoveries (see e.g. **?** for further explanations on the ring-recovery model). Thus, the model for $\mathbf{m}$ is then specified as

$$\mathbf{w}_t|R_t, \theta \sim \text{Multinomial}(R_t, \mathbf{q}_t).$$

Here, R_t denotes the number of herons that are ringed as chicks and released in the tth

time period. The multinomial cell probabilities $\mathbf{q}_t \coloneqq \{q_{t,s} \colon t_1 < s \leq t_2 + 2\}$ are given by

$$q_{t,s} \coloneqq \begin{cases} 0, & \text{if } t_1 < s \leq t, \\ (1 - \phi_{\min\{s-t,\mathrm{A}\},s-1})\lambda_{s-1} \prod_{a=1}^{s-t-1} \phi_{\min\{a,\mathrm{A}\},t+a-1}, & \text{if } t < s \leq t_2 + 1, \\ 1 - \sum_{s=t_1+1}^{t_2+1} q_{t,s}, & \text{if } s = t_2 + 2. \end{cases}$$

6.3 Parametrisation

6.3.1 Parameters common to all models

We consider additional covariate information to explain temporal variability. The recovery probabilities are assumed to be logistically regressed on the normalised covariate $time_t$ which represent the normalised (bird) year t:

$$\operatorname{logit} \lambda_t = \alpha_0 + \beta_0 time_t, \quad t = t_1, \ldots, t_2 - 1.$$

We specify the survival probabilities to be logistically regressed on the normalised covariate $fdays_t$ which represents the (normalised) number of days in (bird) year t on which the mean daily temperature fell below freezing in Central England:

$$\operatorname{logit} \phi_{a,t} = \alpha_a + \beta_a fdays_t, \quad t = 1, \ldots, T - 1. \tag{7}$$

The free parameter in the negative-binomial observation equation is parametrised as $\kappa = \operatorname{logit}^{-1}(\omega) \in (0, 1)$ with $\omega \in \mathbb{R}$.

6.3.2 Models for the productivity rate

We specify a set of models for which we perform model comparison on the productivity rates. The unknown parameters are given by $\theta = \{\omega, \alpha_0, \beta_0, \alpha_1, \ldots, \alpha_\mathrm{A}, \beta_1, \ldots, \beta_\mathrm{A}, \vartheta\}$, where ϑ represents the additional model parameters needed for one of the following models for the productivity rate.

Constant. We set $\log \rho_t = \psi$. Thus, $\vartheta = \{\psi\}$.

Regressed on frost days. We set $\log \rho_t = \gamma_0 + \gamma_1 fdays_{t-1}$. Thus, $\vartheta = \{\gamma_0, \gamma_1\}$.

Direct density dependence. We set the log-productivity to be a linear function of abundance, $\log \rho_t = \varepsilon_0 + \varepsilon_1 \tilde{y}_t$, where $\tilde{y}_t$ denotes the tth normalised observation. Thus, $\vartheta = \{\varepsilon_0, \varepsilon_1\}$. This is one of the models considered by **?**.

Threshold dependence. **?** also investigate models in which the productivity is a step function with K levels and hence $K-1$ thresholds (K itself may be unknown) which is defined in terms of the observations. More specifically, the productivity rates are constant between the change-points and monotonically decreasing with increasing population size, i.e. assuming that $K > 1$,

$$\rho_t = \begin{cases} \nu_1, & \text{if } y_t < \tau_1, \\ \nu_k, & \text{if } \tau_{k-1} \leq y_t < \tau_k \text{ for } 1 < k < K, \\ \nu_K, & \text{if } \tau_{K-1} \leq y_t, \end{cases}$$

where $\nu_1 > \nu_2 > \cdots > \nu_K$ and $\tau_1 < \tau_2 < \cdots < \tau_{K-1}$. Thus, it is assumed that larger population sizes induce lower productivity rates. For example, this may be due to an exhaustion of high quality breeding sites leading to a reduction in the quantity/quality of young. To ensure these inequalities we set $\nu_K = \exp(\zeta_K)$ and

$$\nu_k = \textstyle\sum_{l=k}^{K} \exp(\zeta_l); \quad \tau_k = y_{\min} + (y_{\max} - y_{\min}) \dfrac{\sum_{l=1}^{k} \exp(\eta_l)}{\sum_{m=1}^{K} \exp(\eta_m)},$$

for $k \in \{1, \ldots, K-1\}$, where $y_{\min} = \min\{y_1, \ldots, y_T\}$ and $y_{\max} = \max\{y_1, \ldots, y_T\}$. In this case, $\vartheta = \{\zeta_k, \eta_k : 1 \leq k \leq K\}$.

Regime switching dynamics. To constract a more flexible model for the productivity rate, we extend the latent states $\mathbf{x}_t$ by including an additional (unobserved) regime indicator variable r_t which takes values in $\{1, \ldots, K\}$. Conditional on r_t, the productivity rate ρ_{t-1} is then defined as $\rho_{t-1} = \nu_{r_t}$, where $\nu_1, \ldots, \nu_K$ are specified as in the threshold model, above. The evolution of the latent regime indicator r_t is assumed to be a Markov chain with transition equation

$$r_t | r_{t-1}, \theta \sim \text{Multinomial}(K, \mathbf{P}_{r_{t-1}}),$$

where $\mathbf{P}_k = (P_{k,1}, \ldots, P_{k,K})$ with $P_{k,l} = \exp(\varpi_{k,l}) / \sum_{m=1}^{K} \exp(\varpi_{k,m})$, for $1 \leq l \leq K$,

is the kth row of the (K, K)-transition matrix for the regime indicator variable. In this case, $\vartheta = \{\zeta_k, \varpi_{k,l} \colon (k, l) \in \{1, \ldots, K\}^2\}$.

Finally, we note that we also vary the number of levels, K, and the number of age groups, A, so that the number of models to be compared is much larger than the five specifications for the productivity rate summarised above.

6.3.3 Prior specification

We assume that all the model parameters in θ are independent a-priori with Normal$(0, 1)$ priors, except that $\omega \sim$ Normal$(-2, 4)$. The motivation for this choice of priors is the same as in Subsection 5.3.

6.4 Results

Estimates of the evidence for the models can be found in Figure 3. The fit of the different models for the productivity rates is illustrated in Figure 4 below (see also Web Figure 2 in Web Appendix B). Due to the increased flexibility of the productivity rates, the regime-switching model leads to a smaller measurement error. In addition, the evidence for the regime-switching model is much higher than the other models in Figure 3.

Figure 3 supports the finding from **?** that modelling the herons using four age groups is appropriate (though the results with three age groups are similar). However, using only two age groups drastically reduces the model evidence across all specifications for the productivity rate. The results also support the findings from **?** that the first three models (with productivity rate constant, regressed on the number of frost days, or density-dependent) do not explain the data well.

The posterior distribution of the productivity rate (under any of the models) must be interpreted with care. Indeed, note the sharp decline of the productivity rate in the years immediately preceding the severe winters of 1946–47 and 1962–63 in Figure 4b. This indicates that the linear model for the survival rates in (7) may not be flexible enough to accommodate the drop of the heron population in subsequent years.

We also implemented all of the above-mentioned models using a continuous (linear-Gaussian) approximation to the state-space model for the count data. For the regime-switching model, a PF is then still necessary to sample the latent regime indicators. However, as these take values in a small finite set, this can be done highly efficiently using the *discrete* PF from **?**. The results (omitted here) are relatively similar to the results obtained for the original models, i.e. the approximation did not affect the ordering of the models in terms of the evidence. However, the regularising effect of the continuous (linear-Gaussian) approximation artificially increased the evidence for all models by roughly the same amount. In other words, such linear-Gaussian approximations lead to an overestimation of the model fit.

7 Conclusion

We have proposed an efficient Monte Carlo methodology for Bayesian parameter estimation and model comparison for integrated population models which have a state-space model for the noisily observed population sizes as one of their constituent parts. Utilising PMCMC techniques, our approach can be generally applied to such models, requiring neither (a) approximate linear or Gaussian modelling assumptions which introduce a bias that is often difficult to quantify nor (b) data-augmentation schemes which can lead to poor mixing in MCMC algorithms if highly correlated states or parameters are updated separately. Incorporating these ideas into an SMC sampler also yields estimates of posterior model probabilities allowing us to perform model comparison, e.g. for the number of age groups. Finally, we have proposed two extensions which enhance the efficiency of our methodology by exploiting the structure of integrated population models.

We have demonstrated the methodology on two different applications: (1) little owls and (2) grey herons. For the owls, we found no evidence in favour of some of the more complex model specifications proposed in the literature, e.g. for the dependence of immigration on the abundance of voles proposed in **?**. For the herons, we showed that existing models, including the state-of-the-art threshold model for the productivity from

?, do not explain the data well. To remedy this problem, we proposed a novel regime-switching model and demontrated that it is very strongly favoured over the other models in terms of the Bayes factor. Our methodology is related to the SMC^2 algorithm from **?**. However, even in low-dimensional settings (i.e. 3-4 unknown model parameters) **?** had to combine SMC^2 with another importance-sampling algorithm to obtain evidence estimates accurate enough for model comparison in some examples (and, as pointed out by **?**, this importance sampling scheme may not be applicable in higher dimensions). In contrast, in all applications considered in this work, the evidence estimates provided by our methodology were accurate enough to directly identify the best-performing models despite the relatively large number of unknown model parameters (i.e. 6–58 for the owls; 8–31 for the herons).

Acknowledgements

A.B. and A.F. were funded by a Leverhulme Trust Prize. This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1

Supplementary Materials

Web Appendices and Figures, referenced in Sections 3–6, are available at xxx.

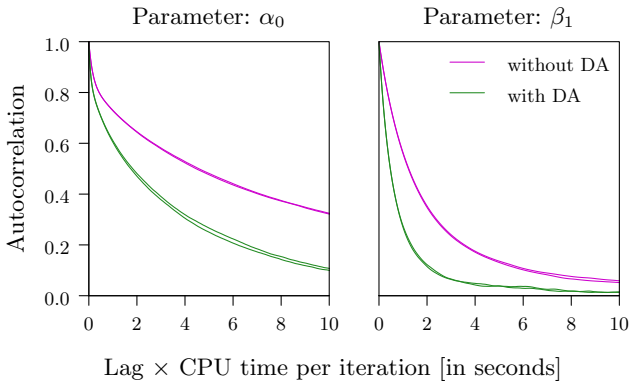


Figure 1. Autocorrelation (rescaled by computation time) of the estimates of the parameters α_0 and β_1 in the little-owls model (with the productivity rates assumed to be constant, i.e. $\gamma_1 = \ldots = \gamma_T$) and immigration independent of the abundance of voles, i.e. $\delta_1 = 0$. The results are based on two independent repeats (each comprised of 10^7 iterations) of the MCMC algorithms with and without delayed-acceptance.

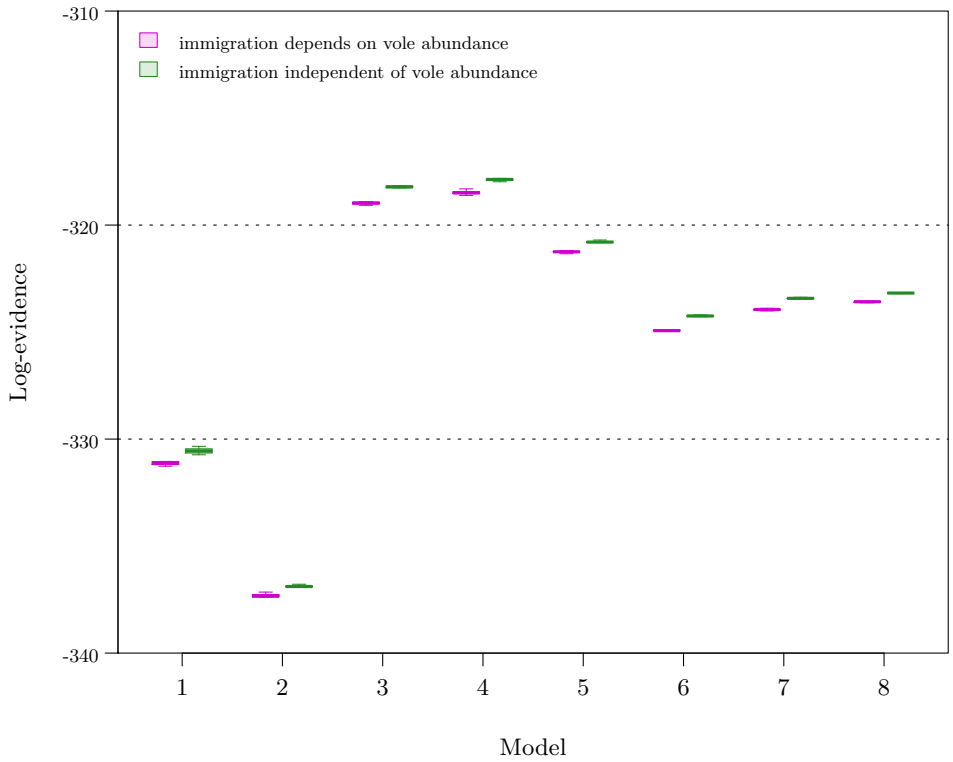


Figure 2. Logarithm of the estimates of the evidence for the eight models for the little owls with or without dependence of the immigration rate on the abundance of voles. The results were obtained from 20 independent runs of the adaptive SMC sampler using $10,000$ particles; the particle filters used to approximate the marginal likelihoods use $1,000$ particles. The average computation time for each SMC sampler was around 9–18 hours on a single core (we stress such a relatively large number of particles was only used to gain accurate evidence estimates in the large models (in terms of the number of parameters), i.e. in Models 1–5; for the smaller models, i.e. Models 6–8, quite similar results could have been obtained in 30 minutes by using only 500 particles.

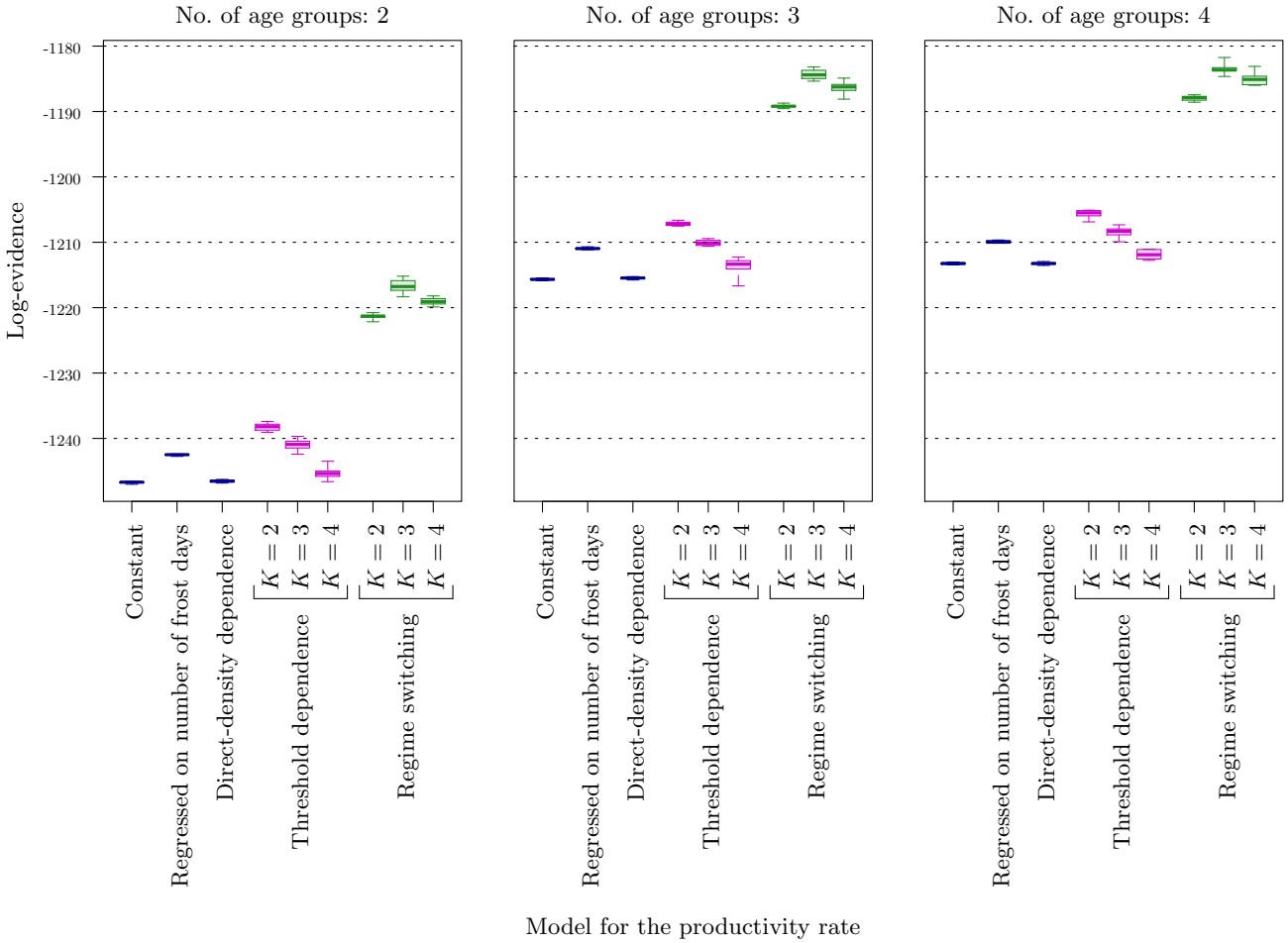
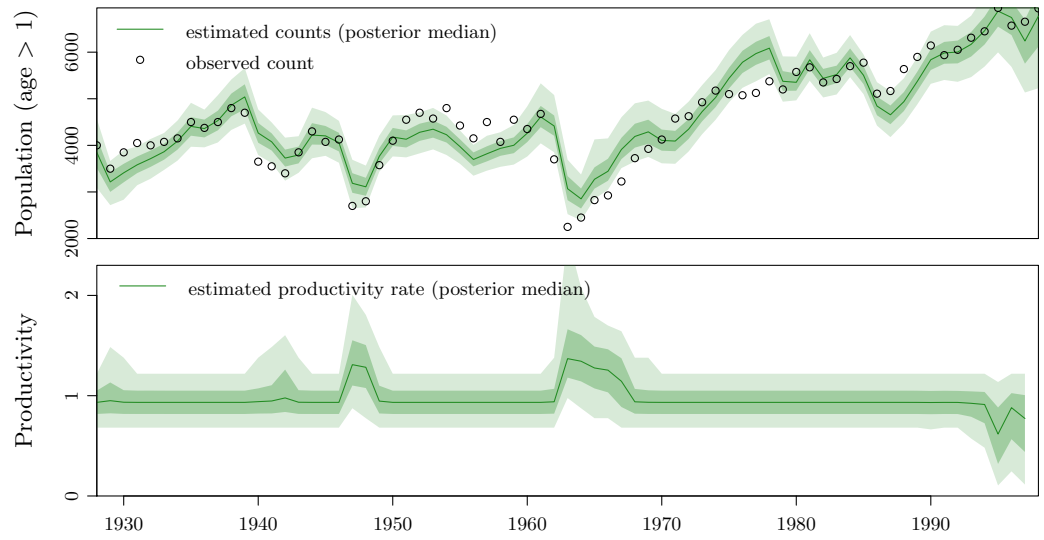
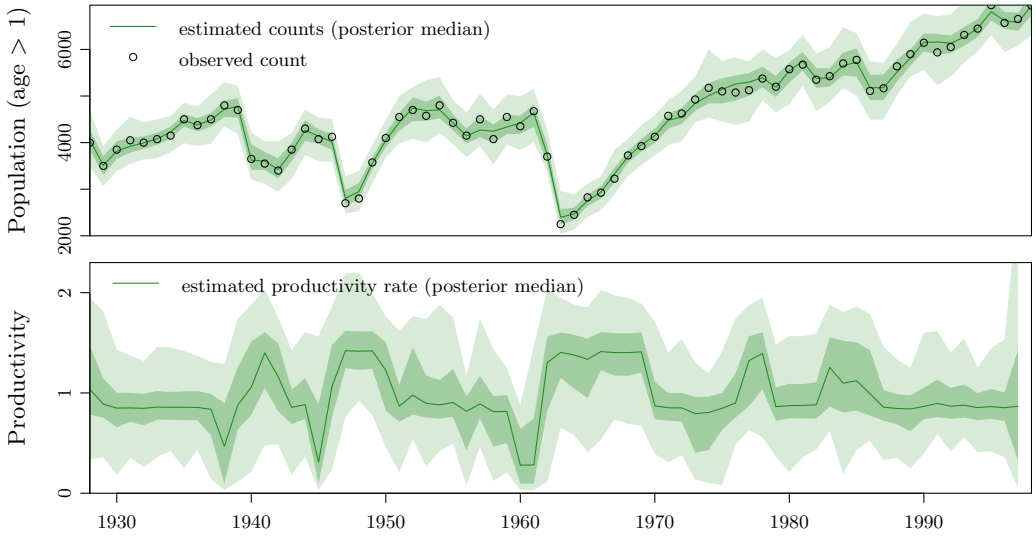


Figure 3. Logarithm of estimates of the evidence for different models for the grey herons. Shown are results for the different models for the productivity rate and different numbers of distinctly modelled age categories (A). For the threshold and regime-switching models, we also investigate different values for the number of thresholds/regimes (K). Obtained from 10 independent runs of the adaptive SMC sampler using $1,000$ particles; the PFs used to approximate the count-data likelihood employed $4,000$ particles. The average computation time was 42–61 hours for the threshold models, 32–45 hours for the regime-switching models and 29–48 for the remaining models, the lower numbers corresponding to A $= 2$ age categories and the higher numbers to A $= 4$ age categories.



a. Threshold dependence ($K = 4$ levels, i.e. 3 thresholds).



b. Regime switching ($K = 4$ regimes).

Figure 4. Marginal posterior distributions of the estimated heron counts (top rows) and productivity rates (bottom rows) for the threshold model from **?** and the novel regime-switching model (results for other models are shown in Web Appendix B) with A = 4 distinct age categories. The shaded areas represent, respectively, the 90 % quantile and range of all encountered realisations. The shown results display the average over 10 independent repeats of the adaptive SMC sampler (each using $1,000$ particles). The PFs used to approximate the count-data likelihood use $4,000$ particles.

Table 1. Average efficiency gain (as defined in Equation (6)) of the refined likelihood tempering scheme (see Section 4.3 of the main manuscript) over standard likelihood tempering for different numbers of particles (M). To simplify the presentation, we only show results for each of the eight models in the case that $\delta_1 \neq 0$, i.e. we allow for dependence of immigration on the abundance of voles.

Model	1	2	3	4	5	6	7	8
efficiency gain ($M = 1,000$)	14.0	4.7	2.8	2.3	2.9	0.9	0.9	1.2
efficiency gain ($M = 10,000$)	19.1	4.5	2.3	2.3	2.5	0.8	1.2	0.6