

Prior distributions for stratified capture-recapture models

J. A. DUPUIS, *Laboratoire de Statistique et Probabilités, University Paul Sabatier, France*

ABSTRACT *We consider the Arnason-Schwarz model, usually used to estimate survival and movement probabilities from capture-recapture data. A missing data structure of this model is constructed which allows a clear separation of information relative to capture and relative to movement. Extensions of the Arnason-Schwarz model are considered. For example, we consider a model that takes into account both the individual migration history and the individual reproduction history. Biological assumptions of these extensions are summarized via a directed graph. Owing to missing data, the posterior distribution of parameters is numerically intractable. To overcome those computational difficulties we advocate a Gibbs sampling algorithm that takes advantage of the missing data structure inherent in capture-recapture models. Prior information on survival, capture and movement probabilities typically consists of a prior mean and of a prior 95% credible confidence interval. Dirichlet distributions are used to incorporate some prior information on capture, survival probabilities, and movement probabilities. Finally, the influence of the prior on the Bayesian estimates of movement probabilities is examined.*

1 Introduction

When information about the movement of animals among populations is provided by capture-recapture data, multi-strata models have been developed to estimate movement probabilities (Arnason, 1973; Schwarz *et al.*, 1993; Brownie *et al.*, 1993). Based upon observations only, frequentist statistical analyses do not take into account the knowledge that biologists have on the studied species. For open populations, Seber (1992) and Pollock (1991) suggest we can adopt the Bayesian viewpoint, integrating the prior knowledge of experts, as has been done for closed populations (Casteldine, 1981; George & Robert, 1992). In all these papers, prior modelling has

Correspondence: Laboratoire de Statistique et Probabilités, University Paul Sabatier, Bat. 1R1, 118 Route de Narbonne, 31062 Toulouse, France. E-mail: dupuis@cict.fr

focused on the size of the population, which was the parameter of interest. Since these recommendations, Bayesian procedures have been developed in the setting of open populations (Vounatsou & Smith, 1995; Brooks *et al.*, 2000a); however, in most studies (apart from Brooks *et al.*, 2000b), the authors only contemplate non-informative distributions, although information about capture, survival, and movement parameters is often available, as pointed out, for instance, by Pollock (1991). In most situations, biologists know or can appreciate some qualitative or/and quantitative elements that are liable to affect the value of some parameters (Breslow, 1990). For instance, capture parameters will depend: upon the studied animal species, upon the protocol implemented to capture animals, upon the time spent to capture them during a given capture session, upon the dexterity of the experimenter, upon time-dependent environmental covariates, etc. In this paper, we deliberately adopt an informative point of view. In Sections 4 and 5, we show how to incorporate some prior information on capture, survival or movement probabilities. In Section 6, we examine to what extent an informative prior distribution on a given movement parameter can affect the Bayesian estimation of the corresponding parameters. To reach this goal we need efficient algorithms for computing Bayesian estimations of these probabilities. Because of the complexity of the posterior distribution, we have proposed in Dupuis (1995) a Gibbs sampling algorithm that takes advantage of the missing data structure of the Arnason-Schwarz model. This specific structure is reviewed in Section 2. In Section 3, extensions of the Arnason-Schwarz model are considered and a Bayesian statistical analysis of these complex models is outlined.

2 The Arnason-Schwarz model

The experimental protocol is standard, and is not reviewed in this paper (see Schwarz *et al.*, 1993 or Dupuis, 1995). We denote by T the number of capture occasions, including the first tagging period. The capture-recapture data thus include $T - 1$ cohorts. We denote by n the number of tagged animals. We assume there is no loss on capture, no mark loss and that no observation can be collected from dead animals. The study zone denoted by K , has been divided in $k \geq 2$ strata. Moreover, the marked sample is assumed to be representative; in particular, we assume that tagging does not influence movement.

2.1 The missing data structure

The description of the AS (Arnason-Schwarz) model we develop in this paper stresses the missing data structure of this model. The key idea is to model the two processes that underlie the data: that is the capture process and the movement process (which constitutes the process of interest). This formulation, first proposed in Dupuis (1995), gives a new light to usual formulations of the AS model, while providing an appropriate framework to implement the algorithms we develop to obtain the posterior quantities of interest. We define $1 \leq \tau_i \leq T - 1$ as the time at which animal i has been marked. We denote by $z_{(i,t)} \in K_{\dagger} = K \cup \{\dagger\}$ the state of animal i at time $t \geq \tau_i$, where $z_{(i,t)} = r \in K$ means that animal i is alive at time t in stratum r , and where $z_{(i,t)} = \dagger$ means that it is dead at time t (or outside K). We denote by $\mathbf{y}_i = (y_{(i,t)}; t \geq \tau_i)$ the capture-recapture history related to the animal i ; for instance, a possible occurrence when $T = 8$, and $k = 2$ is:

$$\mathbf{y}_i = 1 \ 2 \ 2 \ . \ 2 \ 1 \ . \ . \ . \quad (1)$$

This sequence means that the i th animal, marked at time $t = 1$ in stratum $r = 1$, has been recaptured at times $t = 2, 3, 5$ in stratum $r = 2$, and has been recaptured, at time $t = 6$, in stratum 1. It has not been recaptured at times $t = 4, 7, 8$. At these times, states $z_{(i,t)}$ are not available from the data y_i . The notation $y_{(i,t)} = .$ means that $z_{(i,t)}$ is missing: animal i is, at time t , either dead (or outside K), or alive (in K) but it has not been captured. Note that, although state $z_{(i,t)}$ at time $t = 4$ is not available from the data y_i , we know that $z_{(i,t)} \in K$, since animal i has been later recaptured.

For equation (1), we define the corresponding capture process:

$$\mathbf{x}_i = 1 \ 1 \ 1 \ 0 \ 1 \ 1 \ 0 \ 0$$

where $x_{(i,t)} = 1$ if the animal i is captured at time t and $x_{(i,t)} = 0$ otherwise. The capture-recapture history y_i can be viewed as the stacking of the movement process z_i , and of the capture process \mathbf{x}_i . A possible z_i for the above y_i is:

$$\mathbf{z}_i = 1 \ 2 \ 2 \ 1 \ 2 \ 1 \ 1 \ \dagger$$

where we have indicated in bold face the missing $z_{(i,t)}$. Given y_i , there are 14 possibilities for z_i , which are: $1 \ 2 \ 2 \ 1 \ 2 \ 1 \ 1 \ 1$; $1 \ 2 \ 2 \ 1 \ 2 \ 1 \ 1 \ 2$; $1 \ 2 \ 2 \ 1 \ 2 \ 1 \ 2 \ 2$; $1 \ 2 \ 2 \ 1 \ 2 \ 1 \ 2 \ 1$; $1 \ 2 \ 2 \ 1 \ 2 \ 1 \ 1 \ \dagger$; $1 \ 2 \ 2 \ 1 \ 2 \ 1 \ 2 \ \dagger$; $1 \ 2 \ 2 \ 1 \ 2 \ 1 \ \dagger \ \dagger$; the other seven possibilities are obtained from $z_{i,4} = 2$.

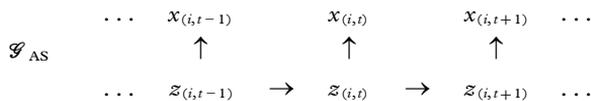
By considering the two underlying processes \mathbf{x}_i and z_i we can formulate biological assumptions concerning capture and movement, as well as some possible dependencies between those two processes (see below, Section 2.2).

2.2 Assumptions and parameters

The (usual) following biological assumptions are made:

- (i) the n individual processes (\mathbf{x}_i, z_i) are independently and identically distributed; thus we assume that animals behave independently with respect to capture, and movement;
- (ii) the probability of being at time t , in $r \in K_+$ depends upon the capture and movement history until $t - 1$, only through the location at time $t - 1$; thus, movement is modelled by a first-order Markov chain;
- (iii) the probability of being captured at time t depends upon the capture and movement history up to t only through $z_{(i,t)}$; thus we assume that there is no trap-response;
- (iv) movements among strata between sampling times are unconstrained and migrations beyond the study strata are permanent.

Assumptions (ii) and (iii) characterize the dependence structure between \mathbf{x}_i and z_i ; they are summarized in the directed graph \mathcal{G}_{AS} below:



This representation allows a visual and clear formulation of the conditional independence assumptions between the different random variables present in the model; see for example, Whittaker (1990) for details. Such a representation is the signature of the AS model. The presence (or absence) of arrows between variables that constitute the graph has a translation in biological terms (see further). Directed

graphs thus appear as an attractive tool to summarize current capture-recapture models or to devise new ones (see Section 3 for illustration).

Consistent with the previous assumptions, we introduce the following parameters.

The capture probability $\Pr\{x_{(i,t)} = 1 \mid z_{(i,t)} = r\}$, where $r \in K$, is denoted by $p_t(r)$. In this paper the state \dagger is assumed to be not observable (thus $p_t(\dagger) = 0$), but our approach can be easily extended to the situation where $p_t(\dagger) \in [0, 1]$ (see Dupuis *et al.*, 2001). The capture probability depends on $z_{(i,t)}$. This is consistent with the graph, since $x_{(i,t)}$ and $z_{(i,t)}$ are connected by an arrow. The absence of an arrow (for all t) between the $x_{(i,t)}$ s is interpreted as an absence of any trap-response. The absence of an arrow (for all t) between $x_{(i,t)}$ and $z_{(i,t+1)}$ is interpreted as an absence of effect of capture upon movement.

For $1 \leq t \leq T - 1$, $r \in K_\dagger$, $s \in K_\dagger$ we denote by $q_t(r, s)$ the transition probability $\Pr\{z_{(i,t+1)} = s \mid z_{(i,t)} = r\}$. In order to obtain expressions in terms of quantities of biological interest (see Brownie *et al.*, 1993, and Schwarz *et al.*, 1993, for biological justifications), the transition probability $q_t(r, s)$ is, for r and s in K , decomposed as the product of a survival probability $\phi_t(r)$ and an interstratum movement probability $\psi_t(r, s)$, namely: $q_t(r, s) = \phi_t(r)\psi_t(r, s)$ where $\phi_t(r) = \sum_{s \in K} q_t(r, s) = 1 - q_t(r, \dagger)$. Note that $\sum_{s \in K} \psi_t(r, s) = 1$, while $\sum_{s \in K_\dagger} q_t(r, s) = 1$. Let $\psi_t(r) = (\psi_t(r, 1), \dots, \psi_t(r, s), \dots, \psi_t(r, k))$. We denote by $\theta = (p, \phi, \psi)$ the parameters of the model, where: $\psi = (\psi_t(r); t = 1, T - 1; r \in K)$; $p = (p_t(r); t = 2, T; r \in K)$; and $\phi = (\phi_t(r); t = 1, T - 1; r \in K)$. Note that the model associated to this parameterization is not identifiable, as it is often the case in missing data models (e.g. Robert & Casella, 1999). From a Bayesian point of view, this is not a problem as long as we use proper prior distributions, which ensures the existence of the posterior distribution of θ .

2.3 Prior distributions

The density of the prior distribution on θ is denoted by $\pi(\theta)$. We assume that:

$$\psi_t(r) \sim \mathcal{D}_k(e_t(r, 1), \dots, e_t(r, k)), \quad p_t(r) \sim \mathcal{B}e(a_t(r), b_t(r)), \quad \phi_t \sim \mathcal{B}e(\alpha_t(r), \beta_t(r)),$$

all independently, where $(e_t(r, 1), \dots, e_t(r, k))$, $(a_t(r), b_t(r))$, and $(\alpha_t(r), \beta_t(r))$, are determined by the prior information. The choice of Beta and Dirichlet distributions is justified by practical and computational considerations, and their use allows us to incorporate easily the expert knowledge (see Sections 4 and 5).

2.4 Estimation via a Gibbs sampling algorithm

In this paper, the only Bayesian estimator of θ we consider is the posterior mean of θ ; it is denoted by $\mathbb{E}[\theta \mid \mathbf{y}]$. Owing to missing data, the likelihood, $L(\theta \mid \mathbf{y})$, is complex (see Schwarz *et al.*, 1993). As pointed out in Dupuis (1995), the posterior distribution of parameters $\pi(\theta \mid \mathbf{y})$, which is proportional to $L(\theta \mid \mathbf{y}) \times \pi(\theta)$, is numerically intractable. To overcome these computational difficulties, we advocate a Gibbs sampling algorithm that takes advantage of the missing data structure. This strategy is now very common in Bayesian analysis of missing data models (e.g. Robert & Casella, 1999). The alternative would consist of implementing the Gibbs sampling on each component of θ , but this approach is more difficult to implement than the missing data approach, because specific algorithms have to be developed to simulate the conditional distributions that appear in this implementation of the Gibbs algorithm (Vounatsou & Smith, 1995). In our case, the Gibbs sampling algorithm produces two chains: one is $(\theta^{(l)}, l \geq 1)$ and the other one is $(\mathbf{z}_m^{(l)}, l \geq 0)$,

where \mathbf{z}_m denotes the set of the missing $\mathbf{z}_{(i,t)}$ s. The missing data simulation phase and the parameter simulation phase are reviewed in Dupuis (1995). Nevertheless, we would like to make clear what we mean by: ‘to take advantage of the missing data structure’. One first takes advantage of the fact that the likelihood of the complete data, we denote by $L(\theta|\mathbf{y}, \mathbf{z}_m)$, where $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_i, \dots, \mathbf{y}_n)$, is very simple. Taking into account assumptions (i), (ii), and (iii) we have shown, in Dupuis (1995), that $L(\theta|\mathbf{y}, \mathbf{z}_m)$ is equal to:

$$\prod_{\substack{2 \leq t \leq T \\ 1 \leq r \leq k}} p_t(r)^{u_t(r)} \{1 - p_t(r)\}^{v_t(r)} \times \prod_{\substack{1 \leq t \leq T-1 \\ 1 \leq r \leq k}} \left\{ \phi_t(r)^{w_t(r, \cdot)} \{1 - \phi_t(r)\}^{w_t(r, \dagger)} \prod_{s \in K} \psi_t(r, s)^{w_t(r, s)} \right\}$$

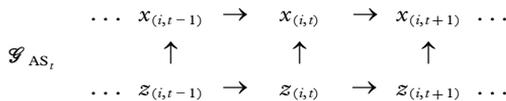
where $w_t(r, s)$ denotes the number of transitions from $r \in K$ to $s \in K_{\dagger}$, between times t and $t + 1$, counted on the complete data $(\mathbf{y}, \mathbf{z}_m)$, and where $w_t(r, \cdot) = \sum_{s \in K} w_t(r, s)$. Moreover $u_t(r)$ denotes the number of animals captured in stratum r at time t , and $v_t(r)$ denotes the number (counted on the complete data $(\mathbf{y}, \mathbf{z}_m)$) of animals such that $x_{(i,t)} = 0$ and $\mathbf{z}_{(i,t)} = r$. Then, taking into account that $\pi(\theta|\mathbf{y}, \mathbf{z}_m) \propto L(\theta|\mathbf{y}, \mathbf{z}_m)\pi(\theta)$, it is very easy to simulate according to $\pi(\theta|\mathbf{y}, \mathbf{z}_m)$, since $\pi(\theta)$ is conjugate for the complete likelihood, i.e. $p_t(r)$ and $\phi_t(r)$ are simulated according to Beta distributions, and $\psi_t(r)$ is simulated according to a Dirichlet distribution.

Since the sequence $(\theta^{(l)}, l \geq 0)$ produced by the Gibbs sampling algorithm converges to $\pi(\theta|\mathbf{y})$, (Dupuis, 1995), we can use it to construct posterior 95% credible intervals for each of the parameters. In addition, by applying the ergodic theorem, we can approximate the Bayes estimator of $p_t(r)$, $\phi_t(r)$ and $\psi_t(r, s)$. For instance, $\mathbb{E}[\psi_t(r, s)|\mathbf{y}]$ is approximated, for ‘large’ L , by $(1/L)\sum_{l=1}^L \psi_t^{(l)}(r, s)$. The rate of convergence of $\theta^{(l)}$ to $\pi(\theta|\mathbf{y})$ is geometric; in practice, it ensures a good rate of convergence of the Gibbs sampling algorithm.

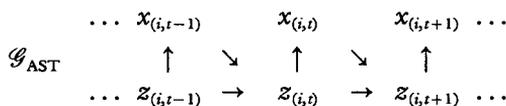
3 Extensions of the AS model

As pointed out in Section 2.2, a directed graph representation is an appealing tool to visualize most assumptions of biological interest. We now provide the directed graph of some extensions of the AS model.

The directed graph \mathcal{G}_{AS} , below assumes the presence of a (first-order) trap-response. This extension of the AS model, is denoted AS_t



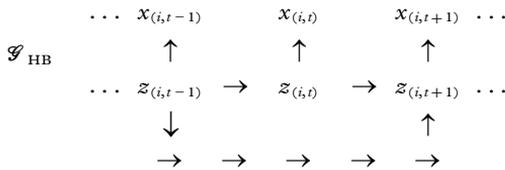
Conditionally on the $\mathbf{z}_{(i,t)}$ s, the $x_{(i,t)}$ s are no longer independent, unlike the AS model. The $x_{(i,t)}$ s now constitute a first-order Markov chain, conditionally on the $\mathbf{z}_{(i,t)}$ s. The directed graph \mathcal{G}_{AST} below assumes a possible influence of capture at time t on the transition between times t and $t + 1$.



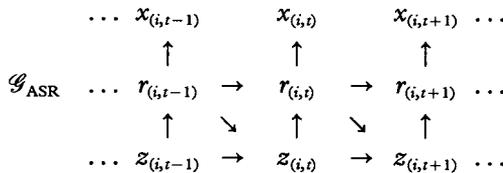
An immediate extension of this model would be to consider that $x_{(i,t)}$ represents the treatment administered at the animal i , at time t (after it has been captured).

This extension of the AS model is denoted AST. Such an extension has been considered by Doligez *et al.* (this issue) to assess responses to clutch size manipulation according to dispersal status.

The Hestbeck-Brownie model (Hestbeck *et al.*, 1991; Brownie *et al.*, 1993) assumes that the movement behaviour includes a memory effect, i.e. a second-order Markov chain. The directed graph \mathcal{G}_{HB} below summarizes the main assumptions of this model.



More complex models can be contemplated. Danchin *et al.* (1998) were interested in investigating some lagged effects between reproduction and fidelity in black-lagged kittiwake (*Rissa tridactyla*). In their article, $r_{(i,t)}$ denotes the breeding status of animal i at time t and $z_{(i,t)}$ represents its breeding site. A simple biological framework is: $r_{(i,t)} \in \{0, 1\}$ where $r_{(i,t)} = 0$ means a failed reproduction. This extension of the AS model is denoted ASR. Biological considerations leads these authors to propose a modelling that can be summarized in the graph \mathcal{G}_{ASR} below.



For simplicity, we have omitted in this graph the possible interactions between bird i and birds that nest in the same location as bird i (see Danchin *et al.*, 1998 and Section 7).

For each of the above extensions, the posterior distribution of parameters is numerically intractable, and we advocate a Gibbs sampling algorithm that takes advantage of the missing data structure. We believe that the missing data formulation is the only feasible approach in the HB and ASR models, on account of the complexity of the likelihood of these two extensions. As in the AS model, the complete likelihood of all those extensions is very simple, which makes the simulation parameters phase of the Gibbs sampling algorithm very easy to implement. Moreover the directed graph representation constitutes a valuable tool to calculate the conditional distributions that appear in the missing data phase. The Bayesian analysis of the AS, AST, HB and ASR models is presented in Dupuis & Clobert (2000).

4 Arguments for the Beta distribution

This section is organized as follows. We first provide arguments for the Beta distribution to incorporate some prior information on survival and capture parameters, as well as on movement parameters (when the study zone includes two strata). Then, computational arguments are produced; the latter concern the AS model and its extensions.

To develop our arguments it is convenient to consider an abstract parameter $\theta \in [0, 1]$, which can be thought of as the parameter of a Bernoulli distribution $b(\theta)$.

Recall that $\theta \sim Be(a, b)$ if its density $\pi(\theta)$ is equal to:

$$\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1} \mathbb{I}_{[0,1]}(\theta) \tag{2}$$

where a and b are two strictly positive real numbers.

Argument 1

Prior information on a proportion θ typically consists of a prior 95% credible confidence interval $I =]g, h[$ and of a prior mean μ (so that $\mu \in I$); or of a prior variance v and of a prior mean μ . This is exactly the type of prior information that can be very easily incorporated via a Beta distribution. The coefficients a and b are easily obtained from μ and v (Berger, 1985). The other type of prior information, which is composed of a prior mean μ and a prior 95% CI, has been proposed in Dupuis (1995). As pointed out by Berger (1985), prior information composed of quantiles is much more robust than one based on a variance v . Moreover, in practice, it is much more convenient to manipulate prior confidence intervals than prior variances. To determine a and b from μ and I we consider a new parameterization of the Beta distribution, by $\lambda > 0$ and $\mu \in]0, 1[$, such that

$$a = \lambda\mu \quad \text{and} \quad b = \lambda(1 - \mu) \tag{3}$$

The Beta distribution is re-parameterized by two parameters, its mean $\mu = a/(a + b)$ and $\lambda = a + b$. We now search for a and b so that $\mathbb{E}[\theta] = \mu$ and $\pi(\theta \in I) \leq 0.95$, where $\mu \in I$ and $I =]g, h[$ are given. Using the parameterization (3), we only have to find λ so that $\pi_\lambda(\theta \in I) = 0.95$ where $\theta \sim Be(\lambda\mu, \lambda(1 - \mu))$. We assume that $\lambda \geq 2$, which corresponds to informative situations (see argument 6). Our procedure includes two steps.

Step 1. Determine $\lambda \in \mathbb{N}$ so that: $\pi_\lambda(\theta \in I) \leq 0.95$ and $\pi_{\lambda+1}(\theta \in I) > 0.95$. To find such a λ , start with $\lambda = 2$, and increase λ until this condition is satisfied. Such a λ always exists since $\lambda \rightarrow +\infty$, implies $\text{Var } \theta \rightarrow 0$, and $\hat{\theta}_\pi \rightarrow \mu(a.s)$. Therefore $\pi_\lambda(\theta \in I \leq 0.95) \rightarrow 1$ (when $\lambda \rightarrow +\infty$), since I is such that $\mu \in I$.

Step 2. Determine (by dichotomy) $\lambda \in]\tilde{\lambda}, \tilde{\lambda} + 1[$, where $\tilde{\lambda}$ is the value found in Step 1, so that: $|\pi_\lambda(\theta \in I) - 0.95| \geq \varepsilon$; where ε denotes the wanted precision.

To illustrate this approach we give two examples. For $\mathbb{E}[\theta] = 0.9, I = [0.7, 1]$ and $\mathbb{E}[\theta] = 0.4, I = [0.2, 0.6]$, the parameters (a, b) of the Beta distribution are respectively equal to: (13.05, 1.45) and to (8, 12).

Argument 2

As pointed out by Box & Tiao (1973), the Beta distribution is suitable for a wide range of situations. These authors provide the different shapes of the density of a Beta distribution, depending upon the values of a and b . When the parameterization (3) is used, the Beta distribution is particularly suitable to incorporate prior information on θ . First note that the parameter λ is a measure of the dispersion (or of the concentration) of θ around the mean μ , since $\text{Var } \theta = \mu(1 - \mu)/(1 + \lambda)$ is a decreasing function of λ . The parameter λ can thus be interpreted as a measure of the precision of the prior information: the larger is λ , the more precise is the prior information. Let $\mathbf{y} = (y_1, \dots, y_i, \dots, y_n)$, where the y_i s are i.i.d. and where $y_i \sim b(\theta)$. It is easy to check that:

$$\hat{\theta}_n = \frac{n}{\lambda + n} \hat{\theta}_{ml} + \frac{\lambda}{\lambda + n} \mathbb{E}[\theta] \tag{4}$$

where $\hat{\theta}_{ml} = (1/n)\sum_i^n y_i$ denotes the maximum likelihood estimator of θ , based upon \mathbf{y} . Let us consider the two extreme situations: $\lambda \rightarrow 0$ and $\lambda \rightarrow +\infty$. If $\lambda \rightarrow 0$, then $\text{Var } \theta \rightarrow \mu(1-\mu)$ and $\hat{\theta}_\pi \rightarrow \hat{\theta}_{ml}$ which does not depend on μ . If $\lambda \rightarrow +\infty$, then $\text{Var } \theta \rightarrow 0$, and $\hat{\theta}_\pi \rightarrow \mu(a.s)$, which does not depend on the data.

As pointed out by many authors (e.g. Gelman *et al.*, 1995), $\hat{\theta}_\pi$ appears as an average of $\hat{\theta}_{ml}$ and $\mathbb{E}[\theta]$, where weightings are respectively proportional to the size of the sample (that is n) and to the precision of the prior (that is λ). Note that, when a and b are integer, λ can be thought of as the size of a virtual sample that would support the prior information. More precisely, if we denote by \mathbf{x} a virtual sample including $a = \lambda\mu$ successes, then the MLE of θ , based upon (\mathbf{x}, \mathbf{y}) is equal to the Bayesian estimation of θ based upon \mathbf{y} and such that $\theta \sim \mathcal{B}e(a, b)$.

Argument 3

We can generalize the Beta distribution to take into account a prior constraint between two parameters that take their values in $[0, 1]$. For instance, the following distribution on $(\theta_1, \theta_2) \in [0, 1] \times [0, 1]$

$$\pi(\theta_1, \theta_2) \propto g_1(\theta_1)g_2(\theta_2)\mathbb{1}_E(\theta_1, \theta_2) \quad (5)$$

where $E = \{(\theta_1, \theta_2) \mid 0 \leq \theta_1 < \theta_2 \leq 1\}$ and where the functions $g_r(\theta_r); r = 1, 2$ are densities of Beta distributions, allows us to take into account the *a priori* constraint $\theta_1 < \theta_2$. If θ_r represents the capture probability $p_i(r)$ where $r \in \{1, 2\}$ and if $g_1(p_i(1)) = g_2(p_i(2)) = 1$, then the prior distribution (5) simply expresses that the only available prior information we have on the capture probabilities during the capture session t , is that: $p_i(1) < p_i(2)$. Such a prior can be justified, for instance, that during the session t , the catch effort was more intensive in stratum 2 than in stratum 1. From a computational point of view, the simulation parameter phase of the Gibbs sampling associated to the prior distribution (5) is especially easy to implement. Of course, constraints can also include survival and movement parameters.

Argument 4

The use of a Beta distribution on θ leads to a closed expression for the Bayes estimator of $\theta \in [0, 1]$ (see equation (4)). In a capture-recapture set-up the use of Beta distributions on $p_i(r)$ and $\phi_i(r)$ allows us easily to implement the Gibbs sampling algorithm based on (θ, \mathbf{z}_m) , because those distributions are conjugate for the complete likelihood $L(\theta \mid \mathbf{y}, \mathbf{z}_m)$ (see Section 2.4 or Dupuis, 1995).

Argument 5

In multi-strata capture-recapture models, classical numerical optimization methods encounter very serious difficulties in obtaining the MLE, especially when the number of strata exceeds 3 (see Brownie *et al.*, 1993; Schwarz *et al.*, 1993; Lebreton & Pradel, this issue). The effectiveness of the Gibbs sampling algorithm implemented in Dupuis (1995) is not at all affected by the number of strata in K . It is well known that, for a large enough sample size n , the non-informative Bayes estimators are very close to the MLE (Berger, 1985). Note that Beta distribution includes the uniform distribution ($\mu = 0.5$ and $\lambda = 2$) and the Jeffrey distribution ($\mu = 0.5$ and $\lambda = 1$). The improper Haldane distribution $\pi(\theta) = [\theta(1-\theta)]^{-1}\mathbb{1}_{[0,1]}(\theta)$ corresponds to $\lambda \rightarrow 0$. Therefore, if the MLE is of interest, the non-informative Bayesian approach turns out to be an attractive alternative to numerical methods when the latter encounter computational difficulties in obtaining the MLE (Dupuis,

1997). For moderate sample size n , the likelihood associated to the AS model, can exhibit several local maxima, as is often the case in missing data models (Diebolt, 1997). In that situation, non-informative Bayesian estimators can constitute an ideal starting point for classical optimization numerical methods.

5 Arguments for the Dirichlet distribution

We now provide arguments for the Dirichlet distribution to incorporate some prior information on movement parameters (when the study zone includes $k \geq 3$ strata). It is convenient to consider with an abstract parameter $\theta = (\theta_1, \dots, \theta_j, \dots, \theta_k)$ where $\theta_j \in]0, 1[$ and where $\sum_{j=1}^k \theta_j = 1$, which can be thought as the parameter of a Multinomial distribution. The Dirichlet distribution generalizes the Beta distribution, and arguments in favour of the Beta distribution could also be used in favour of the Dirichlet distribution. Recall that a random variable $\theta \sim \mathcal{D}_k(a_1, \dots, a_j, \dots, a_k)$, where the parameters a_j are strictly positive, if the density of θ is such that: $\pi(\theta) \propto \prod_{j=1}^k \theta_j^{a_j-1}$. The particular case $k = 2$ reduces to the Beta distribution. We can reparameterize the Dirichlet distribution by $\lambda > 0$ and $\mu = (\mu_1, \dots, \mu_j, \dots, \mu_k)$ where $\mu_j \in]0, 1[$ and $\sum_{j=1}^k \mu_j = 1$; the new parameters $\lambda > 0$ and μ being linked to the a_j s by $a_j = \lambda \mu_j$. It is easy to verify that:

$$\lambda = \sum_{j=1}^k a_j \quad \mu_j = \mathbb{E}[\theta_j] \quad \text{Var } \theta_j = \frac{\mu_j(1 - \mu_j)}{1 + \lambda} \tag{6}$$

As in the Beta distribution, the parameter λ is a measure of the concentration of the random variable θ around its mean $\mu = (\mu_j; j = 1, \dots, k)$. The parameter λ can be interpreted as a measure of the precision of the prior information. It is easy to check that the posterior mean of θ is:

$$\hat{\theta}_\pi = \frac{n}{\lambda + n} \hat{\theta}_{ml} + \frac{\lambda}{\lambda + n} \mathbb{E}[\theta] \tag{7}$$

where $\mathbb{E}[\theta] = \mu$, and $\hat{\theta}_{ml} = (N_j/n; j = 1, \dots, k)$ where N_j denotes the count associated with the category j in the multinomial experiment. Thus, the formula (4) also applies, and the consequences that we have outlined from equation (4) remain valid.

Let θ be a random variable with density $\pi(\theta)$, such that $\theta \sim \mathcal{D}_k(a_1, \dots, a_j, \dots, a_k)$. It is of great practical interest to notice that the parameters a_j s are completely determined by the mean vector μ and by an interval $I =]g, h[$ so that $\pi_j(\theta_j \in I) = 0.95$, where j represents any one of the k components of θ , and $\pi_j(\cdot)$ denotes the density of θ_j . Starting from the parameterization (λ, μ) , the condition $\pi_j(\theta_j \in I) = 0.95$ allows us to determine λ . Because $\theta_j \sim \mathcal{B}e(\lambda \mu_j, \lambda(1 - \mu_j))$, the problem reduces to that in Section 4, and the algorithm given in Section 4 can be used to determine λ (see Dupuis, 1995, for examples). It should be noted that a Dirichlet distribution cannot incorporate prior information that would consist, for each θ_j , of a prior mean μ_j and of a prior variance $v_j = \text{Var}[\theta_j]$, because the v_j s are constrained with the terms $\text{Var } \theta_j / [\mu_j(1 - \mu_j)]$ constant (equal to $1/\lambda$). Moreover, a Dirichlet distribution cannot incorporate prior information that would consist, for each θ_j , of a prior mean μ_j and of a prior 95% credible interval I_j , because λ is a one-dimensional parameter that, only globally, characterizes the dispersion of θ round μ . In practice, we recommend, once the parameters a_j have been determined, examining whether the resulting 95% credible intervals of the component θ_j are reasonable.

6 Influence of prior on Bayesian movement parameters estimations

In this section we examine to what extent an informative prior distribution on a given movement probability can influence the Bayes estimate of this parameter, and how it can affect the Bayes estimate of the other parameters. Moreover, in this section we aim to rediscover, in a simple capture-recapture set-up, some general observations we have made in Section 4 concerning the Bayes estimator of a proportion. Nevertheless, let us underline that the formulae (4) and (7) are no more valid in capture-recapture set-ups. As an illustration, we assume that survival parameters are equal to 1, and we consider an artificial, but meaningful, data set. For an illustration on a real data set, see Dupuis (1995).

6.1 The data set and methods

The data set was constructed as follows. Starting with a given θ , the data set \mathbf{y} , of size n , is such as $n_h = \mathbb{E}[N_h]$ where n_h denotes the observed count associated to the capture-recapture history h , and the expectation of N_h is taken under the (closed) AS model and for the given θ . Note that $\sum_h n_h = n$ and that $n_h = np_h$ if we denote $p_h = \Pr(\mathbf{y}_i = h | \theta)$. Actually this procedure is not necessary for our purpose and we could have chosen any data set; its interest is that it allows us to consider a data set that is plausible under the AS model and for the selected θ .

We assume that K includes two strata 1 and 2, and that the experimental protocol includes $T = 3$ capture-recapture sessions (including the tagging session); for convenience, we assume that tagging has been carried out only at time $t = 1$ and in stratum 2. The number of marked animals is $n = 40$. The parameter of the AS model is: $\theta = (\psi_1(2, 1), p_2(1), p_2(2), \psi_2(1, 1), \psi_2(2, 1), p_3(1), p_3(2)) = (0.75, 0.4, 0.8, 0.5, 0.5, 0.5, 0.5)$. It is easy to check that the expected counts are for each history: 222(2), 221(2), 220(4), 211(3), 212(3), 210(6), 201(5), 202(5), 200(10). For example, for $h = 201$, we have: $p_h = \psi_1(2, 1)(1 - p_2(1))\psi_2(1, 1)p_3(1) + \psi_1(2, 2)(1 - p_2(2))\psi_2(2, 1)p_3(1) = 0.8$. Thus $n_h = \mathbb{E}[N_h] = Np_h = 40 \times 0.8 = 5$.

To facilitate our analysis, we have focused on only one parameter, $\psi_1(2, 1)$, but similar results occur with the other movement parameters. We consider different informative prior distributions on $\psi_1(2, 1)$, while putting uniform prior distribution on all the other parameters. Then, we compare the Bayesian estimates of all the parameters for those different prior distributions. We also compare these estimates with those obtained by putting a non-informative prior on $\psi_1(2, 1)$.

We have used a single run of $L = 10^4$ iterations of the Gibbs sampler. The convergence of the algorithm has been visually appreciated; it has been based on a stabilization of all the empirical averages $(1/L)\sum_{i=1}^L \psi_i^{(j)}(r, s)$ and $(1/L)\sum_{i=1}^L p_i^{(j)}(r)$; see for example Robert & Casella (1999) for discussion about this criteria.

6.2 Results analysis and discussion

In Table 1, we have reported the posterior means and posterior 95% credible intervals of $p_2(1)$, $p_2(2)$ and $\psi_1(2, 1)$, for three non-informative prior distributions on $\psi_1(2, 1)$ (uniform, Jeffrey and Haldane). The (Bayesian) estimations of $\psi_2(2, 1)$, $\psi_2(1, 1)$, $p_3(1)$ and $p_3(2)$, as well as the corresponding posterior credible intervals, have not been reported since they were not affected by the choice of the non-informative prior on $\psi_1(2, 1)$. The Bayesian estimations and the posterior 95% confidence intervals of $\psi_1(2, 1)$, $p_2(1)$ and $p_2(2)$ are similar for the three non-

TABLE 1. Posterior mean and posterior 95% credible interval of $\psi_1(2, 1)$, $p_2(1)$ and $p_2(2)$

	Uniform	Jeffrey	Haldane
$\psi_1(2, 1)$	0.57 [0.27, 0.83]	0.57 [0.26, 0.84]	0.59 [0.26, 0.85]
$p_2(1)$	0.56 [0.29, 0.95]	0.56 [0.28, 0.95]	0.55 [0.28, 0.95]
$p_2(2)$	0.52 [0.21, 0.95]	0.53 [0.21, 0.95]	0.55 [0.21, 0.96]

TABLE 2. Posterior mean and posterior 95% credible interval of $\psi_1(2, 1)$, $p_2(1)$ and $p_2(2)$

	Be(0.7, 5) [0.30, 0.97]	Be(0.7, 10) [0.40, 0.93]	Be(0.7, 20) [0.48, 0.87]	Be(0.7, 30) [0.54, 0.84]	Be(0.75, 10) [0.46, 0.96]	Be(0.8, 10) [0.52, 0.97]
$\psi_1(2, 1)$	0.64 [0.35, 0.85]	0.66 [0.42, 0.83]	0.68 [0.49, 0.82]	0.69 0.53, 0.81]	0.69 [0.45, 0.85]	0.72 [0.49, 0.87]
$p_2(1)$	0.50 [0.27, 0.88]	0.48 [0.27, 0.80]	0.46 [0.27, 0.72]	0.45 [0.27, 0.69]	0.46 [0.27, 0.76]	0.44 [0.26, 0.70]
$p_2(2)$	0.59 [0.24, 0.96]	0.61 [0.26, 0.97]	0.63 [0.29, 0.96]	0.64 [0.30, 0.97]	0.61 [0.28, 0.97]	0.68 [0.30, 0.98]

informative distributions. Therefore, the Bayesian estimations of these parameters is not very sensitive to the choice of the non-informative prior put on $\psi_1(2, 1)$. Note that the MLE of $\psi_1(2, 1)$, obtained by program MARK, is 0.56; it is very near to the Bayesian estimate obtained with a uniform prior on $\psi_1(2, 1)$. Moreover, the 95% CI yielded by program MARK is $[0.17 \times 10^{-10}, 1]$, which is not at all informative (contrary to the Bayesian CIs).

In Table 2 we have reported the posterior mean and posterior 95% credible interval of $\psi_1(2, 1)$, $p_2(1)$ and $p_2(2)$ for different informative prior distributions on $\psi_1(2, 1)$. We have parameterized the Beta distributions by μ and λ in order to examine the impact of modifications involving only μ , the parameter λ being fixed (or conversely). We also have indicated, for each prior distribution on $\psi_1(2, 1)$, a prior 95% credible interval for this parameter determined by implementing Monte Carlo simulation methods.

The Bayesian estimations of $\psi_2(2, 1)$, $\psi_2(1, 1)$, $p_3(1)$ and $p_3(2)$, as well as the corresponding posterior credible intervals, have not been reported since they were practically not affected by the choice of the prior distribution on $\psi_1(2, 1)$. Therefore, the Bayesian estimation of these parameters is not sensitive to the different priors on $\psi_1(2, 1)$. However, the Bayes estimates of $p_1(2)$ and $p_2(2)$ are significantly affected by the choice of the priors on $\psi_1(2, 1)$ (as is clear by comparing the results in Tables 1 and 2). Most of the observations we have deduced from formula (4) can be rediscovered in the framework of our capture-recapture data set. First, whatever the prior, the Bayes estimate of $\psi_1(2, 1)$ is always located between the MLE of $\psi_1(2, 1)$ and $\mathbb{E}[\psi_1(2, 1)]$. Second, when $\mathbb{E}[\psi_1(2, 1)] = 0.7$ and λ varies, the Bayes estimate of $\psi_1(2, 1)$ is closer and closer to 0.7, as λ increases. The range of the posterior credible intervals $\psi_1(2, 1)$ becomes narrower, as λ increases. Third, when $\lambda = 10$ and μ varies we observe that the Bayes estimate of $\psi_1(2, 1)$ increases with μ ; note that the range of the credible intervals is stable when μ varies.

The Bayes estimates of $\psi_1(2, 1)$, $p_2(2)$ and $p_2(2)$ appear very sensitive to the prior we put on $\psi_1(2, 1)$. In particular, it is notable that poor prior information on $\psi_1(2, 1)$ (such as $\lambda = 5$ or $\lambda = 10$), significantly affects the Bayes estimate of these parameters (compare Tables 1 and 2). Moreover, for moderate values of λ (≤ 10), and a moderate increase in μ ($= 0.05$), the corresponding increase in the Bayes estimate of $\psi_1(2, 1)$ is significant ($= 0.03$). We also put a uniform distribution on $(\psi_1(2, 1), \psi_1(2, 2))$ so that these two parameters are constrained with $\psi_1(2, 1) > \psi_1(2, 2)$. Again, this poor prior significantly affects $\psi_1(2, 1)$. Its Bayes estimate is 0.64 and its posterior 95% *CI* is [0.40, 0.85]. For this prior, the estimations and *CI*s of $p_2(1)$ and of $p_2(2)$ are very close to those corresponding to the prior $Be(0.75, 10)$.

As noted before, poor prior information on $\psi_1(2, 1)$ can significantly affect the estimates and the posterior *CI* of $\psi_1(2, 1)$, as well as estimates and *CI* of $p_2(1)$ and $p_2(2)$. Instead of reviewing some general guidelines in the case of strong sensitivity of certain parameters to the choice of the prior (this point of view is well discussed by Brooks *et al.*, 2000b), we prefer to conclude this section in an orthogonal direction. Not surprisingly, we have observed that informative prior distributions on $p_2(1)$ and on $p_2(2)$ affect the precision of Bayesian estimators of $\psi_1(2, 1)$ (these results have not been reported in this paper). Such an observation suggests examining to what extent some prior information on some nuisance parameters (namely capture parameters) can significantly perform the Bayesian estimators of interest (namely movement parameters). Those important issues are the subject of a subsequent paper.

7 Conclusions

Implementing a Gibbs sampling algorithm that takes advantage of the missing data structure of the Arnason-Schwarz model allows a complete Bayesian analysis of this model. Furthermore, complex extensions of this model can be investigated that were out of reach before. We think that our strategy is probably necessary to approach models that will relax the assumption (i) of the Arnason-Schwarz model: 'animals behave independently with respect to movement'. As such, the modelling proposed by Danchin *et al.*, 1998, highlights the computational challenge we have to accept, when the assumption (i) has been relaxed. When parameters of interest are capture, survival and movement probabilities, we have shown that the use of conjugate prior distributions (here Beta and Dirichlet distributions) proves to be particularly attractive. Of course the sensitivity of the estimators to this class of prior distributions will have to be investigated, but our paper confirms Schwarz & Seber (1999) who considered that the Gibbs sampler now allows the use of realistic prior.

Acknowledgements

Suggestions made by the two referees as well as the comments of Byron Morgan led to this improved version of the paper.

REFERENCES

- ARNASON, A. N. (1973) The estimation of population size, rates and survival in a stratified population, *Researches in Population Ecology*, 13, pp. 97-113.
- BERGER, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*, 2nd edn (New York, Springer-Verlag).

- BOX, G. & TIAO, G. (1973) *Bayesian Inference in Statistical Analysis* (Reading, Addison-Wesley).
- BRESLOW, N. (1990) Bio-Statistics and Bayes, *Statistical Science*, 5(3), pp. 269-298.
- BROOKS, S. P., CATCHPOLE, E. A. & MORGAN, B. J. T. (2000a) A Bayesian approach to animal survival estimation, *Statistical Science*, 15, pp. 357-376.
- BROOKS, S. P., CATCHPOLE, E. A., MORGAN, B. J. T. & BARRY, S. C. (2000b) On the Bayesian analysis of ring-recovery data, *Biometrics*, 56, pp. 951-956.
- BROWNIE, C., HINES, J. E., NICHOLS, J. D., POLLOCK, K. H. & HESTBECK, J. B. (1993) Capture-recapture studies for multiple strata including non-Markovian transition probabilities, *Biometrics*, 49, pp. 1173-1187.
- CASTLEDINE, B. (1981) A Bayesian analysis of multiple capture-recapture sampling for a closed population, *Biometrika*, 45, pp. 343-359.
- DANCHIN, E., BOULINIER, R. & MASSOT, M. (1998) Conspecific reproductive success and breeding habitat selection: implications for the study of the coloniality, *Ecology*, 79(7), pp. 2415-2428.
- DIEBOLT, J. (1997) Discussion of the paper of Meng and Van Dyk: 'The E.M. algorithm: an old folk-song sung to a fast new tune', *Journal of the Royal Statistical Society, B*, 57(3), pp. 545-546.
- DOLIGEZ, B., CLOBERT, J., PETTIFOR, R. A., ROWCLIFFE, M., GUSTAFSSON, L., PERRINS, C. M. & MCCLEERY, R. H. (2002) Costs of reproduction: assessing responses to brood size manipulations on life-history and behavioural traits using multi-state capture-recapture models, *Journal of Applied Statistics*, this issue.
- DUPUIS, J. A. (1995) Bayesian estimation of movement and survival probabilities from capture-recapture data, *Biometrika*, 82(4), pp. 761-772.
- DUPUIS, J. A. (1997) Discussion of the paper of Meng and Van Dyk: 'The E.M. algorithm: an old folk-song sung to a fast new tune', *Journal of the Royal Statistical Society, B*, 57(3), p. 553.
- DUPUIS, J. A., BADIA, J., MAUBLANC, M. L. & BON, R. (2001) Survival and spatial fidelity of mouflon (*Ovis gmelini*): a Bayesian analysis of an age-dependent capture-recapture model, *Journal of Agriculture, Biological, and Environmental Statistics*, to appear.
- GELMAN, A., CARLIN, J. B., STERN, H. S. & RUBIN, D. B. (1995) *Bayesian Data Analysis* (Chapman & Hall).
- GEORGE, E. I. & ROBERT, C. P. (1992) Capture-recapture estimation via Gibbs sampling, *Biometrika*, 79, pp. 677-683.
- HESTBECK, J. B., NICHOLS, J. D. & MALECKI, R. A. (1991) Estimation of movement and site fidelity using mark-resight data of wintering Canada Geese, *Ecology*, 72, pp. 523-533.
- LEBRETON, J. D. & PRADEL, R. (2002) Multistate recapture models: modelling incomplete individual histories, *Journal of Applied Statistics*, this issue.
- POLLOCK, K. H. (1991) Modeling capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present, and future, *Journal of the American Statistical Association*, 86, pp. 225-238.
- ROBERT, C. P. & CASELLA, G. (1999) *Monte-Carlo Statistical Methods* (New York, Springer).
- SCHWARZ, C. G., SCHWEIGERT, J. F. & ARNASON, A. N. (1993) Estimating migration rates using tag-recovery data, *Biometrics*, 49, pp. 177-193.
- SCHWARZ, C. G. & SEBER, G. A. F. (1999) Estimating animal abundance III, *Statistical Science*, 14, pp. 427-456.
- SEBER, G. A. F. (1982) *The Estimation of Animal Abundance and Related Parameters*, 2nd edn (New York, Macmillan).
- SEBER, G. A. F. (1992) A review of estimating animal abundance II. *Workshop on Design of Longitudinal Studies and Analysis of Repeated Measures Data*, 60, 129-166.
- VOUNATSOU, P. & SMITH, A. F. M. (1995) Bayesian analysis of ring-recovery data via MCMC simulation, *Biometrics*, 51, pp. 687-708.
- WHITTAKER, J. (1990) *Graphical Models* (Wiley).