Hierarchical Bayesian modelling with habitat and time covariates for estimating riverine fish population size by successive removal method

Etienne Rivot, Etienne Prévost, Anne Cuzol, Jean-Luc Baglinière, and Eric Parent

Abstract: We present a hierarchical Bayesian modelling (HBM) framework for estimating riverine fish population size from successive removal data via electrofishing. It is applied to the estimation of the population of Atlantic salmon (Salmo salar) juveniles in the Oir River (France). The data set consists of 10 sampling sites sampled by one or two removals over a period of 20 years (1986–2005). We develop and contrast four models to assess the effect of temporal variations and habitat type on the density of fish and the probability of capture. The Bayes factor and the deviance information criterion are used to compare these models. The most credible and parsimonious model is the one that accounts for the effects of the years and the habitat type on the density of fish. It is used to extrapolate the population size in the entire river reach. This paper illustrates that HBM successfully accommodates large but sparse data sets containing poorly informative data for some units. Its conditional structure enables it to borrow strength from data-rich to data-poor units, thus improving the estimations. Predictions of the population size of the entire river reach can be derived, while accounting for all sources of uncertainty.

Résumé : Nous proposons un cadre de modélisation bayésien hiérarchique (HBM) pour estimer l’abondance d’une population de juvéniles de saumon atlantique (Salmo salar) dans la rivière Oir (France) par la méthode des retraits successifs par pêche électrique. Le jeu de données est composé de 10 sites d’échantillonnage, chacun ayant été échantillonné par un ou deux passages par année entre 1986 et 2005. Quatre modèles sont développés pour introduire les variations inter-annuelles et les effets du type d’habitat sur la densité et sur la probabilité de capture. Ces modèles sont comparés à l’aide du facteur de Bayes et d’un critère d’information basé sur la déviance. Le modèle retenu est celui qui prend en compte l’effet de l’année et du type d’habitat sur la densité de juvéniles de saumons. Il est utilisé pour extrapoler la population de saumon à l’ensemble du cours d’eau. Cet article illustre que les HBM permettent de traiter des jeux de données de grande taille dont l’information portée par chaque unité échantillonnée est hétérogène. La structure conditionnelle permet d’améliorer les estimations car elle organise un transfert d’information entre les unités. Le modèle permet d’obtenir des prédictions de l’abondance sur l’ensemble du cours d’eau, tout en prenant en compte les différentes sources d’incertitude.

Introduction

For fish populations, the quantitative assessment of their abundance is the basis for both scientific research and management. Successive removal by electrofishing is the most commonly used method for deriving estimates of abundance of riverine fish such as salmonids (Bohlin et al. 1989). In Atlantic salmon (Salmo salar), the estimation of freshwater juvenile (i.e., parr) abundance is of primary interest. It is critical for the analysis of stock (i.e., egg deposition by spawners) and recruitment (i.e., juvenile production of the subsequent generation) relationships. The transition from the egg to the age-0+ juvenile is a major bottleneck for Atlantic salmon populations (Elliott 2001; Milner et al. 2003). From a management perspective of salmon stocks, the freshwater juvenile is the earliest, the most widely used, and often the sole development stage that can be monitored prior to and independently from fisheries.

Models providing a fair appraisal of the uncertainty about population size estimates are most needed because this uncertainty is relevant to the proper assessment of the uncertainty about population dynamics parameters and ultimately...
the risk associated with management decisions (Harwood and Stokes 2003). State–space modelling is increasingly recommended as a comprehensive and flexible framework for applied population dynamics (Buckland et al. 2004; Clark and Bjørnstad 2004). In a state–space life cycle model like the one developed by Rivot et al. (2004) for Atlantic salmon, a model linking successive removal data to parr abundance can be used as an observation process to update the hidden population renewal process. The present work falls within the perspective of building such an observation model.

For estimating riverine fish population size, the successive removal method is typically used within the context of a two-stage sampling scheme: first, a number of sites are selected from the river stretch of interest and then the fish are sampled from each site by the successive removal method (Hankin 1984; Bohlin et al. 1989). The estimation can be viewed as having two steps: (i) estimate population size (or density) at the sampling site level and (ii) predict the population on the whole river stretch using these sampling site estimates. In many cases, because of logistic constraints, the proportion of the wetted area sampled is low. As a result, the uncertainty due to the extrapolation process may represent the most important part of the overall uncertainty (Hankin 1984; Mitro and Zale 2000), yet it is often ignored.

Hankin (1984) showed how the data collection should be conducted and how the sampling effort should be allocated in order to optimise the accuracy and precision of the estimation. However, one will often need to retrospectively analyse historical data, possibly collected for a diverse set of objectives. These data rarely fulfil all the conditions required to perform estimation by means of classical methods. Fairly large but sparse data sets are commonplace. The data are often gathered over a rather large number of units, a unit being defined as a sampling site at a given time during the survey. Data of each individual unit may not be informative enough to provide reliable estimates. When the number of fish caught is too small, classical methods may fail or give results with poor precision (Hirst 1994). When only a single pass removal is available, the data do not allow for estimating the catchability parameter. To circumvent these problems, the analyst often has recourse to ad-hoc methods such as crude pooling considering the capture probability is equal for all units or fixing the capture probabilities that cannot be estimated to some values previously estimated from other units (Mitro and Zale 2000). These methods make use of all of the data to confine parameter estimates on the troublesome units, but they ignore, at least partially, the between-unit variability in the unknown quantities and the results of the estimation may be inaccurate or unduly precise (Wyatt 2002). There is a need for a generic, consistent, and robust statistical framework that allows large but sparse data sets to be accommodated, away from “ad-hoceries”.

Wyatt (2002, 2003) effectively addressed this issue by a hierarchical Bayesian modelling (HBM) approach applied to the analysis of successive removal data via electrofishing of young-of-the-year (age-0+) juveniles of brown trout (Salmo trutta). Hierarchical modelling is one of the most important features of Bayesian data analysis (Gelman et al. 2004). It has been widely applied to the treatment of ecological data (Link et al. 2002; Wikle 2003; Clark 2005), with many applications in the context of fisheries ecology (Harley and Myers 2001; Rivot and Prévost 2002; Michielsen and McAllister 2004). The two companion papers of Wyatt (2002, 2003) explain how the HBM methodology, applied to the combined catch data from all sampling sites in a river stretch, provides better estimates for both the sampled sites and the reach as a whole. The HBM methodology explicitly accounts for the between-site variability in density and capture probability. The founding hypothesis of exchangeable HBM is that the density and the capture probability of each observation unit are drawn, conditionally on covariates, from a common prior probability distribution function (PDF) governed by some unknown parameters. This conditional structure enables information to be shared between units, thus helping to improve the estimation for data-poor units by borrowing strength from data-rich units. Under a Bayesian hierarchical model, obtaining posterior predictive probability distributions for nonsampled units that incorporate the uncertainty resulting from both the sampling error within the sampled sites and the uncertainty due to the variability across units is straightforward. Therefore, estimates of the population size of an entire river reach can be derived while accounting for all sources of uncertainty.

Following the same line of thought, we developed a HBM approach to estimate salmon juvenile population size and apply it the age-0+ parr population of the Oir River (Lower Normandy, France). The available data set is large: 7–10 electrofishing sites sampled over a series of 20 years (1986 to 2005), with typically two successive removals by site. Yet, it is a sparse data set, as the number of fish caught is often very low and there are missing data. Our approach enhances that of Wyatt (2002, 2003) in three significant respects. The first is that we extend it to the joint treatment of multiple years by adding a supplementary level in the hierarchical structure of the models. The second is the introduction of a categorical habitat covariate (instead of a continuous one) to explain the variability of fish density among sampling sites. The third is the assessment of the value of year and habitat type to explain variations in the probability of capture by means of model comparison. We elaborate a baseline model, grounded on well-established prior knowledge about Atlantic salmon juvenile ecology, and further propose three variants of the hierarchical structure by introducing the effects of the year and (or) the habitat type on the probability of capture. We use two Bayesian criteria, the Bayes factor and the deviance information criterion, to compare the four hierarchical structures. These criteria are first used to discard little likely models. We then compare the most credible ones regarding their ultimate objective, i.e., the posterior prediction of total population size. Lastly, we propose a method to extrapolate the results from the sampled sites to the sites not sampled in order to derive an estimation of the population size on the whole stretch of river.

Material and methods

Study site, sampling design, and data

The Oir River is an index river for Atlantic salmon population dynamics and stock assessment in France and Europe
It is a spawning tributary of the Sélune River, which flows into the English Channel (Fig. 1). It is 19.5 km long with a drainage basin of 85 km². A habitat survey covered all the areas colonized by salmon in the Oir River network. Habitat was classified into three categories: rapids–riffles, runs, and pools (Baglinière et al. 1993). Poools were neglected because of the near absence of age-0+ salmon in this type of habitat (Baglinière and Champigneulle 1986) and their marginal occurrence on the Oir River (3.5% of the water surface area). Rapids–riffles and runs are identified based on a combination of depth (<25 cm, between 25 and 60 cm, respectively), water velocity (>40 cm·s⁻¹, between 20 and 40 cm·s⁻¹, respectively), and bottom substrate (a mixture of sand, gravel, and pebbles with a higher proportion of coarse material in the rapid–riffles compared with the runs). Because of the habitat preferences of Atlantic salmon (Bardonnet and Baglinière 2000; Armstrong et al. 2003), this classification has been shown to explain a significant part of the spatial variability of the density of age-0+ juveniles in French rivers (Baglinière and Champigneulle 1986; Dumas and Prouzet 2003).

Since 1986, the age-0+ juvenile production of the main stem was surveyed every year in autumn over a 12.3 km long stretch extending from a trapping facility (the Cerisel Station) to an impassable dam (the Buat Watermill) (Fig. 1). The data are collected according to a two-stage sampling scheme.

For the first stage, depending on the year, 7–10 sampling sites (or inventory sites) were selected within the area of interest (Table 1). Their location is the same every year during the study period, but their surface area may vary between years. Each site is a section of the river associated with a unique habitat type, i.e., rapid–riffle or run. Each sampling unit is identified by three indices i, h, and k: i = 1,...,20 for the years 1986 to 2005, respectively; h = 1, 2 for the habitat type (rapid–riffle and run, respectively); k stands for the repetition per stratum (i, h) (the number of repetitions varies among strata).

For the second stage, in each unit (i, h, k), the age-0+ salmon population was sampled by electrofishing with two successive removals. From 1986 to 2005, the survey was conducted with a similar operating protocol and essentially the same staff. A direct-current generator (200 W) was used. The fishing crew consisted of three people: the anode operator and two dip-netters. The sites were swept by progressing upstream to surprise the fish, the age-0+ salmon being oriented head upstream. The second pass was realized shortly after the first one. During the experiment, the sampling sites were not closed with barrier nets. However, given the length of the sampling sites (>50 m) and the short lapse of time between consecutive passes, it is considered that immigration into or emigration out of the sampling sites was negligible.

The complete data set consists of the numbers of fish captured at the first and the second pass for 190 sampling units (Table 2). It is a sparse data set because often very small numbers of fish were caught and there are some missing data for the second pass. In the following, it is assumed that the missing data mechanism is ignorable in the sense defined by Gelman et al. (2004). Inferences about the unknowns of the model and the missing data are then carried out without further modelling of the missing data process. The ignorability hypothesis is discussed in detail in Appendix A.

The sampling rate (measured as the proportion of the wetted area sampled) in each stratum (i, h) varies between 2.9% and 12.2%.

**Baseline model (M₀)**

We denote random variables by Greek letters and observed values (i.e., data) and constants by capital letters. \( p(\bullet | \theta) \) denotes a conditional probability distribution function knowing \( \theta \). \( \bullet | \theta \sim f(\theta) \) means that conditionally on \( \theta \), the random variable of interest, has the density \( f \) with parameters \( \theta \). \( E(\bullet) \) is the mean.

**Sampling distribution**

In electrofishing removal experiments, fish caught at the first pass are removed and are not available to the catch at the second pass. Adhering to the classical assumptions of the removal method (Carle and Strub 1978), for each unit (i, h, k) we assume that (i) the population is closed between the first and the second pass, i.e., no emigration, immigration, recruitment, or natural mortality occur during the experiment; (ii) for each removal event, all the fish have the same probability of capture and are independent from each other with regard to the capture process; and (iii) the probability of capture is constant between the two successive passes. A more realistic model accounting for a decline between the successive pass could have been proposed here. However, because a maximum of two passes is available per site, there would not be much information under our data collection setting to separate out the parameters of such a model, and we would rather choose a more parsimonious option. Based on these assumptions, a binomial model with a constant probability of capture is set to mimic the removal experiments. Catch data of the first and second passes are denoted \( C_{i1}(i, h, k) \) and \( C_{i2}(i, h, k) \), respectively; \( \nu(i, h, k) \) is the probability of capture; and \( \nu(i, h, k) \) is the initial population size. The sampling distributions of the catch data are

\[
(1a) \quad C_{i1}(i, h, k) | \nu(i, h, k), \pi(i, h, k) \sim \text{Binomial}(\nu(i, h, k), \pi(i, h, k))
\]

\[
(1b) \quad C_{i2}(i, h, k) | C_{i1}(i, h, k), \nu(i, h, k), \pi(i, h, k) \sim \text{Binomial}(\nu(i, h, k) - C_{i1}(i, h, k), \pi(i, h, k))
\]

The population size \( \nu(i, h, k) \) depends on the expected fish density \( \delta(i, h, k) \) (fish·m⁻²) and the surface area \( S(i, h, k) \) of the sampling sites. We assume that the fish are randomly distributed in space within an infinitely large surface from which the site with finite surface \( S(i, h, k) \) is randomly picked. Then, \( \nu(i, h, k) \) is assumed to be Poisson-distributed with parameter \( \lambda(i, h, k) = \delta(i, h, k) \cdot S(i, h, k) \):

\[
(2) \quad \nu(i, h, k) | \lambda(i, h, k) \sim \text{Poisson}(\lambda(i, h, k))
\]
given year, the spatial variability of the density is high and correlated with riverine physical habitat (Amiro 1990; Bardonnet and Baglinière 2000; Armstrong et al. 2003). In French rivers, it has been demonstrated that age-0+ Atlantic salmon densities in autumn where notably higher in rapids–riffles compared with runs (Baglinière and Champigneulle 1986; Dumas and Prouzet 2003; Baglinière et al. 2005). We incorporate this prior knowledge about the Atlantic salmon ecology by setting a hierarchical structure on the density conditioned by year and habitat type. The densities $\delta(i, h, k)$ are assumed to be partially exchangeable. They are modelled as identically and independently distributed (iid) variables drawn from a lognormal prior PDF. The influence of the covariates year and habitat type is introduced via additive effects on mean log(density) (multiplicative on density). No spatial dependence of the density between adjacent sites is introduced in this model. We denote $\mu$ the overall mean of log(density), $\alpha(i)$ and $\beta(h)$ the effect of the year $i$ and the habitat $h$, respectively, and $\sigma^2$ the between-unit variance that is common to all the $(i, h)$ strata:

$E(\log(\delta(i, h, k))) = \mu + \alpha(i) + \beta(h)$

$\log(\delta(i, h, k))|\mu, \alpha(i), \beta(h), \sigma^2 \sim \text{iid Normal}(E(\log(\delta(i, h, k))), \sigma^2)$

The years of observation are seen as a sample drawn from an infinite set of years that are exchangeable regarding their effect on density. Thus, conditionally upon between-year variance $\sigma_{\alpha}^2$, the $\alpha(i)$s are modelled as random effects, drawn from a common distribution:

$\alpha(i) | \sigma_{\alpha}^2 \sim \text{iid Normal}(0, \sigma_{\alpha}^2)$

In contrast, the habitat types are a finite and exhaustive set of two mutually exclusive categories and the $\beta(h)$s are modelled as fixed effects. Consistently with the available knowledge, the $\beta(h)$s must a priori keep to the constraint $\beta(1) > \beta(2)$. To meet this requirement, $\beta(1)$ is drawn from a half positive normal prior with large variance $V_{\beta} = 100$, and a
sum-to-zero constraint is imposed to avoid confusion with the overall mean $\mu_\pi$:

\[
\beta_\delta(1) \sim \text{Normal}(0, V_{\beta_0} = 100) I(0, +\infty)
\]

and

\[
\beta_\delta(2) = -\beta_\delta(1)
\]

### Exchangeable hierarchical structure on $\pi$

The probability of capture $\pi(i,h,k)$ can also vary among units. In stream-dwelling salmonids, many influential factors have been identified (e.g., river width, depth, water velocity, temperature, water conductivity, fish size, habitat complexity), but the results are not fully consistent across studies (Randall 1990; Riley et al. 1993; Speas et al. 2004). In our case study, it is not clear whether the year or the habitat type would a priori cause systematic variations of the probability of capture. Consequently, in the baseline model, a hierarchical structure was used for the probabilities of capture assuming they were fully exchangeable among the sampling units. The log-transformed probabilities of capture were modelled as identically and independently drawn from a common normal prior PDF with unknown mean $\mu_\pi$ and variance $\sigma_\pi^2$:

\[
\logit(\pi(i,h,k)) | \mu_\pi, \sigma_\pi \sim \text{Normal}(\mu_\pi, \sigma_\pi^2)
\]

This baseline modelling option is flexible to accommodate variability in the probability of capture, while being parsimonious and allowing a transfer of information between sampling units.

### Directed acyclic graph and prior specification

A representation of the baseline model $M_0$ using the directed acyclic graph (DAG) conventions is provided (Fig. 2). A DAG is a useful metaphor of a Bayesian model that represents the conditional relationship between variables (Spiegelhalter et al. 1996). All the free parameters of the model, i.e., those not conditioned by any quantity, are assigned weakly informative and independent prior PDFs (Table 3). In agreement with Gelman (2006) statement, “we follow the approach of Bernardo (1979) and consider so-called noninformative priors as reference models to be used as a standard of comparison or starting point in place of the proper, informative prior distributions that would be appropriate for a full Bayesian analysis”. It is also assumed that the $\pi(i,h,k)$ and the $\delta(i,h,k)$s are a priori uncorrelated. A sensitivity analysis to the choice of priors is presented in Appendix B.

### Extended models with effects of the year and the habitat type on $\pi$

Because of the uncertainty about systematic effects of the year or the habitat type on the probability of capture, we assessed the value of these covariates to explain variations of the $\pi(i,h,k)$ between sampling units by means of a model comparison. Three variant models were elaborated from $M_0$ by modifying the hierarchical structure on $\pi$.

In the most comprehensive model $M_3$, the effects of year and habitat type are introduced by means of a partially exchangeable hierarchical structure similar to the one adopted for the densities. We denote $\mu_{\pi_A}$ the overall mean of logit($\pi$) and $\alpha_{\pi_A}(i)$ and $\beta_{\pi_A}(h)$ the effects of year $i$ and habitat $h$, respectively:

### Table 2. Number of fish captured on the first ($C_1$) and second ($C_2$) passes.

<table>
<thead>
<tr>
<th>Year</th>
<th>Habitat type 1</th>
<th></th>
<th>Habitat type 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>1986</td>
<td>7</td>
<td>0</td>
<td>73</td>
<td>7</td>
</tr>
<tr>
<td>1987</td>
<td>35</td>
<td>2</td>
<td>23</td>
<td>NS</td>
</tr>
<tr>
<td>1988</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>1989</td>
<td>20</td>
<td>5</td>
<td>56</td>
<td>3</td>
</tr>
<tr>
<td>1990</td>
<td>19</td>
<td>1</td>
<td>54</td>
<td>4</td>
</tr>
<tr>
<td>1991</td>
<td>28</td>
<td>3</td>
<td>39</td>
<td>5</td>
</tr>
<tr>
<td>1992</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1993</td>
<td>16</td>
<td>7</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1994</td>
<td>5</td>
<td>1</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>1995</td>
<td>13</td>
<td>6</td>
<td>23</td>
<td>6</td>
</tr>
<tr>
<td>1996</td>
<td>32</td>
<td>11</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>1997</td>
<td>79</td>
<td>13</td>
<td>70</td>
<td>17</td>
</tr>
<tr>
<td>1998</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>1999</td>
<td>30</td>
<td>5</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>2000</td>
<td>27</td>
<td>7</td>
<td>21</td>
<td>4</td>
</tr>
<tr>
<td>2001</td>
<td>74</td>
<td>27</td>
<td>33</td>
<td>12</td>
</tr>
<tr>
<td>2002</td>
<td>21</td>
<td>8</td>
<td>25</td>
<td>15</td>
</tr>
<tr>
<td>2003</td>
<td>31</td>
<td>4</td>
<td>33</td>
<td>11</td>
</tr>
<tr>
<td>2004</td>
<td>24</td>
<td>5</td>
<td>65</td>
<td>12</td>
</tr>
<tr>
<td>2005</td>
<td>40</td>
<td>9</td>
<td>45</td>
<td>12</td>
</tr>
</tbody>
</table>

**Note:** Habitat types: 1, rapid–riffle; 2, run. Data are presented as $C_1$, $C_2$; NS, second pass not completed. Dash (—) indicates that the site was not sampled at all.

*Repetitions ($k$) 1–4 for habitat 1 and 1–6 for habitat type 2.
The two other variant models are nested into $M_3$ as they only account for the effect of either year ($M_1$), by setting $\beta_{\pi}(h) = 0$ for all the habitat types, or habitat ($M_2$), by setting $\alpha_{\pi}(i) = 0$ for all the years. Note that $M_0$ is also nested within these three variants as it amounts to setting both $\beta_{\pi}(h) = 0$ and $\alpha_{\pi}(i) = 0$. For model $M_0$, diffuse prior PDFs are used on all parameters (Table 3; a sensitivity analysis to the priors is presented in Appendix B).

**Model checking**

We used the omnibus $\chi^2$ discrepancy as a summary measure to assess a posteriori the consistency between the model and the data (Gelman et al. 2004, p. 175). We computed the $\chi^2$ statistics for the catch at the two successive passes $C_1$ and $C_2$:

$$\chi^2(C_1;\nu,\pi) = \sum_{\text{all units}(i,h,k)=1,...,n} \frac{(C_1(i,h,k) - E(C_1(i,h,k)|\nu(i,h,k),\pi(i,h,k)))^2}{\text{Var}(C_1(i,h,k)|\nu(i,h,k),\pi(i,h,k))}$$

$$\chi^2(C_2;\nu,\pi) = \sum_{\text{all units}(i,h,k)=1,...,n} \frac{(C_2(i,h,k) - E(C_2(i,h,k)|C_1(i,h,k),\nu(i,h,k),\pi(i,h,k)))^2}{\text{Var}(C_2(i,h,k)|C_1(i,h,k),\nu(i,h,k),\pi(i,h,k))}$$

For each set of parameters $(\nu(i,h,k), \pi(i,h,k))_{i,h,k=1,...,n}$ drawn in their joint posterior distribution, we compared the realized discrepancies $\chi^2(C_1;\nu,\pi)$ and $\chi^2(C_2;\nu,\pi)$ with the observed values of $C_1$ and $C_2$ with the predicted $\chi^2$ discrepancies $\chi^2(C_1^{\text{rep}};\nu,\pi)$ and $\chi^2(C_2^{\text{rep}};\nu,\pi)$ computed with posterior predictive replicates of $C_1$ and $C_2$. $C_1^{\text{rep}}(i,h,k)$ ~ Binomial$(\nu(i,h,k),\pi(i,h,k))$ and $C_2^{\text{rep}}(i,h,k)$ ~ Binomial$(\nu(i,h,k) - C_1(i,h,k),\pi(i,h,k))$. The Bayesian $p$ value is the probability that $\chi^2(C_1^{\text{rep}};\nu,\pi) \geq \chi^2(C_2^{\text{rep}};\nu,\pi)$ (pass = 1 or 2) estimated over a sample of $(C_1^{\text{rep}};\nu,\pi)$ values. A $p$ value reasonably near 0.5 indicates consistency between the model and the data, whereas very high (near 0.95) or very
Table 3. Prior distributions of free hyperparameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu_\delta )</td>
<td>( \sim ) Normal(0, ( V_\delta = 100 ))</td>
</tr>
<tr>
<td>( \sigma_\delta )</td>
<td>( \sim ) Uniform on [0, ( \sigma_{\max} = 10 )]</td>
</tr>
<tr>
<td>( \mu_\pi )</td>
<td>( \sim ) Normal(0, ( V_\pi = 100 ))</td>
</tr>
<tr>
<td>( \sigma_\pi )</td>
<td>( \sim ) Uniform on [0, ( \sigma_{\max} = 10 )]</td>
</tr>
<tr>
<td>( \sigma_{\pi_1} ) (( M_1 ) and ( M_3 ) only)</td>
<td>( \sim ) Uniform on [0, ( \sigma_{\max} = 10 )]</td>
</tr>
</tbody>
</table>

Note: All standard deviation parameters \( \sigma_\nu \), \( \sigma_\alpha \), \( \sigma_\beta \), and \( \sigma_\pi_2 \) are assigned uniform prior on a sufficiently large range, as recommended by Gelman (2006) for hierarchical models.

The deviance information criterion (DIC; Spiegelhalter et al. 2002) is a measure of complexity and fit designed to compare hierarchical models of arbitrary structure. Its rationale and interpretation are different than those of the BF and are more analogue to the frequentist Akaike information criterion (AIC): the smaller the DIC, the more favoured is a model. It combines a measure of the goodness of fit, defined as the posterior mean of the deviance \( \text{Dev}(\theta) \) (where the deviance is \(-2\) times the log-likelihood), with a measure of the model complexity \( pD \), acting as a penalty term:

\[
(10) \quad \text{DIC} = \text{Dev}(\theta) + pD
\]

\( pD \) is defined as \( pD = \text{Dev}(\hat{\theta}) - \text{Dev}(\theta) \), where \( \hat{\theta} \) is a point estimate of the parameter vector. The posterior mean is the natural choice for \( \theta \). However, following Spiegelhalter et al. (2002), we used the posterior medians instead because they are better point estimates when the posterior densities of the parameters are highly skewed. The DIC is easily calculated by sampling techniques and its computation avoids the instability observed when estimating BF from eq. 9c.

The BF is known to be sensitive to the prior choice. To assess this sensitivity, we fit and compared the four models \( M_0, \ldots, M_3 \) under alternative prior configurations (see Appendix B).

Extrapolation to the whole river stretch

The ultimate goal of our study is to provide estimates of the total age-0+ salmon population for each year \( i \), denoted \( v(i) \). It is the sum of the population size \( v(i) \) estimated on the sampling sites with the population size \( v(i) \) predicted on the nonsampled area, where \( v(i) \) and \( v(i) \) are calculated as a sum over the habitat types:

\[
(11) \quad v(i) = \sum_{h=1}^{2} v(i, h) + \sum_{h=1}^{2} v(i, h)
\]

Each term \( v(i, h) \) is directly issued from the model under concern (\( M_0 \) to \( M_3 \)). It is the sum of the estimated population size \( v(i, h, k) \) over all the sampling units with habitat type \( h \). The \( v(i, h) \) are the extrapolated population sizes for the \((i, h)\) strata.

The relative importance of the extrapolation term \( v(i) \) depends on the sampling rate. At one extreme, if the whole water surface area was sampled, then \( v(i) = 0 \). At the other extreme, if a very small proportion of the water surface area was sampled, then \( v(i) \) makes up a very high proportion of \( v(i) \). Two sources of uncertainty, corresponding to the two stages of the sampling scheme, must be taken into account in the estimation of \( v(i) \): that arising from the binomial sampling errors associated with the removal method within the sampling units and that arising from the sampling variance between the sampling units. The second source of uncertainty concerns the extrapolation term \( v(i) \) only. It depends on the sampling rate and on the spatial variability of the density. If the sampling rate is low and the density is highly variable, it can represent the greatest source of uncertainty in the estimation of \( v(i) \).

To explicitly account for spatial variability in density within the \((i, h)\) strata, we mimic the first stage of the sampling scheme: the whole river is regarded as being divided into a very large number of sites (sufficiently large to be considered as infinite) from which the sampled sites were drawn. The nonsampled surface area \( S(i, h) \) (Table 4) is considered to consist of a number \( n_s(i, h) \) of sites of equal surface area \( S_s(i, h) \). \( n_s(i, h) \) is set so that \( S_s(i, h) \) is approxi-
chain Monte Carlo (MCMC) sampling. MCMC methods now enable the Bayesian estimation of complex highly dimensional hierarchical models, which could not be treated otherwise (Brooks 2003). MCMC sampling was carried out by means of the WinBUGS® software (Spiegelhalter et al. 2003) and the BRugs package of R (www.r-project.org). It was used to obtain a sample from the joint posterior distribution of all unknowns, i.e., model parameters, missing data, and predictions. The convergence of the MCMC chains to their ergodic distribution was tested for the model parameters via the Gelman–Rubin (GR) diagnostics as implemented by WinBUGS®. This diagnostic assesses the mixing of several independent chains starting at contrasted initial values spread over the variable space. We used three independent chains, and the first 10 000 iterations were discarded as an initial burn-in period. Then, 100 000 further iterations were performed and convergence was checked using the GR diagnostic. CPU time needed to run three parallel chains is 1 h with a 2.8GHz Intel P4 processor. All inferences were derived using the sample obtained from pooling the three chains (i.e., size = 300 000). To smooth the numeric instability in the computation of the BF, a sample of size 600 000 was used to compute BF.

Results

The outcomes of the model comparison and the posterior estimation are robust to the choice of the prior configuration (see Appendix B). The results presented in the following are those obtained under the baseline prior configuration described in the Material and methods section.

Model comparison

Both models $M_2$ and $M_3$ are clearly discarded by the BF (Table 5). Based on the guidelines proposed by Kass and Raftery (1995) for interpreting BF values, the evidence against $M_2$ and $M_3$ is very strong and there is slightly less support for $M_1$ compared with $M_0$. These results are robust against the instability of the BF approximations calculated by the Kass and Raftery (1995) method (Appendix B). The differences between the DIC values of the four models are much less contrasted (Table 5). To interpret DIC values, Spiegelhalter et al. (2002) suggest adopting the rule of thumb proposed by Burnham and Anderson (2002) for AIC. Following this, none of the four models can be clearly discarded. Given these results, we retained $M_0$ as the preferred model, with $M_1$ as a valuable alternative candidate.
Table 5. Alternative competing models and criterions for model selection.

<table>
<thead>
<tr>
<th>Models</th>
<th>Density</th>
<th>Probability of capture</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_0$</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>$M_1$</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>$M_2$</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>$M_3$</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\text{Dev}(\theta)$</th>
<th>$pD$</th>
<th>DIC</th>
<th>Log(BF $M_0$ vs. $M_i$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>576</td>
<td>239</td>
<td>815</td>
<td>0</td>
</tr>
<tr>
<td>586</td>
<td>231</td>
<td>817</td>
<td>1.2</td>
</tr>
<tr>
<td>574</td>
<td>240</td>
<td>814</td>
<td>3.7</td>
</tr>
<tr>
<td>584</td>
<td>230</td>
<td>814</td>
<td>5.6</td>
</tr>
</tbody>
</table>

Note: Values are computed under the baseline prior configuration as described in the Material and methods section. ×, the effect of the corresponding covariate is introduced; —, no effect is introduced. $\text{Dev}(\theta)$, the posterior mean of the deviance; $pD$, measure of the model complexity; DIC, deviance information criterion; BF, Bayes factors. BF's are estimated from eqs. 9a–9c by considering the whole Markov chain Monte Carlo (MCMC) sample (size = 600 000); $M_i$, model $i$, $i = 1–3$; $M_0$, baseline model.

Table 6. Main statistics of the marginal posterior probability distribution functions (PDFs) of the free parameters in model $M_0$.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>2.5%</th>
<th>5%</th>
<th>25%</th>
<th>Median</th>
<th>75%</th>
<th>95%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_\delta$</td>
<td>0.87</td>
<td>0.25</td>
<td>0.37</td>
<td>0.46</td>
<td>0.71</td>
<td>0.87</td>
<td>1.03</td>
<td>1.27</td>
<td>1.35</td>
</tr>
<tr>
<td>$\sigma_\delta$</td>
<td>1.09</td>
<td>0.09</td>
<td>0.93</td>
<td>0.95</td>
<td>1.03</td>
<td>1.09</td>
<td>1.15</td>
<td>1.25</td>
<td>1.29</td>
</tr>
<tr>
<td>$\mu_\pi$</td>
<td>1.62</td>
<td>0.11</td>
<td>1.41</td>
<td>1.44</td>
<td>1.55</td>
<td>1.62</td>
<td>1.70</td>
<td>1.81</td>
<td>1.85</td>
</tr>
<tr>
<td>$\sigma_\pi$</td>
<td>0.61</td>
<td>0.13</td>
<td>0.37</td>
<td>0.41</td>
<td>0.52</td>
<td>0.60</td>
<td>0.69</td>
<td>0.84</td>
<td>0.89</td>
</tr>
<tr>
<td>$\sigma_{\alpha_\pi}$</td>
<td>0.97</td>
<td>0.22</td>
<td>0.63</td>
<td>0.67</td>
<td>0.82</td>
<td>0.95</td>
<td>1.10</td>
<td>1.36</td>
<td>1.47</td>
</tr>
</tbody>
</table>

Note: Posterior PDFs are computed under the baseline prior configuration as described in the Material and methods section. All statistics are computed from a Markov chain Monte Carlo (MCMC) sample (size = 300 000). SD, standard deviation.

Estimation

Model $M_0$

The estimation of the overall mean density in the log scale ($\mu_\delta$) is fairly precise (Table 6). The posterior PDFs of the $\alpha_\delta(i)$s (Fig. 3a) illustrate the high between-year variability of the age-0+ salmon density without any particular trend. The variability between the $\alpha_\delta(i)$s (the mean of the posterior PDF of $\alpha_\delta$) is 0.97; Table 6) is much higher than the within-year uncertainty of estimated effects (the standard deviations of the posterior PDF of the $\alpha_\delta(i)$s range from 0.42 to 0.58 according to the year). In the natural scale of the density, the mean of the posterior predictive PDFs of the density on a rapid–riffle habitat in a given year ranges from 0.018 (year 1998) to 0.38 fish·m$^{-2}$ (year 2001). The effect of the habitat type on the density is very strong (Fig. 3b). The mean of the posterior PDF of the density ratio in the habitat type 1 versus type 2, calculated as $e^{\alpha_\delta(i)}$ or $e^{\beta_\delta(i)}$, is 6.5. Even after accounting for the effects of year and habitat type, the residual spatial variability of the density $\sigma_\pi$ remains high when compared with the overall mean $\mu_\pi$ or with the between-year variability $\alpha_\pi$ (Table 6).

The overall mean probability of capture in the logit scale, $\mu_\pi$, is estimated with good precision (Table 6). In the natural scale, the mean of the posterior predictive PDF of the probability of capture is 0.82. The between-unit variability of the probability of capture ($\sigma_\pi$) is rather low compared with the overall mean $\mu_\pi$ (Table 6). Consistently, even for units with missing or little informative data, the probability of capture is rather precisely estimated (the $\pi(i,h,k)$ have dome-shaped posterior densities close to 0 around 0 and 1) and rather high (the mean of the posterior PDF of the probability of capture ranges from 0.64 to 0.90 according to the sampling unit).

Model $M_1$

The between-year variability of the probability of capture, as illustrated by the posterior PDFs of the $\alpha_\pi(i)$s (Fig. 3c), is significant but with no particular pattern. In the natural scale of the probability of capture, the mean of the posterior predictive PDF of the probability of capture in a given year ranges from 0.68 (2002) to 0.9 (1986). The posterior PDFs of the $\alpha_\pi(i)$s (not shown) are very similar to those obtained from the model $M_0$. There is no evidence of confusion between the effects of year on the density and on the probability of capture. The patterns of the between-year variations of the estimated effects $\alpha_\delta(i)$ or $\alpha_\pi(i)$s are different, and the joint posterior PDFs of each couple ($\alpha_\delta(i),\alpha_\pi(i)$) (not shown) do not reveal any noteworthy correlation pattern.

Models $M_2$ and $M_3$

For both models there is no significant effect of habitat type on the probability of capture (Figs. 3d and 3f). The year effect on the probability of capture from the model $M_3$ (Fig. 3e) is similar to that obtained from the model $M_1$ (Fig. 3c).

Model checking

We only report the results of the posterior checking obtained for our favourite model $M_0$. The plots of the $\chi^2$ discrepancy statistics do not reveal any significant inconsistency between $M_0$ and the data (Fig. 4). The $p$ values of 0.28 (first pass) and 0.55 (second pass) indicate that the discrepancies
Fig. 3. Summaries of the marginal posterior probability distribution functions (PDFs) of (a) the $\alpha_i$s under the baseline model $M_0$, (b) the $\beta_h$s under $M_0$, (c) the $\alpha_i$s under model $M_1$, (d) the $\beta_h$s under model $M_2$, (e) the $\alpha_i$s under model $M_3$, and (f) the $\beta_h$s under model $M_3$. Posterior PDFs are computed under the baseline prior configuration as described in the Material and methods section. Median (solid circle), 25% and 75% percentiles (solid triangles), and 5% and 95% percentile (+) are computed from a Markov chain Monte Carlo (MCMC) sample (size = 300 000).
Fig. 4. Discrepancy plot obtained from the baseline model $M_0$. $\chi^2(C_1, \nu, \pi)$ and $\chi^2(C_2, \nu, \pi)$ are the realized discrepancies computed with the observed values of $C_1$ and $C_2$. $\chi^2(C_{1}^{\text{rep}}, \nu, \pi)$ and $\chi^2(C_{2}^{\text{rep}}, \nu, \pi)$ are the predicted discrepancies computed with posterior predictive replicates $C_{1}^{\text{rep}}$ and $C_{2}^{\text{rep}}$. The corresponding $p$ values are 0.28 and 0.55.

Extrapolation to the whole river stretch

Given the results of the model comparison, both $M_0$ and $M_1$ were used to estimate the population size on the whole river stretch. The posterior predictive PDFs of the $\nu(i)$s (eqs. 11a–11b) are almost identical in terms of both expected values and precision, whether model $M_0$ or $M_1$ is used (Fig. 5). We then eventually retain the most parsimonious one, $M_0$, as our preferred model.

The between-year variability of the estimated population size is high with no particular trend (Fig. 5). The posterior mean of $\nu(i)$ ranges from 309 (year 1998) to 6762 (year 2001). The contrast in the uncertainty about the yearly population size estimates is large as well. As expected when using lognormal distributions for the densities, the uncertainty about the population size estimates increases with their means.

The estimates of the age-0+ population size (posterior means of the $\nu(i)$s, years 1986–2003) were compared with those of the smolt run (from capture–mark–recapture data) of the corresponding cohorts provided by Baglinière et al. (2005) (Fig. 6). The correlation between the two estimates is consistent with moderate yearly fluctuations of the age-0+–smolt survival rate in the Oir River.

Discussion

Critical examination of modelling hypotheses

The accuracy of the model used for the removal method in each sampling unit depends on the fulfilment of its basic assumptions: (i) change in population size only occurs through the capture process; (ii) during each removal pass, the fish are identically and independently distributed with regard to the probability of capture; and (iii) the probability of capture is constant between the first and the second removal pass. Certainly the last hypothesis is the most questionable. Indeed, the probability of capture is likely to decrease in successive removals (Bohlin and Cowx 1990; Peterson et al. 2004). This has been suggested to arise from different
causes: (i) heterogeneous catchability among individuals, leading to the most easily captured individuals being caught first (Dolan and Miranda 2003; Mäntyniemi et al. 2005); (ii) a decrease in catchability between successive passes due to changes in the fish behaviour or the reduced susceptibility to electricity for fish already having experienced an electric shock (Bohlin et al. 1990). Schnute (1983) proposed a more realistic approach that allows for a decline in the probability of capture between passes. Wang and Loneragan (1996) developed an overdispersed model where the probability of capture varies randomly and independently among passes. Mäntyniemi et al. (2005) proposed a model explicitly accounting for a random between-individual variability in the probability of capture. Following these authors, more realistic models could have been applied to the data. However, there is not much information about the parameters of such models when only two passes are available. Using such a model would have required the use of informative priors to avoid confusion among parameters in the posterior inferences. The use of informative priors is fine in theory but has to be taken with great caution in practice. Informative priors should be derived from existing information (i.e., data or expert opinion) by formal means (e.g., meta-analysis or formal methods for prior elicitation from expert opinion) and by a thorough examination of all the sources of information available. Otherwise, there is a great risk that the practitioner, even unintentionally, favours his own beliefs over alternatives. In addition, when the modeller also has a good knowledge of the observed data available for the Bayesian updating, it is very difficult to elaborate informative priors independently from this knowledge, with the undesired consequence of using part of the information of the data twice. Hence, without dismissing the ultimate interest in using informative priors and complex models, an analysis based on parsimonious models with vague reference priors was our preferred default approach. As a first step, it is useful — would it only be for an illustrative purpose — to ensure that the posterior estimates reflect primarily the information contained in the data (Gelman et al. 2004; Gelman 2006). For this reason, we built a simplified model where the capture probability is constant between the two successive passes. Still, our results depend on this hypothesis. Our simplification is likely to cause not only underestimation of the population size (Peterson et al. 2004; Mäntyniemi et al. 2005; Rosenberg and Dunham 2005), but also underestimation of the uncertainty and overestimation of spatiotemporal variability (Mäntyniemi et al. 2005). In our case, this problems may not be too acute. Indeed, the probability of capture is high (on average from 0.68 to 0.9 according to the year). In addition, the age-0+ salmon juveniles are mainly found in open shallow waters in the middle of the river channel and their size range is limited. These conditions favour robustness against the violation of the constant catchability hypothesis (Mäntyniemi et al. 2005).

Fish densities have been assumed to be lognormally distributed. We did not formally test for the relevance of this modelling hypothesis, which is both convenient and classical. It allows the introduction of multiplicative effects of covariates on the density by means of standard linear-normal modelling of the log(density). The commonly held belief that population abundance follows a lognormal distribution has been critically examined by Halley and Inchausti (2002). These authors recommend the gamma PDF as the most valuable alternative. Our HBM approach could readily be adapted to gamma-distributed fish densities. Using the scale–shape parameterisation of the gamma PDF, it is easy to introduce multiplicative effects of covariates on the mean.

For the sake of simplicity, no spatial dependence between densities at adjacent sites was introduced. We focussed our modelling effort on the joint treatment of several years and the accounting for habitat effects. Further development of spatially structured models is certainly warranted as it has been demonstrated that juvenile salmon densities are spatially autocorrelated (De Oliveira et al. 2001). The spatially explicit model coupled with a GIS developed by Wyatt (2003) is an interesting starting point in that direction.

Under our HBM approach, priors have to be assigned to the hyperparameters at the top of the multilevel structure. One could have looked for hyperpriors that would better reflect our prior knowledge about the spatial and temporal variability of the density and the probability of capture. However, we instead used weakly informative priors. This choice was motivated by arguments already discussed. Still, strong prior knowledge about the spatial variability of the densities is introduced in the hierarchical model structure through (i) the definition of habitat covariates and (ii) the constraints that we imposed on the habitat effects for the density (we imposed $\beta_{gh} = 1 > \beta_{gh} = 2$, where $\beta_{gh} = 1$ and $\beta_{gh} = 2$ represent the effects of rapid–riffle habitat and run habitat on the density, respectively).

**Bayesian tools for model comparison**

We jointly used the BF and the DIC to compare our four competing models. In the Bayesian framework, the BF is a natural tool to compare the relative credibility of alternative models, yet it is not often implemented. It suffers from two main drawbacks: it may be difficult to calculate, and it may be highly sensitive to the prior. Yet, it revealed a valuable operational tool in our case. Among the various MCMC methods available to compute the BF, the “product space search” method or the “reversible jump” algorithm (reviewed in Han and Carlin 2001) are more recommended but are difficult to put into practice when more than two competing models are compared. We estimated the BF values from a simple method based on a MCMC integration of the likelihood (Kass and Raftery 1995). The numerical instability of this technique did not preclude the identification of the most credible model because the magnitude of the instability was much lower than the BF itself (Appendix B). We contrasted various prior configurations to assess the sensitivity of the BF. Because our data set is informative enough, the conclusion drawn from the BF was revealed to be robust to these prior hypotheses (Appendix B).

The DIC is more recent and its application for model selection is still being discussed (Spiegelhalter et al. 2002; Celeux et al. 2006). This criterion is appealing because of its straightforward computation from posterior samples and its numerical stability, but its rationale and interpretation are different than those of the BF. The differences between the DIC values of the four models were much less contrasted than BF values, and none of the four models could clearly be discarded based on this criterion. The results seem to con-
firm the general tendency of the BF to opt for more parsimonious models compared with other model selection tools such as AIC, an analogue to the DIC in its construction and its interpretation (Kass and Raftery 1995).

In a full Bayesian approach for estimating the population in the entire river stretch, we could have undertaken Bayesian model averaging (BMA) to account for model uncertainty (Hoeting et al. 1999; King and Brooks 2001). Posterior estimations would have been obtained by averaging the posterior PDFs derived from the alternative models, with the posterior relative credibility of each model as weights. The posterior relative credibility is easily derived from the BF. In our case study, the clearly discarded models $M_2$ and $M_3$ are not worth considering in a BMA approach as their weights would be so small that they would not effectively contribute to the averaging procedure. Under the hypothesis of a prior odd of 1 between the two most likely models $M_0$ and $M_1$, their relative credibilities are 0.77 and 0.23, respectively. However, our results proved that both models lead to very similar estimations (Fig. 4). Implementing a time-consuming and computationally intensive BMA approach would have been a distraction, and we recommend the single use of the most parsimonious model $M_0$.

A powerful and flexible framework for integrating successive removal data

The HBM framework proved powerful and flexible for processing multiyear and multisite successive removal data. It successfully enabled joint treatment of several years of data and integration of a habitat-categorical covariate explaining part of the spatial variability of the density of salmon juveniles. Our modelling approach takes advantage of the “borrowing of strength” between units and the additional information brought by the habitat covariate to improve the estimation of the total salmon population size every year.

The HBM framework successfully accommodated our large but sparse data set, which contained poorly informative or even missing data for some units. The conditional probability structure enables the transfer of information from data-rich to data-poor units, while explicitly accounting for the cross-units variability of the density and probability of capture. The posterior predictive PDF of the probability of capture conveys all the posterior information that can be used to infer the density at any site where only one pass is carried out. This methodology is more comprehensive compared with others (e.g., the mean capture probability method; Mitro and Zale 2000) that only address the transfer of information issue in a rather ad-hoc manner. This method could be taken advantage of to enhance the precision of the yearly population size estimates without augmenting the sampling effort. The sampling area could be increased by reallocating part of the sampling effort to additional single removal sites. Such strategy would be of special interest when the between-site variability of the density represents the main source of uncertainty, as is often the case (Hankin 1984; Mitro and Zale 2000).

Our HBM approach provides estimates of the population size on a whole river stretch, while accounting for the two sources of uncertainty associated with the two-stage sampling scheme: (i) the sampling variance due to the removal method in sampling sites and (ii) the uncertainty arising from the across-site variability of the density. The resulting posterior PDF of the population size can be used in further stock assessment work or population dynamics studies.

Improving the assessment of population size and dynamics

The results obtained from our case study are sensible, consistent, and insightful for the dynamics and the management of Atlantic salmon populations. They highlight that the probability of capture varies among units. Several factors, both abiotic (e.g., river width, depth, water velocity, temperature, water conductivity, habitat complexity) and biotic (e.g., fish size, fish density), could explain this variability (Randall 1990; Riley et al. 1993; Speas et al. 2004). Systematic variations of these variables may translate into an effect of the year and the habitat type, but our results indicate that once the influence of these covariates on the density has been taken into account, there is no significant effect of habitat type on the probability of capture. This can be explained by two facts: (i) in spite of the variability of the salmon habitat, many physical characteristics of the stream remain relatively homogeneous along the stretch of interest and (ii) the electrofishing crew is trained to operate in a consistent manner to maximise the capture efficiency whatever the habitat type. Conclusions regarding the effect of the year on the probability of capture are less clear-cut. Indeed, the BF is in favour of $M_0$ (no effect of years), whereas the DIC is in favour of $M_1$ (effect of year), and the posterior estimates of the $\alpha_p(i)$ suggest that year effects may be significant. Variations of the hydrological conditions between years during the electrofishing campaigns may explain part of the observed pattern.

The age-0+ salmon density fluctuates widely between years. This is a typical result for juvenile salmonid recruitment, which is highly sensitive to environmental fluctuations (Elliott 2001; Milner et al. 2003; Baglinière et al. 2005). On average, the density of age-0+ juveniles on rapid–riffle habitat is 6.5 times higher than on run habitat. These figures are in agreement with previous findings on the habitat preferences of Atlantic salmon (Bardonnet and Baglinière 2000; Armstrong et al. 2003) and with interhabitat ratios established for French rivers (Baglinière et al. 2005). We also show that our estimates of the age-0+ population size are well correlated with those of the smolt run of the corresponding cohorts, which is consistent with previous knowledge about the low yearly fluctuations of the age-0+–smolt survival rates in French rivers.

Fish density was extrapolated over 95% of the study area. Because the spatial variability of the density is high, a high price is paid in terms of uncertainty for the very low sampling rate, and our estimates of the population size have poor precision. By the sole use of the available electrofishing data, only very large variations in the population size can be detected, but our HBM framework for the estimation of age-0+ population size can also serve to incorporate information into a population dynamic model. Rivot et al. (2004) illustrated the coupling of capture–mark–recapture observation models with a population dynamics life cycle model for Atlantic salmon, within the Bayesian state–space modelling framework. This flexible methodology can accommodate an additional observation model to assimilate the successive removal data on juveniles into the population dynamics model.
and, therefore, improve the estimation of demographic parameters and the predictions about the future of the population.

Acknowledgements

We are grateful to all the staff of the Unité Expérimentale d’Ecologie et d’Ecotoxicologie Aquatique (INRA Rennes), especially Frédéric Marchand, for their electrofishing field work and their help in extracting and cleaning the data. We also thank the staff of the Conseil Supérieur de la Pêche for their contribution to the electrofishing operations.

References


© 2008 NRC Canada
Appendix A. About the ignorability hypothesis of the missing data generating process

Following Gelman et al. (2004, p. 204), two conditions are sufficient to ensure ignorability: (i) data are missing at random (MAR); and (ii) parameters are distinct (DP). For the MAR condition to hold, the missing data mechanism can depend on the observed covariates and the data but it must be independent from the missing data values. The DP condition is verified when the parameters of the missing data process and the data generating process are a priori independent.

In our case study, we assumed that the missing data process is ignorable. To justify our view, it is useful to come back to the way in which the field operations are conducted and how the missing data process could be modelled accordingly. Having been frequently involved in the electrofishing operations, we contend that the field crew rely primarily on the available observations in hand, i.e., the number of fish caught at the first pass, to decide whether or not a second pass should be undertaken: the probability of missing data at the second pass is higher when the catch at the first pass is low. This may be modulated according to the year (for instance, see 1987 in Table 2), although it is difficult to retrospectively know which were the exact reasons leading the field crew to be less prone to undertake second passes in some years. The habitat type may influence the decision as well, e.g., it is possible that the field crew will be more willing to undertake a second pass in a riffle–rapid, the preferred habitat for juvenile salmon, than in a run. Based on this, a sensible modelling of the missing data process could be inferred.

(A1) \[ I(i,h,k) \sim \text{Bernoulli}(\xi(i,h,k)) \]
\[ \logit(\xi(i,h,k)) = \phi + \gamma \cdot C(i,h,k) + \eta_i + \lambda_h \]

where \(I(i,h,k)\) is the indicator of missing data for the unit \((i,h,k)\) as defined in the paper, \(\xi(i,h,k)\) is the probability of missing data, \(C(i,h,k)\) is capture at the first pass, and \(\phi = (\gamma, \eta_i, \lambda_h)\) is the parameter vector of the missing data process, including year \((\eta_i)\) and habitat type \((\lambda_h)\) effects. In this model, the data “missingness” depends only on the observed data \(C(i,h,k)\), the observed covariates (year and habitat), and the parameters of the missing data model \(\phi\). The data MAR condition is thus verified. There is no reason to hypothesize that the parameters \(\phi\) are not a priori independent from the parameters of the data-generating process, i.e., densities, capture probabilities, habitat and year effects on the densities and capture probabilities, and the associated hyperparameters (grand means and variances). We thus assumed by default that the DP condition holds as well.
By looking at the observed data in Table 2, one may note that the decision to undertake a second pass (i.e., data missingness) shall not be (fully) independent of the density. However, this does not contradict the assumption of ignorability of the missing data mechanism by indicating that either the data MAR or the DP condition would be violated. In our case, it is just the consequence of the dependence of the missing data mechanism on the observed data $C_i$ as acknowledged in eq. A1.

**Appendix B. Sensitivity analysis of model comparison criteria (BF and DIC) to alternative prior configurations**

The influence of the prior on the BF may be high when using weakly informative priors as in the models described in the Material and methods section (Kass and Raftery 1995; Gelman et al. 2004). Marginal likelihoods $P_M$ (DATA) and therefore BFs may also be sensitive to the choice of the system of constraints on the qualitative covariates effects (Fernandez et al. 2002). To assess these sensitivities, we fit and compared the four models $M_0$,...,$M_3$ under several prior configurations denoted $PC_{i,j}$. The indices $i$ and $j$ denote the hyperparameter prior option and the alternative specifications of constraints, respectively, on the fixed effects of the habitat, $\beta(h)$s and $\beta_\delta(h)$s.

The baseline hyperparameter prior configuration, referred as $PC_{0,0}$, is the one described in the Material and methods section. It is based on vague prior distributions on all the hyperparameters of the model. In $PC_{1,0}$, we instead used the informative prior on the hyperparameters related to the catchability that might reflect expert knowledge (Table B1). Priors were set so that the marginal prior on $\beta_\delta$ in each sampled unit is dome-shaped with a mode around 0.75 and small between-site and between-year variability. The priors on the hyperparameters for density remained unchanged compared with $PC_{0,0}$. In $PC_{2,0}$, informative hyperpriors were assigned to the hyperparameters related to catchability (the same as in $PC_{1,0}$) and density. For the later, priors were set to confine the between-year variability of the age-0+ density in a range that is compatible with the between-year variability of the smolt number estimated in the Oir River by Baglinière et al. (2005).

### Table B1. Alternative hyperparameter prior configurations tested for the four competing models $M_0$ to $M_3$.

<table>
<thead>
<tr>
<th>Prior configurations</th>
<th>Hyperparameters related to the probability of capture</th>
<th>Hyperparameters related to the density</th>
</tr>
</thead>
<tbody>
<tr>
<td>$PC_{1,0}$</td>
<td>$\mu_\pi \sim \text{Normal}(1.5, V_{\beta_\pi} = 0.1)$</td>
<td>Baseline $\mu_\pi \sim \text{Normal}(0.5, V_{\beta_\pi} = 0.02)$</td>
</tr>
<tr>
<td></td>
<td>$\sigma_\pi \sim \text{Uniform on } [0, \sigma_{\pi\max} = 1]$</td>
<td>$\sigma_\delta \sim \text{Uniform on } [0, \sigma_{\delta\max} = 2]$</td>
</tr>
<tr>
<td></td>
<td>$\sigma_{\alpha_\pi} \sim \text{Uniform on } [0, \sigma_{\alpha\max} = 1]$</td>
<td>$\sigma_{\alpha_\delta} \sim \text{Uniform on } [0, \sigma_{\alpha\max} = 3]$</td>
</tr>
<tr>
<td></td>
<td>$V_{\beta_\pi} = 0.1$</td>
<td>$V_{\beta_\delta} = 0.02$</td>
</tr>
</tbody>
</table>

Note: The baseline configuration $PC_{0,0}$ is that described in the Materials and methods section.

The three prior configurations ($PC_{i,j}$, $i = 0,1,2$) were crossed with three alternative specifications of the constraints on the fixed effects of the habitat, $\beta(h)$s and $\beta_\delta(h)$s. $PC_{0,0}$ corresponds to the system of constraints described in the Material and methods section. The system of constraints in $PC_{1,0}$ is $\beta_\delta(h = 1) \sim \text{Normal}(0, V_{\beta_\delta})$, $\beta_\delta(h = 2) = 0$ and $\beta_\delta(h = 1) \sim \text{Normal}(0, V_{\beta_\delta})$, $\beta_\delta(h = 2) = 0$. The system of constraints in $PC_{2,0}$ is $\beta_\delta(h = 1) = 0$ and $\beta_\delta(h = 2) \sim \text{Normal}(0, V_{\beta_\delta}((\rightarrow,0))$ and $\beta_\delta(h = 1) = 0$ and $\beta_\delta(h = 2) \sim \text{Normal}(0, V_{\beta_\delta}).$

### Results

The estimations of all unknown quantities are not sensitive to the choice of prior (results not shown). In terms of model comparison, the BF (Fig. B1) and DIC (Table B2) show some sensitivity to the prior configurations, but the main conclusions are unchanged. Indeed, apart from one case (BF with $PC_{1,2}$, where $M_1$ appears as a good candidate although $M_0$ remains preferred), both models $M_2$ and $M_3$ are always clearly discarded by BF. From the BF, the model $M_0$ is always the most credible, given the data for all prior configuration, and in most cases, $M_0$ clearly outcompetes the other models. According to the DIC, none of the four models can be clearly discarded because the differences between DIC values rarely exceed 5.

The BFs of $M_0$ versus all other models computed from the harmonic mean of the likelihood values in eq. 9c show some instability (Fig. B1), which results from the occurrence of a
few very low values of the likelihood along the MCMC sampling that significantly influence the harmonic mean. However, in our case study, this instability is limited and never changes the results of the model comparison whatever the prior configuration used.