Information dynamics and optimal sampling in capture-recapture

BY T. SCHWEDER

Department of Economics, University of Oslo, PO Box 1095, Blindern, 0317 Oslo, Norway
tore.schweder@econ.uio.no

AND D. SADYKOVA

Centre for Ecological and Evolutionary Synthesis, Biological Department, University of Oslo,
PO Box 1066, Blindern, 0316 Oslo, Norway
dinara.sadykova@bio.uio.no

SUMMARY

The build up of information in a continued capture-recapture experiment of simple random sampling of an open population is studied by predicting the conditional approximate Fisher information for abundance in data from one survey given the previous data. By neglecting the stochasticity in survival, a simple approximate likelihood is obtained. Optimal temporal allocation of a given total effort is found by numerical optimization for various objective functions based on the approximate Fisher information. For aerial photographic surveys of bowhead whales, the performance of estimates of abundance and of demographic parameters is compared between constant yearly survey effort and nominally optimal sampling by simulating a realistic model over 50 years.

Some key words: Abundance estimation; Bowhead whale; Fisher information; Optimal temporal sampling; Photo-identification.

1. INTRODUCTION

Long-lived animals such as whales are suitable for sequential capture-recapture studies. Small and well-marked populations are regularly monitored by photographic capture-recapture methods (Larsen & Hammond, 2004; Bradford et al., 2008). Attempts have also been made to monitor the abundance of larger whale populations by aerial photographic surveys (da Silva et al., 2000; Schweder et al., 2010), and by other marking methods, e.g., biopsy sampling and genetic marking (Stevick et al., 2001). Continued capture-recapture studies are also in wide use for other mammals, birds and fish (White, 2008; Cooch & White, 2010).

Monitoring by capture-recapture is nonlinear in the sense that the information in a sample depends on previous data and is generally not proportional to the sampling effort. In a sequential capture-recapture programme, the first surveys will mainly function as investment in an increasing stock of marked individuals while the yield in the form of information will come later.

The pattern of information build-up is investigated in a simple stylized model chosen to allow easy computation of the Fisher information for abundance in a set of new data, conditional on previously obtained data. The population is assumed open with recruitment through births and removal by deaths. Mortality is assumed constant. Captures are assumed to be made by simple random sampling at the beginning of the year. The model is simplified by neglecting the stochasticity in the survival of marked individuals. This simplification is deemed adequate in our context of survey design, but the simple model might be inadequate for inference.
Should survey effort in terms of the number of captures be constant over future years, or should the temporal allocation of effort have another pattern? We study this by optimizing the sum of predicted conditional Fisher information for abundance over \( T \) future surveys for a given total budget. We also consider optimal sampling strategies for some other objectives.

2. Model and approximation

For simplicity we discuss yearly surveys. In the survey in year \( t \), \( n_t \) individuals are randomly sampled and released after identification over a period short enough to disregard mortality and recruitment during the survey. Each capture is linked with the capture history of the individual if it is a between-year recapture, and otherwise starts a new capture history. Before the survey that year the archive contains \( M_{st} \) capture histories of individuals last seen in year \( s = 0, \ldots, t - 1 \). In sample \( t \) let there be \( X_{st} \) recaptures of individuals last seen in year \( s \), and let \( X_t = n_t - \sum_{s=0}^{t-1} X_{st} \) be the number of new captures. With \( M_{st} \) the number of individuals last captured in year \( s \) surviving to year \( t \), the probability that a capture is a recapture of one of the \( M_{st} \) individuals is \( p_{st} = M_{st}/N_t \), conditional on the binomially distributed number of survivors, where \( N_t \) is the population size in year \( t \). The survival probability from year \( s \) to year \( t \), \( q_{st} \), is assumed known and

\[
E(X_{st}) = n_t E(p_{st}) = n_t \frac{m_{st} q_{st}}{N_t} \quad (s = 0, \ldots, t - 1).
\]

The abundance \( N_t \) is the only free parameter. Since sampling is random and survival is independent, \((X_{0t}, \ldots, X_{t-1}, X_t)\) has a multivariate hypergeometric distribution, conditional on the number of survivors. The conditional likelihood given the data from previous surveys is

\[
L = E \left\{ \prod_{s=0}^{t-1} \left( \frac{M_{st}}{X_{st}} \right) \left( \frac{N_t - M_t}{X_t} \right) \left/ \frac{N_t}{n_t} \right. \right\},
\]

where \( M_t = \sum_{s=0}^{t-1} M_{st} \) and expectation is with respect to the binomial survivals. Replacing \( M_{st} \) with its expectation and assuming \( X_t \) to have a binomial rather than a hypergeometric distribution, the approximate conditional loglikelihood

\[
l_t = A + X_t \log \left( \frac{N_t - \sum_{s=0}^{t-1} m_{st} q_{st}}{n_t} \right) - n_t \log(N_t)
\]

is obtained. Here \( A \) is a term involving data and known survival probabilities.

The approximate score function for \( N_t \), conditional on previous data is \( S_t = d l_t / d N_t = X_t / (N_t - m_t) - n_t / N_t \) where \( m_t = \sum_{s=0}^{t-1} m_{st} q_{st} \) is the expected number of marked individuals at the start of survey \( t \). The approximate conditional Fisher information is

\[
I_t = E \left( \frac{d^2 l_t}{d N_t^2} \right) = \frac{n_t}{N_t^2} \frac{m_t}{N_t - m_t}.
\]

With constant yearly survival probability \( q \), \( m_{t+1} = q (m_t + X_t) \). In our planning context, the predicted approximate conditional Fisher information is found from the predicted expected number of marked individuals, \( \hat{m}_{t+1} = q \{ \hat{m}_t + n_t (N_t - \hat{m}_t) / N_t \} \). With \( \hat{m}_0 = 0 \),

\[
\hat{m}_t = \sum_{j=0}^{t-1} n_j q^{t-j} \prod_{s=j+1}^{t-1} \left( 1 - \frac{n_s}{N_s} \right).
\]

For a stable population, \( N_t = N \), and a fixed sample size, \( n_t = n \), we have

\[
\hat{m}_t = n q \frac{1 - q'(1 - n/N)^t}{1 - q (1 - n/N)},
\]

where \( q' \) is the fitted survival probability. The approximate conditional score function is

\[
\hat{s}_t = E \left( \frac{d l_t}{d N_t} \right) = n_t \frac{n_t}{N_t} \frac{m_t}{N_t - m_t}.
\]
which increases in \( t \). The predicted information for \( N \) contained in the data from survey \( t \) thus in this case increases monotonically to its limit \( I_\infty = n^2/N^3q/(1-q) \).

For other schedules of sample size or population size, the predicted information might peak or show other patterns.

For a closed stable population, \( q = 1 \), the expected number of individuals never captured in the first \( t \) surveys is \( N - m_t = N(1 - n/N)^t \). Consequently \( I_t = n\{(1 - n/N)^{-t} - 1\}/N^2 \), which agrees with Fewster & Jupp (2009, Theorem 1). The expected Fisher information in sample \( t \) is seen to grow exponentially in this case, as opposed to the open case where it levels off, even when mortality is low; see Fig. 1(a).

Simulations show that the approximation to the conditional loglikelihood is good. The true likelihood, with stochastic survival, is in general slightly more concentrated than the approximate likelihood. The two likelihoods peak at about the same abundance. The true Fisher information is around 10% above \( I_t \) for small values of \( t \), 3–5% above for \( t = 10 \) and slightly less for larger values of \( t \). The conditional likelihoods are positively correlated through the correlation in the latent survival process. The loglikelihood function based on the data from all the surveys is markedly different from the sum of the approximate conditional loglikelihoods.

The survival probability has been assumed known, which might be appropriate in the context of planning with abundance as the key parameter. The simple model might however also be of help in the inference context, at least in finding starting values when estimating abundance and other parameters by numerical optimization.

### 3. Optimum Sampling

Consider a population to be monitored by a programme of capture-recapture surveys. Subject to a budget constraint of total survey cost not exceeding \( C_{\text{tot}} \), the problem is to find an optimal temporal allocation, i.e., to determine \( n_1, n_2, \ldots \) to optimize information on abundance in a specific way.

Let the cost of survey \( t \) be \( c_t = c + n_t \), where \( c \) is the initial cost of setting up a particular survey, measured in the unit cost of making a capture, which is assumed constant. The budget constraint is \( \sum_{t=1}^T (cS_t + n_t) \leq C_{\text{tot}} \), where \( T \) is the number of possible survey occasions, and \( S_t \) is an indicator for a survey being carried out on occasion \( t \).

Consider first the objective function

\[
I_{\text{sum}}(n_1, \ldots, n_T) = \sum_{t=1}^T I_t, \tag{3}
\]
where $I_t$ is the predicted version of (1) for an assumed abundance trajectory $\{N_t\}$. Due to the dependence in the conditional loglikelihoods caused by the correlations in the latent survival process, $I_{\text{sum}}$ is not a good approximation to the total information on abundance over the $T$ surveys. The total information is hard to calculate, and by maximizing $I_{\text{sum}}$ the average new information in the abundance estimates over the surveys is optimized.

Assuming the above approximate model, with expected numbers given by (2) for $m_i$ in (1), the problem is one of dynamic programming. We will however use direct optimization, which has become easy with modern computing software such as ADMB (ADMB Foundation, 2010).

Our numerical results suggest that in stable and in growing populations it is optimal to spend much of the total effort in the early surveys, and have effort rapidly decreasing, possibly after a peak. The pattern depends on $m_0$, the expected number of marked individuals before the survey in year $t = 0$. When this number is larger than some level, the whole optimal effort trajectory is monotonic, while it peaks for $m_0$ smaller than that level.

We also studied how effort should be allocated when an abundance estimate based on a fresh sample is required every $r$th year. Consider first the problem of allocating the total effort $c_\text{tot}$ over the first $r$ surveys. The objective function is $I_r$, which depends on $n_1, \ldots, n_{r-1}$ through $\hat{m}_r$. Due to mortality exposure $\hat{m}_r(n_1, \ldots, n_{r-1}) \leq \hat{m}_r(0, \ldots, 0, n_1 + \cdots + n_{r-1})$. Saving initial costs of setting up surveys in the first $r - 2$ years will also allow larger samples in years $r - 1$ and $r$. There should thus be no surveys in years $1, \ldots, r - 2$. The optimum is most easily found by calculating $I_r$ for all possible values of $n_{r-1}$ and $n_r$.

For long-term monitoring the process might be assumed stationary. If the strategy also is stationary, i.e., periodic with $I_{jr+i} = n_i$ ($i = 1, \ldots, r$), abundance can be assumed constant and $\hat{m}_i$ periodic. Under the constraint of at most $c_\text{tot}$ in each period, the periodic strategy that is optimal in the first period will then also maximize $I_{jr+r}$ for every period $j = 0, \ldots, r$.

When mortality varies from year to year and must be estimated, this periodic strategy with no effort in mid period is poor. If however mortality is constant but unknown, the strategy should perform well. When the primary concern is also for other parameters, such as carrying capacity or survival probabilities, the simple one-parametric model we have considered is inadequate. In addition to broadening the model, an optimization criterion then needs to be specified in order to approach the problem of optimal sampling.

4. Bowhead whales off Alaska

Bowhead whales are subject to airborne photographic surveys when they migrate eastwards past Barrow in the spring (Rugh, 1990; Angliss et al., 1995; Koski et al., 2006; Rugh et al., 2008). The degree of natural marking varies considerably among the whales. Schweder et al. (2010) found that only 1-3% of the captures made before 2004 had natural marks that make the probability of successful matching more than 0.5.

Due to variability in natural marking, identification is difficult, and to prevent false positive matches the protocol for declaring a capture a recapture of a previously captured individual is stringent, at a cost of probably getting many false negatives. Captures are assumed to be made by random sampling. The surveys are simulated as in Schweder et al. (2010), and abundance and other parameters are estimated from the simulated data by maximizing a simplified likelihood as in Schweder et al. (2010), but without employing their regression approach and parametric bootstrapping to handle the uncertainties in matching.

Our aim is to compare estimation performance between the optimal strategy and a strategy with the same total effort spread evenly over 50 years.

The optimal strategy was found by scaling down the population by a factor of 1-3% to a virtual population of perfectly identifiable individuals. By maximizing (3) with $I_t$ given by (1) for the virtual population an optimal strategy for surveying these individuals is found. The fixed cost of setting up a survey is assumed to be $c = 0.156$ in cost units of one virtual capture. The optimal strategy for virtual individuals was then scaled up to what is taken as the optimal strategy. Most effort is used early on in the optimal strategy, see Fig. 1(b).

The total population is assumed to follow the deterministic model of Pella & Tomlinson (1969)

$$N_{t+1} = N_t\{1 + r(1 - (N_t/K)^\gamma)\} - C_t$$

(4)
Table 1. Mean nominal coefficients of variation over 200 replicated simulations, for abundance and parameters of population dynamics (4) and mortality for bowhead whales

<table>
<thead>
<tr>
<th>Year</th>
<th>N_t</th>
<th>r</th>
<th>µ</th>
<th>K</th>
<th>z</th>
<th>N_t</th>
<th>r</th>
<th>µ</th>
<th>K</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>0.19</td>
<td>0.35</td>
<td>0.27</td>
<td>—</td>
<td>—</td>
<td>0.19</td>
<td>0.35</td>
<td>0.27</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>2010</td>
<td>0.06</td>
<td>0.15</td>
<td>0.12</td>
<td>—</td>
<td>—</td>
<td>0.15</td>
<td>0.30</td>
<td>0.24</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>2015</td>
<td>0.04</td>
<td>0.08</td>
<td>0.07</td>
<td>—</td>
<td>—</td>
<td>0.08</td>
<td>0.19</td>
<td>0.17</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>2020</td>
<td>0.03</td>
<td>0.07</td>
<td>0.06</td>
<td>0.06</td>
<td>0.05</td>
<td>0.06</td>
<td>0.14</td>
<td>0.13</td>
<td>0.13</td>
<td>0.11</td>
</tr>
<tr>
<td>2030</td>
<td>0.03</td>
<td>0.06</td>
<td>0.05</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
<td>0.10</td>
<td>0.09</td>
<td>0.08</td>
<td>0.07</td>
</tr>
<tr>
<td>2040</td>
<td>0.02</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
<td>0.08</td>
<td>0.07</td>
<td>0.07</td>
<td>0.06</td>
</tr>
<tr>
<td>2060</td>
<td>0.02</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
</tr>
</tbody>
</table>

The coefficients of variation for 2009 are based on historic data only. Simulations start in 2010. The first five columns are for the optimal strategy, while the rest are for constant yearly effort.

over the years, with \( C_t \) being the aboriginal harvest in year \( t \). Following Brandon & Wade (2006), Zeh et al. (2002) and Zeh & Punt (2005), parameter values were taken to be \( r = 0.054 \), \( K = 1400 \), \( z = 4.86 \) and mortality \( \mu = -\log q = 0.01 \) for the whole population.

The 10 historic surveys previous to 2004, with 4936 captures of which 42 are recognized as recaptures, were used together with simulated data to obtain yearly maximum likelihood estimates and standard errors from the Hessian of the loglikelihood for abundance and each of the other 12 model parameters.

Mean nominal coefficients of variation are given in Table 1 both for the optimal strategy and the constant effort strategy. The mean is over 200 replicate simulations for both strategies. The nominal coefficient of variation is the standard error from the Hessian matrix divided by the assumed value of the parameter. Due to the higher effort in the early years for the optimal strategy, parameters are estimated much more precisely in those years, while at the end the precision is nearly the same for both strategies.

That performance in 2060 mainly depends only on the total effort must be due to the population dynamics being deterministic with fixed parameters and the mortality being low. The optimal strategy might thus not be acceptable for long-term monitoring, since survey effort would then be zero from 2038. A stationary strategy with estimates based on surveys, say every fifth year, might be more attractive since the purpose of monitoring would be to learn about changes in the population dynamics or surprising changes in abundance. The performance under a periodic strategy is similar to that under constant yearly effort when the total effort is the same in the long run.

Acknowledgement

We thank an anonymous referee for helpful comments.

References


T. Schweder and D. Sadykova


[Received March 2011. Revised December 2011]