Web-based Supplementary Materials for Maximum likelihood estimation for model $M_{t,\alpha}$ for capture-recapture data with misidentication,

by R. T. R. Vale, R. M. Fewster, E. L. Carroll, N. J. Patenaude

Web Appendix A: Likelihood computation

The method of computation described below is based on reformulating sums as matrix products, and organising the computation so that it only involves a sequence of vectors multiplied by matrices. This is efficient in minimizing the size of objects that need to be created, but involves some redundant computation of quantities that cancel from one product to the next. Although further work might uncover more efficient computation methods, the procedure below gives fast computation within the likely sphere of application of the model.

Let ℓ be the log-likelihood gained from equation (2) in the main text, ignoring terms involving the data only. Write $r_\bullet = r_1 + \ldots + r_T$ for any selection of $(r_1, \ldots, r_T)$. We have:

$$\ell = C \log(\alpha) + \sum_{t=1}^{T} \{n_t \log(p_t) + (N - n_t) \log(1 - p_t)\} + \log(N!) + \log(S), \tag{A.1}$$

where

$$S = \sum_{r_1=0}^{u_1} \ldots \sum_{r_T=0}^{u_T} \mathcal{I}\{r_\bullet \leqslant N - D\} \; \frac{\alpha^{r_\bullet} (1-\alpha)^{U - r_\bullet}}{(N - D - r_\bullet)!} \; \prod_{t=1}^{T} \frac{1}{r_t!} \binom{N - n_t + u_t - r_t}{u_t - r_t}. \tag{A.2}$$

Define index variables $v_t = r_1 + \ldots + r_t$, and $y_t = u_1 + \ldots + u_t$, for $t = 1, \ldots, T$, and define constants $v_0 = y_0 = 0$. Note that $v_T = r_\bullet$ and $y_T = U$. Also define $\binom{k}{l} = 0$ if $k < 0$, $l < 0$, or $k < l$ to enforce the conditions that $r_\bullet \leqslant N - D$ and $v_{t-1} \leqslant v_t \leqslant v_{t-1} + u_t$ for all t. We reindex the sums in (A.2) using $\{v_t\}$ and $\{y_t\}$, to give after some algebra:

$$S = \frac{1}{(N-D)!} \sum_{v_T=0}^{y_T} \cdots \sum_{v_1=0}^{y_1} \alpha^{v_T}(1-\alpha)^{U-v_T} \prod_{t=1}^{T} \binom{N - n_t + u_t - (v_t - v_{t-1})}{u_t - (v_t - v_{t-1})} \binom{N-D-v_{t-1}}{v_t - v_{t-1}} \quad \text{(A.3)}$$

Define the $1 \times (y_1 + 1)$ row vector B_1 with entry v_1 given by $B_1(v_1) = \binom{N-n_1+u_1-v_1}{u_1-v_1}\binom{N-D}{v_1}$ for $v_1 = 0, \ldots, y_1$. For $t = 2, \ldots, T$, define the $(y_{t-1} + 1) \times (y_t + 1)$ matrix B_t with entries (v_{t-1}, v_t) for $v_{t-1} = 0, \ldots, y_{t-1}$ and $v_t = 0, \ldots, y_t$ given by

$$B_t(v_{t-1}, v_t) = \binom{N - n_t + u_t - (v_t - v_{t-1})}{u_t - (v_t - v_{t-1})} \binom{N-D-v_{t-1}}{v_t - v_{t-1}}.$$

For time T, further define the $(y_{T-1} + 1) \times (y_T + 1)$ matrix B_T^*, with entries (v_{T-1}, v_T) for $v_{T-1} = 0, \ldots, y_{T-1}$ and $v_T = 0, \ldots, y_T$ given by

$$B_T^*(v_{T-1}, v_T) = \alpha^{v_T}(1-\alpha)^{U-v_T} B_T(v_{T-1}, v_T).$$

Then (A.3) reduces to:

$$S = \frac{1}{(N-D)!} \sum_{v_T=0}^{y_T} \left(B_1 B_2 \ldots B_{T-1} B_T^* \right)_{v_T}. \quad \text{(A.4)}$$

Because B_1 is a vector, every product of the form $(B_1 \ldots B_t) \times B_{t+1}$ is the product of a vector with a matrix. We evaluate the product in (A.4) sequentially so that all computations are of this form.

Matrix entries are computed as logarithms to avoid numerical problems. Finding the logarithm of the scalar product of two vectors $\boldsymbol{b}$ and $\boldsymbol{c}$, which are each stored as their logarithms, is a computation of the form $\log\left(\sum_i b_i c_i\right) = \log\left\{\sum_i \exp\left(\log b_i\right) \exp\left(\log c_i\right)\right\}$. These computations are performed as $m + \log\left\{\sum_i \exp\left(\log b_i + \log c_i - m\right)\right\}$, where $m = \max_i\left(\log b_i + \log c_i\right)$, otherwise large exponentials can cause numerical errors.

Inserting (A.4) into (A.1) completes the computation.

Web Appendix B: photographic survey of New Zealand southern right whales

We use photo-identification data from 54 survey days in 1998, described in Fewster and Patenaude (2009) and Carroll et al. (2011). Capture occasions correspond to days when weather conditions were suitable for sampling. There were $T = 9$ successful capture occasions in the 18 days between 18th July and 4th August 1998, and this period corresponds to a central part of the season during which numbers appeared reasonably stable. Regular counts were conducted in addition to capture-recapture sampling, aimed at counting all adults present; these can be interpreted as minimum numbers present. The counts for 18th July, 26th July, and 4th August were respectively 112, 117, and 109. The high level of agreement between these counts supports the assumption of a closed population, but complete closure cannot be guaranteed over the period.

The error rate for matching photographs of the same animal has previously been estimated by counting discrepancies between different experts reviewing the same photo catalogue. On this basis the error rate was found to be less than 3% (Fewster and Patenaude, 2009), corresponding to a value of α of 0.97 or above. The true α for this study could be either higher or lower than this value: the estimate of 0.97 is based on between-year photographs which may be harder to match than within-year photographs; but on the other hand the agreement of different experts does not guarantee correctness as all experts may be misled by the same features of a photograph.

The data consist of 93 capture histories, with 64 unit histories, and respectively 24, 3, and 2 histories with 2, 3, and 4 entries. Applying model $M_{t,\alpha}$ yields $\widehat{N} = 144$ with 95% confidence interval $(49, 419)$; $\widehat{\alpha} = 0.94$ with standard error 0.24; and estimates of capture probability $\widehat{p}_1, \ldots, \widehat{p}_9$ from 0.05 to 0.18 with mean 0.10. Standard errors of $\widehat{p}_t$ ranged from 0.03 to 0.11 with mean 0.06. The estimated CV of $\widehat{N}$ is 58.9%. The results are identical if we optimize only over (N, α).

Applying model M_t, we gain $\widehat{N} = 166$ with 95% confidence interval (131, 211); and estimates of $p_1, \ldots, p_9$ from 0.04 to 0.16 with mean 0.086. Standard errors of $\widehat{p}_t$ ranged from 0.02 to 0.03 with mean 0.02. The estimated CV of $\widehat{N}$ is 12.2%.

According to our simulation study, for capture probabilities in this range with $T = 9$ we should expect maximum likelihood estimates for $\widehat{N}$ to have low precision, reduced CI coverage, and some negative bias. We conducted further simulations assuming that the true values of $(N, \alpha, p_1, \ldots, p_9)$ are given by their estimates for these data. These show that the MLEs for $\widehat{N}$ exhibit -11% bias, 87% CI coverage, 31% RMSE, and 29% empirical CV with model $M_{t,\alpha}$. The estimated CV from the real data lies within the distribution of estimated CVs from the simulated results, at the 85th percentile. For comparison, we also fitted model M_t to the same simulated data, to mimic Scenario B in Section 3 of the main text. Under model M_t, ignoring errors in photo-matching, the MLEs for $\widehat{N}$ have 15% bias, 86% CI coverage, 21% RMSE and 14% empirical CV.

Although the $M_{t,\alpha}$ estimates are plausible based on our prior knowledge of the population and photo-matching errors, they suffer from extremely high variance. The confidence interval of (49, 419) for N does not provide a useful conclusion for management purposes. Despite model $M_{t,\alpha}$ being the correct model for the simulations, the high variance introduced by the parameter α means that it yields a substantially higher average error in $\widehat{N}$ than that obtained by ignoring photo-matching errors and fitting model M_t (RMSE 31% for $M_{t,\alpha}$ versus 21% for M_t), and there is no significant improvement in confidence interval coverage (86.7% for $M_{t,\alpha}$ versus 86.2% for M_t out of 500 simulations).

We note that the poor results from model $M_{t,\alpha}$ using maximum likelihood estimates for these data do not necessarily apply to other modes of inference. In particular, a Bayesian approach using informative priors for α may resolve the problems caused by weak identifiability of α. Informative priors may be based on discrepancies counted when multiple

experts review the same photo catalogue. Additionally, other estimation methods such as least-squares may perform better than maximum likelihood estimation for small samples arising from situations like this. Variance can be estimated by bootstrap methods. For future statistical development, other alternatives are to direct efforts towards developing photo-matching software (e.g. *BigFish:* Pirzl, Murdoch, and Lawton, 2006), or to devise a model in the same vein as the genetic model of Wright et al. (2009) that can incorporate direct information on matching discrepancies from independent experts.

Web References

Fewster, R. M. and Patenaude, N. J. (2009). Cubic splines for estimating the distribution of residence time using individual resightings data. In *Modeling Demographic Processes in Marked Populations,* D. L. Thomson, E. G. Cooch, and M. J. Conroy (eds), pp. 393–415. Berlin: Springer.

Pirzl, R., Murdoch, G., and Lawton, K. (2006). *BigFish: computer assisted matching software and data management system for photo-identification.* Skadia Pty Ltd. www.skadia.com.au.