Maximum likelihood estimation for model $M_{t,\alpha}$ for capture-recapture data with misidentification

R. T. R. Vale[1], R. M. Fewster[2,*], E. L. Carroll[3], and N. J. Patenaude[4]

[1]IRD, Asteron Centre, 55 Featherston St, Wellington, New Zealand

[2]Department of Statistics, and [3]School of Biological Sciences,

The University of Auckland, Private Bag 92019, Auckland, New Zealand

[4]Collégial International Sainte-Anne, 1300 Boulevard Saint-Joseph, Montréal, Quebec, Canada H8S 2M8

**email:* r.fewster@auckland.ac.nz

SUMMARY: We investigate model $M_{t,\alpha}$ for abundance estimation in closed-population capture-recapture studies, where animals are identified from natural marks such as DNA profiles or photographs of distinctive individual features. Model $M_{t,\alpha}$ extends the classical model M_t to accommodate errors in identification, by specifying that each sample identification is correct with probability α and false with probability $1 - \alpha$. Information about misidentification is gained from a surplus of capture histories with only one entry, which arise from false identifications. We derive an exact closed-form expression for the likelihood for model $M_{t,\alpha}$ and show that it can be computed efficiently, in contrast to previous studies which have held the likelihood to be computationally intractable. Our fast computation enables us to conduct a thorough investigation of the statistical properties of the maximum likelihood estimates. We find that the indirect approach to error estimation places high demands on data richness, and good statistical properties in terms of precision and bias require high capture probabilities or many capture occasions. When these requirements are not met, abundance is estimated with very low precision and negative bias, and at the extreme better properties can be obtained by the naive approach of ignoring misidentification error. We recommend that model $M_{t,\alpha}$ be used with caution and other strategies for handling misidentification error be considered. We illustrate our study with genetic and photographic surveys of the New Zealand population of southern right whale (*Eubalaena australis*).

KEY WORDS: Genetic capture-recapture; Latent multinomial; Mark recapture; Misidentification; Natural tags; Photo-identification.

1. Introduction

Capture-recapture studies for estimating the size of animal populations increasingly rely upon natural individual marks instead of physical tags that require the animal to be caught and handled. Identification by natural marks is usually based on DNA samples or photographs. Photographs may focus on skin or coat patterns, scars, callosity patterns, or fin shapes, and are widely used for surveys of large cats and cetaceans (Karanth et al., 2006; Carroll et al., 2011). Genetic sampling may take place by biopsy darts that glance off the thick skin of a cetacean (Carroll et al., 2011) or by DNA extracted from hair snags or faeces (Wright et al., 2009). Individuals are typically identified from DNA samples using a set of 5–20 microsatellite loci (Pompanon et al., 2005).

Identification using natural marks is inherently error-prone. Photographs are affected by visibility and angle (Morrison et al., 2011). For genetic samples, allelic dropout and other errors are common (Wright et al., 2009): for example, the single-locus genotype AB may be misreported as AA, because allele B ‘dropped out’. The frequency of genetic errors depends upon DNA quality, with errors more common in samples collected from hair or faeces than from high-quality tissue samples obtained from biopsy darts (Pompanon et al., 2005).

The most likely identification error is failure to link two captures of the same animal, so they are wrongly recorded as captures of two different animals. Even a small level of error in identifying recaptures can lead to marked overestimation of abundance (Lukacs and Burnham, 2005; Yoshizaki et al., 2011). For this reason, there has been considerable recent interest in incorporating identification errors into capture-recapture models (Lukacs and Burnham, 2005; Wright et al., 2009; Link et al., 2010; Yoshizaki et al., 2011). Much of the recent work has focused on variants of model $M_{t,\alpha}$ for estimating the size of a closed population.

Model $M_{t,\alpha}$ is based upon the classical model M_t (Otis et al., 1978) which specifies that each of N animals in a population has probability p_t of being detected on capture occasion t ($t = 1, 2, \dots, T$). Maximum likelihood estimation is straightforward for the parameters N and $p_1, \dots, p_T$. Model $M_{t,\alpha}$ extends this to incorporate identification error by specifying that each sample is correctly identified with probability α. A misidentified sample is assumed to create a 'ghost' capture history with precisely one sighting. Information to estimate the parameter α comes from the surplus of capture histories with only one sighting.

Model $M_{t,\alpha}$ was first introduced in a slightly different formulation by Lukacs and Burnham (2005), who provided a likelihood-based approach. Yoshizaki et al. (2011) pointed out errors in their formulation, showing that misidentifications on first capture were not correctly handled and there was no allowance for the dependence between multiple histories generated by a single animal and its 'ghosts'. These problems might be responsible for a breakdown in confidence interval performance reported by Lukacs and Burnham (2005).

Link et al. (2010) and Yoshizaki et al. (2011) commented that the exact likelihood appeared incomputable given the impossibility of summing over all ways of assigning capture histories to animals, where each animal can be responsible for multiple ghost histories. Yoshizaki et al. (2011) provided an approach based on least-squares, but did not include a variance estimator. Link et al. (2010) showed that the observed data may be expressed as a transformation of a multinomial random variable, and coined the name $M_{t,\alpha}$ for the model. They gave an elegant formulation in which the true, unobservable, capture histories are expressed with error code '2': for example, for $T = 4$ capture occasions the capture history 0102 denotes an animal captured on the second and fourth occasions but misidentified on the fourth. This unobservable true history spawns two unlinked observations: 0100 and 0001.

Using a Bayesian approach, Link et al. (2010) treated counts of the unobservable true histories as latent variables which could be sampled with an MCMC sampler rather than enumerated. This formulation is particularly appealing in its generality. Subsequent authors have adapted the approach to solve other capture-recapture problems involving latent identity information (Bonner and Holmberg, 2013; Higgs et al., 2013; McClintock et al., 2013).

Wright et al. (2009) developed a different approach to genetic misidentification, specifically for allelic dropout. Instead of estimating error rate indirectly through a surplus of single-entry capture histories, they require all samples to be genotyped at least twice, gaining a direct estimate of error rate by discrepancies between repeat attempts. This approach is suitable for low-quality DNA in hair or faeces, where repeat genotyping is standard protocol (Taberlet and Luikart, 1999). However, it might not be cost-effective to repeat all genotypes when using high-quality DNA obtained from tissue. An analogous approach based on repetition is not available for photographic surveys, although it might be possible to devise a scheme with direct error information based upon photo-matchings made by multiple experts.

In this paper we focus on the issues facing researchers dealing with a small level of error in genetic or photographic surveys. We formulate an exact closed-form likelihood for model $M_{t,\alpha}$, and show how to compute it efficiently. Computation time is typically reduced to a few seconds on a customary laptop. This fast fitting time enables us to explore model $M_{t,\alpha}$ by extensive simulation studies, which are not possible with the more time-consuming Bayesian approach. We investigate the finite-sample performance of the maximum likelihood estimates, and investigate cost-effective strategies for researchers dealing with misidentification. We apply the model to genetic and photographic surveys of wintering southern right whales (*Eubalaena australis*) in the subantarctic Auckland Islands, New Zealand.

2. Model $M_{t,\alpha}$

We describe model $M_{t,\alpha}$ following Link et al. (2010). A closed population of N animals is observed at times $t = 1, 2, \ldots, T$. Associated with each time t is a capture probability p_t. At time t, there are three possible outcomes for each animal:

- Not captured: probability $1 - p_t$. The outcome of non-capture is given code 0.

- Captured and correctly identified: probability αp_t. We call this a *sound capture* (code 1).
- Captured and misidentified: probability $(1-\alpha)p_t$. We call this a *faulty capture* (code 2).

An identification may be defined as 'correct' or 'sound' if an accurate genotype is obtained, or if a photograph adequately captures the individual's distinguishing features. In practice it will not be known whether the identification is sound or not. The outcome at time t is independent of outcomes at other times and for other animals.

The outcomes for times $t = 1, \dots, T$ form the true, unobservable capture history for each animal. We describe these as *latent histories* and define the set of possible latent histories as $\{\lambda_1, \dots, \lambda_J\}$, where $J = 3^T$. For example, if $T = 5$, the latent history $\lambda_j = 10122$ denotes an animal with sound captures at times 1 and 3, faulty captures at times 4 and 5, and non-capture at time 2. We define λ_{jt} to be the entry of history λ_j corresponding to time t. The latent variable underlying the observed data is the vector of frequencies, $\boldsymbol{x} = (x_1, \dots, x_J)$, where x_j is the number of the N animals whose true capture history is λ_j, and $\boldsymbol{x}$ is multinomial with index N.

The latent histories are not observed. Instead, model $M_{t,\alpha}$ makes the assumption that each faulty capture generates a unique error, so it spawns an observed history with exactly one entry. Thus, an animal with latent history $\lambda = 10122$ generates three observed histories: 10100, 00010 and 00001. This assumption, discussed by Lukacs and Burnham (2005), may be questionable in some circumstances, but it provides a reasonable platform for us to investigate the statistical properties of the indirect approach to identity-error detection.

We call any observed history with exactly one entry a *unit history.* A unit history whose single entry derives from a faulty capture (a 2 in the latent history) is a *ghost history.* An observed history with more than one entry is termed a *duplicate history.* Excluding the unobservable zero history, there are $K = 2^T - 1$ possible observed histories, which we write as as $\{\omega_1, \dots, \omega_K\}$. The observed data form the vector of frequencies, $\boldsymbol{f} = (f_1, \dots, f_K)$, where f_k is the number of times history ω_k is observed. Note that the f_k do not refer to numbers of animals, but to numbers of observed histories, because each animal may be responsible for multiple histories. Link et al. (2010) showed that $\boldsymbol{f}$ is a linear transformation of the multinomial random variable $\boldsymbol{x}$: that is, $\boldsymbol{f} = A^T\boldsymbol{x}$ for a known matrix A.

Throughout, we will index latent histories by $j = 1, \ldots, J$, observed histories by $k = 1, \ldots, K$, and capture occasions by $t = 1, \ldots, T$. The aim is to estimate N, given the data $\boldsymbol{f}$.

The likelihood formulated by Link et al. (2010) sums over the multinomial probabilities of all $\boldsymbol{x}$ consistent with the observed data $\boldsymbol{f}$ and parameter N:

$$\mathcal{L}(N, p_1, \ldots, p_T, \alpha \; ; \; \boldsymbol{f}) = \sum_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{f}, N)} \frac{N!}{x_1! \ldots x_J!} \, \pi_1^{x_1} \ldots \pi_J^{x_J} \, , \tag{1}$$

where $\mathcal{X}(\boldsymbol{f}, N)$ is the feasible set of all $\boldsymbol{x}$ vectors such that $\sum_{j=1}^{J} x_j = N$ and $A^T\boldsymbol{x} = \boldsymbol{f}$, and π_j is the probability of latent history λ_j for $j = 1, \ldots, J$:

$$\pi_j = \prod_{t=1}^{T} p_t^{\mathcal{I}\{\lambda_{jt}>0\}} \, (1-p_t)^{\mathcal{I}\{\lambda_{jt}=0\}} \, \alpha^{\mathcal{I}\{\lambda_{jt}=1\}} \, (1-\alpha)^{\mathcal{I}\{\lambda_{jt}=2\}} \, ,$$

where $\mathcal{I}\{\cdot\}$ is the usual indicator. Link et al. (2010) and Yoshizaki et al. (2011) stated that computation of the likelihood would be difficult, because enumerating the feasible set $\mathcal{X}(\boldsymbol{f}, N)$ would involve a high-dimensional search even for the smallest values of T.

2.1 *New likelihood formulation for model $M_{t,\alpha}$*

The formulation (1) sums over all latent frequency vectors $\boldsymbol{x}$ that can create the observed data. Here we show that the likelihood can be reformulated by summing over the number of sound captures among the unit histories. This reformulation enables us to compute the likelihood efficiently.

First we note that the only uncertain captures in the data reside in unit histories: the unit histories are an unknown composition of animals with only one sound capture, and ghosts resulting from faulty captures. For observed history ω_k, let $|\omega_k| = \sum_{t=1}^{T} \omega_{kt}$ be the number of 1-entries in ω_k: for example if $\omega_k = 10100$ then $|\omega_k| = 2$. Define the following quantities:

- $n_1, \ldots, n_T$ are the numbers of captures at times $1, \ldots, T$; thus $n_t = \sum_{k\,:\,\omega_{kt}=1} f_k$.

- U is the number of unit histories in the observed data; $U = \sum_{k\,:\,|\omega_k|=1} f_k$.
- $u_1, \ldots, u_T$ are the numbers of the U unit histories whose single entry occurs at times $1, \ldots, T$ respectively. We have $\sum_{t=1}^{T} u_t = U$.
- D is the number of duplicate histories in the observed data: $D = \sum_{k\,:\,|\omega_k|\geqslant 2} f_k$. All captures in these histories are sound, by definition. However, animals contributing to these histories may also contribute ghost histories.
- $d_1, \ldots, d_T$ are the numbers of captures within the D duplicate histories that occur at times $1, \ldots, T$; so $d_t = \sum_{k\,:\,|\omega_k|\geqslant 2} f_k\, \omega_{kt}$.
- C is the total number of captures in the D duplicate histories: $C = \sum_{t=1}^{T} d_t$, or alternatively $C = \sum_{t=1}^{T} n_t - U$. These C captures are all known to be sound, by assumption.

We calculate the probability of the data by partitioning over choices for the number of sound captures among the u_t uncertain captures at time t. Define the latent variable $\boldsymbol{r} = (r_1, \ldots, r_T)$, such that r_t is the number of the u_t uncertain captures that are sound. For any $\boldsymbol{r}$, let $r_{\bullet} = \sum_{t=1}^{T} r_t$ be the total number of animals with exactly one sound capture, noting that any of these animals may have additional faulty captures at other times. Let $\mathcal{R}(\boldsymbol{f}, N)$ be the feasible set of all $\boldsymbol{r}$ compatible with the data $\boldsymbol{f}$ and parameter N: then $\mathcal{R}(\boldsymbol{f}, N) = \{\boldsymbol{r} \,:\, r_{\bullet} \leqslant N - D \text{ and } r_t \in \{0, \ldots, u_t\} \text{ for } t = 1, \ldots, T\}$.

Direct combinatorial calculation gives the following likelihood, which is explained in detail in Figure 1. For $N \geqslant \max\{D, n_1, \ldots, n_T\}$, and $(\alpha, p_1, \ldots, p_T) \in (0,1)^{T+1}$,

$$\begin{aligned}\mathcal{L}(N, p_1, \ldots, p_T, \alpha\,;\, \boldsymbol{f}) \quad &= \quad \left\{ \alpha^C \prod_{t=1}^{T} p_t^{n_t} (1 - p_t)^{N - n_t} \right\} \quad \times \\ &\left\{ \sum_{\boldsymbol{r} \in \mathcal{R}(\boldsymbol{f}, N)} \frac{N!\, \alpha^{r_{\cdot}} (1-\alpha)^{U - r_{\cdot}}}{\left(\prod_{k\,:\,|\omega_k|\geqslant 2} f_k!\right) r_1! \ldots r_T!\, (N - D - r_{\bullet})!} \prod_{t=1}^{T} \binom{N - d_t - r_t}{u_t - r_t} \right\}. \qquad (2)\end{aligned}$$

The limit of (2) as $\alpha \uparrow 1$ is the likelihood for model M_t. The first braced factor of (2) arises because there are n_t captures and $N - n_t$ non-captures at time t, regardless of the number of faulty captures. Further, the C captures in the duplicate histories are known to be sound.

The second braced factor of (2) sums over all $\boldsymbol{r}$ in the feasible set $\mathcal{R}(\boldsymbol{f}, N)$. For a particular choice of $\boldsymbol{r}$, the N animals are partitioned into D with two or more sound captures; $r_{\bullet}$ with exactly one sound capture; and $N - D - r_{\bullet}$ with no sound captures, these being a mix of animals with no captures and animals with only faulty captures. This accounts for the multinomial coefficient. The term $\alpha^{r_{\cdot}} (1-\alpha)^{U-r_{\cdot}}$ arises from $r_{\bullet}$ sound and $U - r_{\bullet}$ faulty captures for this $\boldsymbol{r}$. It remains to allot the $u_t - r_t$ faulty captures among all available animals at each time t. The number of animals available for faulty capture at time t is $N - d_t - r_t$ (see Figure 1), accounting for the $\binom{N-d_t-r_t}{u_t-r_t}$ ways of partitioning these among faulty captures and non-captures. Note that $d_t + u_t = n_t$, so an alternative way of expressing the final product of binomial coefficients in (2) is $\prod_{t=1}^{T} \binom{N-n_t+u_t-r_t}{u_t-r_t}$, which makes explicit the apportioning of animals unaccounted for at time t between $N - n_t$ non-captures and $u_t - r_t$ faulty captures.

[Figure 1 about here.]

Parameters are estimated by maximizing the log-likelihood gained from (2). It can readily be shown that the MLEs satisfy $\widehat{p}_t = n_t/\widehat{N}$ for $t = 1, \ldots, T$. This suggests that we might be able to substitute $p_t = n_t/N$ in (2) and optimize only over (N, α) instead of over $(N, p_1, \ldots, p_T, \alpha)$. Optimizing over (N, α) is considerably faster, and gives the same results for $(\widehat{N}, \widehat{\alpha})$ and their standard errors in cases we have examined. However, it might occur that the optimizer in the constrained parameter space gets trapped in a local extremum that does not occur in the unconstrained space, so we optimize over $(N, p_1, \ldots, p_T, \alpha)$ in all results shown. The constrained optimization might however be useful for handling large problems.

We estimate variances by inverting the estimated Hessian at the maximum likelihood estimates. We use log-Normal confidence intervals for N, and Normal confidence intervals for all other parameters. Fewster and Jupp (2009) show that these are the appropriate asymptotic distributions for multinomial models; we assume that the same distributions apply to the transformed multinomial of this model. The 95% confidence interval for N is therefore $(\widehat{N}/A,\ \widehat{N} \times A)$, where $A = \exp\left[z_{.025}\sqrt{\log\left\{1 + \widehat{\text{var}}(\widehat{N})/\widehat{N}^2\right\}}\right]$, and where $z_{.025}$ is the upper 0.025 point of the Normal(0, 1) distribution. For other parameters $\theta \in \{\alpha, p_1, \ldots, p_T\}$, 95% confidence intervals are $(\widehat{\theta} - B_\theta, \widehat{\theta} + B_\theta)$ where $B_\theta = z_{.025}\sqrt{\widehat{\text{var}}(\widehat{\theta})}$.

Efficient computation of (2) is described in Web Appendix A. We maximize the likelihood using ADMB (Fournier et al., 2012) or R (R Development Core Team, 2012). ADMB provides fast, stable optimization using automatic differentiation, which is as accurate as symbolic differentiation but does not require analytic likelihood derivatives (Fournier et al., 2012). Computer code for the likelihood is broken down algorithmically into its simplest component functions, which can each be differentiated symbolically, then these are combined into the overall likelihood gradient using repeated application of the chain rule.

2.2 *Application to simulated data in Link et al. (2010)*

To verify our procedure and demonstrate computation speed, we apply our maximum likelihood method to the data in Section 4 of Link et al. (2010), which they generated using $T = 5$, $N = 400$, $\alpha = 0.9$, and $(p_1, \ldots, p_5) = (0.3, 0.4, 0.5, 0.6, 0.7)$. We obtain $\widehat{N} = 397.93$ with 95% CI $(365.9, 432.8)$, $\widehat{\alpha} = 0.910$ $(0.878, 0.942)$, and $\widehat{p} = (0.302, 0.407, 0.500, 0.598, 0.706)$. The results are identical if we optimize only over (N, α). These results correspond very closely with those in Table 2 of Link et al. (2010), where the posterior mean for N is 399.4 with central 95% posterior interval $(370, 432)$, and for α is 0.910 $(0.879, 0.940)$. Our estimates of p are in every case within 0.002 of the posterior means given by Link et al. (2010).

The computation time for our analysis on a 1.73GHz laptop is 1.2 seconds using ADMB, and 10.0 seconds using R. This fast computation enables us to explore statistical properties of the model via simulation, which has not been done by previous authors.

3. Simulation study

We conducted extensive simulations to study the performance of maximum likelihood estimates in model $M_{t,\alpha}$. We investigate bias, precision, and CI coverage in three scenarios:

A. Simulate data from model $M_{t,\alpha}$, and fit model $M_{t,\alpha}$;

B. Simulate data from model $M_{t,\alpha}$, and fit model M_t;

C. Simulate data from model M_t (i.e. $\alpha = 1$), and fit model M_t.

The purpose of scenario A is to investigate finite-sample properties of the maximum likelihood estimates within the likely range of application of the model. Although we expect the usual asymptotic MLE properties to apply as $N \to \infty$, good performance is not guaranteed for finite N. Scenario B represents the naive approach of ignoring potential errors and fitting model M_t, and we are interested in a comparison between the naive approach and the correctly specified scenario A. Scenario C represents the case where identification errors can be corrected before the model is applied, for example by repeat genotyping. We are interested in the cost of reaching a desired coefficient of variation (CV) or root-mean-square error (RMSE) for $\widehat{N}$ under Scenario C compared with that under Scenario A.

We used 500 simulation replicates for every combination of inputs T and $(N, \alpha, p_1, \ldots, p_T)$. We focused on T from 4 to 12, $N \in \{400, 1000\}$, $\alpha \in \{0.90, 0.95, 0.97, 0.99\}$, and $p_1 = p_2 = \ldots = p_T$ with value varying from 0.05 to 0.50, although we also conducted numerous simulations outside these ranges. When $p_1 = \ldots = p_T$, they are estimated as T free parameters. Simulations were created in R, and we used the R package `R2admb` (Bolker and Skaug, 2012) to interface with ADMB for computing MLEs and standard errors.

We calculate percentage bias as $(\widehat{N}_{\text{mean}} - N)/N \times 100$, where $\widehat{N}_{\text{mean}}$ is the mean $\widehat{N}$ from the 500 simulations, and N is the known generating value. Percentage root-mean-square error is $\text{RMSE} = \left\{ \sqrt{\frac{1}{500} \sum_{s=1}^{500} \left(\widehat{N}_s - N \right)^2} \right\} / N \times 100$, and it is approximately equal to the empirical CV when $\widehat{N}$ is unbiased. Confidence interval coverage is the percentage of nominal 95% confidence intervals to contain the true generating value of the parameter, using the expressions in section 2.1. Convergence of the optimization in ADMB is confirmed by checking that the maximum gradient component is sufficiently close to zero. Where boundary estimates were obtained ($\widehat{\alpha} = 1$, $\widehat{N} = D$, or $\widehat{N} = \max_t\{n_t\}$), these were retained for all calculations.

3.1 *Results*

Figure 2 shows the distribution of MLEs obtained from $T = 8$ capture occasions with $N = 400$, with $\alpha = 0.97$ representing a moderate error rate of 3%, and with very high capture probabilities $p_1 = \ldots = p_8 = 0.4$. With these parameters, we expect over 98% of animals to be captured at least once, and an average of about 430 capture histories in the data. Scenario A in Figure 2 shows that model $M_{t,\alpha}$ performs well, with no discernible bias, low RMSE (approximately equal to the CVs), and approximately nominal confidence interval coverage. Ignoring the 3% error rate and fitting model M_t, as in scenario B, leads to a 10% bias in $\widehat{N}$ and CI coverage of zero. The failure to match a small number of samples and subsequent surplus of unit histories creates negative bias in the capture probabilities, $\widehat{p}_t$, and leads to a positive bias in $\widehat{N}$. This shows that ignoring just a small amount of error can have detrimental consequences, as the overestimated abundance and very tight precision in scenario B could lead to poor management decisions for an endangered population.

[Figure 2 about here.]

[Figure 3 about here.]

Figure 3 shows equivalent results when capture probability is reduced from 0.4 to 0.1, and all other parameters remain the same. With these parameters, we expect about 57% of animals to be captured at least once, and an average of 233 capture histories in the data. For these parameters neither scenario A nor scenario B performs well, but arguably scenario B is better, with 12% RMSE and 89% CI coverage for $\widehat{N}$, compared with 23% RMSE and 90% CI coverage under scenario A. The reason for the poor performance of scenario A is the large variance introduced by the parameter α. Estimation of α is extremely diffuse with these settings, because the predominant unit histories in the data could arise either from high $\{p_t\}$ and low α, or from low $\{p_t\}$ with any α. The diffuse distribution of $\widehat{\alpha}$ is reflected in the high average error, negative bias, and reduced CI coverage for $\widehat{N}$. These parameter settings are realistic for many situations, for example cetacean surveys, and the conclusion is notable that substantially better RMSE and equivalent CI coverage for N are obtained by ignoring errors rather than by trying to correct for them.

Figure 4 shows the pattern of results for bias and RMSE in $\widehat{N}$ as capture probability ranges from $p_t = 0.05$ to $p_t = 0.50$, constant for each of $T = 8$ capture occasions. As before, $\alpha = 0.97$, and $N = 400$ (top row) or 1000 (bottom row). For $N = 400$, the RMSE in scenario A improves upon that in scenario B for approximately $p_t > 0.2$; for $N = 1000$ this is slightly better at approximately $p_t > 0.15$. However, the RMSE (equivalently, the CV) remains much higher for scenario A than for scenario C. If errors can be corrected before the model is applied, as in scenario C, the RMSE and CV are reduced by a factor of 3 or more.

[Figure 4 about here.]

[Figure 5 about here.]

Inspection of numerous results shows that bias and precision in model $M_{t,\alpha}$ improve as either T or p_t increases. Because $M_{t,\alpha}$ gains information about α from the surplus of unit capture histories compared with multiple-entry histories, we conjecture that the probability of a multiple-entry history may be a key predictor for good performance of the model. In particular, all histories with at least two entries are assumed to be sound, so they provide direct information about α, and for high T or p_t a unit history may be unlikely unless it is an error. In Figure 5 we plot bias and RMSE against the probability that an animal included in the sample is caught at least twice: that is, $P(Y \geqslant 2 \,|\, Y \geqslant 1)$ where $Y \sim \text{Binomial}(T, p_t)$. We find that for given values of α and N, results for multiple different combinations of T and p_t lie close to a smooth curve. For $\alpha = 0.97$ and $N = 400$, for MLE performance to be no worse than 15% RMSE, 15% CV, and 5% bias, this conditional probability should be about 0.4 or above. For $N = 1000$, the equivalent threshold is about 0.33. These results may help to plan surveys where it is possible to increase either capture probability or the number of capture occasions to meet these thresholds.

3.2 *Cost of achieving a desired precision for genetic surveys*

In view of the greatly decreased precision of model $M_{t,\alpha}$ (scenario A) compared with model M_t (scenario C), it is natural to ask whether model $M_{t,\alpha}$ is a cost-effective approach to error correction. In genetic surveys, especially those using high-quality tissue samples, repeat genotyping offers a way of correcting most misidentification errors before the model is applied, enabling researchers to use model M_t instead of $M_{t,\alpha}$. To reach a target precision, model $M_{t,\alpha}$ requires much higher capture probabilities than model M_t, and therefore requires a larger number of samples to genotype. On the other hand, applying model M_t would involve repeat-genotyping of uncertain samples in order to eliminate errors, and this will also inflate genotyping costs.

Here, we investigate which of these approaches is more cost-effective to reach a target RMSE. We take account only of genotyping costs, and disregard sampling costs. In some cases, such as shipboard cetacean surveys, sampling costs may far exceed genotyping costs, so the greater sample sizes needed for model $M_{t,\alpha}$ may be unachievable. In studies using hair or faeces, by contrast, sampling may be relatively cheap and genotyping costs may dominate. To avoid making assumptions about the relative costs of sampling and genotyping, we consider only the numbers of samples that must be genotyped under the different strategies.

The following strategies are simplistic but nonetheless provide a useful comparison.

(1) Genotype all samples once. Regard all unit capture histories as suspicious, so repeat the genotyping for all unit-history samples once. After the repeats, re-match all samples using concensus genotypes and incomplete matching criteria as appropriate. Assume all errors are eliminated, and apply model M_t.

(2) As (1), but with two repeats for all samples initially in unit histories. Apply M_t.

(3) Genotype all samples once only, and apply model $M_{t,\alpha}$.

The results are shown in Figure 6, where the cost of strategies (1) to (3) is assumed proportional to the number of genotypes required. The capture probabilities needed to reach a desired RMSE are estimated by interpolation from the traces on Figure 4, and are thenceforth used to find the mean number of genotypes required under the different strategies. Figure 6 shows $\alpha = 0.97$ with $T = 8$ and $N = 400$. The features of the graphic are very similar for all combinations of $T \in \{4, 8\}$ and $N \in \{400, 1000\}$, and also if RMSE is replaced by CV.

If errors can be corrected by genotyping effort equivalent to repeating all unmatched genotypes once (strategy 1), Figure 6 shows it is always substantially cheaper to apply model M_t than model $M_{t,\alpha}$ to reach a desired precision. If the effort required is equivalent to repeating all unmatched genotypes twice (strategy 2), model $M_{t,\alpha}$ is better for higher target RMSEs, but becomes worse as the target RMSE becomes more ambitious.

[Figure 6 about here.]

We note that even strategy 1 might overstate the amount of error-correcting needed to apply model M_t when genetic samples are obtained from high-quality tissue, as the majority of unmatched samples are sufficiently distinct from all other samples to be safely assumed to be sole captures of an individual based on established methods. However, from our laboratory experience, the few samples over which there are doubts may take two or three repeats, and finalising this phase can be very time-consuming.

4. Application to southern right whales

We apply model $M_{t,\alpha}$ to genetic and photographic surveys of the endangered New Zealand population of the southern right whale (*Eubalaena australis*). Southern right whales were common in New Zealand waters prior to 19th-century whaling. After severe population depletion, only an outpost of the original population remained in the subantarctic Auckland Islands ($50^\circ\ 32'$ S), and the NZ population is still largely restricted to this region (Carroll et al., 2011). The population can be surveyed at the Auckland Islands during annual winter calving congregations. However, the expense and difficulty of working in subantarctic waters during the austral winter means that survey opportunity is limited, despite the national significance of the population.

4.1 *1995-1998 genetic survey*

Genetic samples were collected annually in the $T = 4$ austral winters of 1995–1998, using small biopsy darts deployed from a crossbow. Samples were genotyped at 13 microsatellite loci and sexed. We restrict our analysis to males, because for females there is evidence of a cyclic pattern of capture in the calving grounds due to their 3-year calving intervals (Carroll et al., 2013). Treating the male population as a closed population over the survey period is an approximation, and open-population estimates can be found in Carroll et al. (2011; 2013). We include this example because it illustrates the level and type of genetic error obtained from biopsy samples, and the practicalities of applying model $M_{t,\alpha}$.

The data consist of 132 genetic samples and 13 loci. This creates 1716 single-locus results, of which 139 are missing (8%). Pairwise comparison of all samples comprises 8646 pairs in total, of which the number of pairs with exact matches at 0-4 loci is 8621; at 5-6 loci is 5; at 7-8 loci is 0; and at 9-13 loci is 20. These matching loci exclude any with missing data. The results show a clear break between samples matching at 6 or fewer loci, and samples matching at 9 or more loci.

Using the least variable 9 loci in the data, the probability that two individuals have the same genotype by chance is PID $= 6.0 \times 10^{-11}$ (Paetkau and Strobeck, 1994), or for closely related individuals, PIDsib $= 1.5 \times 10^{-4}$ (Evett and Weir, 1998). For the most variable 6 loci, PID $= 1.8 \times 10^{-9}$ and PIDsib $= 1.3 \times 10^{-3}$, so a few 6-locus matches are reasonable by chance in a population of a few hundred animals that includes siblings. These figures support the use of the break between 6 and 9 matching loci as a matching criterion. We shall assume that two samples belong to the same individual if they have at least 9 matching loci.

Using the 9-loci match rule, there are 60 non-matching loci among samples assumed to come from the same individual. Of these, 50 are due to missing data; 6 could be due to allelic dropout; 3 have a single allele substitution; and in the remaining case the non-matching loci have no alleles in common. Thus, every type of non-match appears in the data.

Applying model $M_{t,\alpha}$ with the 9-loci match rule gives a boundary estimate: $\widehat{N} = 306$ with 95% confidence interval (212, 443); $\widehat{\alpha} = 1.000$ with standard error 0.004; and $(\widehat{p}_1, \ldots, \widehat{p}_4) = (0.09, 0.07, 0.10, 0.17)$ with standard errors $(0.02, 0.02, 0.03, 0.04)$. Identical results are obtained by applying model M_t.

For every other match rule we examined, we also obtained boundary estimates: either $\widehat{\alpha} \simeq 1$, or $\widehat{N} = \max_t\{n_t\}$. If we apply the strict rule that samples must share 13 matched loci to be attributed to the same animal, there is only one recapture in the data-set, and both models $M_{t,\alpha}$ and M_t give implausible results: $\widehat{N} = \max_t\{n_t\} = 51$ and $\widehat{\alpha} = 0.09$ from model $M_{t,\alpha}$, and $\widehat{N} = 6148$ with 95% CI (1226, 31 386) from model M_t. In every case, optimizing over (N, α) instead of $(N, \alpha, p_1, \ldots, p_4)$ gives the same results for $M_{t,\alpha}$.

We conclude that model $M_{t,\alpha}$ does not provide a helpful mechanism for correcting genetic errors in our example. The common-sense 9-loci matching criterion probably eliminates all errors, and model $M_{t,\alpha}$ gives the same results as model M_t. Under the exact-match 13-loci criterion, the data have 130 unit histories out of 131 observed in total, whereas the probable truth under the 9-loci criterion is 95 unit histories out of 113 total. The model does not succeed in correcting the balance, but instead drives $\widehat{\alpha}$ as low as possible to give $\widehat{N} = 51$. We suggest that these conclusions are likely to apply to the majority of genetic studies with high-quality tissue samples, and scrutiny of the data and matching criteria before applying the model is likely to be a more effective strategy than applying model $M_{t,\alpha}$.

4.2 *1998 photographic survey*

In 1998 the field season was 54 days long, enabling a within-season estimate of abundance. Analysis of photo-identification data from this survey is provided in Web Appendix B.

5. Discussion

Model $M_{t,\alpha}$ is an elegant formulation for modeling misidentification in capture-recapture studies. The latent-multinomial formulation of Link et al. (2010) also creates a framework for dealing with other latent-identity problems. However, we have found that the indirect method of detecting errors, relying on a surplus of histories with only one entry, places heavy demands on the data. The model can be used with confidence when there are high capture probabilities or many capture occasions, and from Figure 5 we suggest that the probability of a multiple-entry history is of particular importance for predicting MLE performance.

We have focused here on maximum likelihood estimation. The weak identifiability of α is a general problem for model $M_{t,\alpha}$, but it may be addressed to some extent by using a different mode of inference. A Bayesian approach using informative priors for α may be a good choice, and computation speed could be considerably improved by our likelihood reformulation (2). Informative priors for α can be based on experience from replicate genotyping or from multiple experts matching photo catalogues. We also ran some simulations using a least-squares approach similar to that in Yoshizaki et al. (2011), and found some improvement in RMSE over the MLEs for small samples in which information was sufficient to draw the MLEs away from the boundaries but insufficient to provide strong identifiability of α. Variance for the least-squares approach can be estimated by bootstrap resampling.

We had difficulty in applying model $M_{t,\alpha}$ to real data. In our genetic example, incomplete matching of genetic samples was almost ubiquitous and the identification process was not well described by the simplistic model of $M_{t,\alpha}$, which assumes every identification is either right or wrong. Applying a strict complete-match criterion led to an implausible boundary estimate, whereas a common-sense matching criterion eliminated errors to the point that there was no need for model $M_{t,\alpha}$. In our photo survey, on the other hand, N was estimated with 95% confidence interval (49, 419) which is not adequate for management purposes. Our difficulty reflects that of other authors: none of the methodological treatments to date have provided a fully satisfactory analysis of real data. Overall, we suggest researchers should be cautious in applying model $M_{t,\alpha}$, and should first test its properties by simulation and consider whether the model for identification error is suitable for their situation.

In cases with high error rate, we expect a model similar to that of Wright et al. (2009) might have better performance than $M_{t,\alpha}$, because information on error rate is gained directly from discrepancies between repeated attempts to genotype the same sample, resolving issues of weak identifiability of α. Wright et al.'s (2009) model is specific to genetic surveys, and does not extend obviously to photographic surveys. A similar formulation based on multiple judgments from independent experts might have potential in the photographic case.

Our formulation of the exact likelihood for $M_{t,\alpha}$ enables very fast computation of MLEs, especially when using ADMB. Computation speed depends upon the combinatorial problem in Figure 1 and does become slow at times, for example taking 13 minutes with $\alpha = 0.6$, $T = 12$, $p_t = 0.1$, and $N = 1000$, although this reduces to 2 minutes if optimizing only over (N, α). Link et al. (2010) indicate that their Bayesian approach takes over 30 minutes for $T = 5$, contrasting with our 1.2 seconds to obtain MLEs for the same data. A reformulation of the Bayesian approach using our likelihood computation would enable practitioners to exploit the joint advantages of computation speed and informative priors for α.

A disadvantage of our formulation is that it is very specific to model $M_{t,\alpha}$. We treat the capture occasions individually, partitioning captures into sound and faulty captures. This would not be possible for models that examine each animal's complete capture history, such as the behavioral-response model M_b, or model M_h which incorporates individual heterogeneity (Otis et al., 1978). Unlike the latent-history approach of Link et al. (2010), our formulation cannot be extended to these cases. However, in view of the high variance inherent in model $M_{t,\alpha}$, it might be unwise to combine error estimation with models M_b or

M_h, which themselves suffer from high variance, and confounding of α with other parameters seems likely. More of a problem is the inability to generalise our approach to open-population models, in which whole capture histories must be considered because an animal may be born or die part-way through the study. Despite this lack of generality, our device in equation (2) of switching the sum from latent frequency vectors $\boldsymbol{x}$ to a different latent variable might be applicable in other situations. Our conclusions about the low precision associated with indirect error estimation are also likely to hold more generally.

6. Supplementary Materials

Code for fitting $M_{t,\alpha}$ in R and ADMB, and the Web Appendices referenced in Sections 2.1 and 4.2 are available with this paper at the Biometrics website on Wiley Online Library.

Acknowledgements

We thank Scott Baker, Simon Childerhouse, Debbie Steel, and all fieldworkers involved in the Auckland Island surveys. We thank the associate editor and referees for their helpful comments which greatly improved the article, and David Gauld for insight on constrained optimization. Field and laboratory work was funded by the Whale and Dolphin Conservation Society, the US Department of State (Program for Cooperative US/NZ Antarctic Research), the Auckland University Research Council, the NZ Marsden Fund, the NZ Department of Conservation, the Australian Antarctic Division, the Heseltine Trust and OMV NZ Ltd. ELC was supported by a Tertiary Education Commission Top Achiever Scholarship.

References

Bolker, B. and Skaug, H. (2012). *R2admb: ADMB to R interface functions.* R package 0.7.5.3.

Bonner, S. J. and Holmberg, J. (2013). Mark-recapture with multiple, non-invasive marks. *Biometrics* **69,** 766–775.

Carroll, E. L., Patenaude, N. J., Childerhouse, S. J., Kraus, S. D., Fewster, R. M., and Baker, C. S. (2011). Abundance of the New Zealand subantarctic southern right whale population estimated from photo-identification and genotype mark-recapture. *Marine Biology* **158,** 2565–2575.

Carroll, E. L., Childerhouse, S. J., Fewster, R. M., Patenaude, N. J., Steel, D., Dunshea, G., Boren, L., and Baker, C. S. (2013). Accounting for female reproductive cycles in a superpopulation capture-recapture framework: application to southern right whales (*Eubalaena australis*). *Ecological Applications* **23,** 1677–1690.

Evett, I. and Weir, B. (1998). *Interpreting DNA evidence: statistical genetics for forensic scientists.* Sunderland: Sinauer.

Fewster, R. M. and Jupp, P. E. (2009). Inference on population size in binomial detectability models. *Biometrika* **96,** 805–820.

Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., Nielsen, A., and Sibert, J. (2012). AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software* **27,** 233–249.

Higgs, M. D., Link, W. A., White, G. C., Haroldson, M. A., and Bjornlie, D. D. (2013). Insights into the latent multinomial model through mark-resight data on female grizzly bears with cubs-of-the-year. *Journal of Agricultural, Biological, and Environmental Statistics* **18,** 556–577.

Karanth, K. U., Nichols, J. D., Kumar, N. S., and Hines, J. E. (2006). Assessing tiger population dynamics using photographic capture-recapture sampling. *Ecology* **87,** 2925–2937.

Link, W. A., Yoshizaki, J., Bailey, L. L., and Pollock, K. H. (2010). Uncovering a latent multinomial: analysis of mark–recapture data with misidentification. *Biometrics* **66,** 178–185.

Lukacs, P. M. and Burnham, K. P. (2005). Estimating population size from DNA-based closed capture-recapture data incorporating genotyping error. *Journal of Wildlife Management* **69,** 396–403.

McClintock, B. T., Conn, P. B., Alonso, R. S., and Crooks, K. R. (2013). Integrated modeling of bilateral photo-identification data in mark-recapture analyses. *Ecology* **94,** 1464–1471.

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs* **62,** 3–135.

Paetkau, D. and Strobeck, C. (1994). Microsatellite analysis of genetic variation in black bear populations. *Molecular Ecology* **3,** 489–495.

Pompanon, F., Bonin, A., Bellemain, E., and Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics* **6,** 847–859.

R Development Core Team. (2012). R: *A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Morrison, T. A., Yoshizaki, J., Nichols, J. D., and Bolger, D. T. (2011). Estimating survival in photographic capture-recapture studies: overcoming misidentification error. *Methods in Ecology and Evolution* **2,** 454–463.

Taberlet, P. and Luikart, G. (1999). Non-invasive genetic sampling and individual identification. *Biological Journal of the Linnean Society* **68,** 41–55.

Wright, J. A., Barker, R. J., Schofield, M. R., Frantz, A. C., Byrom, A. E., and Gleeson, D. M. (2009). Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples. *Biometrics* **65,** 833–840.

Yoshizaki, J., Brownie, C., Pollock, K. H., and Link, W. A. (2011). Modeling misidentification errors that result from use of genetic tags in capture-recapture studies. *Environmental and Ecological Statistics* **18,** 27–55.

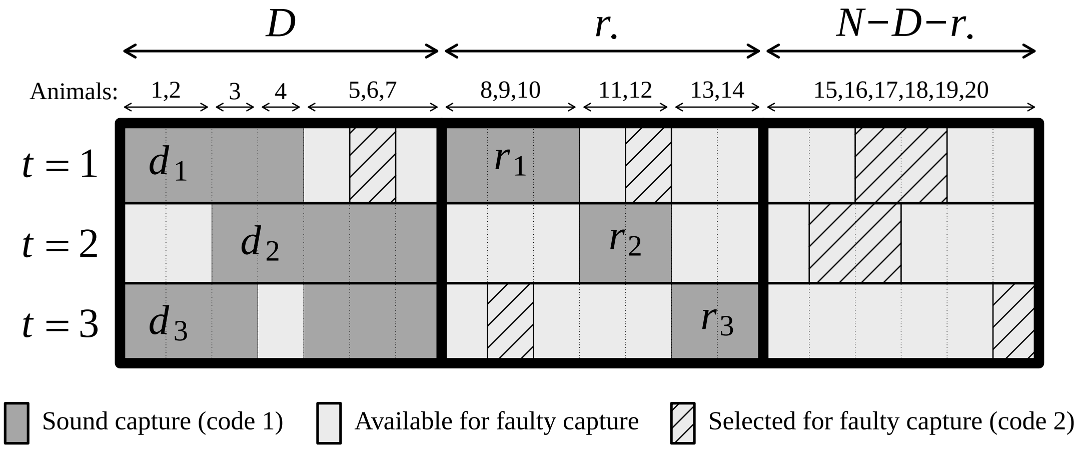

Figure 1. Diagram showing $N = 20$ latent capture histories at $T = 3$ times, for a particular choice of $\boldsymbol{r} = (r_1, r_2, r_3)$. Each latent history is a vertical slice through the diagram, shaded according to its capture code at times $t = 1, 2, 3$, and there is one latent history for each of the N animals in the population. The N histories are partitioned into: D with two or more sound captures; $r_\bullet = r_1 + r_2 + r_3$ with exactly one sound capture, of which r_t have their sound capture at time t; and the remaining $N - D - r_\bullet$ with no sound captures. Firstly we fix the pattern of sound captures denoted by dark shading: these are pre-determined by the observed duplicate histories and the choice of $\boldsymbol{r}$. Secondly we assign animals to these 'sound' histories. For each group of identical sound histories, we list the animals allocated to the group in numerical order across the columns. One such ordered assignment is marked on the diagram, showing for example that the two copies of sound history 101 are allocated to animals 1 and 2. In total, there are $N! \,/\, \left\{ \left(\prod_{k\,:\,|\omega_k| \geqslant 2} f_k! \right) \left(\prod_{t=1}^{T} r_t! \right) (N - D - r_\bullet)! \right\}$ possible assignments. Thirdly we assign faulty captures to animals. For each row t, there are $N - (d_t + r_t)$ animals available for faulty capture (pale shading), of which $u_t - r_t$ must be selected (hatched shading). Choices for faulty capture can be made independently in each of the T rows. The diagram shows one assignment of faulty captures to animals; there are $\prod_{t=1}^{T} \binom{N-d_t-r_t}{u_t-r_t}$ assignments available. The completed diagram shows one way of allocating latent histories to animals consistent with the observed data. For a given $\boldsymbol{r}$, the selections in the second and third steps generate all possible allocations exactly once, and each allocation has the same probability $\left\{ \alpha^{C} \prod_{t=1}^{T} p_t^{n_t} (1 - p_t)^{N - n_t} \right\} \alpha^{r_\bullet} (1 - \alpha)^{U - r_\bullet}$.

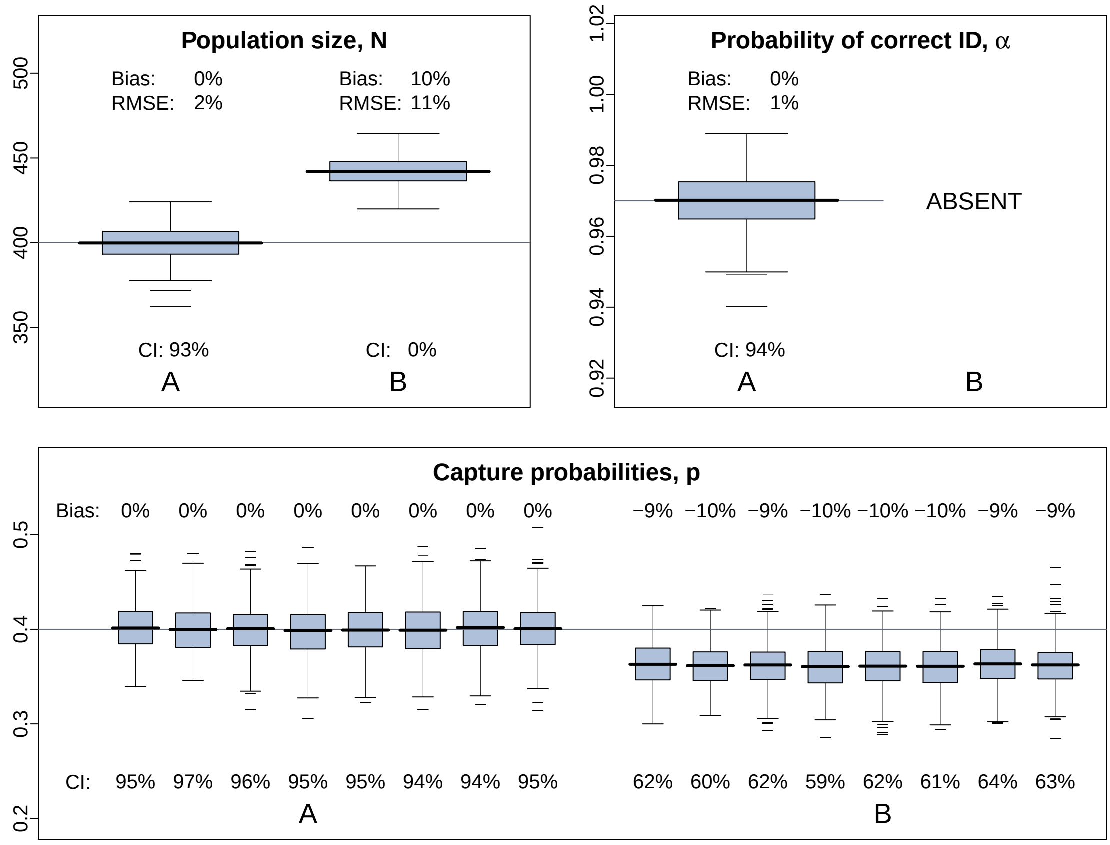

Figure 2. Distributions of parameter estimates from 500 simulations with $T = 8$, $N = 400$, $\alpha = 0.97$, and $p_1 = \ldots = p_8 = 0.4$, under scenarios A and B. Scenario A: model $M_{t,\alpha}$ used for both simulation and estimation. Scenario B: model $M_{t,\alpha}$ used for simulation; model M_t used for estimation. Bold horizontal lines on the boxes show the mean of the estimates; numbers above and below show percentage bias, root-mean-square error (RMSE), and confidence interval coverage for nominal 95% confidence intervals. Thin horizontal lines across each plot show the true values of the parameters. Boxes are drawn between the upper and lower quartiles of estimates; whiskers extend to the last estimate within 1.5 times the interquartile range from the quartiles; and estimates beyond the whiskers are marked as outlying points. The mode of the distribution of $\widehat{N}$, estimated by R function `density`, is 400.6 for Scenario A and 441.5 for Scenario B.

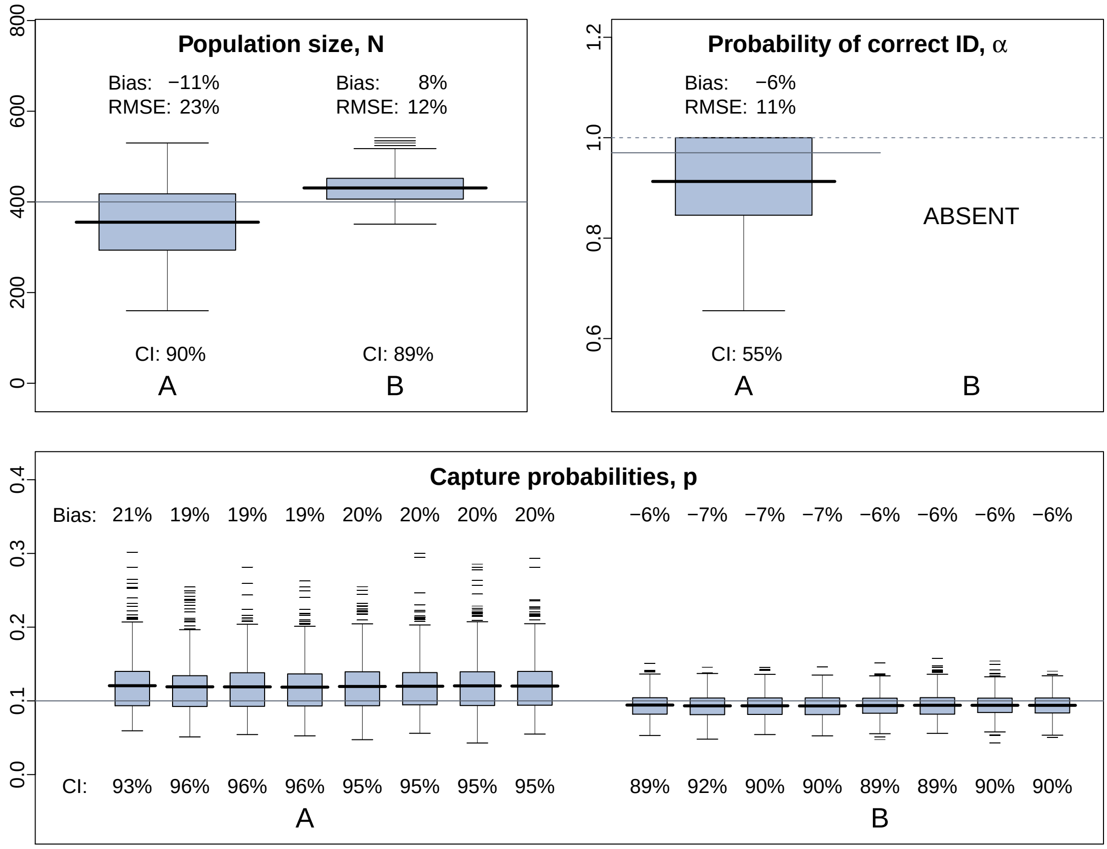

Figure 3. Distributions of parameter estimates under scenarios A and B when capture probability p is reduced to $p_1 = \ldots = p_8 = 0.1$. All other details are identical to Figure 2. The estimated mode of the distribution of $\widehat{N}$ is 406.6 for Scenario A, and 415.0 for Scenario B.

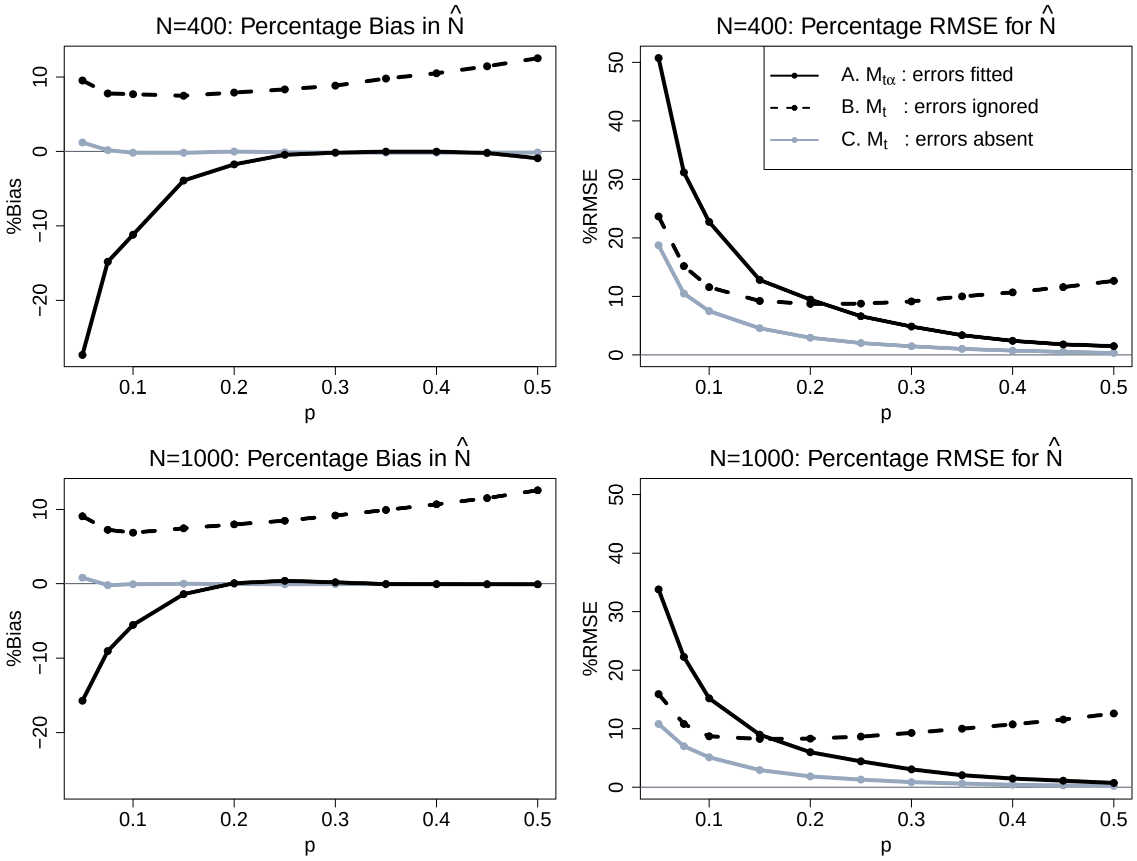

Figure 4. Percentage bias and root-mean-square error (RMSE) in estimates of $\widehat{N}$ when $T = 8$, $\alpha = 0.97$, and $N = 400$ (top row) or $N = 1000$ (bottom row), as capture probability ranges from $p = 0.05$ to $p = 0.50$. Scenario A (solid black lines): model $M_{t,\alpha}$ used for both simulation and estimation. Scenario B (dashed black lines): model $M_{t,\alpha}$ used for simulation; model M_t used for estimation. Scenario C (grey lines): model M_t used for both simulation and estimation. Plotted points show results obtained and lines interpolate between them.

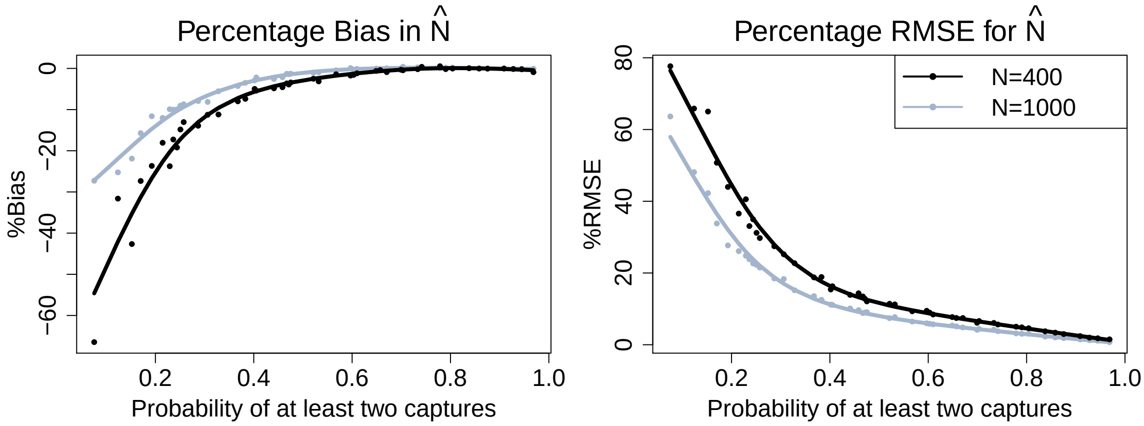

Figure 5. Percentage bias and root-mean-square error (RMSE) in estimates of $\widehat{N}$ under scenario A when $\alpha = 0.97$, shown against the probability that a sampled animal is caught at least twice. Black points show results for $N = 400$, grey points for $N = 1000$, and smooth curves are fitted to results for each N. Results for a single value of N fall broadly on a smooth curve for a spectrum of values of T between 4 and 12, and capture probabilities p_t between 0.05 and 0.50.

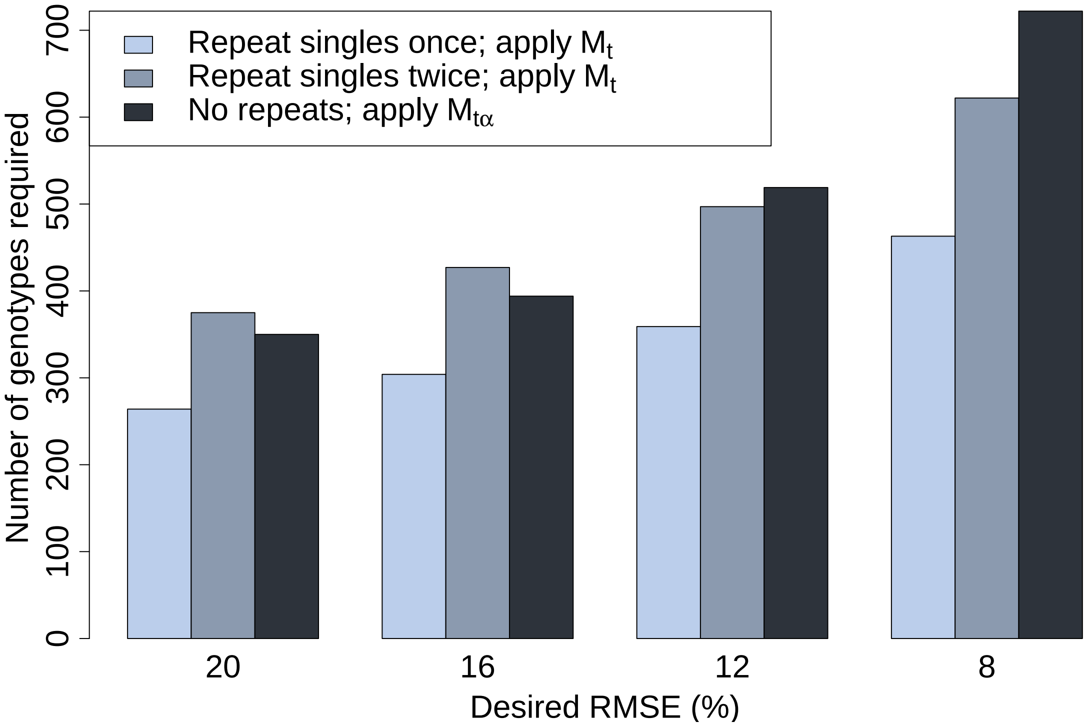

Figure 6. Cost of reaching specified thresholds of root-mean-square error (RMSE) under different strategies, when $T = 8$, $N = 400$, $\alpha = 0.97$, and capture probability can be selected. The bar heights show the number of samples that must be genotyped, including repeats, under each of the strategies.