

Evaluation of some random effects methodology applicable to bird ringing data

KENNETH P. BURNHAM¹ & GARY C. WHITE², ¹ Colorado Cooperative Fish & Wildlife Research Unit, USGS-BRD, Colorado, USA, and ² Department of Fishery and Wildlife Biology, Colorado State University, USA

ABSTRACT Existing models for ring recovery and recapture data analysis treat temporal variations in annual survival probability (S) as fixed effects. Often there is no explainable structure to the temporal variation in S_1, \dots, S_k ; random effects can then be a useful model: $S_i = E(S) + \varepsilon_i$. Here, the temporal variation in survival probability is treated as random with average value $E(\varepsilon^2) = \sigma^2$. This random effects model can now be fit in program MARK. Resultant inferences include point and interval estimation for process variation, σ^2 , estimation of $E(S)$ and $\text{var}(\hat{E}(S))$ where the latter includes a component for σ^2 as well as the traditional component for $\bar{\text{var}}(\hat{S}|\hat{S})$. Furthermore, the random effects model leads to shrinkage estimates, \tilde{S}_i , as improved (in mean square error) estimators of S_i compared to the MLE, \hat{S}_i , from the unrestricted time-effects model. Appropriate confidence intervals based on the \tilde{S}_i are also provided. In addition, AIC has been generalized to random effects models. This paper presents results of a Monte Carlo evaluation of inference performance under the simple random effects model. Examined by simulation, under the simple one group Cormack-Jolly-Seber (CJS) model, are issues such as bias of $\hat{\sigma}^2$, confidence interval coverage on $\hat{\sigma}^2$, coverage and mean square error comparisons for inference about S_i based on shrinkage versus maximum likelihood estimators, and performance of AIC model selection over three models: $S_i \equiv S$ (no effects), $S_i = E(S) + \varepsilon_i$ (random effects), and S_1, \dots, S_k (fixed effects). For the cases simulated, the random effects methods performed well and were uniformly better than fixed effects MLE for the S_i .

1 Introduction

The objective of this paper is to evaluate, by simulation, the basic operating characteristics of some simple random effects inference methodology applicable to

Correspondence: K. P. Burnham, USGS-BRD, Colorado Cooperative Fish and Wildlife Research Unit, 201 Wagar Building, Colorado State University, Fort Collins, CO 80523, USA.

open model capture-recapture data. The reader is assumed to have a basic knowledge of dead-recovery (e.g. Brownie *et al.*, 1985) and live-recapture (e.g. Lebreton *et al.*, 1992) models, referred to here generically just as capture-recapture (for a current review see Schwarz & Seber, 1999). Basic random effects concepts and models are well established in general statistical theory under various names, such as variance components, random effects, random coefficient models, or empirical Bayes estimates (see, for example, Efron & Morris, 1975; Casella, 1985; Searle *et al.*, 1992; Longford, 1993; Carlin & Louis, 1996). There has been only modest application of random effects in ecology (see, for example, Johnson, 1981, 1989; Burnham *et al.*, 1987; Link & Nichols, 1994; Ver Hoef, 1996; Link, 1999; Franklin *et al.*, this issue) despite that, these methods are needed, for example, to estimate correctly process variation in survival probabilities over space or time (White, 2000). We assume the reader has some familiarity with basic concepts of random effects, such as process variation versus sampling variation.

The specific random effects methodology for capture-recapture evaluated here is implemented in program MARK (White & Burnham, 1999; White *et al.*, 2002). In this paper we first summarize the relevant theory (i.e. methods evaluated). Next, we give the specific inferential aspects of random effects methods evaluated here. Then we give the design of the Monte Carlo simulation study used for this evaluation. Finally, there are summarized results, such as on bias, efficiency, confidence interval coverage, and model selection.

This is perhaps the first performance evaluation of a random effects model for capture-recapture. Consequently, we keep the evaluation simple by restricting it to simulated data based only on the Cormack-Jolly-Seber (CJS) time-specific model $\{S_t, p_t\}$ wherein estimated survival and capture probabilities are allowed to be time varying for $k + 2 = t$ capture occasions, equally spaced in time. However, this should entail no loss of generality because both relevant theories are quite general. First, the random effects theory used here is general; it applies to any set of maximum likelihood estimates, $\hat{S}_1, \dots, \hat{S}_k$ regardless of the type of capture-recapture data analysed. Secondly, time-specific models for recovery and recapture data stem from the same deep, unified theory (Burnham, 1991; Barker, 1997).

2 Inference methods

2.1 Random effects

Our context here is open models capture-recapture wherein k survival probabilities are estimable, corresponding to k equal length time periods, often years. The basic time-specific CJS model is considered as being conditional on the underlying estimable survival probabilities, S_1, \dots, S_k . Hence, the MLEs for this model can be considered in the context of a linear model (conditional by definition on \underline{S}) represented as $\hat{\underline{S}} = \underline{S} + \underline{\delta}$ (these are $k \times 1$ column vectors). Here, \underline{S} represents the structural parameters and $\underline{\delta}$ is the model's stochastic component. Given \underline{S} , the (large sample) expected value of $\underline{\delta}$ is zero, hence we can take $E(\hat{\underline{S}} | \underline{S}) = \underline{S}$. Also, $\underline{\delta}$ has conditional (on \underline{S}) sampling variance-covariance matrix \underline{W} , which will be a complicated function of \underline{S} and other parameters, such as capture probabilities, p , or ring recovery probabilities, f .

For a random effects model, \underline{S} is modelled as a random vector with expectation $X\beta$ and variance-covariance matrix $\sigma^2 I$; β has r elements. By assumption, the process residuals, $\varepsilon_i = S_i - E(S_i)$, are independent with homogeneous variance σ^2 .

Also, we assume mutual independence of sampling errors δ and process errors ε . We envision fitting a capture-recapture model (that does not constrain \hat{S}) to get the MLE \hat{S} and the usual likelihood-based estimator of W . Only the simplest case of this random effects theory is used here in the simulations, but we present the general theory. The special case used here, in the simulations, is the means model; hence, $r = 1$, X is a column vector of k ones, and β is just the scalar parameter $\mu = E(S)$.

The basic, unconditional, random effects model is

$$\hat{S} = X\beta + \delta + \varepsilon, \quad VC(\hat{\delta} + \varepsilon) = D = \sigma^2 I + E_S(W)$$

Here, we let VC denote a variance-covariance matrix. The obvious inference issues are to estimate β, σ^2 , an unconditional variance-covariance matrix for $\hat{\beta}$, and compute a confidence interval on σ^2 , on $k - r = df$ degrees of freedom. A non-obvious inference issue is the opportunity to use the shrinkage estimator of S, \hat{S} . In a random effects context the shrinkage estimator has a smaller mean square error than the maximum likelihood estimator (Efron & Morris, 1975), and hence is to be preferred on that basis.

Shrinkage estimators are neither intuitive, nor easy to explain thoroughly. Shrinkage estimators are also called empirical Bayes estimators (Carlin & Louis, 1996). For a simple model with independent MLEs, \hat{S}_i , and only one population-level structural parameter, $E(S)$, \tilde{S}_i lies between $\hat{E}(S)$ and \hat{S}_i . The extent of this shrinkage towards $\hat{E}(S)$ depends upon the variance components proportion

$$\frac{\sigma^2}{\sigma^2 + E_S\{\text{var}(\hat{S}_i|S_i)\}}$$

If this proportion is 1, no shrinkage occurs; if it is 0 then for all i , $\tilde{S}_i \equiv \hat{E}(S)$. An individual \tilde{S}_i may not improve upon the corresponding \hat{S}_i in the sense of being nearer to S_i in a given case. However, overall the shrinkage estimators as a set are to be preferred as being closer to the true S_i if the random effects model applies with $\sigma^2 > 0$ (Efron & Morris, 1975; Casella, 1985).

From generalized least squares theory, for σ^2 given, the best linear unbiased estimator of β is

$$\hat{\beta} = (X'D^{-1}X)^{-1}X'D^{-1}\hat{S} \tag{1}$$

Assuming normality of \hat{S} (approximate normality suffices) then from the same generalized least squares theory the weighted residual sum of squares $(\hat{S} - X\hat{\beta})'D^{-1}(\hat{S} - X\hat{\beta})$ has a central chi-squared distribution on $k - r$ degrees of freedom. Therefore, a method of moments estimator of σ^2 is obtained by solving the equation

$$k - r = (\hat{S} - X\hat{\beta})'D^{-1}(\hat{S} - X\hat{\beta}) \tag{2}$$

where $\hat{\beta}$ comes from equation (1). Note that the critical issue of what is $\hat{E}_{\hat{S}}(W)$ is dealt with below.

An aside about the quantity $RSS(\sigma^2) = (\hat{S} - X\hat{\beta})'D^{-1}(\hat{S} - X\hat{\beta})$, as a function of σ^2 , is in order. It can be shown that in the limit as σ^2 goes to ∞ , $RSS(\sigma^2)$ goes to 0. It can also be shown that $RSS(\sigma^2)$ is monotonically decreasing in σ^2 . Furthermore, the mathematically admissible range of σ^2 is $-\lambda_1 < \sigma^2 < \infty$, where λ_1 is the smallest eigenvalue of $\hat{E}_{\hat{S}}(W)$. For $\sigma^2 = -\lambda_1$, matrix D is singular, hence $RSS(-\lambda_1)$ is, roughly speaking, ∞ . For $\sigma^2 < -\lambda_1$, matrix D is not a valid variance-covariance

matrix. These properties of $RSS(\sigma^2)$ mean there is a unique $\sigma^2 > -\lambda$ as the solution to any equation $y = RSS(\sigma^2) = (\hat{S} - X\hat{\beta})'D^{-1}(\hat{S} - X\hat{\beta})$.

Solving equation (2) for $\hat{\sigma}^2$ requires only a 1-dimensional numerical search; a unique numerical solution always exists, but may be negative. If $\hat{\sigma}^2 < 0$ occurs, truncate it to 0, i.e. take $\hat{\sigma}^2 = 0$. The theoretical unconditional sampling variance-covariance of $\hat{\beta}$ is

$$VC(\hat{\beta}) = (X'D^{-1}X)^{-1} \tag{3}$$

To get a $(1 - \alpha)$ 100% confidence interval on σ^2 we solve for σ_L^2 and σ_U^2 , respectively, from

$$\chi_{df,1-\alpha/2}^2 = (\hat{S} - X\hat{\beta})'D^{-1}(\hat{S} - X\hat{\beta}) \tag{4a}$$

$$\chi_{df,\alpha/2}^2 = (\hat{S} - X\hat{\beta})'D^{-1}(\hat{S} - X\hat{\beta}) \tag{4b}$$

Here, $\chi_{df,p}^2$ is the p th percentile of the central chi-squared distribution on df degrees of freedom. Unique solutions exist to equations (4a) and (4b), although the lower confidence limit can be negative. In fact, even the upper confidence limit can be negative. In practice, any negative solutions are replaced by zero.

The shrinkage estimator, \tilde{S} , used in this study requires the matrix

$$H = \sigma D^{-1/2} = \sigma(\sigma^2 I + \hat{E}_S(W))^{-1/2} = (I + \frac{1}{\sigma^2} \hat{E}_S(W))^{-1/2}$$

evaluated at $\hat{\sigma}$. Then $\tilde{S} = H(\hat{S} - X\hat{\beta}) + X\hat{\beta}$. An alternative formula uses the projection matrix

$$G = H + (I - H)AD^{-1} \tag{5}$$

where $A = X(X'D^{-1}X)^{-1}X'$. Then

$$\tilde{S} = G\hat{S} \tag{6}$$

The theoretical, conditional, variance-covariance matrix of the shrinkage estimator is

$$VC(\tilde{S}|\mathcal{S}) = GE_S(W)G'$$

its diagonal elements are $\text{var}(\tilde{S}_i|\mathcal{S})$. Because the shrinkage estimator is conditionally biased we based confidence intervals on

$$\hat{rmse}(\tilde{S}_i|\mathcal{S}) = \sqrt{\hat{var}(\tilde{S}_i|\mathcal{S}) + (\tilde{S}_i - \hat{S}_i)^2} \tag{7}$$

The shrinkage estimator (6) used here is such that the sum of squares of the shrunk residuals (i.e. $\tilde{S} - X\hat{\beta}$), divided by $k - r$, equals $\hat{\sigma}^2$. This coherent relationship does not hold for the usual shrinkage estimator found in the statistical literature (Morris, 1983; Louis, 1984). The shrinkage estimator defined here is central to being able to obtain a useful, simple extension of AIC for this random effects model, because $\hat{\beta}$ and $\hat{\sigma}^2$ are, essentially, computable from \hat{S} . Therefore, the likelihood value to associate with this random effects model can be obtained based on the fixed effects likelihood evaluated at \tilde{S} without the need to compute (as via numerical integration) the proper marginal likelihood of the random effects model.

An important uncertainty about these random effects methods is that we do not have formulae for the elements of $E_S(W)$; hence, this optimal weight matrix cannot be used. We cannot take the exact expectations needed. We will have only an

estimator of $E_{\mathcal{S}}(W)$, and it may have inherent biases. In simple cases, approximate expectations over \mathcal{S} can be found for $\text{var}(\hat{S}_i|\mathcal{S})$ and $\text{cov}(\hat{S}_i, \hat{S}_j|\mathcal{S})$. However, for a general method we want to just use -1 times the matrix of second partial derivatives of the log-likelihood function, say F , which estimates the Fisher information matrix, and then $\hat{W} \equiv \hat{E}_{\mathcal{S}}(W) = F^{-1}$. We have here used this general method to obtain $\hat{E}_{\mathcal{S}}(W)$, used in place of $E_{\mathcal{S}}(W)$. Given such approximations, and the overall complexities, of this random effects methodology, Monte Carlo simulation evaluation is required to determine actual inference performance.

2.2 AIC for random effects

We will have started with a likelihood for a model at least as general as full time variation on all the parameters, say $\mathcal{L}(\mathcal{S}, \theta) = \mathcal{L}(S_1, \dots, S_k, \theta_1, \dots, \theta_\ell)$. For the CJS model, the additional parameters would be just the capture probabilities, p_2, \dots, p_k . However, the formulae given here are meant to apply also to other types of models, such as band recoveries, which might be parameterized in terms of S_i and f_i . Therefore, for generality we denote the additional model parameters generically as θ_i . And in general under the model, $\{S_i, \theta\}$, we have the MLEs, \hat{S} and $\hat{\theta}$, and the maximized log-likelihood, $\log \mathcal{L}(\hat{S}, \hat{\theta})$ based on $K = k + \ell$ parameters. Thus, for large sample size, n , AIC for the fixed effects model is $-2\log \mathcal{L}(\hat{S}, \hat{\theta}) + 2K$ (Burnham & Anderson, 1998).

The log-likelihood value for the fitted random effects model on \mathcal{S} comes from re-optimizing over θ at the value of \hat{S} . We denote this random effects log-likelihood value as

$$\log \mathcal{L}(\hat{\mathcal{S}}, \hat{\theta}^{\times}) \equiv \log \mathcal{L}(\hat{\mathcal{S}}, \hat{\theta}(\hat{\mathcal{S}})) = \max_{\theta} [\log \mathcal{L}(\hat{\mathcal{S}}, \theta)]$$

where $\hat{\mathcal{S}}$ essentially ‘contains’ $\hat{\beta}$ and $\hat{\sigma}^2$. The dimension of the parameter space to associate with this random effects model is K_{re} , where

$$K_{re} = \text{tr}(G) + \ell$$

G is the projection matrix (6) mapping \hat{S} into $\hat{\mathcal{S}}$, and $\text{tr}(\cdot)$ is the matrix trace function ($\text{tr}(G) =$ the sum of the diagonal elements of G ; see, for example, Schott, 1997, p. 4).

The large-sample AIC for the random effects model on \mathcal{S} is

$$-2 \log \mathcal{L}(\hat{\mathcal{S}}, \hat{\theta}^{\times}) + 2K_{re}$$

The small sample corrected version, AICc, for this random effects model is

$$-2 \log \mathcal{L}(\hat{\mathcal{S}}, \hat{\theta}^{\times}) + 2K_{re} + 2 \frac{K_{re}(K_{re} + 1)}{n + (K_{re} - 1)} \tag{8}$$

In these simulations we used equation (8) with effective sample size, n , as the total number of animal-release events (Burnham *et al.*, 1994). However, sample sizes in these simulations were always large enough that the small sample term in equation (8) was irrelevant.

Results such as equation (8) are, in the literature, giving AIC generalized to semi-parametric smoothing applications; see, for example, Hurvich & Simonoff (1998) and Shi & Tsai (1998). However, those papers are not about random effects models. Instead, those papers note a generalized AIC where the effective number

of parameters is the trace of a smoothing matrix. In fact, the mapping $G\hat{\mathcal{S}} = \tilde{\mathcal{S}}$ is a type of generalized smoothing. It is known that the effective number of parameters to associate with such smoothing is the trace of the smoother matrix (see, for example, Hastie & Tibshinari, 1990, section 3.5; also see pp. 48-49).

3 Study design

The questions one might ask about inference performance under random effects models, and information one could desire, are quite general, certainly more so than the limited simulation evaluation presented here. To clarify the information we sought, we first present the questions we asked. Then we give the design aspects of the particular simulation study reported here.

3.1 Inference issues

- (Q1) What is the bias of $\hat{\sigma}^2$ (i.e. the solution of equation (2))? We address this in two parts because the truncation of negative values of $\hat{\sigma}^2$ back to zero will induce bias, especially when process variation is zero, even if the signed variance estimator is unbiased (which can be the case in variance components estimation).
- (Q1a) What is the bias of the signed estimator $\hat{\sigma}$?
- (Q1b) What is the bias of the zero-truncated process variance estimator? To explore this question we actually use $\hat{\sigma}$, i.e. the square root of the zero-truncated estimator (not possible to do with the signed estimator).
- (Q2) What is the relative frequency of cases wherein $\hat{\sigma}^2 < 0$?
- (Q3) What is the coverage of the nominal 95% confidence interval on estimated process variance? This can be done on the scale of σ as well as σ^2 ; we focus on σ . A secondary question is also considered: what is the frequency of coverage failures above and below the confidence interval? For example, here a coverage failure is said to be ‘above’ if the confidence interval is entirely above true σ , hence σ is below the lower confidence limit.

The next set of questions relate to inference about the individual S_i . These inferences are computed as conditional on the actual survival probabilities of a given case, but we are then evaluating their performance over the Monte Carlo trials used.

- (Q4) What is achieved coverage of the nominal 95% confidence interval on S_i , $i = 1, \dots, k$? The focus is on average coverage over all k survival probabilities (and whether coverage varies much by occasion, i). Based on the MLE under the time-specific CJS model, the interval used is $\hat{S}_i \pm 1.96\hat{se}(\hat{S}_i|\mathcal{S})$. For the shrinkage estimator the interval uses \hat{rmse} from equation (7), hence is $\tilde{S}_i \pm 1.96\hat{rmse}(\tilde{S}_i|\mathcal{S})$.
- (Q5) What are the relative lengths of these two confidence intervals? It suffices to compare ratios of average $\hat{rmse}(\tilde{S}_i|\mathcal{S})$ to average $\hat{se}(\hat{S}_i|\mathcal{S})$. An overall ratio is based on first computing the average of each quantity over all simulation trails, by occasion, for a set of conditions (occasions, releases, $E(S)$, p , and σ), then looking at

$$\text{grand average length ratio} = \frac{\sum_{i=1}^k \bar{r}mse(\tilde{S}_i | \mathcal{S})}{\sum_{i=1}^k \bar{s}e(\hat{S}_i | \mathcal{S})} \tag{9}$$

If needed, we can also look at length ratio by occasion:

$$\text{occasion } i \text{ average length ratio} = \frac{\bar{r}mse(\tilde{S}_i | \mathcal{S})}{\bar{s}e(\hat{S}_i | \mathcal{S})} \tag{10}$$

If the shrinkage estimator is superior, these ratios will be less than one while confidence interval coverage remains at, or near, the nominal level.

(Q6) What are the relative mean square errors of the MLE and shrinkage estimators? Theory says the shrinkage estimator has the smaller mean square error (MSE). For R Monte Carlo trials (under set conditions; r indexes trial), the estimated MSEs by occasion are

$$MSE(\hat{S}_i) = \frac{\sum_{r=1}^R (\hat{S}_{r,i} - S_{r,i})^2}{R}$$

$$MSE(\tilde{S}_i) = \frac{\sum_{r=1}^R (\tilde{S}_{r,i} - S_{r,i})^2}{R}$$

The ratio of interest is

$$\frac{\sum_{i=1}^k MSE(\tilde{S}_i)}{\sum_{i=1}^k MSE(\hat{S}_i)} \tag{11}$$

and less so, the by-occasion- i ratios,

$$\frac{MSE(\tilde{S}_i)}{MSE(\hat{S}_i)} \tag{12}$$

If the ratios given by equation (12) are stable over occasions we can focus on equation (11).

(Q7) On average, within trial, is the MLE or the shrinkage estimator closer to the survival probabilities, $S_{r,i}$, of that trial? The summary statistics used to investigate this question are

$$SSE_r(\hat{S}) = \sum_{i=1}^k (\hat{S}_{r,i} - S_{r,i})^2$$

$$SSE_r(\tilde{S}) = \sum_{i=1}^k (\tilde{S}_{r,i} - S_{r,i})^2$$

Of interest is the ratio

$$RSSE_r = \frac{SSE_r(\tilde{S})}{SSE_r(\hat{S})} \tag{13}$$

which will be less than one if the shrinkage estimator is better than the MLE. Questions (6) and (7) involve similar, but not identical summaries of results.

The final area of investigation concerns the performance of AIC model selection. The most fundamental issue was whether AIC for random effects (i.e. formula (8)) would perform satisfactorily. To answer the question we tabulate the performance

of AICc as regards selection among three models fit to each simulated data set. The models fit were $\{S_t, p_t\}$, $\{S, p_t\}$ (i.e. S_i constant), and the random effects model, which is properly considered intermediate between these two fixed effects models. Hence, the question is simply

(Q8) What is the performance here of AICc as regards these three models?

3.2 Simulation design

We simulated single-group live capture-recapture data of the CJS type (Lebreton *et al.*, 1992) as the basis to address the above questions (other choices are possible, such as dead recoveries). The Monte Carlo simulations were done as a factorial treatment design using five factors:

| | |
|--|--------------------------------|
| capture occasions ($t = k + 2$), | 4 levels (7, 15, 23, 31) |
| releases of new animals (u) on each occasion, | 2 levels (100, 400) |
| constant capture probability (p) on each occasion, | 2 levels (0.6, 0.8) |
| mean survival probability, $E(S)$, | 2 levels (0.6, 0.8) |
| process variation, σ , | 4 levels (0, 0.025, 0.05, 0.1) |

All 128 combinations ($= 4 \times 2 \times 2 \times 2 \times 4$) of these levels defined the points used in the design space. At each design point we simulated 500 independent data sets. For cases of $\sigma > 0$ the S_1, \dots, S_k were generated as a random sample from a beta distribution with mean $E(S)$ and variance σ^2 (such as was done in Burnham *et al.*, 1995, for random capture probabilities). On each occasion a fixed number of new 'animals' ($u_i \equiv 100$ or 400) were released into the population. Data sets were generated one at a time in SAS (SAS Institute Inc, 1985) and passed directly to MARK where three models were fit to each data set: $\{S_t, p_t\}$, $\{S, p_t\}$, and the random effects model, which also required the re-optimization of the model $\{S_t, p_t\}$ likelihood over p , at fixed \tilde{S} , as noted in Section 2.2.

A final comment on design, in case the reader wonders why the levels for occasion (t) were 7, 15, 23, 31. For this factor we focused on the number of estimable S_t , which is $k = t - 2$, under the time-specific model, $\{S_t, p_t\}$. We first decided on 5 as our minimum for k ; and we wanted about 30 for our maximum k . Next we decided to use four levels. Given these choices, an increment to k of 8 was selected. Hence, levels on k are 5, 13, 21, 29, or in terms of occasions t , levels are 7, 15, 23, 31.

3.3 Maximum likelihood estimation

A critical point here is that all the MLEs were computed without being constrained to be ≤ 1 . This was done simply by using the identity link in MARK along with the standard likelihood for the CJS time-specific model. If the logical constraint $\hat{S} \leq 1$ is imposed it will have very undesirable effects on the numerically estimated Fisher information matrix, F , if any instance of $\hat{S}_j = 1$ occurs. The basic problem is that now the partial derivatives with respect to S_j end up being very wrong (usually 0 if a logit link is used). The resultant numerically estimated sampling variances are biased low. Of course now the variation in the set of MLEs is also reduced by the imposed constraint. One might think everything would work out all right in the end, but exploratory results showed this hope to be false.

Applying the variance-components methods given here to MLEs is not recommended when they are bounded at 1 and any of them fall on this bound. Note, however, that if all MLEs are well below 1 (on the scale of their standard errors), then doing the basic fitting for MLEs bounded, as with use of a logit link function, or unbounded gives the same results for random effects inferences.

4 Results

To present results, our strategy is to give an overall performance result first (e.g. over all 128 cases, if appropriate), by question given their ordering in Section 3.1. Then we partition the inferential statistic(s) by any important factors that affect the result. Major influences can be, for example, the value of process variation, σ , number of capture occasions, and number of releases at each occasion (100 or 400). The total number of Monte Carlo trials (64 000) was large enough that trivially small differences can be detected as ‘statistically significant’. Therefore, we usually do not present measures of precision on the summary means. In the spirit of first giving the big-picture result we note that the inference procedures examined (i.e. the questions in Section 3.1) performed very well.

First, we present information on the expected value of $\hat{\sigma}^2$, the signed estimator of σ^2 , which is the solution of equation (2). Table 1 gives simulation results for $\hat{E}(\hat{\sigma}^2)$ by σ^2 and occasions, t , the only factors having any noteworthy effects on the results. There is distinct positive bias when process variation is 0, but little or no bias in this study for $\sigma \geq 0.025$ (except at $k = 5$ and $\sigma = 0.025$). We had hoped the signed $\hat{\sigma}^2$ would be more nearly unbiased when true σ was at or near 0; instead it has positive bias.

However, the estimator to use in practice for a single data set would be $\max\{0, \hat{\sigma}^2\}$, for which there must be positive bias, at least at small values of σ . Therefore, we present in Table 2 results on bias of $\hat{\sigma} = [\max\{0, \hat{\sigma}^2\}]^{1/2}$. From Table 2, $\hat{\sigma}$ has little bias, in this study, when true $\sigma \geq 0.025$. An unbiased $\hat{\sigma}$ when true $\sigma = 0$ probably cannot be achieved.

Comparison of results in Tables 1 and 2, to the extent this is meaningful given the non-linearities involved, suggests zero-truncation is not a major cause of bias in $\hat{\sigma}$. This observation motivated us to produce Table 3, which gives the proportion of trials in which $\hat{\sigma}^2$ was negative. Especially when $\sigma = 0$, we thought perhaps the proportion of negative signed estimates would be much less than 0.5; it was not. However, the distribution $\hat{\sigma}^2$ is very asymmetric: there are not a lot of negative

TABLE 1. Mean value of the signed estimator of process variance, $\hat{\sigma}^2$, by true σ and number of capture occasions, $t = k + 2$, from simulation; means by σ^2 and t are based on 4000 trials; column means (i.e. by σ^2 only) are based on 16 000 trials

| t | True σ^2 (σ in parentheses) | | | |
|------|--|----------|---------|---------|
| | (0) | (0.025) | (0.05) | (0.1) |
| | 0 | 0.000625 | 0.00250 | 0.01 |
| 7 | 0.0001078 | 0.000761 | 0.00249 | 0.01015 |
| 15 | 0.0000446 | 0.000620 | 0.00253 | 0.01002 |
| 23 | 0.0000299 | 0.000637 | 0.00251 | 0.00999 |
| 31 | 0.0000273 | 0.000645 | 0.00251 | 0.00994 |
| mean | 0.0000524 | 0.000666 | 0.00251 | 0.01002 |

TABLE 2. Mean value of the $\hat{\sigma} = [\max(0, \hat{\sigma}^2)]^{1/2}$, by true σ and number of capture occasions, t , from simulation; means by σ and t are based on 4000 trials; column means (i.e. by σ only) are based on 16 000 trials

| t | True σ | | | |
|------|---------------|--------|--------|--------|
| | 0 | 0.025 | 0.05 | 0.10 |
| 7 | 0.00916 | 0.0275 | 0.0499 | 0.1007 |
| 15 | 0.00591 | 0.0249 | 0.0502 | 0.1001 |
| 23 | 0.00467 | 0.0252 | 0.0501 | 0.0999 |
| 31 | 0.00462 | 0.0254 | 0.0501 | 0.0997 |
| mean | 0.00609 | 0.0257 | 0.0501 | 0.1001 |

TABLE 3. Mean proportion of signed estimates, $\hat{\sigma}^2$, that are negative, by true σ and number of capture occasions, t , from simulation; means by σ and t are based on 4000 trials; column means (i.e., by σ only) are based on 16 000 trials

| t | True σ | | | |
|------|---------------|-------|-------|-------|
| | 0 | 0.025 | 0.05 | 0.10 |
| 7 | 0.592 | 0.339 | 0.159 | 0.026 |
| 15 | 0.516 | 0.144 | 0.022 | 0.001 |
| 23 | 0.510 | 0.076 | 0.004 | 0.000 |
| 31 | 0.484 | 0.047 | 0.002 | 0.000 |
| mean | 0.526 | 0.152 | 0.047 | 0.007 |

values that have large absolute values, where ‘large’ is with respect to the sampling distribution of the positive values of $\hat{\sigma}^2$. For the record we note here that the proper lower bound to enforce on $\hat{\sigma}^2$ is the negative of the smallest eigenvalue of $\hat{E}_S(W)$.

Next, we consider question (3) in Section 3.1: actual coverage of the nominal 95% confidence interval on σ^2 , or equivalently, on σ . The endpoints of this interval are computed based, on σ^2 , from formulae (4a) and (4b). Results are then zero-truncated; hence, we can take the square root of those endpoints, and coverage is the same whether considered for σ^2 or σ . Evaluated over all 64 000 computed confidence intervals on σ , coverage was 94.8%. Coverage averaged by σ is shown below (each mean is based on 16 000 trials), and separately by capture occasions, t :

| σ | %coverage | t | %coverage |
|----------|-----------|-----|-----------|
| 0.000 | 94.6 | 7 | 94.9 |
| 0.025 | 95.0 | 15 | 94.9 |
| 0.050 | 95.1 | 23 | 94.8 |
| 0.100 | 94.7 | 31 | 94.8 |

The other question about coverage is whether the 5% of intervals that failed to cover did so with equal ‘tails’. We therefore looked at the percentage of intervals that failed to cover because σ was below the lower limit of the interval (denoted as missed ‘above’), or σ was above the interval upper end point (missed ‘below’). The answer is simple, there was symmetry: 2.63% of misses were below and 2.53% of misses were above. Even for the 16 000 trials at $\sigma = 0$ the results were 2.95% of misses were below and 2.47% of misses were above.

The next set of results are about inference on individual survival rates, S_i (questions (4) to (7) in Section 3.1). We believed, and results support, that shrinkage estimates have smaller mean square error than MLEs under a random effects model. However, did we then suffer big losses in confidence interval coverage? Thus, coverage (question (4)) was paramount in our thinking, so we first present results for coverage of intervals computed, based on use of equation (7), as $\tilde{S}_i \pm 1.96 \hat{rmse}(\tilde{S}_i | \mathcal{S})$, and for MLE-based intervals as $\hat{S}_i \pm 1.96 \hat{se}(\hat{S}_i | \mathcal{S})$ (from the time-specific CJS model). Conditional on a given trial and occasion ($i = 1, \dots, k$), an interval ‘covers’ if true S_i for that trial is in the confidence interval. These are the type of conditional intervals one wants for the individual survival probabilities.

Confidence interval coverage results, averaged over all estimable survival probabilities (k) within case (500 trials) and then over all 128 cases, are 95.5% for the shrinkage estimator and 95.0% for the MLE under the time-specific CJS model. No design factors affected the coverage of the MLE-based interval; therefore, we do not look closer at its coverage results. Coverage of the confidence interval based on \tilde{S} was close to the nominal 95% except when $\sigma = 0$, wherein coverage was too high. Coverage was relatively lower for seven occasions, $k = 5$, compared to $k \geq 13$. Numbers of releases (u) had only trivial effects on coverage. As Table 4 shows, coverage based on the shrinkage estimator was essentially 95% for the cases examined here when $\sigma > 0$ and $k > 5$.

The only striking result is that coverage based on \tilde{S} is higher than nominal when $\sigma = 0$ (Table 4). This result is understandable from formula (7), wherein if $S_i \equiv \tilde{S}$ then shrinkage to the common mean is correct and the term $(\tilde{S}_i - \hat{S}_i)^2$ is neither needed nor appropriate. However, when $\sigma > 0$ (0.025 or more here), this added term is needed, even when the analysis in a given trial produces $\hat{\sigma} = 0$. For real data we do not know σ and observing $\hat{\sigma} = 0$ does not reliably mean $\sigma = 0$. Hence, we have here an issue in data analysis strategy. We could always use equation (7); this is probably a good strategy if true $\sigma > 0$. Or we could use equation (7) unless $\hat{\sigma} = 0$, in which case we then use the traditional $\hat{se}(\tilde{S}_i | \mathcal{S})$. We will comment more on this matter below.

Given that confidence interval coverage is good, we address question (5): what are the relative lengths of these two confidence intervals? First we look at results for the 128 cases based on formula (9), which is the ratio of overall average interval lengths (shrinkage versus MLE based). A ratio less than one favours the shrinkage-based method. The mean ratio over all 128 simulation cases is 0.842. However, this ratio in equation (9) is strongly affected by the value of σ :

TABLE 4. Average percentage coverage for nominal 95% confidence intervals on survival probabilities, S_i , based on the shrinkage estimator, \tilde{S}_i ; results here are averaged over i and are based on 4000 independent trials by combination of true σ and number of capture occasions, t , and on 16 000 trials for each column mean

| t | True σ | | | |
|------|---------------|-------|------|------|
| | 0 | 0.025 | 0.05 | 0.10 |
| 7 | 97.8 | 93.1 | 92.1 | 93.7 |
| 15 | 99.5 | 94.0 | 94.3 | 94.8 |
| 23 | 99.7 | 94.8 | 94.7 | 94.7 |
| 31 | 99.8 | 95.3 | 95.0 | 94.8 |
| mean | 99.2 | 94.3 | 94.0 | 94.5 |

| confidence intervals ratio | |
|----------------------------|-------------|
| σ | formula (9) |
| 0.000 | 0.739 |
| 0.025 | 0.799 |
| 0.050 | 0.879 |
| 0.100 | 0.949 |

There are no other factors that have effects worth noting on this ratio of average confidence interval lengths. There is a striking result here. Fitting the CJS model $\{S_i, p_i\}$ to data wherein the S_i did not vary over time ($\sigma = 0$), the confidence interval based on the shrinkage estimator had coverage $> 95\%$, the MLE-based confidence interval covered at 95% , and yet the confidence interval based on the shrinkage method was, on average, about 25% shorter than the interval based on the MLE. For the cases having $\sigma > 0$, coverage of both intervals was about 95% ; however, the confidence interval based on the shrinkage method was always shorter on average than that from the MLE analysis. Thus, the analysis strategy issue raised above is easily resolved as regard to the use of MLE versus shrinkage when a random effects model is a proper model to consider: random-effects based inference uniformly beat MLE in these simulations.

A few more analyses are useful. We considered the stability of the ratio of average confidence interval lengths by occasion, within the simulation case. Using formula (10) we computed this ratio of average lengths (from 500 trials). Then we computed the coefficient of variation of these k ratios, hence getting one number for each of the 128 simulation cases. If these CVs are small (5% would be small, we feel) then our analysis of confidence lengths based on equation (9) is sufficient to apply to all occasions, $i = 1, \dots, k$. The average of the 128 CVs is 2.0% , and the largest four CVs are (as %) $3.6, 3.7, 4.2$ and 6.0 . We conclude that the analysis of the ratio of average confidence interval lengths based on equation (9) is sufficient.

Next, we consider question (6): what are the relative mean square errors of the MLE and shrinkage estimators? First we used equation (11) to obtain ratios of average mean square errors over occasions within case; hence, we get one ratio for each of the 128 cases simulated. A ratio < 1 favours the shrinkage estimator. The average of those 128 ratios is 0.618 . The factor that has the biggest effect, by far, on this ratio is process variation, σ . Average results of equation (11) by σ are

| average MSE ratios | |
|--------------------|--------------|
| σ | formula (11) |
| 0.000 | 0.192 |
| 0.025 | 0.570 |
| 0.050 | 0.790 |
| 0.100 | 0.919 |

Because averages conceal as well as reveal we also used formula (12) to compute this mean square error ratio for all 2176 combinations of factors and occasions within case. Only three of these 2176 ratios exceeded 1: $1.003, 1.004$ and 1.02 . All other such ratios were less than 1. This is additional evidence of the superiority of the shrinkage estimator as compared with the MLE in this random effects context.

TABLE 5. The proportion of simulation trials that had $RSSE_r < 1$ (formula (13)); if proportion > 0.5 the shrinkage estimator is preferred to the MLE; results are based on 4000 independent trials by combination of true σ and number of capture occasions, t , and 16 000 trials for each column mean

| t | True σ | | | |
|------|---------------|-------|-------|-------|
| | 0 | 0.025 | 0.05 | 0.10 |
| 7 | 1.000 | 0.848 | 0.725 | 0.648 |
| 15 | 1.000 | 0.946 | 0.854 | 0.753 |
| 23 | 1.000 | 0.971 | 0.914 | 0.786 |
| 31 | 1.000 | 0.985 | 0.935 | 0.824 |
| mean | 1.000 | 0.937 | 0.857 | 0.752 |

The final question about shrinkage versus MLE is question (7): how do these two estimators compare within a data set, on average, in terms of the by-trial ratio of their SSE, $RSSE_r$ (formula (13))? The average of such ratios can be unstable, so for a basis of comparison we tabulated the proportion of the 500 simulation trials, by case, that had $RSSE_r < 1$. This is the same as the proportion of trials where $SSE_r(\tilde{S}) < SSE_r(\hat{S})$. A proportion between 0.5 and 1 favours the shrinkage estimator, and the closer the proportion is to 1, the more favoured is the shrinkage estimator. Averaged over all 128 design points, the proportion of cases wherein $SSE_r(\tilde{S}) < SSE_r(\hat{S})$ occurred was 0.887, and the minimum proportion over all 128 cases was 0.558. Results are strongly dependent on true σ and t (occasions), hence results averaged by these factors are given in Table 5. In terms of this measure of closeness of the estimator to true S_t , the shrinkage estimator wins handily as compared to the MLE under CJS model $\{S_t, p_t\}$.

The final area of inference explored here is question (8): what is the performance of AIC for random effects models (formula (8)) when the set of three models contains the random effects model and the two fixed effects models, wherein S_i either varies by occasion or is constant. For all three models, capture probability estimates were allowed to be fully time varying. Use of the random effects model for S means we use the shrinkage estimator, which we expect, and results here show, to be more parsimonious than the MLE under model $\{S_t, p_t\}$. Our most striking finding quite surprised us: AIC for the random effects model was *always* smaller than AIC for the ‘parent’ fixed effects model, $\{S_t, p_t\}$. ‘Always’, literally means here in all 64 000 simulated trials. (We expended great effort to be sure this was not the result of any programming error.)

We can denote the three AICc values computed here by $AIC(\cdot)$ for the constant S model, $AIC(RE)$ for the random effects model, and $AIC(t)$ for the time-varying S model. The difference $\Delta = AIC(t) - AIC(RE)$ had a minimum and maximum over all 64 000 trials of 0.022 and 47.938. Because AIC always selected random effects (i.e. \tilde{S}_t) over full time variation MLEs (i.e. \hat{S}_t) the only other issue is about AIC-based selection of the constant S model, $\{S, p_t\}$, versus the random effects model. Table 6 gives such selection results: how often the random effects model was selected rather than the simple time-constant survival probability model. Average results for this selection relative frequency by $\sigma = 0, 0.025, 0.05, 0.1$ are, respectively, 0.257, 0.736, 0.921, 0.987.

One motivation for wanting a random effects model for capture-recapture was the desire for a parsimonious model intermediate between models $\{S, p_t\}$, and

TABLE 6. The proportion of simulation trials in which AIC selected the random effects model rather than the constant S model (note, AIC always selected random effects over S time varying); results are based on 2000 independent trials for each combination of σ , t and μ , 8000 trials by σ and μ , and 16 000 by σ value (column means)

| t | u | True σ | | | |
|------|-----|---------------|-------|-------|-------|
| | | 0 | 0.025 | 0.05 | 0.10 |
| 7 | 100 | 0.303 | 0.445 | 0.653 | 0.914 |
| 7 | 400 | 0.374 | 0.760 | 0.911 | 0.992 |
| 15 | 100 | 0.228 | 0.550 | 0.892 | 0.994 |
| 15 | 400 | 0.350 | 0.892 | 0.996 | 1.000 |
| 23 | 100 | 0.186 | 0.632 | 0.948 | 1.000 |
| 23 | 400 | 0.255 | 0.954 | 1.000 | 1.000 |
| 31 | 100 | 0.158 | 0.677 | 0.967 | 1.000 |
| 31 | 100 | 0.201 | 0.976 | 1.000 | 1.000 |
| mean | 100 | 0.219 | 0.576 | 0.865 | 0.977 |
| | 400 | 0.295 | 0.896 | 0.977 | 0.998 |
| mean | | 0.257 | 0.736 | 0.921 | 0.987 |

$\{S_t, p_t\}$ when there clearly is time variation in survival probability, but no explainable structure to that variation. In particular, the situation arises where $\text{AIC}(\cdot) \doteq \text{AIC}(t)$ and yet using the simpler model, hence a single $\hat{S} \equiv \hat{\hat{S}}$, means foregoing separate estimates for $k - 1$ estimable survival parameters. The random effects model with shrinkage estimates provides the needed intermediate model.

As expected, there are conditions where $\text{AIC}(\cdot) \doteq \text{AIC}(t)$ but $\text{AIC}(\text{RE})$ is substantially less than these other values. Thus, the random effects model gets not only selected, but convincingly so in the sense that its Akaike weight is nearly 1 in this set of three models. In these same circumstances, when the model choice is restricted to the two fixed effects models, selection relative frequencies tend to be about 0.5 for each of those models. Table 7 gives cases from the simulation study where this scenario was common. These cases tended to be ones where process variance was about the same as mean sampling variance, $\sigma^2 \doteq \bar{\text{var}}(\hat{S} | \mathcal{S})$. A dramatic example is for the case having $t = 31$, $E(S) = 0.6$, $\sigma = 0.025$, $p = 0.6$, $u = 400$: $\bar{\text{AIC}}(\cdot) = 13.82$, $\bar{\text{AIC}}(t) = 13.71$, but $\bar{\text{AIC}}(\text{RE}) = 0.064$.

TABLE 7. Average ΔAIC results, over 500 trials, for some cases where $\Delta \text{AIC}(\cdot) \doteq \Delta \text{AIC}(t)$ but $\Delta \text{AIC}(\text{RE})$ is substantial lower

| t | $E(S)$ | σ | p | u | $\Delta \text{AIC}(\cdot)$ | $\Delta \text{AIC}(\text{RE})$ | $\Delta \text{AIC}(t)$ |
|-----|--------|----------|-----|-----|----------------------------|--------------------------------|------------------------|
| 15 | 0.6 | 0.025 | 0.6 | 400 | 5.33 | 0.084 | 7.24 |
| 15 | 0.6 | 0.050 | 0.6 | 100 | 5.19 | 0.095 | 7.26 |
| 15 | 0.8 | 0.025 | 0.8 | 100 | 5.41 | 0.097 | 7.05 |
| 23 | 0.6 | 0.025 | 0.6 | 400 | 9.62 | 0.070 | 10.29 |
| 23 | 0.6 | 0.050 | 0.6 | 100 | 8.95 | 0.124 | 10.79 |
| 23 | 0.8 | 0.025 | 0.8 | 100 | 10.39 | 0.103 | 9.91 |
| 31 | 0.6 | 0.025 | 0.6 | 400 | 13.82 | 0.064 | 13.71 |
| 31 | 0.6 | 0.050 | 0.6 | 100 | 12.64 | 0.131 | 14.45 |
| 31 | 0.8 | 0.025 | 0.8 | 100 | 15.13 | 0.077 | 12.93 |

5 Discussion

Using random effects as a basis for modelling collections of related parameters is a long-standing approach in statistics and one that can be very effective (see, for example, Link, 1999). Use of the random effects approach in capture-recapture has just started. We strongly support this new dimension of capture-recapture models (which can be likelihood, frequentist, Bayesian, or empirical Bayes). However, we also believe that the methodology needs to be better understood as to any potential pitfalls and as to its operating characteristics. While this thinking (e.g. what is an estimator's bias) is classically frequentist, it should apply as well to Bayesian approaches if our methods are to be considered scientific (see, for example, Dennis, 1996).

In the spirit of evidence, and the potential for dis-proof, we have herein done a small Monte Carlo evaluation of one approach to implementing random effects models. We are not aware of any other such evaluation of random effects modelling specifically in capture-recapture. The results (such as in Tables 1-7), taken as a whole, show the method performed quite well under the conditions of the study, especially when $\sigma^2 \geq 0.025$. Desirable performance characteristics (such as unbiased $\hat{\sigma}^2$) may be harder to achieve if σ^2 is 0, or quite small; this is an area where more methodology research could be done. However, we think it reasonable to believe that for a worthwhile study yielding good data, process variation, σ^2 , will not be too small, relative to average sampling variation and it is for these conditions (of 'good data') that we need effective random effects inference methods.

The simulations generated perfect data; in particular there was no overdispersion (see, for example, Lebreton *et al.*, 1992; Burnham & Anderson, 1998). In practice if there is overdispersion, as measured by a scalar often denoted by c , the estimated sampling variance-covariance must be adjusted by a reliable \hat{c} , to be not $\hat{E}_{\hat{S}}(W) = F^{-1}$, but rather $\hat{c}F^{-1}$. Franklin *et al.* (this issue) illustrates this practice with real data.

A key design feature to focus on for 'good data' when applying random effects is simply k , the number of estimable random effects parameters (could be sites instead of time intervals). The sample size for estimating σ^2 is k . Therefore, one must not have k too small; $k = 5$ ($t = 7$) is too small. Even if we knew all the underlying S_i a sample of size 5 is too small for reliable inference about the variation of these parameters (even if we had a random sample of them, which is not required here). Inference performance here was good when $k \geq 15$. We did not look at $k = 10$, but we would guess inference would be acceptable for $k \geq 10$. The benefits (includes shrinkage estimates) of random effects models become greater as the number of underlying parameters, k , increases.

The other influential basic design feature is animal sample size. Both numbers initially released (u_i) and numbers recaptured (these depended on S_i and p_i) are important to the performance of inferences from random effects models. We have no refined design guidelines here. We initially included in our design-factor for releases a level of $u = 25$. This proved to be too few animals to get useful random effects inferences: sampling variation was relatively so large that all too often the point estimate of σ^2 was 0. We also initially included levels for S and p of 0.4. We quickly found that these factor-levels, even combined with $u = 100$, too often did not lead to useful random effects inferences, again in the sense that too often $\hat{\sigma}^2 = 0$. This will not cause problems with real data; one simply will find that a model with constant S is best. But we wanted to focus on design points where

'interesting' results would occur for the random effects model, hence we eliminated some of the initial design points envisioned.

The situation where inferences from a random effects model are most advantageous seem to be for when σ^2 is about the same as average sampling variance, $\bar{var} = [\Sigma var(\hat{S}_i | S_i)]/k$ (note that sampling variance is much influenced by sample sizes of animals captured and recaptured, or recovered). If one or the other variance component dominates the total variation in the MLEs \hat{S}_i , then the data strongly favour either the simple model $\{S, p_i\}$ (\bar{var} dominates), or the general model $\{S_i, p_i\}$ (σ^2 dominates), rather than the random effects model. However, it is not a problem, as regards inference about σ^2 , to have large sample sizes of animals, hence small sampling variances, so that should be one's design goal. If it then turns out that sampling variance is similar to process variance, the random effects model will be quite superior to model $\{S_i, p_i\}$. Thus, in a sense the random effects model is optimal at the 'intermediate' sample size case. As sample size of animals increases, the random effects model converges to model $\{S_i, p_i\}$.

Our results show that the random effects inference methods evaluated here performed well. Hence, we do not further focus on the given results, but rather on issues we think are problematic. A key such issue is a boundary effect, at least under what is basically a likelihood approach. If one enforces $S \leq 1$ when the unbounded MLE \hat{S} exceeds 1 then standard numerical methods (as in MARK) used to get the observed information matrix fail. As a result, the estimated information matrix is incorrect for any terms concerning the \hat{S} that is at the bound of 1 (and the inverse information matrix is likely wrong in all elements). Experience shows that, in this case, the resultant point estimate of σ^2 can be very different from what one gets when the survival parameter MLEs are allowed to be unbounded. The difference can be substantial. We felt that since the bounded case can lead to biased (low) sampling variances (and biased \hat{S}), we would be better off allowing the MLEs used in the random effects formulae to be the unbiased ones that arise by using an identity link for S . By so doing we have found $\hat{\sigma}^2$ to be unbiased in most cases examined here (Table 1 or 2). Because the simulations are very time-consuming we did not also run them for the bounded case.

Note that we do not suggest routinely accepting final inferences that include survival estimates exceeding 1. In fact, the shrinkage estimates will generally not exceed 1, so using \tilde{S}_i not \hat{S}_i will be the needed improved inference. However, to get to this final inference it may be desirable to pass through an imaginary space ($S > 1$), just as imaginary numbers can facilitate real solutions to real problems (or just as conceptualizing a parameter as a random variable allows the power and beauty of the Bayesian approach); models only need to possess utility, not full reality.

It is common to use the logit link function when fitting, as here, generalized linear models when the parameters are probabilities. Moreover, it is common then to implement the random effects on the logit(S) scale, rather than on S directly. We did not do this for two reasons. First, to do so for this random effects modelling would, in effect, enforce the constraint $S < 1$ on the MLEs. As noted, boundary effect problems then arise with the empirical information matrix. Secondly, and more important, is that biologists want, and need, $\hat{\sigma}$ on the scale of survival probability, S (see, for example, White, 2000) to use in population modelling efforts. Basically, we need inferences about survival, much more than we need inferences about logit (survival). This latter reason is the important one since with good data we rarely observe an unbounded MLE of S that exceeds 1. (Note: MARK will do random effects on logit parameters.)

This boundary effect problem can arise unexpectedly, as with the ring recovery models parameterized as S and r (Seber, 1970; there, $\lambda \equiv r$), as opposed to S and f of Brownie *et al.*, 1985. In this formulation, the likelihood for every multinomial cell involves terms in $1 - S$. This effectively constrains $S < 1$ but can cause undesired boundary effects.

Might the full, proper Bayesian Markov chain Monte Carlo approach to random effects (see, for example, Brooks *et al.*, this issue; Royle & Link, this issue) eliminate such boundary effect biases? We do not know; we doubt it (of course, we are taking this from a frequentist operating characteristic viewpoint, which a Bayesian might disavow). We have programmed the simple MCMC random effects analysis and looked at some real data wherein the unconstrained MLE \hat{S} exceeds 1. It is easy in the MCMC analysis to allow S to have its distribution over an interval such as 0 to 2 (rather than 0 to 1). We did so and there is a strong effect of the upper bound on the point estimate (and entire posterior distribution) for σ^2 , and for that particular S . This has implications about operating characteristics of Bayesian random effects inference.

Since we think the operating characteristics of Bayesian methods should be documented, and we have software to do it, why did we not do the study? The simulations reported here took over 4 months of CPU time on what was then a state-of-the-art PC computer. The Bayesian MCMC analysis of a data set took about 100 times as much CPU time as the simpler MARK implemented analyses. Unless we can greatly speed up the MCMC analysis on a single computer, or use many computers in parallel, we might be looking at years to do the corresponding Bayesian simulation study.

Another issue to be aware of, as regards the parameter σ^2 , is the matter of distinctly unequal, rather than equal length, time intervals (such as 1 month periods in summer, 2 months in spring and fall, 4 months in winter). Let the time interval i have length Δ_i . Then we should parameterize the model as $S_i = (\psi_i)^{1/\Delta_i}$ where now each survival probability ψ_i is on the same unit time basis. It may then make biological sense to consider parameters that are a mean and variation for ψ_1, \dots, ψ_k . But this may just as well not make sense, because the time intervals are intrinsically not comparable as they may be in very different times of the annual cycle. It becomes a subject matter judgement as to whether random effects analysis will be meaningful with unequal time intervals. It might be necessary to use a more general fixed effects structural model, rather than $E(\psi_i) = \mu$, to allow for explainable temporal effects on survival (but for which we have no measurable explanatory covariates).

The name 'random effects' can be misleading in that a person may think it means that underlying years or areas (when spatial variation is considered, rather than temporal) must be selected at random. This is neither true, nor possible, for a set of contiguous years. Variance components is a better name, in that at the heart of the method is the separation of process and sampling variance components. The issue of what inferential meaning we can ascribe to $\hat{\sigma}^2$ is indeed tied to design and subject matter considerations. However, the shrinkage estimators do not depend on any inferential interpretation of $\hat{\sigma}^2$; rather, they can always be considered as improvements, in a MSE sense, over the MLEs based on full time-varying S_i .

The random effects model only requires that the conceptual residuals, $S_i - x_i'\beta$, are exchangeable. Hence, these residuals should appear like an iid sample; there should be no recoverable structural information left in them. There is no required distributional assumption, such as normality. Indeed, in our simulation, the S_i were

beta-distributed. The formulae used here (Section 2.1) are based only on first and second moments, except there is an assumption that $RSS(\sigma^2)$ (see text near formula (2)) has a central chi-square distribution. It seemed to be true enough to lead to good inference results here about σ^2 .

Consider the simple case examined in our simulations: $E(S) = \mu$. Given the weak assumption of exchangeability we are just assuming the time variation in the S_i appears as if it were totally random. Thus, a sufficient summary of this variation can be the single parameter, σ^2 . Essentially, all the information in S_1 to S_k can be collapsed into just two parameters, μ and σ^2 . This is exactly what the random effects model does, but in a heuristic sense, via the intermediary values of \tilde{S}_i . And it is correct to say that the random effects model is intermediate between the models with S constant and S as being unconstrained time-varying, regardless of issues about whether the underlying S_i are from a random sample of any sort. This latter issue is relevant to the meaning of any inference one wants to make about σ^2 (and μ), but these become context and subject matter issues. In general, since we cannot select years at random in capture-recapture studies, we have no option except to use such an estimated σ^2 for some sort of cautious, but assumption-based, inference about temporal variation in survival probabilities.

What of the future of multiple and generalized random effects in capture-recapture models, such as to males and females jointly with correlated random variation? It is bright, but probably not to be much found in moment-type equations such as in Section 2.2. We need easy ways to embed general and flexible random effects into extant capture-recapture models, without each time deriving estimators. There are two candidate approaches: Bayesian hierarchical modelling, and likelihood via h-likelihood (Lee & Nelder, 1996, 'h' stands for hierarchical). The Bayesian approach can be easily used (and has been) to produce results (see, for example, Royle & Link, this issue), but it is computationally intensive, which makes simulation evaluation of the method quite demanding. We think some such evaluations must be done because we have found random effects methods have pitfalls; we do not expect the Bayesian approach as such will sidestep these pitfalls. As for model selection, an AIC-like selection criterion does exist for Bayesian hierarchical models: DIC, deviance information criterion (see Spiegelhalter *et al.*, 1998). Finally, theoretically, the h-likelihood approach works, but it has not been tried yet for capture-recapture. It has the advantages of being much faster computationally and of potentially fitting easily into program MARK.

Meanwhile, this study provides the needed evidence that the simple but useful random effects model implemented in MARK can perform well.

Acknowledgements

The authors appreciate helpful comments on an earlier draft from Drs David Anderson, Doug Johnson, Byron Morgan, J. Andy Royle and an anonymous reviewer.

REFERENCES

- BARKER, R. J. (1997) Joint modeling of live-recapture, tag-resighting and tag-recovery data, *Biometrics*, 53, pp. 666-677.
- BROOKS, S. P., CATCHPOLE, E. A. & MORGAN, B. J. T. (2002) Bayesian methods for analyzing ringing data, *Journal of Applied Statistics*, this issue.

- BROWNIE, C., ANDERSON, D. R., BURNHAM, K. P. & ROBSON, D. S. (1985) *Statistical Inference from Band Recovery Data: a Handbook*, 2nd edn (Washington, DC, US Fish and Wildlife Service Resource Publication 156).
- BURNHAM, K. P. (1991) On a unified theory for release-resampling of animal populations. In: M. T. CHAO & P. E. CHENG (Eds), *Proceedings of the 1990 Taipei Symposium in Statistics*, pp. 11-35 (Taipei, Taiwan, Institute of Statistical Science).
- BURNHAM, K. P. & ANDERSON, D. R. (1998) *Model Selection and Inference: a Practical Information-Theoretical Approach* (New York, Springer-Verlag).
- BURNHAM, K. P., ANDERSON, D. R., WHITE, G. C., BROWNIE, C. & POLLOCK, K. H. (1987) *Design and Analysis of Fish Survival Experiments Based on Release-recapture Data* (American Fisheries Society Monograph 5).
- BURNHAM, K. P., ANDERSON, D. R. & WHITE, G. C. (1994) Evaluation of the Kullback-Leibler discrepancy for model selection in open population capture-recapture models, *Biometrical Journal*, 36, pp. 299-315.
- BURNHAM, K. P., ANDERSON, D. R. & WHITE, G. C. (1995) Selection among open population capture-recapture models when capture probabilities are heterogeneous, *Journal of Applied Statistics*, 22, pp. 611-624.
- CARLIN, B. P. & LOUIS, T. A. (1996) *Bayes and Empirical Bayes Methods for Data Analysis* (London, Chapman & Hall).
- CASELLA, G. (1985) An introduction to empirical Bayes data analysis, *The American Statistician*, 39, pp. 83-87.
- DENNIS, B. (1996) Discussion: should ecologists become Bayesians?, *Ecological Applications*, 6, pp. 1095-1103.
- EFRON, B. & MORRIS, C. (1975) Data analysis using Stein's Estimator and its generalizations, *Journal of the American Statistical Association*, 70, pp. 311-319.
- FRANKLIN, A. B., ANDERSON, D. R. & BURNHAM, K. P. (2002) Estimation of long-term trends and variation in avian survival probabilities using random effects models, *Journal of Applied Statistics*, this issue.
- HASTIE, T. & TIBSHIRANI, R. (1990) *Generalized Additive Models* (London, Chapman & Hall).
- HURVICH, C. M. & SIMONOFF, J. S. (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, *Journal of the Royal Statistical Society, Series B*, 60, pp. 271-293.
- JOHNSON, D. H. (1981) Improved population estimates through the use of auxiliary information, *Studies in Avian Biology*, 6, pp. 436-440.
- JOHNSON, D. H. (1989) An empirical Bayes approach to analyzing recurrent animal surveys, *Ecology*, 70, pp. 945-952.
- LEBRETON, J. D., BURNHAM, K. P., CLOBERT, J. & ANDERSON, D. R. (1992) Modeling survival and testing biological hypotheses using marked animals: case studies and recent advances, *Ecological Monographs*, 62, pp. 67-118.
- LEE, Y. & NELDER, J. A. (1996) Hierarchical generalized linear models, *Journal of the Royal Statistical Society, Series B*, 58, pp. 619-678.
- LINK, W. A. (1999) Modeling pattern in collections of parameters, *Journal of Wildlife Management*, 63, pp. 1017-1027.
- LINK, W. A. & NICHOLS, J. D. (1994) On the importance of sampling variation to investigations of temporal variation in animal population size, *Oikos*, 69, pp. 539-544.
- LONGFORD, N. T. (1993) *Random Coefficient Models* (New York, Oxford University Press).
- LOUIS, T. A. (1984) Estimating a population of parameter values using Bayes and empirical Bayes methods, *Journal of the American Statistical Association*, 79, pp. 393-398.
- MORRIS, C. N. (1983) Parametric empirical Bayes inference: theory and applications, *Journal of the American Statistical Association*, 78, pp. 47-65.
- ROYLE, J. A. & LINK, W. A. (2002) Random effects and shrinkage estimation in capture-recapture methods, *Journal of Applied Statistics*, this issue.
- SAS INSTITUTE INC (1985) *SAS Language Guide for Personal Computers*, Version 6 edn (Cary, North Carolina, SAS Institute).
- SCHOTT, J. R. (1997) *Matrix Analysis for Statistics* (New York, Wiley).
- SCHWARZ, C. J. & SEBER, G. A. F. (1999) Estimating Animal Abundance: Review III, *Statistical Science*, 14, pp. 427-456.
- SEARLE, S. R., CASELLA, G. & MCCULLOCH, C. E. (1992) *Variance Components* (New York, Wiley).
- SEBER, G. A. F. (1970) Estimating time-specific survival and reporting rates for adult birds from band returns, *Biometrika*, 57, pp. 313-318.

- SHI, P. & TSAI, C.-L. (1998) A note on the unification of the Akaike information criterion, *Journal of the Royal Statistical Society, Series B*, 60, pp. 551-558.
- SPEIEGELHALTER, D. J., BEST, N. G. & CARLIN, B. P. (1998) Bayesian deviance, the effective number of parameters and the comparison of arbitrarily complex models. Technical Report, MCR Biostatistics Unit, Cambridge, UK.
- VER HOEF, J. M. (1996) Parametric empirical Bayes methods for ecological applications, *Ecological Applications*, 6, pp. 1047-1055.
- WHITE, G. C. (2000) Population viability analysis: data requirements and essential analysis. In: L. BOITANI & T. K. FULLER (Eds), *Research Techniques in Animal Ecology: Controversies and Consequences*, pp. 288-331 (New York, Columbia University Press).
- WHITE, G. C. & BURNHAM, K. P. (1999) Program MARK—survival estimation from populations of marked animals, *Bird Study*, 46 (Supplement), pp. 120-138.
- WHITE, G. C., BURNHAM, K. P. & ANDERSON, D. R. (2002) Advanced features of program MARK. In: R. FIELDS (Ed), *Integrating People and Wildlife for a Sustainable Future, Proceedings of the Second International Wildlife Management Congress* (Bethesda, Maryland, The Wildlife Society).