Journal of Applied Statistics, Vol. 29, Nos. 1-4, 2002, 329-351

Routledge
Taylor & Francis Group

Random effects and shrinkage estimation in capture-recapture models

J. ANDREW ROYLE[1] **& WILLIAM A. LINK**[2], [1]*US Fish and Wildlife Service, Division of Migratory Bird Management and* [2]*US Geological Survey, Patuxent Wildlife Research Center*

ABSTRACT *We discuss the analysis of random effects in capture-recapture models, and outline Bayesian and frequentists approaches to their analysis. Under a normal model, random effects estimators derived from Bayesian or frequentist considerations have a common form as shrinkage estimators. We discuss some of the difficulties of analysing random effects using traditional methods, and argue that a Bayesian formulation provides a rigorous framework for dealing with these difficulties. In capture-recapture models, random effects may provide a parsimonious compromise between constant and completely time-dependent models for the parameters (e.g. survival probability). We consider application of random effects to band-recovery models, although the principles apply to more general situations, such as Cormack-Jolly-Seber models. We illustrate these ideas using a commonly analysed band recovery data set.*

1 Introduction

Capture-recapture methods are widely used in ecology to estimate important attributes of animal populations including survival, recruitment and mortality rates. Maximum likelihood estimation of parameters in a conventional 'fixed effects' framework, in which parameters are regarded as being unknown constants to be estimated, is standard practice for most capture-recapture problems.

There are many instances in which one would like to consider using random effects in capture-recapture models. First, random effects represent a parsimonious compromise between overly simplistic and more realistic complex models, but which may be difficult to fit adequately. For example, ecologists often consider the extremes of constant survival and year-specific survival models. Random effects lead to a flexible class of intermediate models.

Correspondence: J. A. Royle, US Fish and Wildlife Service, Division of Migratory Bird Management, 11510 American Holly Drive, Laurel MD 20708, USA

ISSN 0266-4763 print; 1360-0532 online/02/010329-23
DOI: 10.1080/02664760120108746

Second, shrinkage estimation is a natural consequence of a random effects assumption on a set of parameters. Shrinkage can lead to improved estimation (in a mean-squared error sense) of a collection of parameters, particularly when there is a large number of parameters in the model; we believe that shrinkage estimation has great potential for many problems, especially those involving small sample sizes and/or highly parameterized models. Third, ecologists are often interested in modelling 'pattern among parameters', for example, modelling trend or weather effects in survival rates. The dependence of parameters on covariates is commonly described using ultra-structural models, which specify deterministic relations among collections of related parameters. Thus, for instance, a collection of temporally varying parameters may be reduced to a pair of parameters, a slope and an intercept. This approach, while possibly reproducing the correct marginal mean structure, often fails adequately to capture variation in the data, leading to overdispersion relative to the multinomial assumption (e.g. Burnham & Anderson, 1998, p. 52). Random effects enable a generalization of ultra-structural models that accounts for this departure. Finally, there are many problems in which parameters are more naturally thought of as being generated from some probability distribution. These include models with spatially or temporally indexed parameters and models for individual heterogeneity. These are typical longitudinal data problems of the type that random effects methodology was developed to address. As their use for modelling longitudinal data suggests, random effects models are well suited for parameterizing correlation among observations.

While some effort has gone into the development of likelihood-based random effects models for capture-recapture data (e.g. Burnham, 2000), the current treatment of this problem is informal and we feel that a more rigorous, model-based framework is needed. In particular, two important issues that arise in random effects models are accounting for uncertainty in estimation of the parameters of the random effect distribution, and analysis of non-normal random effects models such as the multinomial model used in capture-recapture settings. These are both difficult problems within a conventional likelihood-based framework. On the other hand, a Bayesian framework naturally deals with these issues in a concise and unified framework. The Bayesian analysis of data on marked animals has started to receive considerable attention in the literature with the advent of practical computing methods. Two recent examples are Brooks *et al.* (2000a,b).

1.1 Of eggs and omelettes

The distinction between Bayesian and frequentist approaches to dealing with random effects may be summarized with respect to the two defining characteristics of the Bayesian paradigm: (1) the treatment of parameters as random; and (2) inference based on the posterior distribution. In this regard, there is an old metaphor attributed by Morris (1983, rejoinder) to Savage (1961) which refers to considering parameters to be random as 'breaking the Bayesian egg' and posterior inference as 'enjoying the omelette' that results from breaking that egg. Morris summarized major statistical theories in terms of the egg metaphor as follows:

	Enjoys omelette	Does not enjoy omelette
Breaks egg	Bayesian	Empirical Bayesian
Does not break egg	Fiducialist	Frequentist

In practical terms, and in the present context, 'enjoying the omelette' means exploiting the Bayesian framework to aid in dealing formally with uncertainty in prior parameters and the analysis of non-normal models with random parameters. That is, since all unknowns are treated as random, the machinery of probability calculus provides a consistent and rigorous framework within which to convert likelihoods and priors, to posteriors. In contrast, we will see that a frequentist (at least partially) breaks the egg, but because some parameters remain fixed, and because frequentists do not adopt posterior inference, these two problems are difficult. Of the two remaining cells in this table, fiducial inference is not relevant to our discussion. On the other hand empirical Bayes (see Carlin & Louis, 1996, ch. 3) has considerable relevance as classical frequentist methods for analysing random effects may be viewed as such (Laird & Ware, 1982), due to the manner in which the fixed parameters are estimated. We discuss this in Section 3.2.

1.2 Overview

In this paper, we provide a general discussion of random effects and associated technical issues within both Bayesian and classical frameworks.

The first half of this paper is essentially background material, describing various strategies for analysing random effects. Several of these approaches are Frequentist (i.e. treat some or all of the unobservable quantities as 'fixed but unknown' quantities, and base inference on the distribution of the data given the parameters). We also describe the Bayesian approach, in which all unobservable quantities are treated as random variables, and inference is based on the posterior distributions of the parameters given the data. Shrinkage estimators arise naturally as estimators of random effects under either paradigm.

In the second half of the paper, we address random effects models in the context of capture-recapture problems. In Section 5, we specifically deal with random effects in band recovery (or ring recovery) models, which are a special case of a more general class of capture-recapture models. However, the essence of the random effects model—a probability distribution on a collection of parameters such as yearly survival probabilities—is generally applicable. We illustrate these ideas using a waterfowl band-recovery data set in which year specific survival and recovery probabilities are parameterized as normal random effects, on the logit scale. Discussion and conclusions are given in Section 8.

2 Random effects in Bayesian analysis

Random effects are standard operating procedure in a Bayesian analysis since typically all parameters are regarded as random variables. The other important characteristic of Bayesian analysis is that inference is based on the conditional distribution of the parameters given the data, a quantity known as the *posterior distribution*. We will not go into much detail on Bayesian analysis here, instead relying on the simple normal-normal model to make some salient points. In particular, random effects provide a bridge between frequentist and Bayesian methods, in so far as they produce the same estimators when the parameters of the random effects distribution are known. Thus, we will see that interesting posterior quantities discussed here also arise from frequentist considerations in Section 3. The important point of this section is that, by assuming that all unknown quantities in a model are random, Bayesian analysis is able to exploit simple rules of probability calculus in order to provide a unified treatment not only of random effects estimation, but also of prior parameter estimation and accounting for uncertainty in those estimates.

We require some terminology and notation. We will use the notation $[\cdot]$ to refer to the 'distribution of ·', and indicate conditioning with '$|$'; i.e. the conditional distribution of x given y is $[x|y]$. The nominal Bayesian model in which we are interested consists of a data model (i.e. the likelihood) $[\mathbf{y}|\boldsymbol{\alpha},\psi]$, a *prior distribution* for the parameters $\boldsymbol{\alpha}$, $[\boldsymbol{\alpha}|\psi]$, and perhaps a further prior distribution for the remaining parameters ψ, $[\psi]$. For the purposes of this paper, we suppose that primary interest lies in estimating $\boldsymbol{\alpha}$ (these will be the 'random effects' parameters in our model). In our development, ψ will consist of both parameters in the random effects distribution, and additional parameters in the likelihood of $\mathbf{y}$, which are not explicitly accounted for by $\boldsymbol{\alpha}$.

The posterior distribution of the unknown parameters is then, by Bayes rule, equal to:

$$[\boldsymbol{\alpha}\psi|\mathbf{y}] = \frac{[\mathbf{y}|\boldsymbol{\alpha},\psi]\,[\boldsymbol{\alpha}|\psi]\,[\psi]}{[\mathbf{y}]} \tag{1}$$

where $[\mathbf{y}]$ is the marginal distribution of the data. The nature of the likelihood and prior distributions is obviously problem-specific, but the probability machinery required to compute the posterior is identical no matter the precise specification of the likelihood and prior(s) (a subtle but important point).

2.1 The normal-normal model

Consider the simple model $y_t|\alpha_t \sim \mathrm{N}(\alpha_t,\sigma_\varepsilon^2){:}\, t = ,2,\ldots,n$ with random effects α_t assumed to be $\mathrm{N}(\mu,\sigma_\alpha^2)$, Thus, $\psi = (\sigma_\varepsilon^2, \mu, \sigma_\alpha^2)$, and we will assume that ψ is known for the time being. The derived *precision* parameters, $\tau_\alpha = 1/\sigma_\alpha^2$ and $\tau_\varepsilon = 1/\sigma_\varepsilon^2$ will be convenient in some expressions. We might think of this model as applying to data collected over time, and thus α_t are 'year effects', y_t are yearly sample means, and our interest is an estimation of the collection of random effects $\{\alpha_t\}$.

The *posterior distribution* of α_t is proportional to the product of the likelihood and prior, which is easily shown to be:

$$[\alpha_t|y_t] = \mathrm{N}\left(\mu + \left(\frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\varepsilon^2}\right)(y_t-\mu), \frac{\sigma_\alpha^2\sigma_\varepsilon^2}{\sigma_\alpha^2+\sigma_\varepsilon^2}\right) \tag{2}$$

The mean of this posterior distribution is taken as an estimator of α_t and the posterior variance is then used to assess uncertainty about α_t. In general, the posterior is conditional on *all* data, so we might more formally express the posterior as $[\alpha_t|y_1,\ldots,y_n]$. However, with fixed ψ, the posterior of α_t depends only on y_t.

The posterior mean is often referred to as a *shrinkage estimator*, in the sense that the usual estimator of α_t, namely y_t (i.e. the yearly sample mean), is adjusted (or 'shrunk') back towards the prior mean of α_t; if the sample mean is larger than the prior mean, it is adjusted downward, and if it is smaller than the prior mean, it is adjusted upward. Clearly, the relative size of 'among year' versus 'within year' variation controls the amount of shrinkage being done via the *shrinkage weight* $c = \sigma_\alpha^2/(\sigma_\alpha^2+\sigma_\varepsilon^2)$. We discuss shrinkage estimation further in Section 3.3.

An important generalization arises by considering the model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{3}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are vectors of parameters and $\mathbf{X}$ and $\mathbf{Z}$ are known 'design' matrices. The distributional assumption on $\boldsymbol{\alpha}$ is typically made more general by allowing the αs to be correlated: $\boldsymbol{\alpha} \sim \mathrm{N}(\mathbf{0}, \sigma_\alpha^2 \mathbf{D}_\theta)$, where $\mathbf{D}_\theta$ is some correlation matrix with parameter θ. The model given by (3) forms the basis of the classical *linear mixed model* (LMM) (Robinson, 1991) wherein, to a frequentist, $\boldsymbol{\beta}$ are the 'fixed' parameters and $\boldsymbol{\alpha}$ are the 'random effects'. Typically then, the constant μ from (2) would be an element of $\boldsymbol{\beta}$ and regarded as fixed. Although Bayesians do not generally entertain the notion of fixed effects, placing a constant (i.e. uniform) prior on $\boldsymbol{\beta}$ has the effect of yielding results that are often consistent with frequentist procedures. Special cases of (3) are common throughout statistics, and include the usual random effects ANOVA, repeated measures ANOVA, spatial models that form the basis of *kriging* (in which case $\mathbf{D}$ is a spatial variance-covariance matrix), and many others. Under this slightly more general formulation, the posterior distribution of interest (which we omit) is analogous to (2) (see Laird & Ware, 1982).

2.2 Unknown prior parameters

Up to now, we have assumed that ψ is known. Consequently, the resulting posterior distributions were implicitly conditional on ψ. Such posterior distributions are usually called *conditional posterior* distributions. Although this may appear to be a gross simplification of real problems, it turns out that conditional posterior distributions play a very important role in simulation-based analysis of the posterior using Markov chain Monte Carlo (MCMC) methods (Gilks *et al.*, 1996a). Moreover, it is the mean of equation (2) that is often used as the frequentist random effects estimator, although derived from different considerations as we discuss in Section 3. Frequentist applications, however, suffer through various (often ad hoc) procedures to estimate ψ and account for that uncertainty. In the Bayesian framework, when ψ is unknown, it is endowed with a prior distribution, and the rules of probability are applied both to estimate it, and account for that uncertainty in a rigorous manner, as we now discuss.

If ψ is unknown with prior distribution $[\psi]$, then inference about it is based on the *marginal posterior* distribution $[\psi|\mathbf{y}]$. This is easily enough computed in most cases using standard methods. One is often interested in estimation of $\boldsymbol{\alpha}$, the random effects. Then, focus is on the marginal posterior distribution, $[\boldsymbol{\alpha}|\mathbf{y}]$, which is related to the *conditional posterior* distribution $[\boldsymbol{\alpha}|\mathbf{y},\psi]$ as follows:

$$[\boldsymbol{\alpha}|\mathbf{y}] = \int [\boldsymbol{\alpha}|\mathbf{y},\psi]\,[\psi|\mathbf{y}]\,d\psi \tag{4}$$

In words, the conditional posterior is averaged over ψ, weighted according to the posterior distribution of ψ. This integration problem is often analytically intractable and closed form expressions do not generally exist, *even under the normal-normal model* with unknown variance components. Nevertheless, we see from this last expression how the Bayesian approach accommodates 'estimation uncertainty' in prior parameters in a very formal fashion. Computation of interesting features of this marginal posterior distribution is easily accomplished by simulation-based MCMC techniques.

One useful case is to assume that μ in equation (2) is unknown, but not the variance components. With a flat prior on μ, the posterior mean in (2) has $\bar{y}$ in place of μ. This result also applies with $\boldsymbol{\beta}$ as in equation (3). Some difficulty arises when variance components are unknown. In general, a Bayesian analysis simply requires application of equation (4) in more generality, whereas frequentist approaches are less formal (as discussed in Section 3.2).

2.3 Non-normal models

The probability calculus employed in Bayesian analysis also facilitates the analysis of random effects in non-normal models. Obviously, in the general relation between likelihood, priors, and posterior, as expressed by equation (1), there is nothing restricting our likelihood to be normal, nor even our random effects distribution (the prior) to be normal. Models involving Poisson, Binomial, or, in the case of capture-recapture, Multinomial likelihoods, with (e.g. normal) random effects, are common in modern statistical practice. It is almost universally the case that these posterior distributions cannot be analysed directly, but with the recent advances made in numerical and simulation based methods, relevant features of the posterior distribution are easily computed. Very general MCMC algorithms (see Gilks *et al.*, 1996a) are easily applied to these problems. For example, the software package BUGS (Spiegelhalter *et al.*, 1996) implements a wide array of MCMC algorithms for many common classes of statistical models. Indeed, BUGS may be used to analyse data from studies of marked animals as Brooks *et al.* (2000b) illustrate.

3 Frequentist random effects and shrinkage

There are several different approaches for analysing random effects within the Frequentist paradigm. Although these may be perceived as being independent, the resulting random effects estimator under the normal model is a shrinkage estimator similar to that based on the conditional posterior distribution (that is, with fixed prior parameters). In the subsequent development, we will assume that these prior parameters are fixed, and discuss unknown prior parameters in Section 3.2.

3.1 Estimation of random effects

There is a rigorous *model-based* framework for random effects estimation within the frequentist paradigm. This is known as the best unbiased prediction (BUP). The best unbiased *predictor* (we will also use the acronym BUP for that, and let the context determine its meaning) is defined as the unbiased predictor with minimum variance (the frequentist convention is to use the term *predictor* for estimators of *random* effects, while reserving *estimator* for *fixed* effects). BUP is typically appealed to in time-series and spatial analyses, in which inference about future observables is of interest; Frequentists are comfortable regarding these quantities as random.

It is easy to establish the BUP directly. Let y be one or more realizations of a random variable (i.e. data). We wish to predict some quantity z based on y (z is typically an unobserved value of y, but could be a random effect in that model for y). Then, if $\tilde{z}(y)$ is any function of y, the minimizer of $E(z - \tilde{z}(y))^2$ (the mean-squared prediction error) is $\hat{z}(y) = E[z|y]$. In addition, $E[\hat{z}(y)] = EE[z|y] = E[z]$ (unbiased). Thus, the BUP is, in general, the conditional expectation of the thing to predict given the data. To a Bayesian, $E[z|y]$ is the posterior mean, but clearly it is reasonable without regard to one's philosophical beliefs (as the estimator which minimizes the mean squared prediction error). It is also clear that, under the normal-normal model introduced in Section (2.1), the posterior mean given by equation (2) (*with known* ψ) is the conditional expectation of α_t given the data, and hence the BUP. Thus, despite philosophical differences, frequentist and Bayesian alike might agree on the utility of BUP.

The BUP is the 'gold standard' for estimating random effects in classical statistics. As stated in the Introduction, the defining characteristics of the Bayesian paradigm are the treatment of parameters as random variables, and conditioning on data (i.e. posterior inference). Clearly both are present BUP-based estimations of random effects. However, the frequentist adoption of BUP retains the 'fixed but unknown' viewpoint on the prior parameters. Thus, while the Bayesian will integrate ψ from the conditional posterior as discussed in Section 2.2, the frequentist tendency is to use the estimator in its conditional (on ψ) form. This leads to some difficulty in its application, as we discuss in Section 3.2.

Finally, the BUP in the normal-normal case is linear, and consequently is also the best *linear* unbiased predictor (BLUP); i.e. optimal within a slightly less restrictive class of estimators. However, in general, the BUP is *not* linear (i.e. in most non-normal models). Also, if we plug-in MLEs for the fixed parameters (μ or more generally $\boldsymbol{\beta}$), than the 'BUP' is no longer BUP, but it remains BLUP, since the estimator of the mean parameters is linear in the data.

Model-free development. Estimation on the BUP makes use of distributional assumptions in so far as it is the conditional expectation under those assumptions. Alternatively, a model-free development proceeds by minimizing an expression referred to as a *penalized likelihood*, or penalized least-squares criterion. Since this is the classical derivation of random effects estimators in the linear mixed model setting (e.g. Laird & Ware, 1982), we adopt the general statement of the model given in equation (3) for this discussion. The penalized likelihood for the general model is:

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\alpha})'\tau_\varepsilon(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\alpha})+\tau_\alpha\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha} \tag{5}$$

Differentiating leads to the so-called *mixed-model equations* (e.g. see Robinson, 1991), which may be solved for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. In particular, the random effects estimator is a *conditional* posterior mean, as in (2), but under the more general model formulation given by (3).

This penalized least-squares criterion is just a variance weighted sum-of-squares, with the effect of forcing the αs to 0 (i.e. shrinking them). The penalty term $\boldsymbol{\alpha}'\mathbf{D}^{-1}\boldsymbol{\alpha}$ can be thought of as a roughness penalty, which effectively constrains the random effects to be 'smooth'. This penalized likelihood motivation is nice because it unifies a wide array of statistical procedures including thin-plate splines, kriging, ridge regression, mixed models, and others (for some perspective on this, see Nychka, 1998). Although the constrained optimization device provides a non-parametric justification for many techniques, it was originally justified informally from normal likelihood considerations (Henderson, 1950). Discussion of these various derivations can be found in Robinson (1991).

A similar distribution-free derivation of random effects estimators proceeds by explicitly minimizing the *mean squared prediction error* among linear predictors. This is more common in spatial statistics, for development of kriging estimators (for an example of this, see Cressie, 1991, p. 123). Burnham (2000) also presents that development.

Thus, while frequentist random effects estimation may be given a formal model-based development as a conditional expectation, estimators may also be derived within a 'distribution-free' framework (i.e. the solution to an optimization problem). Under the normal model with known variance components, the random effects estimator has a close correspondence to the mean of the (conditional) posterior in a Bayesian analysis.

Unbiased? The random effect estimator is often called the Best Linear *Unbiased* Predictor, but this is somewhat misleading. Strictly speaking, unbiased here means in a marginal sense, so that $E[\hat{\alpha}_t|\mu] = \mu + cE[y_t - \mu] = \mu + c0 = \mu = E[\alpha_t|\mu]$. This is not the same as $E[\tilde{\alpha}_t|\alpha_t] = \alpha_t$, which is the usual interpretation of an unbiased estimator. Clearly, $E[\hat{\alpha}_t|\alpha_t] = \mu + cE[y_t|\alpha_t] \neq \alpha_t$ *unless* $c = 1$. Thus, we see that the shrinkage estimator is, in fact, conditionally biased. This is generally the case for shrinkage estimators. Many statistical procedures accommodate a small amount of bias, in exchange for producing better estimators in terms of mean-squared error. This concept of a 'bias/variance trade-off' appears in smoothing methods (e.g. Hastie & Tibshirani, 1990), model selection (e.g. Burnham & Anderson, 1998, p. 23), and elsewhere.

3.2 Variance component estimation

In estimation of random effects, we have assumed that variance components $\sigma^2_\alpha, \sigma^2_\varepsilon$ are known. Formally, estimation of variance components usually proceeds by maximizing the marginal likelihood of $\mathbf{y}$ (that is, integrating the random effects from the conditional likelihood). Typically then, these estimates are used in the known-variance expressions arising from solving the mixed model equations, BUP considerations, etc. This has led to the terminology, 'plug-in' estimator, which is also often called the *estimated* best linear unbiased predictor, or EBLUP. The same approach also applies to estimation of other parameters in the specification of the random effects distribution. For example, one could assume that the random effects are correlated (the usual context in spatial statistics). That is, $\text{Var}(\boldsymbol{\alpha}) = \sigma^2_\alpha \mathbf{D}_\theta$ where $\mathbf{D}_\theta$ is a correlation matrix with parameter θ.

To illustrate, consider the following normal-normal model:

$$\mathbf{y}|\boldsymbol{\alpha} \sim \text{Normal}(\mathbf{Z}\boldsymbol{\alpha}, \sigma^2_y \mathbf{I})$$

and

$$\boldsymbol{\alpha} \sim \text{Normal}(\mu\mathbf{1}, \sigma^2_\alpha \mathbf{D}_\theta)$$

where $\mathbf{Z}$ is a design matrix as in (3). Then, the marginal distribution of $\mathbf{y}$ is:

$$\mathbf{y} \sim \text{Normal}(\mu\mathbf{1}, \sigma^2_y \mathbf{Z}\mathbf{Z}' + \sigma^2_\alpha \mathbf{D}_\theta)$$

The multivariate normal likelihood may be maximized to obtain estimates of $\psi = (\mu, \sigma^2_y, \sigma^2_\alpha, \theta)$. These may then be used in the known-ψ expression for $\hat{\boldsymbol{\alpha}}$, obtained by minimizing the penalized least-squares criterion. This is essentially the empirical Bayes approach to estimating prior parameters (Laird & Ware, 1982; Carlin & Louis, 1996, ch. 3).

There has been much discussion of this plug-in procedure (for discussion and references, see Laird & Ware, 1982; Christensen, 1991, p. 276; Robinson, 1991; Handcock & Stein, 1993), most having to do with the failure of this procedure properly to account for uncertainty associated with the variance component estimation in the variance of the prediction. Thus, the expected value of the estimated variance of the plug-in predictor will tend to be smaller than its true variance (which must be greater than the variance of the BLUP, by definition). Nevertheless, use of plug-in predictors is often justified as reasonable based on results by Kackar and Harville (1981) showing the plug-in predictor to be unbiased, apparently neglecting the fact that we primarily do statistics on the basis of our ability to quantify uncertainty.

Frequentists have proposed various corrections to account for estimation of the prior parameters (e.g. Kackar & Harville, 1984; Louis, 1984; Laird & Louis, 1987; Link & Hahn, 1996. However, we believe that the Bayesian paradigm provides a much more consistent approach to dealing with this issue, as was discussed in Section 2.

3.3 James-Stein shrinkage

Shrinkage estimators arise throughout statistics, and we have seen that they arise as a consequence of random effects estimation (or prediction) under either Bayesian or frequentist considerations. Such estimators have the desirable property that they perform better in a mean squared error sense, for estimating a collection of parameters. In fact, the frequentist random effects estimator is typically developed explicitly to *minimize* the mean squared prediction error. Consequently, shrinkage estimators tend to be (conditionally) biased, and may not possess a smaller variance than traditional component-wise estimators.

There is another important, 'true frequentist' derivation that leads to essentially the same estimator as that based on the conditional posterior mean, or from BUP considerations. This is known as James-Stein estimation, and is presented here for historical interest, and also because the main result of Stein (1955) provides the primary frequentist justification for shrinkage (i.e. not dependent on assuming parameters to be random). We emphasize that the James-Stein theory does not provide a framework for the analysis of random effects models (indeed, it is wholly unrelated to anything having to do with random effects, hence our labelling it a 'true frequentist' approach). Instead, it merely motivates interest in shrinkage estimation.

As before, assume that $y_t|\alpha_t \sim N(\alpha_t, \sigma_\varepsilon^2)$ for $t = 1, 2, \ldots, p$, with σ_ε^2 known. With no additional model structure imposed among the α_t, Stein (1955) showed that if $p \geqslant 3$, then the obvious estimator of the multivariate normal mean (i.e. the sample means) is *inadmissable* under mean-squared-error loss. Put another way, Stein showed that there exists an estimator that is uniformly better than the usual sample mean. While Stein did not provide such an estimator, this led to the well-known James-Stein estimator of α_t (James & Stein, 1961):

$$\hat{\alpha}_t^{js} = \left(1 - \frac{(p-2)\sigma_\varepsilon^2}{\Sigma y_t^2}\right)\bar{y}_t$$

This estimator looks like the posterior mean in (2), but with $\mu = 0$, and using a reasonable estimate for the 'total' variance, $\sigma_\varepsilon^2 + \sigma_\alpha^2$. To see this, note that $\sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\varepsilon^2) = 1 - \sigma_\varepsilon^2/(\sigma_\alpha^2 + \sigma_\varepsilon^2)$, and further note that $(p-2)/(\Sigma y_t^2)$ is unbiased for $1/(\sigma_\alpha^2 + \sigma_\varepsilon^2)$ under a normal model.

Thus, James & Stein seem to have proposed shrinkage towards 0 as a generally better estimator than the usual sample mean! An obvious extension of this is to shrink towards something other than 0, such as the 'overall' mean (Efron & Morris, 1975), i.e. the mean of the sample means y_t. Casella & Berger (1990, pp. 495-500) is a good introductory reference on 'Stein's Paradox', Stein estimation, and related issues. Carlin & Louis (1996, pp. 84-85) discuss Stein estimation in the context of empirical Bayes procedures.

Note that the James-Stein estimator is 'model-free' in the sense that there are no distributional assumptions imposed on α_t—it is not even random! What's more troubling is that there is nothing about the problem that requires that the y_t even be of the same 'type' of variable. For example, why not apply shrinkage to batting averages and pork belly prices (Efron & Morris, 1997)? This is in contrast to the linear mixed model development, in which the α_j are assumed to be random, typically normally distributed random variables. Because they are *assumed* to be generated from a *common* distribution, this would seem to mitigate concerns about combining batting averages and pork bellies. These ideas were concisely articulated by Robinson (1991):

> [The work by Stein (1955)] ... has led to some theoretical work that I believe to be of little practical value. This work is characterized by a tendency to combine unrelated estimation problems. BLUP helps us to know when to combine estimation problems. Situations where estimation problems ought to be combined are when the parameters to be estimated can be regarded as coming from some distribution. Equivalently, they are 'exchangeable', or are 'random effects'.

And so, batting averages and pork bellies ought not be related to each other, and hence should not be combined into a joint estimation problem. That is, shrinkage only makes sense given a random effects model, and there is little rational basis for one applying to both pork bellies and batting averages.

4 Synthesis

While we have given essentially a frequentist treatment of random effects estimation in Section 3, there are distinctly Bayesian aspects to that development. Clearly, the treatment of parameters as being random is inherently Bayesian ('breaking the egg'). In addition, consideration of the BUP approach invokes an explicit conditioning on data ('eating the omelette'). Indeed, under a normal-normal model, either penalized likelihood or BUP considerations yield estimators that are equivalent to that based on the conditional posterior distribution; i.e. with known prior parameters. Thus, despite philosophical differences between the Bayesian and frequentist schools of thought, there seems to be a correspondence between their respective solutions in normal random effects models.

How then are the Bayesian and frequentist approaches different with respect to random effects? The difference lies primarily in the way that ψ is dealt with. In essence, frequentist shrinkage is Bayesian, with known ψ. Frequentists most often use so-called 'plug-in' predictors and estimators. Bayesian use standard probability calculus and average over the posterior distribution of ψ. Thus, we feel that Bayesian analysis provides a more rigorous framework for accounting for uncertainty in parameter estimation.

Aside from the more formal treatment of ψ that Bayesian analysis permits, Bayesian analysis of random effects allows one to entertain much more complex models (e.g. non-normal) while not having to worry about how to derive estimators, since very general simulation-based (i.e. MCMC) algorithms are easy to adapt to complex problems. The key to Bayesian analysis of random effects is that one must only be able to compute the posterior distribution of the quantity of interest (the particular random effect), under whatever distribution is appropriate. This leads us to consideration of capture-recapture problems in which the likelihood is multinomial.

5 Models for band-recovery data

The field of capture-recapture encompasses a broad class of models for analysing data from marked animals and even a brief introduction is beyond the scope of this paper. Our primary interest is in band-recovery models, (in which there are no formal recaptures) which are a special case of the more general Cormack-Jolly-Seber (CJS) models (e.g. Lebreton *et al.*, 1992). We will briefly review the basic structure of band recovery models here (e.g. Brownie *et al.*, 1985). Ideas pertaining to random effects in band recovery models, which we will develop shortly, also apply to more general capture-recapture models. The common thread is that all of these models contain a collection of *survival* and *capture probability* parameters—the presence of recaptures is generally dealt with trivially.

5.1 Model structure

Band-recovery data may be conveniently summarized in terms of a *recovery matrix* containing the number of band recoveries over time. For example:

		Year								
Cohort	N bands	1	2	3	·	·	·	·	·	T
1	N_1	n_{11}	n_{12}	n_{13}	·	·	·	·	·	n_{1T}
2	N_2		n_{22}	n_{23}	·	·	·	·	·	n_{2T}
3	N_3			n_{33}	·	·	·	·	·	n_{3T}
·					·	·	·	·	·	
·						·	·	·	·	
·							·	·	·	
T	N_T									n_{TT}

In this table, N_i is the number of birds banded in year i (say, cohort i) and n_{ij} is the number of bands recovered from cohort i in year j. In waterfowl applications (such as ours), recoveries are bands returned as a result of *hunting activity* in year j, although there may be formal recaptures (these are simply moved to the next row of the table and treated as initial releases).

The usual assumption is that the vector of recoveries from each cohort is a multinomial random variable, and that the cohorts are independent of one another. For example, for cohort 1 we assume that

$$(n_{11}, n_{12}, \ldots, n_{1T}, N_1 - \Sigma n_{1j}) \sim MN(\pi_{11}, \pi_{12}, \ldots, \pi_{1T}, 1 - \Sigma \pi_{1j}, N_1)$$

where π_{ij} is the probability that a band from cohort i is recovered in year j. We will employ the conventional shorthand notation $[\mathbf{n}_i | \boldsymbol{\pi}_i, N_i]$ to represent this distribution. Assuming conditional independence among cohorts, the joint likelihood is merely the product of T such multinomial likelihoods:

$$[\mathbf{n} | \boldsymbol{\pi}, \mathbf{N}] = \prod_{i=1}^{T} MN(\mathbf{n}_i | \boldsymbol{\pi}_i, N_i)$$

Typically, the π_{ij}s are not of direct interest. Instead, they are assumed to be functions of more relevant parameters such as survival and reporting probabilities. Let λ_t be the *reporting* probability in year t (i.e. the proportion of dead birds that get reported), and ϕ_t be the *survival* probability (the probability that a bird alive at time t survives to time $t+1$). Then, the expected multinomial cell frequencies may be expressed as:

$$\begin{aligned} E[n_{11}] &= N_1\pi_{11} = N_1(1-\phi_1)\lambda_1 \\ E[n_{12}] &= N_1\pi_{12} = N_1\phi_1(1-\phi_2)\lambda_2 \\ E[n_{13}] &= N_1\pi_{13} = \cdot \\ &\quad\vdots \quad \vdots \\ E[n_{1T}] &= N_1\pi_{1T} = N_1\left(\prod_{t=1}^{T-1}\phi_t\right)(1-\phi_T)\lambda_T \end{aligned}$$

Thus, the multinomial cell probabilities are $\pi_{11} = (1-\phi_1)\lambda_1$, etc. One feature of these full year-dependent models is that ϕ_T and λ_T are confounded (see Seber, 1982, p. 241), so that only their product may be estimated. Consequently, the model contains $2T-1$ parameters, when the number of recovery years is equal to the number of cohorts. It is a simple matter to compute maximum likelihood estimates using standard software packages such as MARK (White & Burnham, 1999).

5.2 A posteriori *shrinkage of estimates*

One might consider BLUP-like shrinkage applied to a collection of MLEs, say $\hat\phi_1, \hat\phi_2, \ldots, \hat\phi_{T-1}$ (recall that there are only $T-1$ estimable survival parameters). If we specify the data model as:

$$\hat\phi_t \sim N(\mu_\theta, \sigma^2_\phi)$$

then we are naturally led to consider the estimator (Burnham, 2000):

$$\tilde\phi_t = \hat\mu_\phi + \frac{\sigma^2_\phi}{\sigma^2_\phi + \sigma^2_\varepsilon}(\hat\phi_t - \hat\mu_\phi) \tag{6}$$

where $\sigma^2_\phi = \mathrm{Var}(\phi_t)$ and σ^2_ε is the sampling variance and $\hat\mu_\phi$ is the sample mean of $\hat\phi_t$. This, of course, mimics the conditional posterior mean in (2), or the BLUP (using an estimate of μ_ϕ). Recall too that one may justify this without invoking a normal assumption. Burnham (2000) also considered another form of shrinkage estimator, where the shrinkage weight in equation (6) is replaced by its square-root. See Burnham (2000) for motivation of this form of shrinkage. The software package MARK (White & Burnham, 1999) implements this BLUP-like shrinkage based on MLEs.

To estimate variance components, one might use a number of reasonable choices in a 'plug-in' type estimator. Precisely which estimator to use in the plug-in procedure is a complex issue, since the obvious choice, namely the MLE based on the marginal distribution of the data, is difficult to compute (we illustrate this in Section 6.3), since the likelihood is multinomial, leading to a complex integration problem. Burnham (2000) suggests using moment estimators, or MLEs computed by regarding the $\hat{\phi}$s as 'data'. He also suggests accounting for sampling covariance in the $\hat{\phi}$s, replacing σ_ε^2 with a variance-covariance matrix, and applying the matrix equivalent of equation (6). In addition to some difficulty in estimation of prior parameters under this approach, it is unclear how to accommodate uncertainty associated with that estimation process, which may be particularly important in small sample problems.

This BLUP-like shrinkage approach may lead to reasonable answers, particularly if sample sizes are large. However, it is relatively straightforward to fit random effects models more formally within the multinomial framework, thus negating issues of prior parameter estimation and uncertainty.

6 Bayesian analysis of band-recovery data

Bayesian analysis of capture-recapture data, and band recovery data in particular, has been addressed by George & Robert (1992), Vounatsou & Smith (1995), Dupuis (1995) and recently by Chavez-Demoulin (1999) and Brooks *et al.* (2000a,b). Consequently, we do not feel it necessary to provide in-depth technical detail here. Instead, we present a brief overview, and refer the interested reader to the aforementioned references for further information, and computational details.

The first stage of the Bayesian model consists of the multinomial model given in Section 5 (i.e. the likelihood). The distinction between Bayes and classical modelling approaches is that the Bayesian model requires prior distributions for ϕ and λ, say $[\phi \mid \psi]$ and $[\lambda \mid \psi]$, which depend on parameters ψ.

Since survival and recovery parameters are probabilities, it is natural to model them on the logit scale. In the simplest random effects case, we assume that:

$$\text{logit}(\phi_t) = \alpha_t \qquad \text{and} \qquad \text{logit}(\lambda_t) = \gamma_t$$

where α_t is the survival 'year effect' and γ_t is the reporting rate year effect. One way to parameterize the random effects is to assume they are normally distributed:

$$\alpha_t \sim \mathrm{N}(\mu_\phi, \sigma_\phi^2) \qquad \text{and} \qquad \gamma_t \sim \mathrm{N}(\mu_\lambda, \sigma_\lambda^2) \tag{7}$$

Obviously, these models could be considerably more general, perhaps including covariates in the mean, etc. In this regard, we believe that it is most natural to parameterize variation in survival and recovery rate parameters on the logit scale (as in logistic regression). Note that if one desires estimates on the probability scale, that is $\phi_t = \text{expit}(\alpha_t)$ where $\text{expit}(x) = \exp(x)\,(1 + \exp(x))$, then this can be computed directly from the MCMC output for α_t as we discuss below. Instead of the logit parameterization, a Beta distribution on ϕ_t and λ_t also seems reasonable. We discuss this further in Section 6.2.

We now require prior distributions on the mean and variance parameters in the prior distributions given in (7) (i.e. ψ in our previous notation). Common non-informative priors for the mean parameters are Normal distributions with mean 0 and large variance (say 1000). For variance components, inverse-Gamma priors are used. Equivalently, assign Gamma(a,b) priors to the precisions (i.e. the inverse of the variances). One may reasonably specify non-informative priors by fixing a and b to be small, say 0.01 (using the common parameterization in Gelman *et al.*, 1995).

To accommodate the confounding of ϕ_T and λ_T in a Bayesian analysis, we could fix one or the other, say $\phi_T = 1$, and then specify a prior distribution for the remaining parameter (e.g. a reasonably non-informative prior is logit(λ_T) ~N(0,10)). This acknowledges that it is distinct from those that are assigned the random effects distribution (in essence, this Tth parameter is a pork belly, whereas the remaining are batting averages). One might also consider placing a uniform prior on the product $\phi_T\lambda_T$.

We seek to describe features of the marginal posterior distributions:

$$[\boldsymbol{\alpha}|\mathbf{n}] \propto [\mathbf{n}|\boldsymbol{\alpha},\boldsymbol{\gamma}]\,[\boldsymbol{\alpha}|\boldsymbol{\psi}]\,[\boldsymbol{\psi}]$$

and

$$[\boldsymbol{\gamma}|\mathbf{n}] \propto [\mathbf{n}|\boldsymbol{\alpha},\boldsymbol{\gamma}]\,[\boldsymbol{\gamma}|\boldsymbol{\psi}]\,[\boldsymbol{\psi}]$$

which are products of the multinomial likelihood, normal random effects distribution, and various prior distributions specified for $\boldsymbol{\psi}$. These posterior distributions may be analysed using MCMC techniques, as illustrated by Vounatsou & Smith (1995), and Brooks *et al.* (2000a,b). Gilks *et al.* (1996a) is a good general reference. The details behind MCMC are beyond the scope of this paper, but MCMC for the analyses reported on in the following sections is very easy to implement. We give a brief sketch of the algorithm we employed.

One form of MCMC, known as component-wise Metropolis-Hastings, involves sampling from the *full-conditional distribution* of each parameter using the Metropolis-Hastings algorithm (Gilks *et al.*, 1996b). For example, the full conditional distribution for the first survival year effect, α_1, is the product of the multinomial likelihood with the normal prior distribution:

$$[\alpha_1|\cdot] \propto [\mathbf{n}|\boldsymbol{\alpha},\boldsymbol{\gamma}]\,[\alpha_1|\boldsymbol{\psi}] \tag{8}$$

The parameter α_1 does not appear in any of the expected cell frequencies beyond those of banding cohort 1, and so this reduces to:

$$[\alpha_1|\cdot] \propto [\mathbf{n}_1|\boldsymbol{\alpha},\boldsymbol{\gamma}]\,[\alpha_1|\boldsymbol{\psi}] \tag{9}$$

where $\mathbf{n}_1$ is the vector of returns for the first banding cohort. As it turns out, this distribution is not of a convenient form from which to sample, but use of Metropolis-Hastings is relatively simple and straightforward. This proceeds by drawing a candidate value from some proposal distribution, and accepting that value with some prescribed probability as described in Gilks *et al.* (1996b).

The whole set of full conditionals (one for each unknown, including the recovery rate parameters) is sampled from many times. The resulting output is then used to estimate features of the relevant posterior distribution. For example, the mean of $[\alpha_1|\mathbf{n}]$ is estimated with the mean of the posterior simulated values of α_1 generated from the MCMC algorithm. One nice aspect of MCMC is that posterior quantities of a function of model parameters may be estimated by applying that function to the MCMC samples. For example, the posterior mean of ϕ_1 may be estimated as the mean of expit$(\alpha_1^{(m)})$: $m = 1, 2, \ldots, M$ where $\alpha_1^{(m)}$ is the mth simulated value of α_1.

There are many technical issues having to do with assessing convergence, choosing reasonable proposal distributions, starting values, etc. We are not concerned with these here, although they are important in any analysis. The interested reader should consult Gilks *et al.* (1996a) for details.

Although we did our own programming for the analyses presented in Section 7, we were also able to duplicate our results using the software package BUGS

(Spiegelhalter *et al.*, 1996) quite easily. Although there are many ways to implement these models in BUGS, the most straightforward is to write out the expected cell frequencies directly, which can be easily mechanized for general problems. Consequently, the reader familiar with standard packages (e.g. SURVIV), which operate in this manner, should have little difficulty implementing them in BUGS.

6.1 Model selection and assessment

Model selection and assessment are of great practical importance in any application. Biologists often rely on AIC (Burnham & Anderson, 1998) to choose the best model from among several competing models. AIC is simple to implement and widely understood and accepted among biologists. On the other hand, Bayesian ideas such as Bayes factors and Bayesian model-averaging are not so easily implemented in complex models, and are unfamiliar to practitioners. Brooks *et al.* (2000b) discuss some of these ideas in the context of modelling animal survival.

One tool that shows promise towards simplifying Bayesian model selection is the Deviance Information Criterion (DIC), proposed by Spiegelhalter *et al.* (1998). Essentially, a Bayesian version of AIC, the DIC is easy to compute in most problems using standard MCMC output. The DIC is based on the posterior distribution of minus twice the log-likelihood (i.e. Bayesian deviance):

$$D(\boldsymbol{\phi},\boldsymbol{\lambda}) = -2\log[\mathbf{n}|\boldsymbol{\phi},\boldsymbol{\lambda}]$$

Denote the posterior mean of this quantity as $\bar{D}(\boldsymbol{\phi},\boldsymbol{\lambda})$, which is easy enough to compute from the MCMC output (retaining the simulated values of $\phi_t^{(m)} = \text{expit}(\alpha_t^{(m)})$ and $\lambda_t^{(m)} = \text{expit}(\lambda_t^{(m)})$). In addition to this measure of model fit, we require some measure of model complexity. Spiegelhalter *et al.* (1998) define the effective number of parameters as

$$p_D = \bar{D} - D(\bar{\boldsymbol{\phi}},\bar{\boldsymbol{\lambda}}) \tag{10}$$

where $D(\bar{\boldsymbol{\phi}},\bar{\boldsymbol{\lambda}})$ is minus twice the log-likelihood evaluated at the posterior means of $\boldsymbol{\phi}$ and $\boldsymbol{\lambda}$. They then define the DIC as

$$\text{DIC} = D(\bar{\boldsymbol{\phi}},\bar{\boldsymbol{\lambda}}) + 2p_D$$

which may be applied in a manner analogous to AIC (i.e. small values are better than large values). Spiegelhalter *et al.* (1998) provide the decision-theoretic justification for use of the DIC.

Another particularly simple tool for model assessment, is the *Bayesian p-value*. The basic idea is to compare the distribution of some fit statistic computed from the data to that from simulated data under the model. Similar distributions (a *p*-value near 0.5) suggests consistency of the data with that model. Extreme values suggest otherwise. While we provide both DIC and Bayesian *p*-value results in our analysis below, we refer the interested reader to Brooks *et al.* (2000b) for details on computation of Bayesian *p*-values. We followed the approach outlined by them for our analysis.

6.2 On prior distributions

The Beta model is appealing for capture-recapture problems since the survival and recovery parameters are probabilities. Thus, we might assume $\phi_t \sim \text{Beta}(a,b)$ (Burnham & Overton, 1978; George & Robert, 1992). Of course, one would generally desire prior distributions on a and b, and it is unclear how those may be chosen in general and, in particular, in a manner that facilitates modelling covariate effects (one benefit of the logit-normal parameterization). One possibility is to parameterize covariate effects in the mean, $a/(a+b)$, perhaps on the logit scale. A prior distribution for the variance, or surrogate (such as $a+b$, the *precision*) is also needed. We feel that there is generally less difficulty in modelling the prior parameters using the logit-normal model. In particular, modelling covariates is straightforward. For example, in addition to random year effects, survival might be parameterized to depend on one or more environmental covariates, $x_{1t}, x_{2t}, \ldots, x_{pt}$ as:

$$\text{logit}(\phi_t) = \alpha_t + \sum_{k=1}^{p} \beta_k x_{kt}$$

There are other considerations that motivate consideration of the logit-normal model. One important one is that it is easily generalized to many types of problems making use of a multivariate normal assumption on random effects. One application that we have investigated is the fitting of multispecies models wherein ϕ_{kt} is the survival probability of species k in year t. Then, defining $\text{logit}(\phi_{kt}) = \alpha_{kt}$, and $\boldsymbol{\alpha}_t = (\alpha_{1t}, \alpha_{2t}, \ldots, \alpha_{Kt})$, one may consider interspecies correlations by assuming:

$$\boldsymbol{\alpha}_t \sim N(\boldsymbol{\mu}, \Sigma)$$

Various parameterizations of Σ can be employed, either suggested by biology, or perhaps chosen from among standard parametric families. Such models may have applications in estimation of survival and other vital rates for rare species, when it is easier to mark one or more abundant species that may be related in terms of their variation in model parameters. Also, this model may make it feasible to exploit sparse banding databases that have been largely untapped for reasons of data scarcity (e.g. the vast banding record which results from US Fish and Wildlife Service banding activity). In essence, information on poorly sampled species is 'borrowed' from more abundance species.

Similar models for parameterizing dependence among λ and ϕ, and even a generalized cohort model, with dependence among sexes and age classes, may be employed. One other obvious extension is to accommodate autocorrelation among the year effects, perhaps by assuming that the α_t behave according to an autoregressive process.

6.3 Best unbiased prediction using MCMC

One could argue that the BLUP is reasonable for arbitrary (i.e. non-normal) problems, due to its model-free interpretation (and its convenience is hard to argue). Nevertheless, one might hope to construct better estimators in non-normal problems by employing the appropriate distributional assumptions (e.g. multinomial), particularly in small-sample problems, wherein no sense of normality (asymptotic, approximate, or otherwise) is reasonable. In particular, one might consider attempting to compute the BUP and, following the ideas of Section 3.2, use a 'plug-in' procedure, whereby estimates of prior parameters ψ are used in the known-ψ expressions.

As before, let $[\mathbf{y}|\boldsymbol{\alpha}, \psi]$ be the likelihood of data, $\mathbf{y}$ given *random* effects $\boldsymbol{\alpha}$, and parameters ψ; and, let $[\boldsymbol{\alpha}|\psi]$ be the random effects distribution. Assuming that ψ

is *known*, the distribution of $[\alpha|\mathbf{y},\psi]$ (i.e. the conditional posterior distribution, to a Bayesian), is

$$[\alpha|\mathbf{y},\psi] = \frac{[\mathbf{y}|\alpha,\psi]\,[\alpha|\psi]}{[\mathbf{y}|\psi]}$$

where $[\mathbf{y}|\psi]$ is the marginal distribution of the data. Then, following the frequentist framework for estimating random effects, our estimator of α is taken as the mean of this distribution (i.e. the BUP) but with a reasonable estimate in place of ψ (i.e. the Estimated BUP, or EBUP). Unfortunately, under the multinomial model, known-ψ expressions are unattainable. However, one may evaluate the plug-in BUP indirectly, using MCMC, as we now explain.

To compute the EBUP, we first require an estimator of ψ. The most obvious estimator of ψ is the MLE based on the marginal distribution $[\mathbf{y}|\psi]$. An adequate approximately to the true MLE, may be obtained as the marginal posterior modes (i.e. for each element) under a Bayesian model with flat priors on the elements of ψ. Thus, α is integrated out of the likelihood using MCMC. Although marginal modes are not precisely the MLEs, for low-dimensional ψ they tend to be similar (the point of this exercise is not to compute MLEs but to attempt BUP, and so we will neglect this minor detail). Then, to compute the EBUP, we fix those parameters at their estimated values and rerun the MCMC algorithm to yield estimates of the posterior means say, $[\alpha_1|\mathbf{n},\psi = \psi^{(mle)}]$. In essence, this is a plug-in empirical Bayes procedure—maximum likelihood estimation on prior parameters, followed by computation of the *best unbiased predictor* of the random effects by MCMC simulation. Consequently, this approach is entirely analogous to the usual, frequentist approach of estimating random effects, which was described in Section 3, except that now we are working in the context of a non-normal model, and estimating parameters (and predicting) using MCMC.

We see the difficulty in attempting to retain frequentist notions (i.e. 'fixed' prior parameters) in complex problems. While still failing to account for prior parameter uncertainty, we are faced with two computationally demanding exercises—computation of $\hat{\psi}$, and computation of the random effects estimates. Of course, a simpler unified treatment of the whole problem arises when a fully Bayesian framework is adopted. However, our outlining this approach is intended to demonstrate the impact of failing to account for prior parameter uncertainty. We will compare results computed under this approach to more formally Bayesian and frequentist procedures in Section 7.

7 Illustration: mallards from the San Luis Valley

We consider random effects models for year-specific survival and reporting rates, using band recovery data from adult male mallards banded in the San Luis Valley, CO (taken from Brownie *et al.*, 1985, Example 2.2a). There are 9 years of recoveries from 9 years of banding activity. In addition to the random effects model structure, we consider several other potential models in order to evaluate the utility of the random effects structure. In particular, we considered the sequence of models as follows (in order of increasing complexity, or number of parameters):

Model 0: Constant ϕ and λ.
Model 1: Random effects model on both logit(ϕ_t) and logit(λ_t).
Model 2a: Random effects model on logit(λ_t), flat priors on logit(ϕ_t).

Model 2b: Flat priors on logit(λ_t), random effects on logit(ϕ_t).
Model 3: Flat priors on both logit(ϕ_t) and logit(λ_t).

For the models containing random effects structure, the prior distributions were specified according to (7). Thus, for Model 1, our parameters of interest are the vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\phi}$, which consist of eight reporting-rate and eight survival-rate parameters, respectively. Due to the random effects structure, this model is expected to have fewer than 17 parameters (see Section 6.1). Also recall that the model contains an additional parameter consisting of the product of the last reporting and survival rates. Although present in the model, we omit details concerning that parameter from the following discussion of results. In a classical setting, the full year-dependent model would contain 17 parameters. To evaluate models more complex than that implied by the random effects structure, we assigned flat priors to ϕ and/or λ. For example, we might expect a flat prior on ϕ_t to correspond to a model with eight survival parameters. Flat priors on both ϕ_t and λ_t would produce the full 17 parameter model.

Estimation was by MCMC as outlined in Section 6. Bayesian deviance and p-values (see Section 6.1) were computed from the MCMC output. These are shown in Table 1. Convergence is always an issue in analyses based on MCMC. Based on the Gelman-Rubin (GR) convergence statistic (Brooks & Gelman, 1998), we generally observed rapid convergence *except* for Model 3. For this model, the Markov Chains for the reporting rate parameters were particularly ill-behaved, exhibiting very strong autocorrelation. However, the GR statistic did seem to indicate eventual convergence. To investigate this further, we explored other prior specifications (including highly informative ones), with only slight improvement. One consequence of this poor convergence for Model 3 is that the measure of DIC model complexity, estimated according to (10), was *negative*. Consequently, in calculation of the DIC, we used an effective number of parameters for this model of 17.

The results of Table 1 suggest that the three shrinkage models (1, 2a and 2b) are generally preferred, although there is certainly some ambiguity in the results, particularly among those three models. Owing to the non-linear nature of the model, shrinkage on one component induces some structure (shrinkage) on the other component, perhaps explaining the equivalent complexity of 1 and 2b, and the general similarity among all three random effects models. The effective number of parameters for Model 0 was estimated to be 3.51, which is considerably larger than the nominal number of parameters expected (i.e. 2). Subsequently, we will focus on estimates based on Model 1 to make a couple of salient points.

Since analysis of the model is based on a simulation of all unknowns in the model, one could spend a fair amount of space summarizing the output in various formats, such as history plots and histograms of the simulated values. Instead, we

TABLE 1. Bayesian p-value and DIC results for model set

Model	p-value	deviance	eff. df	DIC
0	0.043	8677.44	3.51	8680.96
1	0.448	8655.81	11.10	8666.91
2a	0.447	8656.43	12.64	8669.07
2b	0.444	8655.80	11.00	8666.80
3	0.362	8658.32	17.00	8675.32

will just present posterior means (point estimates) and standard deviations of the relevant parameters. For the mallard data, the estimated posterior means and standard deviations of the survival and recovery parameters are: $\mu_\phi = 0.547$, $\sigma_\phi = 0.3530$, $\mu_\lambda = -1.468$ and $\sigma_\lambda = 0.322$. One reviewer commented on the difficulty in interpreting parameter estimates on the logit scale. Notwithstanding the rationale for use of the logit parameterization given in Section 6, these estimates may be used to generate results that are interpretable on a probability scale. For example, simulate the sequence of $\alpha^{(i)} \sim \text{Normal}(\mu_\phi^{(i)}, \sigma_\phi^{(m)})$ where $\mu_\phi^{(m)}$ and $\sigma_\phi^{(m)}$ are the mth sample from the posterior distribution (i.e. the output from MCMC analysis). Then, characterize the distribution of $\phi = \text{expit}(\alpha)$ from the sequence $\{\text{expit}(\alpha^{(m)}){:}m = 1, \ldots\}$. In Bayesian parlance, this would constitute an estimate of the *posterior predictive distribution* for the survival rate in an unsampled year. Figure 1 shows this quantity using our MCMC output. The mean and standard deviation (on the probability scale) are 0.627 and 0.091. In this case, the mass of the distribution is not too near the boundary and so summarization by mean and standard deviation seems reasonable (and clearly a normal approximation to Fig. 1 is not unreasonable).

Irrespective of the scale of prior parameter estimates, interest does not typically focus on prior parameters. Instead, the 'random effects' model structure is employed in the name of parsimony, and/or to yield improved estimates of annual survival and reporting rate, which are more often the quantities of interest. Under our logit parameterization, one directly obtains estimates of the annual survival and reporting rate year effects, α_t, and γ_t. However, for practical reasons (e.g. harvest management applications), one would prefer estimates of the probabilities $\text{expit}(\alpha_t)$, $\text{expit}(\gamma_t)$. Thus, we present the estimates on this scale. (As pointed out in Section 6, one may summarize functions of the MCMC output for α_t and γ_t in order to estimate the posterior distribution of functions of those parameters.) Although summary by mean and standard deviation may not be entirely adequate for probabilities, space does not permit more adequate characterization of a large number of posterior distributions.

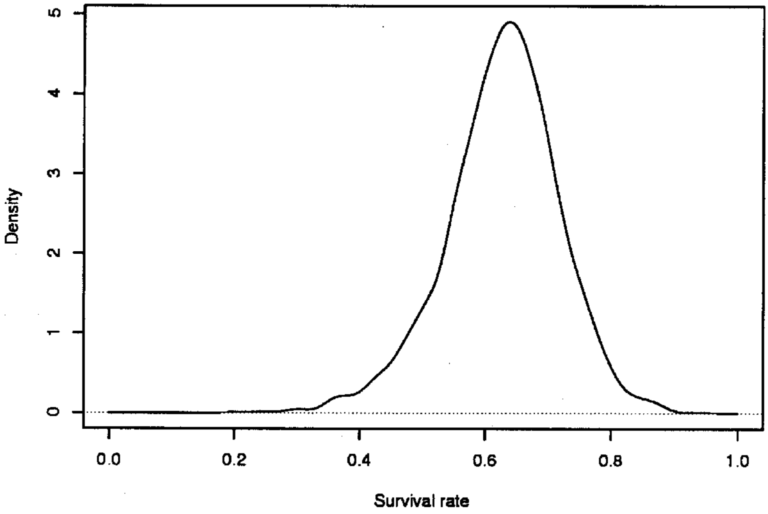

FIG. 1. Posterior predictive distribution of survival probability.

TABLE 2. Survival and reporting probability estimates under Bayesian, MLE and EBUP procedures for the Mallard data

Time	Bayes		MLE		EBUP	
period	mean	std	mle	SE	mean	std
Survival						
1	0.654	0.066	0.579	0.114	0.650	0.054
2	0.605	0.050	0.611	0.078	0.606	0.042
3	0.679	0.042	0.669	0.081	0.677	0.038
4	0.684	0.042	0.785	0.098	0.675	0.036
5	0.654	0.042	0.638	0.074	0.655	0.039
6	0.582	0.045	0.536	0.059	0.589	0.039
7	0.606	0.044	0.590	0.071	0.604	0.042
8	0.584	0.058	0.559	0.136	0.589	0.051
Reporting						
1	0.158	0.036	0.103	0.041	0.162	0.028
2	0.219	0.033	0.233	0.054	0.217	0.027
3	0.188	0.029	0.180	0.050	0.187	0.025
4	0.210	0.033	0.309	0.155	0.202	0.026
5	0.173	0.025	0.150	0.038	0.174	0.023
6	0.166	0.023	0.143	0.025	0.169	0.020
7	0.192	0.026	0.188	0.039	0.190	0.023
8	0.210	0.031	0.202	0.069	0.211	0.029

The posterior means and standard deviations of the year-specific survival rates are shown in Table 2 (columns 2-3). In addition to these Bayesian estimates, we present results from two other analyses. The estimates given in columns 4-5 are the MLEs computed using MARK (White & Burnham, 1999). These results illustrate clearly the shrinkage effect attained by treating the parameters as being random. That is, less variability than exhibited by the MLEs, and larger standard errors of the MLEs as compared with the posterior standard deviations. Since the MLE standard errors are very similar to posterior standard deviations of the estimates under a flat prior distribution, the difference may be interpreted as that induced by the additional model structure (i.e. what is, in essence, an informative prior). Interestingly, the Bayes estimate for ϕ_1 is larger than the mean survival, whereas the MLE is lower than the mean survival. Using the linear BLUP-based shrinkage approach, such behaviour is not possible (which is not to say that it is *desirable*, either). This is a consequence of random effects parameterization within the non-linear multinomial likelihood.

The third set of estimates (columns 6-7) is based on the EBUP procedure described in Section 6.3. While the actual philosophical underpinings of this procedure are somewhat ambiguous, we believe it is essentially a 'frequentist-like' solution to the problem, as it mimics the normal random effects procedures described in Section 3. Note that while the estimates are almost identical to the Bayes estimates, the important result is that the EBUP standard error *estimates* are always smaller—a consequence of failure to account for uncertainty in prior parameter estimation, and so we would tend to overstate our confidence in the results.

For completeness, estimates of λ_t are given in Table 2, although we omit discussion of these results.

8 Discussion and conclusions

Models containing random effects have proven useful in many ecological settings and we believe they have great promise in modelling data from studies of marked animals. Notions of improved estimation through shrinkage, modelling pattern in parameters, and accounting for correlation among model effects are all facilitated by consideration of random effects models.

Informal estimation of random effects within a capture-recapture framework may be carried out by applying BLUP-like estimators to collections of MLEs obtained using conventional methods. While this approach may perform well with large sample sizes, it has several important drawbacks. Notably, it fails to account for uncertainty in prior parameter estimates, and it is not generally suited for analysis of random effects in non-normal models. On the other hand, these are easily and rigorously dealt with within a Bayesian framework. This is because the probability calculus required to compute the posterior distribution of interest is the same no matter the context of the problem.

Adopting the random view of parameters has great promise, not only for conventional modelling applications, but also for the development of models that are largely intractable using conventional techniques, such as parameterization of correlation among effects, as mentioned in Section 6.2. We believe that an enhanced ability to fit models where the secondary model structure is imposed on parameters is the most important benefit of adopting a Bayesian framework. The classical random effects models are only one relatively simple class of such models. One final advantage of Bayesian analysis that we did not discuss in detail, is that Bayesian inference is *not* asymptotic, as are almost all likelihood-based procedures. Therefore, the subjective determination of whether or not asymptotic results apply is unnecessary, and honest accounting of uncertainty is not dependent on that determination.

While we feel that Bayesian analysis has important advantages in capture-recapture settings, the old adage that there is no such thing as a free lunch certainly applies. Importantly, even with recent advances made in Bayesian computation, there are no general software packages available for modelling data from marked animals that compare with those available for applying traditional methods (e.g. MARK; White & Burnham, 1999). However, computation in specific situations is easily accomplished using the popular program BUGS (Spiegelhalter *et al.*, 1996). Brooks *et al.* (2000b) provide the BUGS code required to fit certain types of band-recovery models. It is likely that many useful models may be fit using this software. Model assessment and selection are also important considerations in any application. Classical methods such as AIC (e.g. Burnham & Anderson, 1998) are simple to apply and widely understood among biologists, whereas most Bayesian methods are not (on either count). Nevertheless, there are many Bayesian methods that aid in such activities. Some of these, including DIC, and Bayesian p-values, are relatively straightforward to implement within an MCMC framework, and should facilitate adoption of Bayesian methods.

Acknowledgements

This work began as part of a short course on Bayesian methods presented by the authors at a USGS Cooperative Research Unit quantitative methods workshop in Estes Park, May, 2000. We thank Chris Jennelle for computing assistance. We also thank Professor Byron Morgan and two referees for many helpful comments and suggestions.

REFERENCES

Brooks, S. P. & Gelman, A. (1998) Alternative methods for monitoring convergence of iterative simulation, *Journal of Comp. Graphical Statist*, 7, pp. 434-455.

Brooks, S. P., Catchpole, E. A., Morgan, B. J. T. & Barry, S. C. (2000a) On the Bayesian analysis of ring-recovery data, *Biometrics*, 56, pp. 951-956.

Brooks, S. P., Catchpole, E. A. & Morgan, B. J. T. (2000b) Bayesian animal survival estimation, *Statistical Science*, 15(4), pp. 357-376.

Brownie, C., Anderson, D. R., Burnham, K. P. & Robson, D. S. (1985) *Statistical Inference from Band Recovery Data: A Handbook*, 2nd Edn (US Fish and Wildlife Service Resource Publication 156).

Burnham, K. P. (2000) On random effects models for capture-recapture data, *J. Agr. Biol. and Env. Stats.* (to appear).

Burnham, K. P. & Overton, W. S. (1978) Estimation of the size of a closed population when capture probabilities vary among animals, *Biometrika*, 65, pp. 625-633.

Burnham, K. P. & Anderson, D. R. (1998) *Model Selection and Inference: A Practical Information-Theoteric Approach* (New York, Springer-Verlag).

Carlin, B. P. & Louis, T. A. (1996) *Bayes and Empirical Bayes Methods for Data Analysis* (New York, Chapman & Hall/CRC).

Casella, G. & Berger, R. L. (1990) *Statistical Inference* (Pacific Grove, CA, Brooks/Cole).

Chavez-Demoulin, V. (1999) Bayesian inference for small-sample capture-recapture data, *Biometrics*, 55, pp. 727-731.

Christensen, R. (1991) *Linear Models for Multivariate, Time Series, and Spatial Data* (New York, Springer-Verlag).

Cressie, N. A. C. (1991) *Statistics for Spatial Data* (New York, Wiley).

Dupuis, J. A. (1995) Bayesian estimation of movement and survival probabilities from capture-recapture data, *Biometrika*, 82(4), pp. 761-772.

Efron, B. & Morris, C. (1975) Data analysis using Stein's estimator and its generalizations, *J. Amer. Statist. Assoc.*, 70, pp. 311-319.

Efron, B. & Morris, C. (1977) Stein's paradox in statistics, *Scientific American*, 236(5), pp. 119-127.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (1995) *Bayesian Data Analysis* (London, Chapman & Hall).

George, E. I. & Robert, C. P. (1992) Capture-recapture estimation via Gibbs sampling, *Biometrika*, 79(4), pp. 677-683.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (Eds) (1996a) *Markov Chain Monte Carlo in Practice* (London, Chapman & Hall).

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996b) Introducing Markov chain Monte Carlo. *In:* W. R. Gilks, S. Richardson & D. J. Spiegelhalter (Eds) *Markov Chain Monte Carlo in Practice*, pp. 1-19 (London, Chapman & Hall).

Hastie, T. J. & Tibshirani, R. J. (1990) *Generalized Additive Models* (London, Chapman & Hall).

Handcock, M. S. & Stein, M. L. (1993) A Bayesian analysis of kriging, *Technometrics* 35, 403-410.

Henderson, C. R. (1950) Estimation of genetic parameters, *Ann. Math. Statist.*, 21, pp. 309-310.

James, W. & Stein, C. (1961) Estimation with quadratic loss, *Proceedings of the Fourth Berkeley Symposium on Math. Statist. and Prob.*, 1, pp. 361-379 (Berkeley, CA, University of California Press).

Kackar, R. N. & Harville, D. A. (1981) Unbiasedness of two-stage estimation and prediction procedures for mixed linear models, *Communications in Statistics—Theory and Methods*, A10, pp. 1249-1261.

Kackar, R. N. & Harville, D. A. (1984) Approximations for standard errors of estimators of fixed and random effects in mixed linear models, *Journal of the American Statistical Association*, 79, pp. 853-862.

Laird, N. M. & Louis, T. A. (1987) Empirical Bayes confidence intervals based on bootstrap samples (with discussion), *Journal of the American Statistical Association*, 82, pp. 739-757.

Laird, N. M. & Ware, J. H. (1982) Random-effects models for longitudinal data, *Biometrics*, 38, pp. 963-974.

Lebreton, J. D., Burnham, K. P., Clobert, J. & Anderson, D. R. (1992) Modeling survival and testing biological hypotheses using marked animals: A unified approach with case studies, *Ecological Monographs*, 62(1), pp. 67-118.

Link, W. A. & Hahn, C. (1996) Empirical Bayes estimation of proportions with application to cowbird parasitism rates, *Ecology*, 77, pp. 2528-2537.

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(381), pp. 47-65.

NYCHKA, D. (1998) *Spatial Process Estimates as Smoothers. Smoothing and Regression. Approaches, Computation and Application*, M. G. SCHIMEK (Ed) (New York, Wiley).

ROBINSON, G. K. (1991) That BLUP is a good thing: The estimation of random effects (with discussion), *Stat. Sci.*, 6(1), pp. 15-51.

SAVAGE, L. J. (1961) The foundations of statistics reconsidered. *Proceedings of the Fourth Berkeley Symposium on Math. Statist. & Prob.*, 1, pp. 575-586 (Berkeley, CA, University of California Press).

SEBER, G. A. F. (1982) *The Estimation of Animal Abundance and Related Parameters*, 2nd Edn (London, Griffin).

SPIEGELHALTER, D. J., THOMAS, A. & BEST, N. G. (1996) Computation on Bayesian graphical models, *Bayesian Statistics 5*, pp. 407-425.

SPIEGELHALTER, D. J., BEST, N. G. & CARLIN, B. P. (1998) Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Research Report 98-009, Division of Biostatistics, University of Minnesota.

STEIN, C. (1955) Inadmissability of the usual estimator for the mean of a multivariate normal distribution, *Proceedings of the Third Berkeley Symposium on Math. Statist. & Prob.*, 1, pp. 197-206 (Berkeley, CA, University of California Press).

VOUNATSOU, P. & SMITH, A. F. M. (1995) Bayesian analysis of ring-recovery data via Markov Chain Monte Carlo simulation, *Biometrics*, 51, pp. 687-708.

WHITE, G. C. & BURNHAM, K. P. (1999) Program MARK: Survival estimation from populations of marked animals, *Bird Study*, 46, pp. 120-138 (supplement).