Bayesian Animal Survival Estimation

S. P. Brooks, E. A. Catchpole and B. J. T. Morgan

Abstract. We present the Bayesian approach to estimating parameters associated with animal survival on the basis of data arising from mark recovery and recapture studies. We provide two examples, beginning with a discussion of band-return models and examining data gathered from observations of blue winged teal (*Anas discors*), ringed as nestlings. We then look at open population recapture models, focusing on the Cormack–Jolly–Seber model, and examine this model in the context of a data set on European dippers (*Cinclus cinclus*). The Bayesian procedures are shown to be straightforward and provide a convenient framework for model-averaging, which incorporates the uncertainty due to model selection into the inference process. Sufficient detail is provided so that readers who wish to employ the Bayesian approach in this field can do so with ease. An example of BUGS code is also provided.

Key words and phrases: Band-return, Bayesian analysis, BUGS, capture–recapture, Cormack–Jolly–Seber, model averaging, product-multinomial, ring-recovery.

1. INTRODUCTION

The estimation of animal abundance, survival and capture–recovery parameters has attracted an enormous literature over the last 200 years, with the first such model being described in 1786 by Laplace, who was interested in estimating the population of France. Subsequently, models of this sort have found application in a wide range of settings, from the estimation of census undercount to the number of errors in computer code. However, one of the largest fields of application has been in the estimation of population size and survival probability of animals found in the wild. See, for example, the recent review by Schwarz and Seber (1999).

We concentrate here on survival probabilities, which are important demographic parameters, crucial for understanding and modelling population dynamics. Their behavior over time provides insight

S. P. Brooks is EPSRC Advanced Research Fellow, Statistical Laboratory, University of Cambridge, Cambridge CB30WB, England. E. A. Catchpole is Associate Professor, School of Mathematics and Statistics, University College UNSW, Australian Defence Force Academy, Canberra, ACT 2600, Australia. B. J. T. Morgan is Professor, Institute of Mathematics and Statistics, University of Kent at Canterbury, Canterbury, Kent CT2 7NF, England

into the possible effects of changes in climate and/or land-management practices.

Rather than adopting the classical maximum likelihood method to fit these models, we discuss how the Bayesian approach to model fitting may be used. Related Bayesian work has mainly focused on the estimation of population size; see, for example, Castledine (1981), Underhill (1990), Bolfarine, Leite and Rodriguez (1992), George and Robert (1992), Garthwaite, Yu and Hope (1995), Madigan and York (1997), and Lee and Chen (2000). Early discussion with regard to survival was provided by Janz (1980) and by Freeman (1990). Dupuis (1995) focuses on multiple-site recapture analysis, making use of data augmentation to account for missing records, and employing Gibbs sampling. Burnham (1999) provides an empirical Bayes approach to survival estimation, where appropriate parameters are regarded as random effects. Link and Cam (1999) give an accessible introduction to elementary Bayesian methods for mark-recapture methods, which complements the material of this paper. Finally, the paper by Vounatsou and Smith (1995), discussed critically in Brooks, Catchpole and Morgan (2000), compares Gibbs sampling with the Metropolis–Hastings algorithm for a range of models for band-return data. Vounatsou and Smith give little detail of the updating mechanisms needed for the Markov chain Monte Carlo (MCMC) simulation and no discussion of the more commonly used

capture–recapture modelling. Their work is greatly extended here.

We consider two common classes of models for estimating animal survival and show how the Bayesian approach may be applied in each case. While the basic MCMC methods of Gibbs and Metropolis–Hastings sampling need no description here (see, e.g., Besag, Green, Higdon and Mengersen, 1995; Gelman, Carlin, Stern and Rubin 1995; Brooks, 1997), we provide in Appendix A a description of some recent specific aspects of Bayesian analysis that are important for the modelling applications of this paper. The results which follow were obtained using Gibbs sampling and the ratio method for sampling from nonstandard conditional distributions.

2. BAND-RETURN MODELS

Many wildlife studies involve the analysis of band-return, also called ring-recovery, data. For example, newborn animals might be marked and released into the wild each year for a number of years and a record kept of the recoveries of marked animals that have died in each year. In this section we demonstrate how a Bayesian analysis of such data may be undertaken, focusing upon the estimation of survival and recovery probabilities.

2.1 Model Specification

Let us assume that we have data of the form m_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, $J \geq I$, where m_{ij} denotes the number of animals released at the beginning of year i and subsequently recovered (dead) in the 12 months up to the end of year j. We also have data R_i recording the number of animals marked and released at the beginning of year i.

We assume a model with the following parameters. Let λ_j denote the probability of a particular animal being recovered given that it died in the year up to the end of year j (we assume that recovery is immediate). Also let ϕ_i denote the probability that the animal survives to age i, given that it is alive at age $i - 1$. We allow for time-dependence in the survival rate of an animal in its first year of life and denote the probability of a particular animal surviving its first year, given that it is born in year i, by $\phi_{1, i}$.

This is known as the T/A/T model in the notation of Catchpole and Morgan (1996). Models are described by the triple $x/y/z$, where x, y, z indicate the modelling of first year survival probabilities, adult survival probabilities and reporting probabilities, respectively. Possible values for x, y and z include C, for constant, T, for time-dependent and A, for age-dependent.

We consider the T/A/T model initially. This is a useful and flexible model for band-return data, since many animals experience high first-year mortality which is influenced by time-varying conditions, such as weather. Reporting probabilities can also vary over time for similar reasons. Particular examples and data sets may require alternative models: for example, we may require some age dependence in the reporting probabilities to account for age-dependent behavior, such as breeding. But in many cases, simpler models, resulting from constraining the T/A/T model, will suffice, as we shall see below.

Table 1 provides the expected and corresponding observed values under the T/A/T model.

Given data $\{R_i, m_{ij}: i = 1, \ldots, I, \ j = i, \ldots, J\}$, we obtain the product-multinomial likelihood

$$(1) \qquad L(\boldsymbol{\phi}_1, \boldsymbol{\phi}, \boldsymbol{\lambda}; \mathbf{R}, \mathbf{m}) \propto \Delta \prod_{i=1}^{I} \prod_{j=i}^{J} p_{ij}^{m_{ij}},$$

where $\boldsymbol{\phi} = \{\phi_l, 2 \leq l \leq J\}$, $\boldsymbol{\phi}_1 = \{\phi_{1l}, 1 \leq l \leq I\}$ and $\boldsymbol{\lambda} = \{\lambda_l, 1 \leq l \leq J\}$. Here

$$(2) \qquad p_{ij} = \begin{cases} \lambda_i \tilde{\phi}_{1, i}, & j = i, \\ \lambda_j \phi_{1, i} \tilde{\phi}_{j-i+1} \prod_{k=2}^{j-i} \phi_k, & j = i+1, \ldots, J \end{cases}$$

denotes the probability associated with an observation in entry (i, j) of Table 1, where $\tilde{\phi}_l = 1 - \phi_l$, etc., and Δ denotes the likelihood term associated with unrecovered animals. If we let $q_i = 1 - \sum_{j=i}^{J} p_{ij}$ be the probability of nonrecovery of an animal released at the beginning of year i, either because it was still alive at the end of the experiment or because it died and was not found, and $u_i = R_i - \sum_{j=i}^{J} m_{ij}$ denote the number of animals released at the beginning of year i and never recovered, then $\Delta = \prod_{i=1}^{I} q_i^{u_i}$. Note that Δ is a function of all of the model parameters.

Throughout this paper, we follow the convention that a null sequence has sum 0 and product 1. Thus in formula (2) for p_{ij}, the product term is 1 when $j = i + 1$.

TABLE 1
The data structure for the T/A/T recovery model for a limited study of three years duration[1]

		$j = 1$	2	3	Never recovered
$i = 1$	R_1	$R_1 \tilde{\phi}_{1, 1} \lambda_1$ (m_{11})	$R_1 \phi_{1, 1} \tilde{\phi}_2 \lambda_2$ (m_{12})	$R_1 \phi_{1, 1} \phi_2 \tilde{\phi}_3 \lambda_3$ (m_{13})	$R_1 q_1$ (u_1)
2	R_2		$R_2 \tilde{\phi}_{1, 2} \lambda_2$ (m_{22})	$R_2 \phi_{1, 2} \tilde{\phi}_2 \lambda_3$ (m_{23})	$R_2 q_2$ (u_2)
3	R_3			$R_3 \tilde{\phi}_{1, 3} \lambda_3$ (m_{33})	$R_3 q_3$ (u_3)

[1]The entries denote expected numbers recovered each year. Below in parentheses are the observed values.

Having obtained the likelihood for this model, we now consider the prior distributions for the parameters. For simplicity, we take independent beta priors,

$$\phi_{1,l} \sim \text{beta}(\alpha_1, \beta_1), \quad l = 1, \ldots, I,$$

(3) $$\phi_l \sim \text{beta}(\alpha_\phi, \beta_\phi), \quad l = 2, \ldots, J,$$

$$\lambda_l \sim \text{beta}(\alpha_\lambda, \beta_\lambda), \quad l = 1, \ldots, J.$$

Note that none of these prior distributions depend on l. With relevant expert information, other priors might be more appropriate. For example, different priors could be used for each individual parameter, or we could impose some form of structure on the variables by putting a prior probability of zero on the event that $\phi_{1,i} > \phi_j$ for any i, j, for example. This is an important possibility, as many animals are known to have higher mortality during their first year of life. A simple way to achieve this in MCMC simulations is to discard all outcomes of the chain which do not satisfy the restrictions, thus implicitly redefining the prior.

In choosing between models, we shall consider submodels of the T/A/T model. In order to use Gibbs sampling, we need to calculate the full conditional posterior distributions for the cases in which first-year survival and reporting probabilities are time-varying or constant, and adult survival is age-varying or constant. Note that naïve use of the likelihood in (1) to compute the conditional posterior distributions commonly fails due to numerical underflow. For this reason it is helpful to carry out some of the calculation algebraically, reformulating the conditional posterior distributions as a product of a beta density and a nuisance term, as follows.

Given the likelihood function in (1) and the prior distributions in (3), the full conditional posterior for λ_l is given by

$$\pi(\lambda_l \mid \boldsymbol{\lambda}_{(l)}, \boldsymbol{\phi}_1, \boldsymbol{\phi}, \mathbf{R}, \mathbf{m})$$

$$\propto \lambda_l^{\alpha_\lambda - 1} \tilde{\lambda}_l^{\beta_\lambda - 1} \Delta \prod_{i=1}^{I} \prod_{j=i}^{J} p_{ij}^{m_{ij}}$$

$$\propto \lambda_l^{\alpha_\lambda - 1} \tilde{\lambda}_l^{\beta_\lambda - 1} \Delta \, \lambda_l^{\sum_{i=1}^{l^*} m_{il}}$$

$$\propto \Delta f_{\text{beta}}\left(\lambda_l; \, \alpha_\lambda + \sum_{i=1}^{l^*} m_{il}, \, \beta_\lambda\right), \quad l = 1, \ldots, J,$$

where $\boldsymbol{\lambda}_{(l)}$ denotes the vector $\boldsymbol{\lambda}$ with the lth component omitted, $l^* = \min(l, I)$, and

$$f_{\text{beta}}(x; \alpha, \beta) = x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \le x \le 1.$$

Note that for the submodel with $\lambda_l = \lambda$, for all l,

$$\pi(\lambda \mid \boldsymbol{\phi}_1, \boldsymbol{\phi}, \mathbf{R}, \mathbf{m}) \propto \Delta f_{\text{beta}}\left(\lambda; \, \alpha_\lambda + \sum_{i=1}^{I} \sum_{j=i}^{J} m_{ij}, \, \beta_\lambda\right).$$

Similarly, the full conditional posterior for ϕ_l is given by

$$\pi(\phi_l \mid \boldsymbol{\phi}_{(l)}, \boldsymbol{\phi}_1, \boldsymbol{\lambda}, \mathbf{R}, \mathbf{m}) \propto \Delta f_{\text{beta}}(\phi_l; \, \alpha_\phi + r, \, \beta_\phi + s),$$

$$l = 2, \ldots, J,$$

where $r = \sum_{i=1}^{l^\dagger} \sum_{j=i+l}^{J} m_{ij}$ and $s = \sum_{i=1}^{l^\ddagger} m_{i, i+l-1}$, with $l^\dagger = \min(I, J-l)$ and $l^\ddagger = \min(I, J-l+1)$. For the submodel with $\phi_i = \phi$ for all i,

$$\pi(\phi \mid \boldsymbol{\phi}_1, \boldsymbol{\lambda}, \mathbf{R}, \mathbf{m})$$

$$\propto \Delta f_{\text{beta}}\bigg(\phi; \, \alpha_\phi + \sum_{i=1}^{I^*} \sum_{j=i+2}^{J} (j-i-1)m_{ij},$$

$$\beta_\phi + \sum_{i=1}^{I^\dagger} \sum_{j=i+1}^{J} m_{ij}\bigg),$$

where $I^* = \min(I, J-2)$ and $I^\dagger = \min(I, J-1)$. Finally, the full posterior distribution for $\phi_{1,l}$ is

$$\pi(\phi_{1,l} \mid \boldsymbol{\phi}_{1,(l)}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \mathbf{R}, \mathbf{m})$$

$$\propto \Delta f_{\text{beta}}\bigg(\phi_{1,l}; \, \alpha_1 + \sum_{j=l+1}^{J} m_{lj}, \, \beta_1 + m_{ll}\bigg),$$

and, for the submodel with $\phi_{1,i} = \phi_1$ for all i,

$$\pi(\phi_1 \mid \boldsymbol{\phi}, \boldsymbol{\lambda}, \mathbf{R}, \mathbf{m})$$

$$\propto \Delta f_{\text{beta}}\bigg(\phi_1; \, \alpha_1 + \sum_{i=1}^{I} \sum_{j=i+1}^{J} m_{ij}, \, \beta_1 + \sum_{i=1}^{I} m_{ii}\bigg).$$

Note that each of the conditional posterior distributions is nonstandard, because of the Δ term. The formulation of the posterior conditionals as being proportional to the product of a beta distribution and Δ has computational advantages, since the product of the $p_{ij}^{m_{ij}}$ in (1) is typically very small, which can cause accuracy problems, whereas the forms we have given for the posteriors are much larger, because of the implicit inclusion of the normalization constant of the beta density.

2.2 Example: Blue-Winged Teal

Brownie, Anderson, Burnham and Robson (1985) describe a band-recovery study of blue-winged teal (*Anas discors*) ringed as nestlings in Saskatchewan. The data are given in Table 2. We use this data set since it is sparse, and therefore provides a good illustration of the influence of the prior for certain models, as we shall see later. Freeman and Morgan provide an analysis of these data, and give the maximum likelihood estimates of the parameters for a number of different models.

Figure 1 provides an illustrative example of the MCMC output for three parameters under the C/C/T model. Here, we took independent uniform priors

TABLE 2
Recoveries of blue-winged teal marked as young from 1961 *to* 1973[1]

Year of ringing	Number ringed	Year of recovery (1961+)											
		1	2	3	4	5	6	7	8	9	10	11	12
1961	910	6	2	1	1	0	2	1	0	0	0	0	0
1962	1157		11	5	6	1	1	1	1	0	0	0	1
1963	1394			19	4	4	4	0	0	1	1	0	0
1964	3554				65	25	8	4	2	4	4	1	0
1965	4849					65	17	2	1	6	2	3	1
1966	2555						52	9	8	3	4	2	1
1967	305							3	1	0	1	0	0

[1]Data from Brownie et al. (1985).

for all model parameters and, for illustration and clarity, ran the Gibbs sampling algorithm for only 5000 iterations.

These plots appear to suggest that observations generated at the beginning of the run have a similar distribution to those at the end, indicating that the Gibbs sampler has converged very quickly. In fact this fast convergence seems to be a general property of band-return models. The diagnostic techniques of Brooks and Gelman (1998) were applied to several replications of the MCMC sampler for each of our models and these confirm that convergence is indeed rapid.

Additionally, the autocorrelation plots in Figure 1 illustrate the dependence between successive observations, which appears to die out well before lag
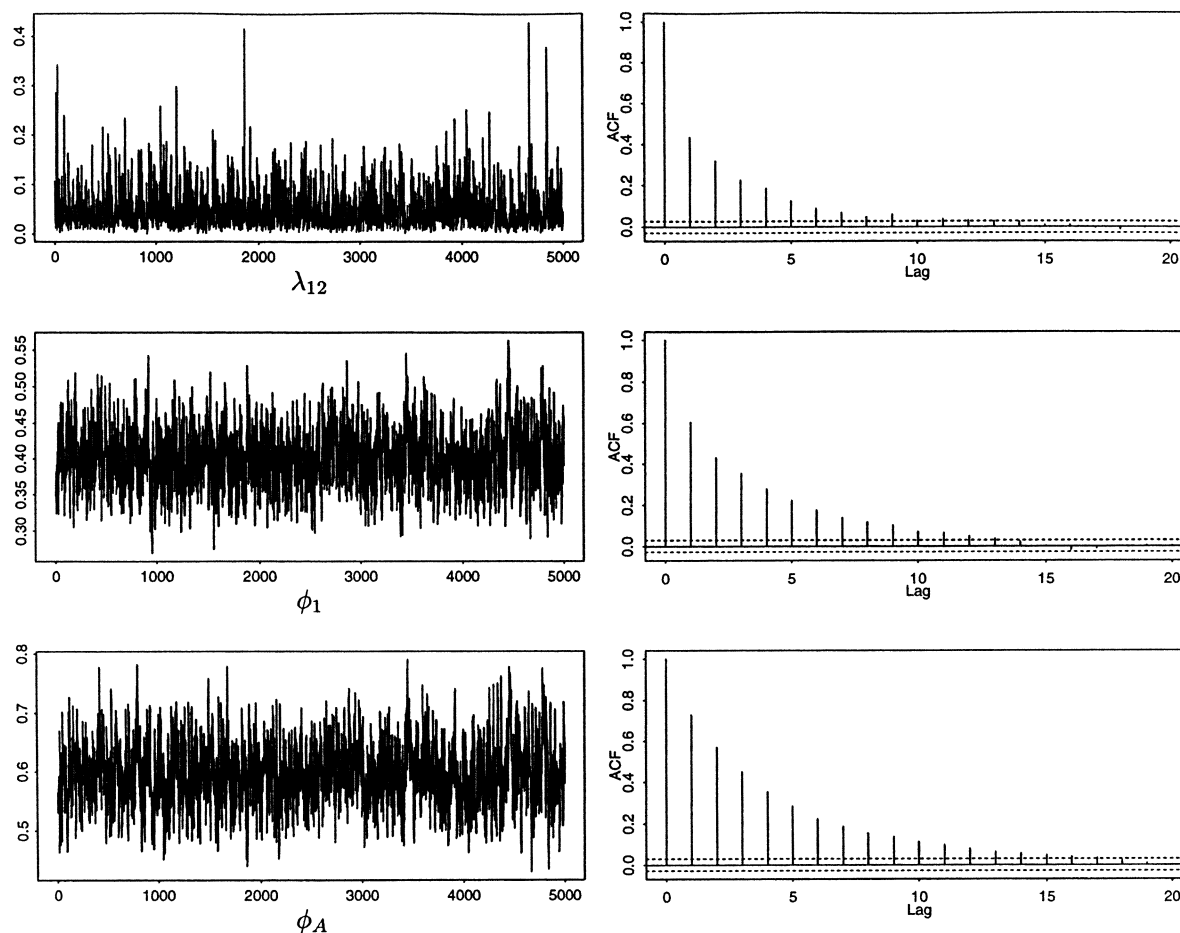

FIG. 1. *Raw trace plots (left) and autocorrelation plots (right) of the MCMC output for parameters* λ_{12}, ϕ_1 *and* ϕ_A *under the C/C/T model for the teal data. Autocorrelations that lie outside of the dotted bands are significantly different from zero at the 5% level.*

TABLE 3

Posterior means and standard deviations (SDs) under independent uniform priors, for four models fitted to the Teal data set, from a sample of 10,000 observations from the posterior distribution, together with the Bayesian p-values and posterior model probabilities, Pr(M | x), associated with each model

Parameter	C/A/C Mean	SD	T/A/C Mean	SD	T/A/T Mean	SD	C/C/T Mean	SD
$\phi_{1,1}$			0.531	0.126	0.525	0.248		
$\phi_{1,2}$			0.588	0.089	0.531	0.114		
$\phi_{1,3}$			0.431	0.081	0.321	0.083		
$\phi_{1,4}$			0.428	0.047	0.484	0.070		
$\phi_{1,5}$			0.396	0.047	0.259	0.055		
$\phi_{1,6}$			0.352	0.052	0.264	0.077		
$\phi_{1,7}$			0.459	0.179	0.158	0.108		
ϕ_1	0.420	0.026					0.403	0.039
ϕ_2	0.604	0.040	0.602	0.038	0.445	0.071		
ϕ_3	0.694	0.047	0.690	0.049	0.538	0.087		
ϕ_4	0.640	0.062	0.636	0.059	0.459	0.094		
ϕ_5	0.690	0.073	0.682	0.070	0.578	0.100		
ϕ_6	0.624	0.093	0.616	0.091	0.545	0.108		
ϕ_7	0.410	0.133	0.397	0.130	0.372	0.134		
ϕ_8	0.494	0.207	0.474	0.206	0.468	0.202		
ϕ_9	0.729	0.217	0.722	0.219	0.749	0.205		
ϕ_{10}	0.688	0.232	0.695	0.230	0.691	0.233		
ϕ_{11}	0.348	0.240	0.352	0.240	0.355	0.240		
ϕ_{12}	0.493	0.285	0.511	0.285	0.504	0.288		
ϕ_A							0.598	0.053
λ_1					0.058	0.137	0.013	0.005
λ_2					0.017	0.008	0.017	0.005
λ_3					0.018	0.005	0.024	0.005
λ_4					0.035	0.006	0.031	0.004
λ_5					0.019	0.003	0.026	0.003
λ_6					0.028	0.004	0.031	0.004
λ_7					0.019	0.005	0.016	0.004
λ_8					0.056	0.022	0.031	0.008
λ_9					0.078	0.040	0.035	0.011
λ_{10}					0.140	0.096	0.051	0.020
λ_{11}					0.149	0.123	0.048	0.025
λ_{12}					0.196	0.181	0.047	0.034
λ	0.026	0.001	0.026	0.001				
p-value	0.27		0.28		0.54		0.50	
Pr(*M* | *x*)	0.009		0.054		0.000		0.937	

20. This indicates fairly rapid mixing and thus good coverage of the parameter space with a reasonably small number of iterations.

The results obtained from a MCMC simulation of 10,000 iterations (after discarding the initial 1000 iterations) with independent uniform priors for each model parameter, under four models, are provided in Table 3. The models selected here are for illustrative purposes only. The Bayesian *p*-values (see Appendix A.1) for goodness of fit are given in the table, and the corresponding discrepancy plots are in Figure 2. They show that models C/C/T and T/A/T fit the data equally well and appreciably better than the other two models.

Note that in Table 3 the parameters which appear least often in the likelihood, such as ϕ_i for i close to 12, have posterior moments close to those expected under the uniform(0, 1) prior. The posterior distribution provides a balance between information provided by the prior and that provided by the likelihood. For parameters where little information is gained from the likelihood, the posterior distribution is dominated by the prior, and hence we see a number of parameters with mean around 0.5 and standard deviation around $\sqrt{1/12}$. Note also that the freedom allowed by the T/A/T model to choose different values of λ for each year has resulted in large changes in the means of some of the $\phi_{1,i}$ as compared with the T/A/C model. This is because a large part of the information on $\phi_{1,i}$ is contained in the diagonal m_{ii} term.

For any Bayesian analysis, it is important to check the sensitivity of the resulting inference upon the choice of prior distribution (see, e.g., Rubin,
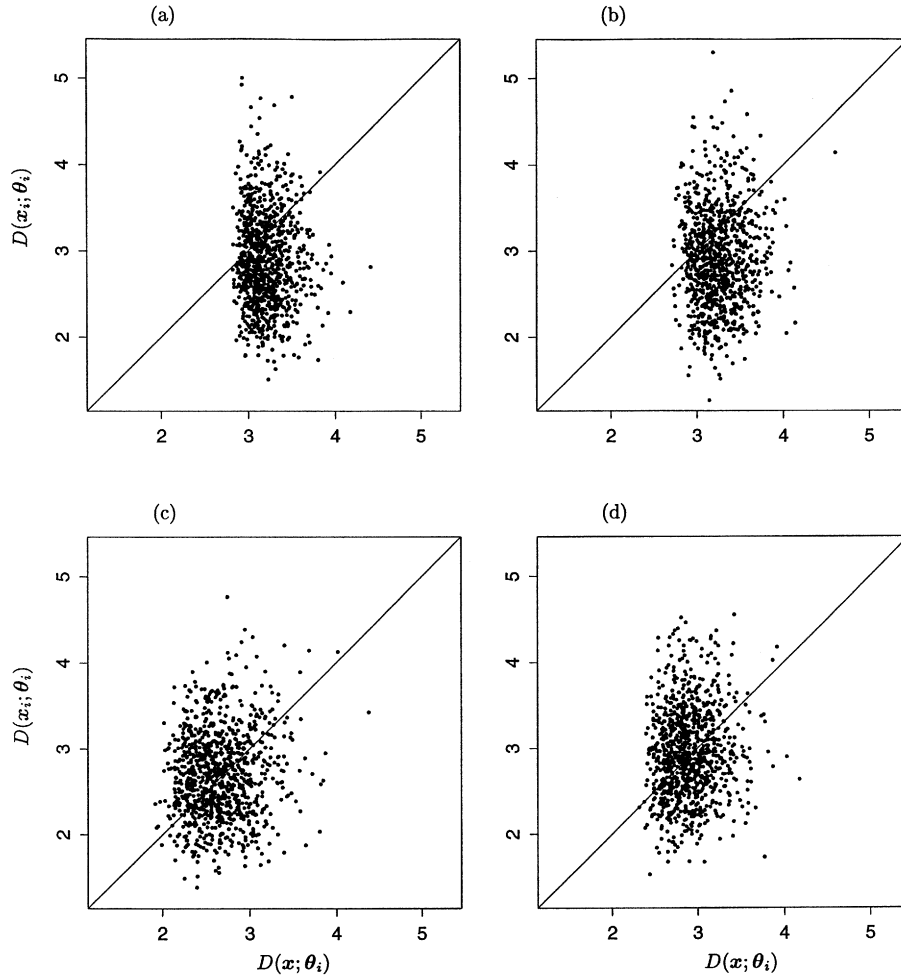
FIG. 2. *Discrepancy plots for the models* (a) $C/A/C$, (b) $T/A/C$, (c) $T/A/T$ *and* (d) $C/C/T$, *from a sample of* 2000 *observations from the posterior distribution corresponding to independent uniform priors on all parameters. Corresponding Bayesian p-values are* 0.27, 0.28, 0.54 *and* 0.50, *respectively.*

1992). Figures 3 and 4 show histograms of posterior samples of selected parameters from the C/C/T and T/A/T models, under various prior distributions.

We can see from Figure 3 that the choice between the beta $(1, 4)$, uniform and beta $(4, 1)$ priors has little effect upon the posterior distributions for the survival parameters of the C/C/T model, reflecting the fact that the likelihood carries strong information concerning these parameters. Note however that an unrealistic beta(4,1) prior for the λ parameters, for many of which there is relatively little information in the likelihood, results not only in a large change in the λ posteriors but also a large shift to the left in the posterior distributions of ϕ_1 and ϕ_A. Exceptionally high values for λ parameters need to be balanced by changes in survival probabilities in order to match the data. If we take the last column of Table 2, for example, each of the cells in this column has a probability which is a multiple of λ_{12}. A large value of λ_{12} is balanced by making

ϕ_a, which appears raised to a power in all the cells in the column, small.

Figure 4 shows that for model T/A/T, the posteriors can be heavily influenced by the priors for the survival parameters, even when the λ parameters have uniform priors. In general, models which contain parameters that are not supported by the data will tend to have a high degree of posterior sensitivity, so that the degree of sensitivity also provides evidence in favor of one model over another.

Also shown in Table 3 are the posterior model probabilities (see Appendix A.2) for the four models considered. These are based on samples of size 100,000 from the prior distribution. The very small posterior probability for the T/A/T model, and the large one for C/C/T, require some explanation in view of the Bayesian p-values, which show both models fitting the data quite well.

Note also that C/C/T is a submodel of T/A/T, and so its maximized likelihood cannot be greater than
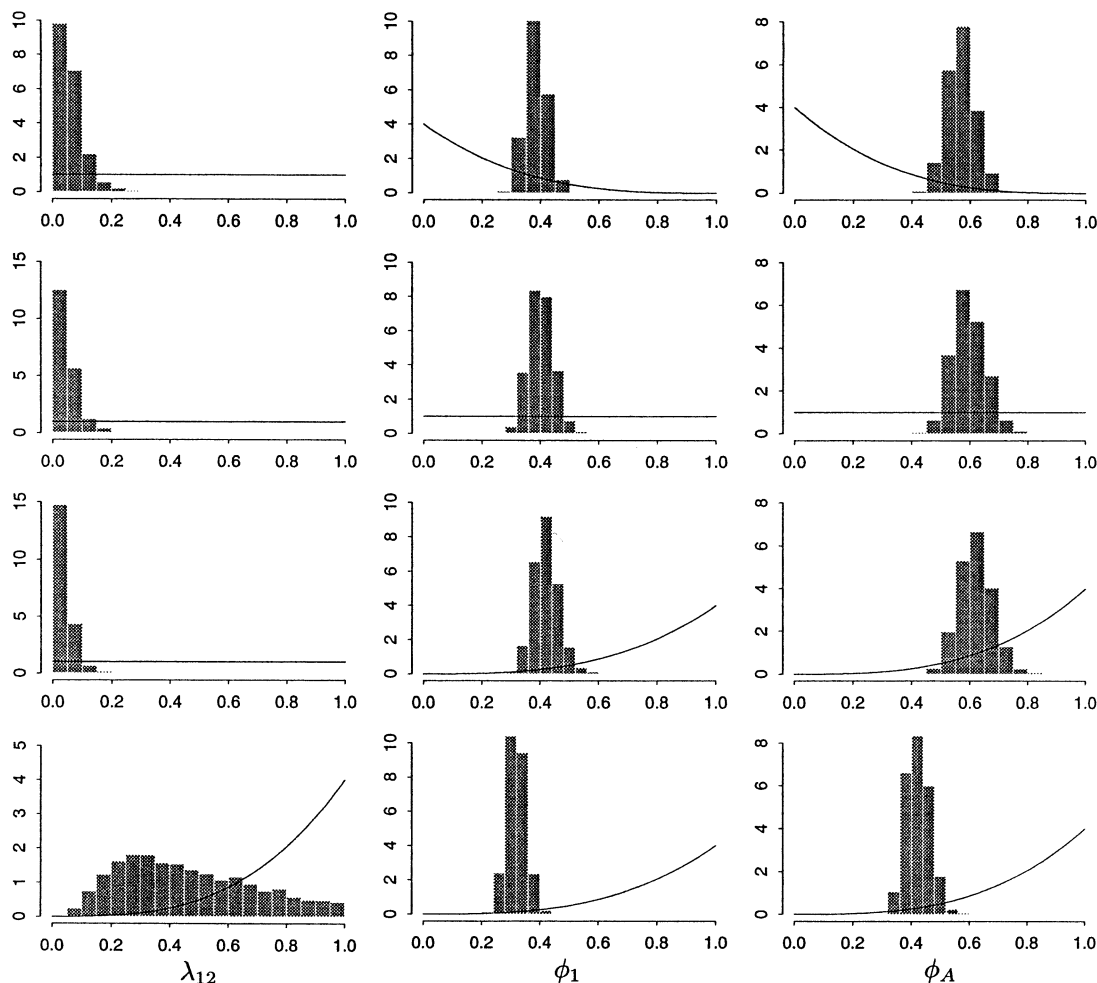
FIG. 3. *Histograms for posterior samples of parameters* λ_{12}, ϕ_1 *and* ϕ_A *under the* $C/C/T$ *model, from a sample of* $10{,}000$ *observations from the posterior distribution, with corresponding prior densities superimposed. The top row corresponds to beta* $(1, 4)$ *priors, the second row to uniform and the bottom two rows to beta* $(4, 1)$ *priors for the two* ϕ *parameters, whereas a uniform prior is adopted for all the* λ *parameters in the first three rows and a beta* $(4, 1)$ *prior for the bottom row.*

that of T/A/T. Since, from Appendix A.2, the relative posterior probabilities of the two models are given by a ratio of averaged likelihoods (given equal prior probabilities), the reader might be surprised that C/C/T ends up with a higher posterior probability than T/A/T. The reason for this is that the averages are taken over the whole parameter space. Depending on the prior, this may involve averaging over some very unlikely regions. The posterior model probabilities given in Table 3 are for independent uniform priors over [0, 1]. With different priors, very different results could occur. If the priors were concentrated sufficiently close to the maximum likelihood estimates, for example, then the T/A/T model would have a higher posterior probability than C/C/T. It is important therefore to compare posterior model probabilities under a range of prior distributions. We do this for the second data set in the next section.

In general, a model with many parameters will have the freedom to explore regions of very small likelihood, and thus will tend to have low posterior probability. Thus, although no explicit penalty for overparameterization is built into the methodology, there is an implicit penalty.

Note also that the discrepancy measure for goodness of fit is averaged over the posterior distribution of the parameters, while the posterior model probabilities are averaged over the prior distribution. This is why the T/A/T and C/C/T models fit the data equally well, while the C/C/T model has much greater posterior probability than T/A/T: the Bayesian p-value does not penalize overparameterization.

Appendix A.2 also describes Bayesian model averaging. In the present example, with the relative posterior probabilities overwhelmingly favoring the C/C/T model, there is little point in doing this. Had we chosen a larger range of models for consideration, this
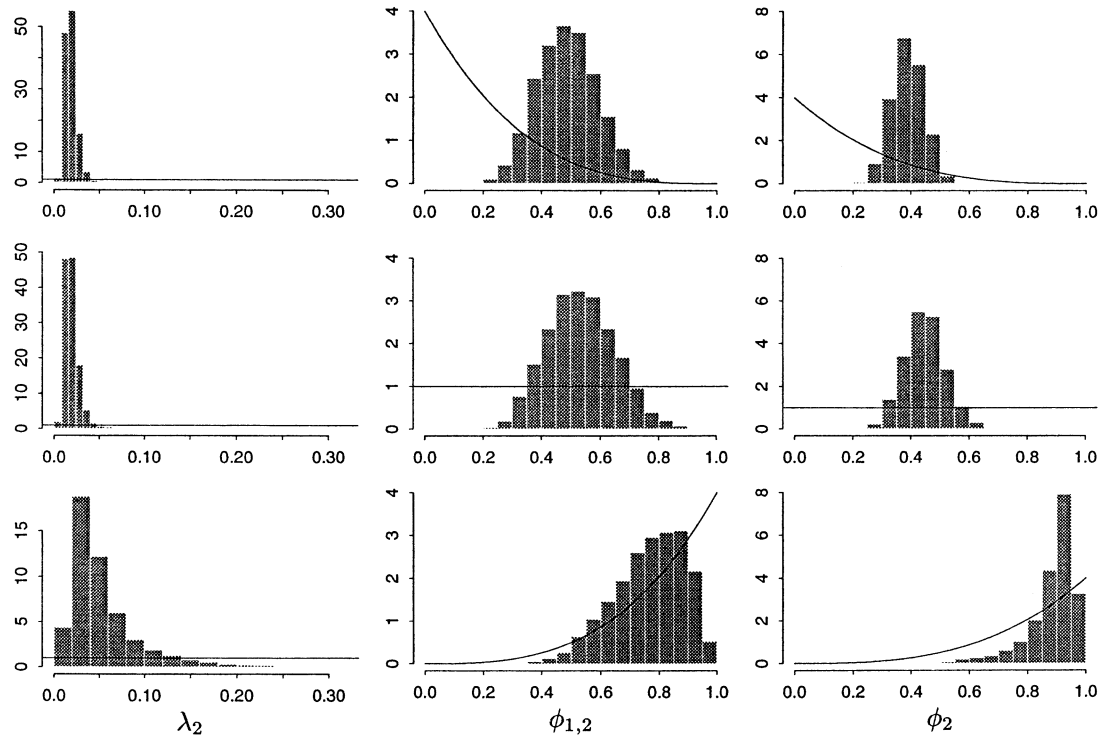
FIG. 4. *Histograms for posterior samples of parameters λ_2, $\phi_{1,2}$ and ϕ_2 under the T/A/T model, from a sample of $10,000$ observations from the posterior distribution, with corresponding prior densities superimposed. The top row corresponds to beta $(1,4)$ priors, the middle row to uniform and the bottom row to beta $(4,1)$ priors for the survival parameters, whereas a uniform prior is adopted for the λ priors throughout.*

might no longer have been the case. However, we shall apply model averaging for the Dipper data in the next section.

We conclude that, of the models considered, the C/C/T and T/A/T models provide the best fit to the blue-wing teal data, and, on the grounds of prior sensitivity, parsimony and posterior model probability, C/C/T is best. This agrees with the classical analysis of Freeman and Morgan (1992), who, using likelihood ratio tests and starting from a submodel of T/A/T, reached C/C/T as their final model. The posterior means also broadly agree with the maximum likelihood estimates from the classical analysis (not shown here).

Note that a reasonable (but more complex) prior might constrain the survival parameters in such a way that adult survival is always greater than first-year survival. We considered this alternative prior for the C/C/T model, where the restriction seems most plausible. However, since the first-year survival probabilities were always less than the adult survival probabilities under the uniform prior, this new prior had no effect on the resulting posterior samples.

In some cases, models for recovery data may need elaboration to allow for different reporting probabilities for young and adult birds, for example. For further investigation of this issue see Freeman and

Morgan (1992) and Catchpole, Freeman and Morgan (1995).

Finally, an interesting comparison between the Bayesian and classical analyses of data of this sort arises when we fit the C/A/C model. In this case, it is well known that the likelihood has a completely flat ridge (Catchpole and Morgan, 1994). Thus, unique maximum likelihood parameter estimates are not available. However, as we see in Table 3, no such problems are encountered in the Bayesian analysis, even with flat priors. In our example, the Bayes estimates for the model parameters are given as the posterior means, which are unique whether or not the multivariate posterior distribution has a ridge. However, Brooks, Catchpole and Morgan (2000), considering this C/A/C model, point out that, since this model is parameter redundant, there is some combination of parameters about which the likelihood contains no information, and the only information is contained in the prior. In such circumstances, posterior inference must proceed with extreme care.

3. OPEN POPULATION CAPTURE–RECAPTURE MODELS

An alternative to the band-return study is the capture–recapture study, which records resightings

and/or recaptures of animals that have been previously marked. In these studies for birds at least, it is quite common for the animals to be marked only as adults; for instance, they may be initially captured when breeding. This is the case with the data we consider, and we therefore describe models with y/z notation, where y describes adult survival as before, but now z refers to recapture rather than reporting probabilities.

We assume that every animal in the population has the same probability of survival between capture times and that they all have the same probability of capture. Here capture and survival rates are of primary interest, but estimates of the population size at the different sampling times can also be obtained.

In this section, we look at the Cormack–Jolly–Seber model, which is the T/T model in our notation, and discuss the Bayesian approach to fitting this and submodels to data on European Dippers (*Cinclus cinclus*).

3.1 The Cormack–Jolly–Seber Model

We observe data of the form m_{ij}, $i = 1, \ldots, I$, $j = i + 1, \ldots, J$, $J > I$, where m_{ij} denotes the number of animals released at time t_i and subsequently recaptured for the first time at time t_j. We also have data R_i recording the number of marked animals released into the population at time t_i (these comprise newly marked animals and those recaptured at time t_i).

We let p_i denote the probability of capturing a particular animal at time t_i, and let ϕ_i denote the probability of the animal surviving the ith time period, that is, the probability that the animal survives to time t_{i+1} given that it is alive at time t_i. Table 4 provides the expected and corresponding observed values under the T/T model.

The resulting likelihood is given by

$$L(\boldsymbol{\phi}, \mathbf{p}; \mathbf{R}, \mathbf{m})$$

(4)
$$\propto \Delta \prod_{i=1}^{I} \prod_{j=i+1}^{J} \left(\phi_i p_j \prod_{k=i+1}^{j-1} \phi_k \tilde{p}_k \right)^{m_{ij}}$$

where $\Delta = \prod_{i=1}^{I} \chi_i^{v_i}$, with $v_i = R_i - \sum_{j=i+1}^{J} m_{ij}$ being the number of animals never recaptured after release at t_i, and χ_i being the probability that an animal, alive at time t_i, is not subsequently captured. This can be calculated from the recursion formula, $\chi_i = 1 - \phi_i(1 - \tilde{p}_{i+1}\chi_{i+1})$, with $\chi_J = 1$ (see, e.g., Lebreton, Burnham, Clobert and Anderson, 1992, page 71).

As before, we adopt independent beta priors for the parameters ϕ_l and p_l,

$$\phi_l \sim \text{beta}(\alpha_\phi, \beta_\phi), \quad l = 1, \ldots, J - 1,$$
$$p_l \sim \text{beta}(\alpha_p, \beta_p), \quad l = 2, \ldots, J.$$

The full conditional posterior for ϕ_l is then given by

$$\pi(\phi_l \mid \boldsymbol{\phi}_{(l)}, \mathbf{p}, \mathbf{R}, \mathbf{m})$$
$$\propto \phi_l^{\alpha_\phi - 1} \tilde{\phi}_l^{\beta_\phi - 1} \Delta \phi_l^r$$
$$\propto \Delta f_{\text{beta}}\left(\phi_l; \alpha_\phi + r, \beta_\phi \right),$$
$$l = 1, \ldots, J - 1,$$

where $r = \sum_{i=1}^{l^*} \sum_{j=l+1}^{J} m_{ij}$ and $l^* = \min(l, I)$. For the submodel with $\phi_l = \phi$ for all l,

$$\pi(\phi \mid \mathbf{p}, \mathbf{R}, \mathbf{m})$$
$$\propto \Delta f_{\text{beta}}\left(\phi; \alpha_\phi + \sum_{i=1}^{I} \sum_{j=i+1}^{J} (j - i)m_{ij}, \beta_\phi \right).$$

Similarly the full conditional posterior distribution for p_l is given by

$$\pi(p_l \mid \mathbf{p}_{(l)}, \boldsymbol{\phi}, \mathbf{R}, \mathbf{m}) \propto \Delta f_{\text{beta}}\left(p_l; \alpha_p + r, \beta_p + s \right),$$
$$l = 2, \ldots, J,$$

where $r = \sum_{i=1}^{l^*} m_{il}$ and $s = \sum_{i=1}^{l^*} \sum_{j=l+1}^{J} m_{ij}$, with $l^* = \min(l - 1, I)$. For the submodel with $p_l = p$ for all l,

$$\pi(p \mid \boldsymbol{\phi}, \mathbf{R}, \mathbf{m}) \propto \Delta f_{\text{beta}}\left(p; \alpha_p + \sum_{i=1}^{I} \sum_{j=i+1}^{J} m_{ij}, \right.$$
$$\left. \beta_p + \sum_{i=1}^{I} \sum_{j=i+1}^{J} (j - i - 1)m_{ij} \right).$$

As in the band-return case, all these conditional posterior distributions are nonstandard.

TABLE 4
The data structure for the T/T recapture model for a limited study[1]

		$j = 2$	3	4	Never recaptured
$i = 1$	R_1	$R_1 p_2 \phi_1$	$R_1 p_3 \tilde{p}_2 \phi_1 \phi_2$	$R_1 p_4 \tilde{p}_3 \tilde{p}_2 \phi_1 \phi_2 \phi_3$	$R_1 \chi_1$
		(m_{12})	(m_{13})	(m_{14})	(v_1)
2	R_2		$R_2 p_3 \phi_2$	$R_2 p_4 \tilde{p}_3 \phi_2 \phi_3$	$R_2 \chi_2$
			(m_{23})	(m_{24})	(v_2)
3	R_3			$R_3 p_4 \phi_3$	$R_3 \chi_3$
				(m_{34})	(v_3)

[1]The entries denote expected values; below in parentheses are the observed values.

3.2 Example: European Dippers

The data given in Table 5 describe the annual capture and recapture of European Dippers in eastern France from 1981–1987. As in Section 2.2, we have deliberately chosen a sparse data set for illustration of how Bayesian methods combine prior information and data. Lebreton et al. (1992) provide the maximum likelihood parameter estimates for a variety of models for these data. These models are the full model T/T and the submodels T/C, C/C and C2/C. The C2/C model has constant recapture probability but two survival rates,

$$\phi_i = \begin{cases} \phi_f, & i = 2, 3, \\ \phi_n, & i = 1, 4, 5, 6 \end{cases}$$

(Lebreton et al. 1992), to allow for possibly different survival caused by a flood in 1983. We consider here only these four models. Note that in the T/T model, the parameters p_7 and ϕ_6 always appear together in the likelihood; they are nonidentifiable, and only their product is estimable via maximum likelihood.

TABLE 5
Capture–recapture data for European Dippers banded in 1981–1986

Year of release	Number released	Year of Recapture (1981+)					
		1	2	3	4	5	6
1981	22	11	2	0	0	0	0
1982	60		24	1	0	0	0
1983	78			34	2	0	0
1984	80				45	1	2
1985	88					51	0
1986	98						52

[1]From Lebreton et al. (1992).

In the submodels all parameters are estimable. A Bayesian analysis for the model T/T can produce estimates of all the model parameters. From the symmetry with regard to ϕ_6 and p_7, their marginal distributions are identical.

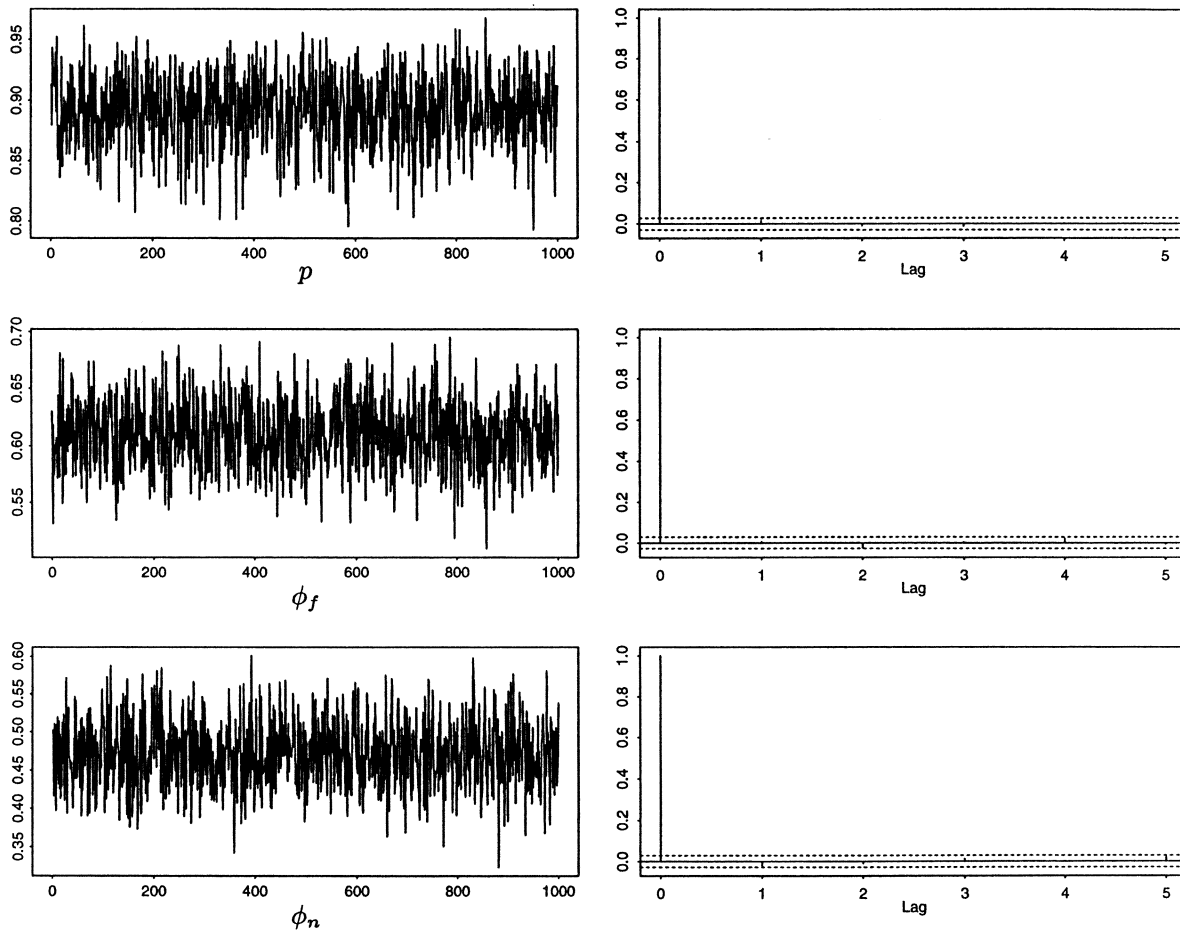Figure 5 provides an illustrative example of the MCMC output for the three parameters of the C2/C

FIG. 5. *Raw trace plots (left) and autocorrelation plots (right) for parameters p, ϕ_f and ϕ_n under the C2/C model. Autocorrelations that lie inside the dotted bands are not significantly different from zero at the 5% level.*

model. Here we took independent uniform priors for all model parameters and ran the Gibbs sampler for just 1000 iterations. As with the Teal example, convergence appears to be rapid, but, unlike the Teal example, the acf plots suggest that successive iterations are uncorrelated. Thus, in this case, the raw MCMC output may be taken as an approximately independent sample from the posterior distribution so that the usual Monte Carlo sample size determination methods may be applied (Brooks, 1997, 1999).

The results from a MCMC simulation of 10,000 iterations under standard uniform priors for all parameters are provided in Table 6 for each of the four models. Note that the standard maximum-likelihood approach to the model T/T is to estimate the product $\phi_6 p_7$. The posterior mean for this product is 0.519, with standard deviation 0.050. We naturally obtain a different result if we treat this product as a single parameter, with a uniform prior distribution.

Lebreton et al. (1992) analyze these data from the classical maximum likelihood perspective, and conclude that the C2/C model provides the best fit to the data. The Bayesian p-values for goodness of fit shown in Table 7 are quite different from the classical p-values provided by Lebreton et al. (1992) (0.30, 0.41, 0.30 and 0.67, respectively), but may be interpreted similarly in that C2/C is clearly the preferred model. We would argue that the Bayesian p-values, which do not require an asymptotic normality argument, are more realistic. However, we do note the very small p-values obtained. It may be

useful to recall that the Bayesian interpretation of these p-values is as a measure of surprise at the data under the assumption of each model. Thus, we might infer that the data appear to be rather surprising (though not significantly so) under all four models under consideration, and might seek further models.

As with the Teal example, we can examine the prior sensitivity of the model parameters. Figure 6 provides the histogram plots corresponding to the MCMC output for the C2/C model for parameters p, ϕ_f and ϕ_n. Superimposed on the histograms is the corresponding prior density. It is clear from these plots that even a fairly strong prior has little effect upon the marginal posterior distributions. In this case the likelihood contains strong information about the parameters so that the posterior is robust to changes in the prior and bears little resemblance to the two prior distributions selected here.

As we note in Appendix A.2, the p-value associated with a model also depends on the prior. To ensure that our p-values are robust, we consider three sets of priors. Prior 1 is a uniform prior on all parameters. Prior 2 puts a beta$(1, 9)$ prior on the ϕ parameters and a uniform prior on the p parameters. Prior 3 is a beta$(1, 9)$ for all parameters. We then consider the four models, T/T, T/C, C/C and C2/C, obtaining the p-values given in Table 7 for each model–prior combination. The p-values change quite substantially, although C2/C appears as the best-fitting model whichever prior is chosen.

Thus, our Bayesian analysis concludes that the C2/C model provides the best fit to the data, agree-

TABLE 6
Posterior means and standard deviations under independent uniform priors, for the Dipper data

Model	T/T Mean	T/T SD	T/C Mean	T/C SD	C/C Mean	C/C SD	C2/C Mean	C2/C SD
ϕ	—	—	—	—	0.561	0.025	—	—
ϕ_1	0.723	0.133	0.622	0.107	—	—	—	—
ϕ_2	0.450	0.071	0.458	0.065	—	—	—	—
ϕ_3	0.482	0.061	0.481	0.058	—	—	—	—
ϕ_4	0.626	0.060	0.623	0.056	—	—	—	—
ϕ_5	0.603	0.058	0.608	0.054	—	—	—	—
ϕ_6	0.728	0.143	0.587	0.057	—	—	—	—
ϕ_f	—	—	—	—	—	—	0.472	0.043
ϕ_n	—	—	—	—	—	—	0.609	0.031
p	—	—	0.893	0.030	0.896	0.029	0.892	0.030
p_2	0.666	0.134	—	—	—	—	—	—
p_3	0.867	0.082	—	—	—	—	—	—
p_4	0.879	0.064	—	—	—	—	—	—
p_5	0.875	0.058	—	—	—	—	—	—
p_6	0.904	0.052	—	—	—	—	—	—
p_7	0.737	0.143	—	—	—	—	—	—

Model	p-values			Pr(M|x)		
	Prior 1	Prior 2	Prior 3	Prior 1	Prior 2	Prior 3
T/T	0.086	0.049	0.000	0.000	0.003	0.000
T/C	0.068	0.078	0.000	0.000	0.003	0.000
C/C	0.069	0.056	0.023	0.205	0.951	1.000
C2/C	0.125	0.153	0.050	0.795	0.004	0.000

[1]Prior 1 is a uniform prior on all model parameters. Prior 2 puts independent Beta$(1, 9)$ distributions on the ϕ parameters and a uniform prior on the p parameters. Prior 3 is a Beta$(1, 9)$ for all model parameters.

ing with the classical analysis of Lebreton et al. (1992). It is clear that the floods of 1983 had a significant impact, dropping the survival rate from around 0.61 to around 0.47. The decision to consider the flood-based model was taken in the light of prior (or data independent) information from the biologists gathering the data, and Lebreton et al. (1992) discuss the usefulness of such information in the classical analysis. However, the Bayesian approach is even more flexible in this respect since other information concerning the relationship between the survival rates in flood and nonflood years might also be incorporated explicitly through the prior. For example, information from previous studies concerning the possibility of emigration during flood years could be used to specify how the survival rate may be altered by this behavior. Thus, the Bayesian analysis provides a more flexible approach to modeling data of this sort, and essentially incorporates the classical maximum likelihood analysis as a special case.

We now consider the posterior model probabilities of Table 7. As before, these are based on samples of 100,000 from the prior distribution. Priors 2 and 3 have mean 0.1, in conflict with the data. As discussed in the previous section, model C2/C, which has one more parameter than model C/C, is penalized more by this, resulting in C/C being the favored model for these unrealistic priors. For the more realistic prior 1, only C/C and C2/C have nonnegligible posterior probabilities, and we can use model averaging to obtain the means and standard errors of parameters common to these two models (see Appendix A.2). For example, the model-averaged estimate of ϕ_n, the survival rate in nonflood years, is, from (6),

$$\overline{\phi_n} = 0.205 \times 0.561 + 0.795 \times 0.609 = 0.599,$$

while, from (7),

$$\overline{\phi_n^2} = 0.205 \times (0.561^2 + 0.025^2)$$
$$+ 0.795 \times (0.609^2 + 0.031^2)$$
$$= 0.3603$$

and therefore $\overline{\text{SD}(\phi_n)} = \sqrt{0.3603 - 0.599^2} = 0.038$. Note from Table 6 that this standard deviation is larger than that for either of the models C/C or C2/C. If a single model is selected, the standard deviation of ϕ_n is underestimated.

4. DISCUSSION

4.1 Relative Merits of Bayesian and Classical Procedures

There are advantages and disadvantages to the modern Bayesian approach to modeling recovery and recapture data.

Bayesian analysis provides a natural way to include expert prior knowledge. In addition to the discussion in the text above, we note that we may wish to impose a bell shape for the relationship between annual survival probabilities and age, to account for the relatively high mortality of both young and old animals, as observed for example in Soay sheep; see Catchpole, Morgan, Freeman, Albon and Coulson (1998). The necessity for prior distributions, and the scope for disagreement amongst experts as to what constitutes the correct prior, could equally be seen as a disadvantage. Possible approaches to this problem include noninformative priors and empirical (data-based) priors (see, e.g., Carlin and Louis, 1996, Chapters 2 and 3), rather than subjective priors. Whichever method is chosen, it is important to check the robustness of the conclusions to the choice of prior. We speculate, on the evidence of the examples of this paper, that prior distributions will frequently have little effect for recapture models, but that this will not be true of recovery models. Relevant here are the findings of Lebreton, Morgan, Pradel and Freeman (1995) and Catchpole, Freeman, Morgan and Harris (1998), who have considered the relative contributions of the recovery and recapture parts of the likelihood in studies involving both recovery and recapture.

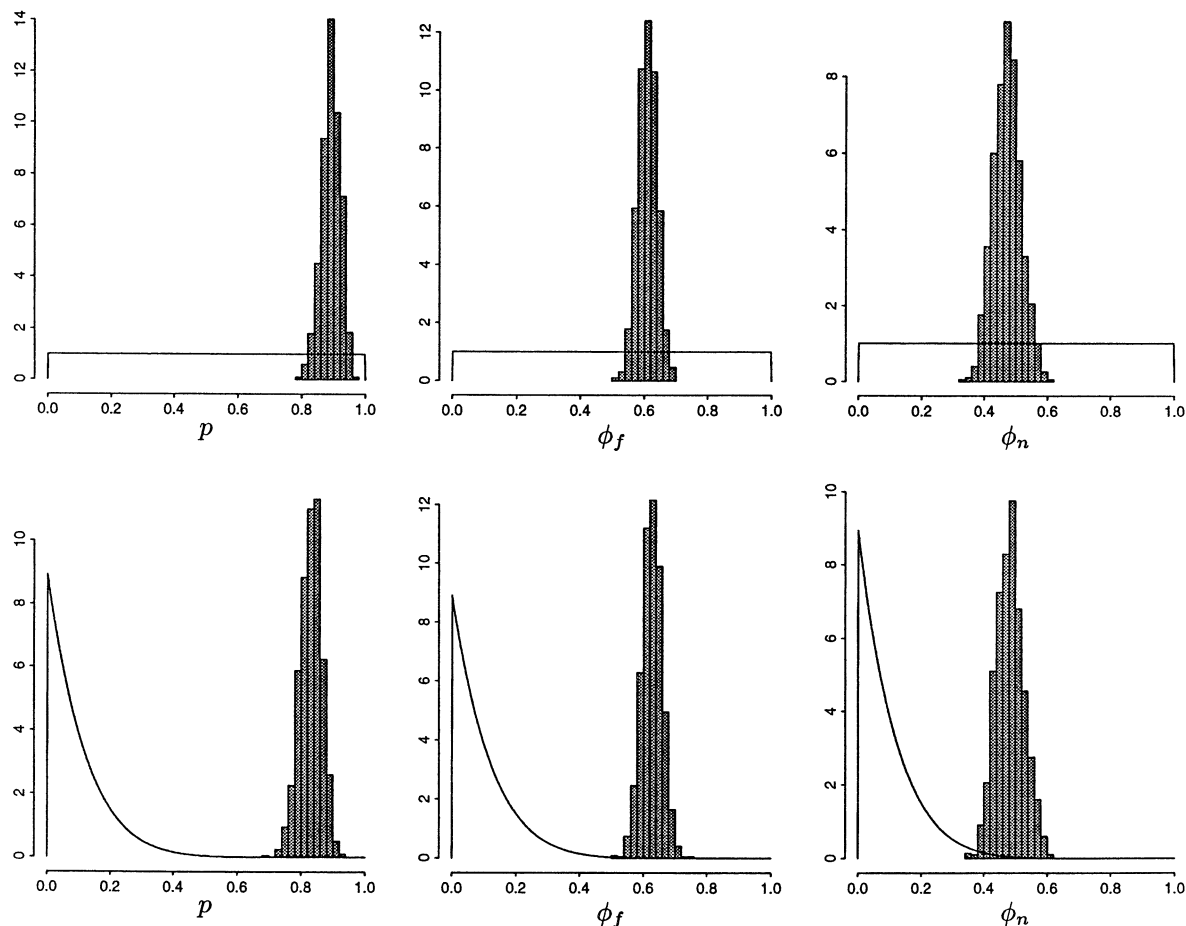Since the Bayesian paradigm stipulates that model parameters are not themselves fixed, but

FIG. 6. *Histograms for posterior samples of parameters under the C2/C model, with corresponding prior densities superimposed. The top row corresponds to uniform and the bottom row to beta*$(1, 9)$ *priors for* p, ϕ_f *and* ϕ_n.

have some unknown fixed distribution, it would be useful to provide more detailed summaries of that distribution than simply recording posterior means and variances. When the posterior conditional distributions are of standard form the normalization constants are necessarily known and Rao–Blackwell density estimates (Casella and Robert, 1996) are available, which provide unbiased estimates of the marginal densities of the model parameters. Unfortunately, none of our posterior conditional distributions are of standard form and so this density estimation technique cannot be easily applied. This problem may be overcome by using data augmentation techniques (Tanner and Wong, 1987) to obtain standard conditional posterior distributions, at the cost of adding the complexity of the data augmentation. This is the focus of current work.

In the Bayesian approach, unlike classical maximum likelihood, one does not need to use asymptotic results to obtain standard errors, as one can estimate marginal distributions by means of MCMC procedures. This is important for sparse data sets

such as those of Tables 2 and 5, which are quite common. In such cases classical inference can proceed via modern procedures such as bootstrapping or Monte Carlo methods. See also Morgan and Freeman (1989) and Cormack (1992), who advocate using profile log-likelihoods for constructing confidence intervals, rather than relying on the asymptotic normality of estimators.

Computational overheads and programming complexities have until recently been a great disincentive to the use of Bayesian methods. The first problem has been greatly ameliorated by the appearance of MCMC algorithms, although for simple models such as those considered here, Bayesian methods still require much more computing power than classical maximum likelihood. The second problem is now being overcome through the appearance of packages such as BUGS (see Appendix B). For recovery and recapture data such as those in this paper, a classical analysis by one of the specialized packages such as MARK (White and Burnham, 1999) is still much simpler to program than a

Bayesian analysis. However, BUGS is relatively simple to learn, and can handle a wide variety of problems. The MCMC simulations performed within this paper are very easily implemented within the BUGS package. The code required for the T/A/T model for the Teal data is provided in Appendix B.1 as an illustration.

A standard problem with MCMC simulations for Bayesian analysis is that it is very difficult to determine exactly the length of the burn-in period. This problem is somewhat akin to the classical problem of trying to determine whether or not an optimization routine has found a global maximum in the likelihood function. However, it should be noted that the models that we discuss in this paper appear to have very favorable convergence properties, mixing well and converging rapidly, so that this appears to be of little concern for models of this sort.

The Bayesian approach provides a simple framework for model averaging, which incorporates the uncertainty due to model selection into inferences on model parameters. This is important as neglecting this uncertainty can lead to spurious precision in the parameter estimates. However, as we have seen in the examples, posterior model probabilities, and therefore model averaging, can be sensitive to the choice of prior distribution.

It is well known that model probabilities can be hard to estimate precisely, and in such cases, it is sometimes necessary to use reversible jump MCMC (Green, 1995). This has not been the case for the examples of this paper. However, it has been true for other examples considered elsewhere (see, e.g., Brooks, Catchpole and Morgan, 2000).

Finally, we note that the Bayesian approach provides a natural framework for modeling parameters as random effects (cf. Burnham, 1999). This is an area of current research.

4.2 Conclusion

Very useful results can be obtained by adopting the Bayesian paradigm. Schwarz and Seber (1999) observe that "with the advent of the Gibbs sampler . . . there will be an upsurge of interest in Bayesian methods as more realistic priors can be used and compared." We have shown how the basic computations can be carried out very simply, using MCMC methods. The BUGS package provides a very useful tool for getting simulations up and running quickly and is ideal for learning. However, most problems require the speed and flexibility of specially written (but still fairly simple) Fortran or C code, for example. In any case, some computer code will always need to be written to supplement any BUGS code. An example of this is the code

required to calculate the Bayesian p-values that we provide in this paper. In addition, many models do not fit within the BUGS framework, for example problems involving age dependence of parameters in capture–recapture studies (see Catchpole, Freeman, Morgan and Harris, 1998). In fact, many more general models than those considered here can be analyzed using straightforward extensions of the Bayesian procedures of this paper. This is a topic of current research.

For both the applications of this paper, we have seen that certain parameters were extremely sensitive to the choice of priors, for certain models. This can be used to help choose between models. For any model–data combination, parameter sensitivity is going to occur to some degree. For the data sets of this paper, the sparseness produced severe examples of parameter sensitivity.

We conclude by noting that in this paper we have not used basic classical tools such as information criteria for model selection and residuals for checking goodness-of-fit. In practice we would advocate a catholic approach, which would employ these and whatever other tools are judged to be useful.

APPENDIX
A. ASPECTS OF BAYESIAN MODELLING

Let $\boldsymbol{\theta}$ denote the full set of parameters in the model that we choose to describe the data $\mathbf{x}$. The prior and posterior distributions are, respectively, denoted by $\pi(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta}|\mathbf{x})$, and $L(\boldsymbol{\theta};\mathbf{x})$ is the likelihood function. Throughout the paper we find it convenient to adopt beta priors for probabilities. Kadane and Wolfson (1998) and O'Hagan (1998) discuss how to translate the beliefs of relevant experts into the form of prior distributions for the parameters.

A.1 Model Assessment

Our preferred criterion for model fit is the Bayesian p-value (Gelman, Meng and Stern, 1996). This assumes that goodness of fit is assessed by some discrepancy measure between the observed data $\mathbf{x}$ and expected values $\mathbf{e}$. We use the Freeman–Tukey statistic (Freeman and Tukey, 1950, Bishop, Fienberg and Holland, 1975)

$$D(\mathbf{x};\boldsymbol{\theta}) = \sum_j \left(\sqrt{x_j} - \sqrt{e_j}\right)^2.$$

There are various alternatives to this discrepancy; for example, Gelman, Meng and Stern (1996) suggest Pearson's X^2. However, in our work, many cells may contain few observations, and the Freeman–Tukey measure removes the need to pool small cells to avoid overweighting. The square root serves to

stabilize the variance in the models we shall be considering.

Typically, the classical approach is to calculate the discrepancy only once, with the parameters equal to the maximum likelihood estimates. However, since, from the Bayesian perspective, θ has not a fixed value but a distribution, an alternative approach is required in which we obtain not a single discrepancy value, but a sample from the posterior distribution of discrepancy values. A similar classical approach is to sample θ values from the asymptotic normal distribution of the maximum likelihood estimator.

Goodness of fit is thus measured by first obtaining a sample θ_i, $i = 1, \ldots, n$, of parameter values from the posterior distribution using an appropriate MCMC sampler. Then, for each θ_i, we calculate $D(\mathbf{x}; \theta_i)$ as a measure of the discrepancy between the data and the corresponding model. For comparison, we also generate a new set of data $\mathbf{x}_i$ by sampling from the model. For each new data set $\mathbf{x}_i$, we then calculate $D(\mathbf{x}_i; \theta_i)$.

These discrepancy values can then be used to obtain a Bayesian p-value, which essentially quantifies the degree of "surprise" associated with the observed data under the assumed model and prior. If the model adequately describes the data, which are themselves not at odds with the prior, then observations sampled from the posterior predictive distribution should be "similar" to the observed data. Thus, the distributions of the discrepancy measures for the observed and simulated data should be similar (unless the data are surprising) and a p-value can be formed by recording the proportion of times $D(\mathbf{x}_i; \theta_i)$ is greater than $D(\mathbf{x}; \theta_i)$.

Note that other definitions of a Bayesian p-value exist. Bayarri and Berger (1999) compare the interpretation of various definitions, based upon the prior and predictive distributions, as well as a compromise between the two based upon the "partial" posterior predictive distribution. Advantages and disadvantages exist for each but they are likely to lead to similar inference in most practical problems.

The Gelman and Meng p-value we use can be illustrated graphically by plotting each of the $D(\mathbf{x}; \theta_i)$ against the corresponding $D(\mathbf{x}_i; \theta_i)$ value. In practice, it is generally best to produce both the graphical summary and Bayesian p-value, since it is possible for the distributions of $D(\mathbf{x}; \theta_i)$ and $D(\mathbf{x}_i; \theta_i)$ to differ, yet still provide an optimal p-value of 0.5. This sort of behavior is easily spotted in the graphical summary. These plots will also reveal the distribution of $D(\mathbf{x}; \theta_i)$, and a model with a smaller mean discrepancy, for example, might be preferred.

It is important to note that changing the prior can have a large effect on the p-value, as in Table 7.

A.2 Model Probabilities and Averaging

Given a set of models $M_1, \ldots, M_k$, say, which a priori we are willing to consider as realistic alternatives for describing a particular data set, it is possible to derive probabilities associated with each model, which may then be used to discriminate between them.

We begin by assigning a prior probability $p_i = \Pr(M_i)$ to each model. Commonly, these prior probabilities are equal, representing the assumption that each model is equally likely. Alternatively, the prior probabilities may be some function of the number of parameters in each model, so that models with large numbers of parameters are explicitly penalized. The choice of prior is entirely at the analyst's discretion, but should be based upon all available information about the problem at hand before the data were collected. For illustrative purposes, we assume throughout the paper that all models under consideration are equally likely a priori.

Once we have observed data $\mathbf{x}$, we can calculate the posterior probabilities associated with each model. Clearly

$$(5) \qquad \Pr(M_i \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid M_i)\, p_i}{f(\mathbf{x})},$$

with obvious notation, and

$$f(\mathbf{x} \mid M_i) = \int L_i(\theta; \mathbf{x})\pi_i(\theta)\, d\theta = c_i(\mathbf{x}),$$

say. Dropping the notational dependence upon the data $\mathbf{x}$, c_i is simply the normalization constant for the posterior distribution of θ under model M_i. Hence, the posterior model probability is proportional to the product of the prior probability p_i and the normalization constant c_i for the model. In practice, the denominator in (5) is unknown, so that typically the posterior model probabilities are known only up to a normalization constant.

The c_i are commonly analytically intractable, but they may be approximated by observing that c_i could be thought of as the expected value of the likelihood under the prior distribution for θ. Thus, by simulating values $\theta_1, \ldots, \theta_n$ from the prior and calculating the corresponding likelihood values, an estimate for c_i is given by

$$\hat{c}_i = \frac{1}{n} \sum_{j=1}^{n} L_i(\theta_j; \mathbf{x}).$$

Various alternative estimators are also available, see Gamerman (1997, page 195).

To compare two competing models, M_1 and M_2, say, we can compute the corresponding model probabilities up to a normalization constant and calculate

the ratio,

$$\frac{\Pr(M_1 \mid \mathbf{x})}{\Pr(M_2 \mid \mathbf{x})} = \frac{p_1 c_1}{p_2 c_2}.$$

If the prior probabilities are equal, this posterior odds ratio is also known as the Bayes factor (Kass and Raftery, 1995). In order to compare several models, the relative posterior model probabilities may be calculated by dividing each by the sum of the probabilities of all models under considerations. Thus,

$$\Pr(M_i \mid \mathbf{x}) = p_i c_i \Big/ \sum_{j=1}^{k} p_j c_j.$$

Note that the posterior model probabilities can be quite sensitive to the priors used; see Table 7.

In many Bayesian applications, these model probabilities are not used to discriminate between models, but to average over them. We write $\pi(\boldsymbol{\theta} \mid M_i, \mathbf{x})$ to denote the posterior distribution for $\boldsymbol{\theta}$ under model M_i. We can then form the averaged posterior distribution $\bar{\pi}(\boldsymbol{\theta} \mid \mathbf{x})$ from

$$\bar{\pi}(\boldsymbol{\theta} \mid \mathbf{x}) = \sum_i \Pr(M_i \mid \mathbf{x}) \, \pi(\boldsymbol{\theta} \mid M_i, \mathbf{x})$$

[Carlin and Louis, 1996, (2.28)]. It then follows that

(6) $\quad \mathrm{E}(\boldsymbol{\theta} \mid \mathbf{x}) = \sum_i \Pr(M_i \mid \mathbf{x}) \, \mathrm{E}(\boldsymbol{\theta} \mid M_i, \mathbf{x}),$

(7) $\quad \mathrm{E}(\boldsymbol{\theta}\boldsymbol{\theta}^T \mid \mathbf{x}) = \sum_i \Pr(M_i \mid \mathbf{x}) \, \mathrm{E}(\boldsymbol{\theta}\boldsymbol{\theta}^T \mid M_i, \mathbf{x}),$

from which it is simple to estimate the model-averaged posterior means and standard deviations of the parameters.

A.3 Sampling Strategies

The implementation of the two most common MCMC samplers is reasonably straightforward. Ideally, when the Gibbs sampler is used, each of the conditionals will be of the form of a standard distribution and suitable prior specification may ensure that this is the case. However, in the cases where one or more of the conditionals is nonstandard, there are many ways to sample from univariate conditionals. The problems commonly encountered when studying data of the sort discussed in this paper are often economically overcome by using the ratio-of-uniforms method to sample from nonstandard conditional distributions; see Wakefield, Gelfand and Smith (1991) or Ripley (1987, page 66).

An alternative to using univariate sampling methods for nonstandard conditionals in the Gibbs sampler is to use a single Metropolis–Hastings accept–reject step instead. Thus, we form a hybrid of the two common MCMC samplers, known as the Metropolis-within-Gibbs sampler. Here, Metropolis–Hastings steps are introduced into the Gibbs sampler, so that components whose conditional distributions are of a standard form may be sampled directly from the full conditional, while those with nonstandard distributions are updated via a univariate Metropolis–Hastings step, as discussed in Tierney (1994). This is sometimes simpler to implement than rejection-based methods such as the ratio method and will generally be quicker per iteration, since only one accept–reject step is performed. However, allowing Metropolis–Hastings updates may sometimes result in a "slow" Markov chain, due to the rejection of Metropolis proposals, restricting movement around the parameter space in these directions. Thus, more iterations are needed to overcome problems associated with strong dependence between successive observations. This is discussed further in Brooks (1999).

In this paper we use the Gibbs sampler, using the ratio-of-uniforms method to sample from nonstandard posterior conditional distributions. This avoids pilot tuning of random walk proposals and appears to be very efficient. Checking was done using the Metropolis-within-Gibbs hybrid procedure provided by BUGS, and identical results were obtained.

A.4 Implementation Issues

Vounatsou and Smith (1995) found, in a limited study, that the Gibbs sampler and Metropolis–Hastings methods had similar convergence rates. This is disputed by Brooks (1999) who finds that the convergence rate of the Gibbs sampler is much faster than that of the Metropolis–Hastings algorithm, though the latter is computationally much faster to implement. Thus, in terms of *effective* samples per second, the two approaches are broadly comparable. In general, we have found that MCMC convergence appears to be very rapid for the simulations that we have performed. This may be partly due to the fact that the models we are using result in posterior distributions that are unimodal (though ridges do sometimes occur as we shall discuss in the next section), so that the MCMC sampler does not have to pass through regions of low density in order to explore all regions of high posterior density in the parameter space. Such an occurrence can slow convergence considerably, since the chain tends to stick around a region of high posterior density. In addition to this, our model parameters are all probabilities and thus constrained to lie between zero and one. This makes the MCMC algorithm fairly robust against the choice of sampling mechanism, since nearly any updating procedure will work fairly well; see Brooks (1999) for example.

In general, a typical Bayesian analysis may be performed in the following steps. Following selec-

tion of a suitable prior in consultation with a relevant expert, the form of the posterior distribution should be examined to determine whether the Gibbs or Metropolis–Hastings algorithms would be easier to implement. The Gibbs sampler is generally preferred when the posterior conditionals are easy to sample from exactly. Brooks (1999) provides further discussion on this point.

We have found that plots of the sample autocorrelation function from runs of the MCMC algorithm provide a good indicator as to how well the sampler is performing and give an idea of the strength of the dependence between successive sampled observations. There are a variety of more formal convergence diagnostics (see Cowles and Carlin, 1996 and Brooks and Roberts 1999). However, the convergence of our MCMC simulations appears to be so rapid that the routine use of such methods seems unnecessary. In general, we have found it necessary to run only a single MCMC replication for around 10 000 iterations, discarding the first 1000 or so as a rather gross overestimate of the length of the burn-in period. Finally, since the MCMC algorithms appear to have little difficulty in traversing the parameter space fairly rapidly, we have found that the choice of starting values is fairly arbitrary. For the simulations reported in this paper, we have started our chains at the mean of the prior distribution.

For further discussion of implementation issues see Kass, Carlin, Gelman and Neal (1998) and Brooks (1997).

B. The BUGS Package

Many people prefer the flexibility of writing their own code to implement MCMC algorithms. However, a very powerful and versatile computer package known as BUGS is also available and can be used to perform the simulations discussed in this paper. The BUGS package implements the Gibbs sampler, see Gilks, Thomas and Spiegelhalter (1992), Spiegelhalter, Thomas and Best (1996) and Spiegelhalter, Thomas, Best and Gilkes (1996), and has been widely adopted by statistical practitioners. Once the command syntax has been understood, the BUGS package is very easy to use and ideal for models of the sort described in this paper.

In order to use the package to perform the simulations associated with any particular model, the user must first define the model and prior structure, using an Splus-type syntax. Typically, this is done by creating a file of commands defining the model structure.

There are a couple of disadvantages associated with the use of BUGS over code written in C or Fortran, for example. First, BUGS tends to be somewhat slower (up to five times slower, depending upon the platform) than C or Fortran code. Second, BUGS decides for itself how to perform the MCMC updates (e.g., adaptive rejection sampling or univariate Metropolis–Hastings steps), so that there is no flexibility in choosing the updating mechanism for the Markov chain. This prevents the user from tailoring the updating mechanism to overcome high serial correlations in the MCMC output, for example. However, BUGS does provide a very user-friendly interface, with the PC versions having various pull-down menus and excellent graphical facilities. Thus, BUGS is ideal for any newcomer to MCMC methods and simulation.

It should be noted that the code shown below may not work in early versions of the BUGS software package, but it is known to work for the Windows version of BUGS (WinBUGS), release 1.0 and later. The WinBUGS package is presently available free of charge from the BUGS website at `http://www.mrc-bsu.cam.ac.uk/bugs/`.

In order to compile and run a model, the code must be saved to a file. This file can then be read into BUGS using the *Open* option on the *File* pull-down menu. Next, the model must be "checked" by using the *Check Model* option on the *Model* menu. The data is read in by highlighting the relevant lines in the code and choosing the *Data* option on the *Model* menu. The code is then ready to be compiled. Once the model has been successfully compiled, the initial values can be read in, in the same way as the data, and then the MCMC simulation begins once the *Updates* option is selected.

Various options are available under the *Statistics* pull-down menu, including real time trace plots of the MCMC simulations, summary statistics and acf plots.

B.1 The BUGS Code

The code shown below is for the T/A/T model for the Teal data. Note that in order to specify the desired model in the form required by BUGS, the likelihood is specified as being the product of a series of multinomial distributions on the rows of the recovery table (given in Table 2.1). Thus, a final column must be added to account for those animals released but never subsequently recovered. Also, since BUGS cannot calculate the product of vector elements, the Δ term from equation (1) is specified as the exponential of the sum of log terms. Beyond that, the code should be self-explanatory.

```
model teal;
# Define the data
list(ni = 7, nj = 12, m = structure(.Data = c(
6, 2, 1, 1, 0, 2,1,0,0,0,0,0, 897,
0,11, 5, 6, 1, 1,1,1,0,0,0,1,1130,
0, 0,19, 4, 4, 4,0,0,1,1,0,0,1361,
0, 0, 0,65,25, 8,4,2,4,4,1,0,3441,
0, 0, 0, 0,65,17,2,1,6,2,3,1,4752,
0, 0, 0, 0, 0,52,9,8,3,4,2,1,2476,
0, 0, 0, 0, 0, 0,3,1,0,1,0,0, 300), .Dim = c(7,13)))

# Define the initial parameter values
list(lambda = c(0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5),
  phi = c(0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5),
  phi1 = c(0.5,0.5,0.5,0.5,0.5,0.5,0.5))

# Define the variables
var
 m[ni,nj + 1],    # Observations
 r[ni],    # Release data
 lambda[nj],    # Recovery probs
 phi[nj − 1],    # Survival probs
 wp[nj,ni,nj],    # Workspace
 p[ni,nj + 1];    # Cell probabilities
 phi1[ni];    # 1st year survival probs

# Begin the model definition
{
# Define the priors
for (i in 1:(nj − 1)){phi[i] ~ dbeta(1,1);}
for (i in 1:nj){lambda[i] ~ dbeta(1,1);}
for (i in 1:ni){phi1[i] ~ dbeta(1,1);}
tt
# Define the likelihood
for(i in 1:ni){m[i, 1:(nj + 1)] ~ dmulti(p[i, ], r[i]);}

# Calculate the no. of birds released each year
for(i in 1 : ni){
 r[i] <- sum(m[i, ])
}
# Calculate the cell probabilities
for(i in 1:ni){
# main diagonal
p[i, i] <- lambda[i] * (1 − phi1[i]);
# above main diagonal
p[i,i + 1] <- lambda[i + 1]*phi1[i]*(1 − phi[1])
# further above
for(j in (i + 2):nj){
 for(k in 1:(j − i − 1)){
 wp[i, j, k] <- log(phi[k])
 }
 p[i, j] <- lambda[j] * phi1[i]*(1 − phi[j − i]) *
  exp(sum(wp[i, j, 1: (j − i − 1)]));
 }
# below main diagonal
for(j in 1:(i − 1)){
 p[i, j] <- 0;
}
# last column: probability of non-recovery
p[i, nj + 1] <- 1 − sum(p[i, 1:nj])
}
}
```

ACKNOWLEDGMENTS

REFERENCES

BAYARRI, M. J. and BERGER, J. O. (1999). Quantifying surprise in the data and model verification. In *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 53–82. Oxford Univ. Press.

BESAG, J., GREEN, P., HIGDON, D. and MENGERSEN, K. (1995). Bayesian computation and stochastic systems. *Statist. Sci.* **10** 3–66.

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press.

BOLFARINE, H., LEITE, J. G. AND RODRIGUEZ, J. (1992). On the estimation of the size of a finite and closed population. *Biometrical J.* **34** 577–593.

BROOKS, S. P. (1997). Markov chain Monte Carlo method and its application. *Statistician* **47** 69–100.

BROOKS, S. P. (1999). Bayesian analysis of animal abundance data via MCMC. In *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 723–731. Oxford Univ. Press.

BROOKS, S. P., CATCHPOLE, E. A. and MORGAN, B. J. T. (2000). On the Bayesian analysis of ring-recovery data. *Biometrics* **56**(3). To appear.

BROOKS, S. P. and GELMAN, A. (1998). Alternative methods for monitoring convergence of iterative simulation. *J. Comp. Graphical Statist.* **7** 434–455.

BROOKS, S. P. and ROBERTS, G. O. (1999). Assessing convergence of Markov chain Monte Carlo algorithms. *Statist. Comput.* **8** 319–335.

BROWNIE, C., ANDERSON, D. R., BURNHAM, K. P. and ROBSON, D. S. (1985). Statistical inference from band recovery data: a handbook. Technical report, U.S. Dept. Interior, Fish and Wildlife Service.

BURNHAM, K. P. (1999). Random effects models in ringing and capture-recapture studies. *Bird Study*. Unpublished manuscript.

CARLIN, B. P. and LOUIS, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, London.

CASELLA, G. and ROBERT, C. P. (1996). Rao-Blackwellization of sampling schemes. *Biometrika* **83** 81–94.

CASTLEDINE, B. (1981). A Bayesian analysis of multiple capture-recapture sampling for a closed population. *Biometrika* **68** 197–210.

CATCHPOLE, E. A., FREEMAN, S. N. and MORGAN, B. J. T. (1995). Modelling age variation in survival and reporting rates for recovery models. *J. Appl. Statist.* **22** 597–609.

CATCHPOLE, E. A., FREEMAN, S. N., MORGAN, B. J. T. and HARRIS, M. P. (1998). Integrated recovery/recapture data analysis. *Biometrics* **54** 33–46.

CATCHPOLE, E. A. and MORGAN, B. J. T. (1994). Boundary estimation in ring recovery models. *J. Roy. Statist. Soc. Ser. B* **56** 385–391.

CATCHPOLE, E. A. and MORGAN, B. J. T. (1996). Model selection in ring recovery models using score tests. *Biometrics* **52** 664–672.

CATCHPOLE, E. A., MORGAN, B. J. T., FREEMAN, S. N., ALBON, S. D. and COULSON, T. N. (1998). An integrated analysis of Soay sheep survival data. Technical Report UKC/IMS/98/32, Univ. Kent, Canterbury, England.

CORMACK, R. M. (1992). Interval estimation for mark-recapture studies of closed populations. *Biometrics* **48** 567–576.

COWLES, M. K. and CARLIN, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *J. Amer. Statist. Assoc.* **91** 883–904.

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

FREEMAN, M. F. and TUKEY, J. W. (1950). Transformations related to the angular and square root. *Ann. Math. Statist.* **21** 607–611.

FREEMAN, S. N. (1990). Statistical analysis of avian breeding and survival. Ph.D. thesis, Univ. Kent.

FREEMAN, S. N. and MORGAN, B. J. T. (1992). A modeling strategy for recovery data from birds ringed as nestlings. *Biometrics* **48** 217–235.

GAMERMAN, D. (1997). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall, London.

GARTHWAITE, P. H., YU, K. and HOPE, P. B. (1995). Bayesian analysis of a multiple-recapture model. *Comm. Statist.* **24** 2229–2247.

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall, London.

GELMAN, A., MENG, X. L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statist. Sinica* **6** 733–759.

GEORGE, E. L. and ROBERT, C. P. (1992). Capture-recapture estimation via Gibbs sampling. *Biometrika* **79** 677–683.

GILKS, W. R., THOMAS, D. and SPIEGELHALTER, D. J. (1992). Software for the Gibbs sampler. *Comput. Sci. Statist.* **24** 439–448.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732.

JANZ, R. J. (1980). Prior knowledge and ornithology. In *Statist. Ornithology* (B. J. T. Morgan and P. M. North, eds.) 303–310. Springer, Berlin.

KADANE, J. B. and WOLFSON, L. J. (1998). Experiences in elicitation. *Statistician* **47** 3–20.

KASS, R. E., CARLIN, B. P., GELMAN, A. and NEAL, R. M. (1998). Markov chain Monte Carlo in practice: a roundtable discussion. *Amer. Statist.* **52** 93–100.

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

LEBRETON, J.-D., BURNHAM, K. P., CLOBERT, J. and ANDERSON, D. R. (1992). Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monographs* **62** 67–118.

LEBRETON, J.-D., MORGAN, B. J. T., PRADEL, R. and FREEMAN, S. N. (1995). A simultaneous survival rate analysis of dead recovery and live recapture data. *Biometrics* **51** 1418–1428.

LEE, S. M. and CHEN, C. W. S. (2000). Bayesian inference of population size for behavioral response models. *Statist. Sinica*. To appear.

LINK, W. and CAM, E. (1999). Of BUGS and birds: an introduction to Markov chain Monte Carlo. Unpublished manuscript.

MADIGAN, D. and YORK, J. C. (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika* **84** 19–31.

MORGAN, B. J. T. and FREEMAN, S. N. (1989). A model with first-year variation for ring-recovery data. *Biometrics* **45** 1087–1101.

O'HAGAN, A. O. (1998). Eliciting expert beliefs in substantial practical applications. *Statistician* **47** 21–36.

RIPLEY, B. D. (1987). *Stochastic Simulation*. Wiley, New York.

RUBIN, D. B. (1992). Using the SIR algorithm to simulate posterior distributions. In *Bayesian Statistics 3* (J. M. Bernardo, M. M. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 395–402. Oxford Univ. Press.

SCHWARZ, C. J. and SEBER, G. A. F. (1999). Review of "Estimating Animal Abundance III." *Statist. Sci.* **14** 427–456.

SPIEGELHALTER, D. J., THOMAS, A. and BEST, N. G. (1996). Computation on Bayesian graphical models. In *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 407–425. Oxford Univ. Press.

SPIEGELHALTER, D. J., THOMAS, A., BEST, N. G. and GILKS, W. R. (1996). *BUGS: Bayesian Inference using Gibbs Sampling. Version 0.50*. MRC Biostatistics Unit, Cambridge.

TANNER, M. and WONG, W. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22** 1701–1762.

UNDERHILL, L. G. (1990). Bayesian estimation of the size of a closed population. *Ring* **13** 235–254.

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

WAKEFIELD, J. C., GELFAND, A. E. and SMITH, A. F. M. (1991). Efficient generation of random variates via the ratio-of-uniforms method. *Statist. Comput.* **1** 129–133.

WHITE, G. C. and BURNHAM, K. P. (1999). Program MARK: survival estimation from populations of marked animals. *Bird Study* **46** (suppl.) 120–139.