HETEROGENEITY IN CAPTURE-RECAPTURE: BAYESIAN METHODS TO BALANCE REALISM AND MODEL COMPLEXITY

by

Simon J. Bonner

B.Sc., McGill University, 2001

M.Sc., Simon Fraser University, 2003

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
in the Department
of
Statistics and Actuarial Science

SIMON FRASER UNIVERSITY
Fall 2008

APPROVAL

Name: Simon J. Bonner

Degree: Doctor of Philosophy

Title of thesis: Heterogeneity in capture-recapture: Bayesian methods to balance realism and model complexity

Examining Committee: Dr. Thomas Loughin
Chair

Dr. Carl Schwarz, Senior Supervisor

Dr. Derek Bingham, Supervisor

Dr. Tim Swartz, Supervisor

Dr. Steven Thompson, SFU Examiner

Dr. Ken Newman, External Examiner
U.S. Fish and Widlife Service

Date Approved: _____

Abstract

Capture-recapture experiments are important for monitoring many endangered animal populations, such as salmon threatened by over-harvesting and migratory songbirds impacted by habitat loss. An important consideration in the analysis of capture-recapture data is potential variation in the probabilities of capture and survival. Failure to account for this variation can lead to incorrect inference, but traditional models incorporating heterogeneity may be very complex. This thesis presents three Bayesian methods that balance realistical modelling of variation in the capture and survival probabilities and increasing model complexity.

In the first project, I consider the analysis of data from two-sample experiments used in estimating the number of juvenile salmon leaving their spawning grounds. These migrations may last for several weeks and standard models may require many parameters to account for variations over time. My solution is to model the population size as a smooth function of time by fitting a Bayesian penalised spline. The method is applied to two datasets from the migration of juvenile salmon and provides more precise estimates of the population size that are less affected by outliers in the data than previous methods.

My second project addresses estimation of the size of an open population when individual capture or survival probabilities are functions of a time-dependent, continuous covariate. The main challenge is that these covariates can only be observed on occasions when an individual is captured. I develop a two-stage Bayesian method that first examines the covariate's effect by analysing the capture of marked individuals, and then applies the results to estimate the total population size. The model is used to study the dynamics of a population of Soay sheep (*Ovis aries*) whose survival is affected by body mass.

Finally, I develop a method to allow more flexibility in modelling the relationship between a covariate and individual survival probabilities. Standard methods assume that the

relationship is linear on some scale. My model incorporates Bayesian adaptive splines to allow smooth but local fitting of the linear predictor. I apply this model to study the effect of body condition on the survival of reed warblers (*Acrocephalus scirpaceus*) breeding in Holland.

Keywords: Adaptive spline; Bayesian inference; Capture-recapture; Hierarchical modelling; Penalized spline; Time-dependent, Continuous covariate

Acknowledgments

I am indebted to Dr. Carl Schwarz who supervised both my M.Sc. and Ph.D. work at Simon Fraser University. Dr. Schwarz afforded me the trust to do my own work, but was always ready and willing to help when needed.

In the spring of 2006, I spent four months working with Dr. Bryon Morgan at the University of Kent at Canterbury and Dr. Ruth King at the University of St. Andrews. This was a very valuable experience for me and I am grateful to Dr. Morgan and Dr. King for making this possible.

During my PhD I was supported by the Scott Paper Ltd. Bicultural Award administered by Simon Fraser University, a CGS-D scholarship from the National Science and Engineering Research Council, and an Internship from the MITACS Accelerate BC program. I would also like to thank the following people and organizations who supplied data for use in the development of the methods in my thesis: Brent Mossop and Caroline Melville from BC Hydro, Dr. David Thompson, Dr. Tim Coulson, and the Hoopla Valley Tribal Fisheries Department.

Alyssa, thank you for all of your support over the past four and a half years. You made many sacrifices, and I could not have done this without you. You are, and always will be, my best friend.

Thanks also for the support and friendship of the students in the Department of Statistics and Actuarial Science who made my time at SFU so enjoyable. Particular mention goes to: Laurie Ainsworth, Kelly Burkett, Wendell Challenger, Carolyn Huston, Elizabeth Juarez-Colunga, Chunfang Lin, Jason Nielsen, Jean Shin, Darby Thompson, and Lihui Zhao.

Finally, I would like to thank Sadika Jungic, Kelly Jay, and especially Charlene Bradbury who have guided me through the process of thesis submission.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Animal populations worldwide face severe threats resulting directly from the impact of human activities. Pacific salmon (*Oncorhynchus spp.*) which migrate between the ocean and their freshwater spawning grounds are threatened by overfishing and by the construction of hydroelectric dams which block their migration routes. Populations of neotropical migrant songbirds have declined due to habitat fragmentation and loss to forestry and agriculture. Many more species are affected by climate change.

To develop management programs to ensure the survival of these species, we must understand their population dynamics and the factors that affect the populations over time. Most commonly, populations of wild animals are studied through capture-recapture experiments in which samples of individuals are repeatedly captured, marked so that they can be identified in subsequent samples, and released back into the population. An important consideration in the statistical analysis of the data collected in these experiments is that individuals in a population never behave in exactly the same way. Individuals may be affected by a variety of factors, and failure to account for the variations may produce incorrect inference including biased estimates of the quantities of interest (such as population size or survival probability) or underestimation of uncertainty. Conventional models may either make assumptions which ignore the variations between individuals or introduce large numbers of parameters to account for these differences, which can lead to further problems in inference. The three projects in this thesis develop methods that balance the need for realistic modelling of the variations in capture-recapture experiments with increasing model complexity.

1.1 Introduction to Capture-Recapture Methods

The statistical analysis of data collected in a capture-recapture experiment depends on three factors: the design of the experiment, the assumed behaviour of the individuals in the population, and the goals of the study. In the simplest capture-recapture experiments, two samples are collected from the population of interest. Individuals captured in the first sample are marked in some way and returned to the population. Fish may be stained with coloured dyes, have their fins clipped or, more recently, be implanted with passive integrated transponder (PIT) tags or acoustic tags that can be detected without physical recapture. Small songbirds are most often marked with uniquely numbered leg bands while larger birds may be marked with coloured leg bands or numbered neck bands which can be read from a distance. New technologies have allowed the use of natural markings such as unique skin patterning which can be identified in photographs (Holmberg et al., 2008) or DNA genotypes that can be identified from spoor or hair samples (Lukacs and Burnham, 2005). After allowing sufficient time for the marked individuals to mix back into the population, a second sample is captured, and the previously marked individuals are identified. A simple estimator of the capture probability is the proportion of marked individuals recaptured (assuming that all marked individuals remained in the population and had the same probability of recapture). Population size can be estimated by dividing the total number of individuals captured in the second sample by the estimated capture probability – the Lincoln-Petersen estimator (Williams et al., 2002, pg. 290).

In more complex capture-recapture experiments, more samples of individuals are captured either at a set of discrete times (the capture occasions) or continuously in time. This thesis considers only the more common practise of discrete capture occasions. When sampling times are discrete, the capture occasions may be spaced equally over time, perhaps daily or yearly, or may be clustered, sampling many days in each year but only during the month that individuals breed. Marks may be applied to the unmarked individuals captured on every capture occasion or only to those captured on the first occasion. All of the individuals captured in a sample may be returned to the population or some may be removed permanently (common in studies of fish populations when individuals may be captured in a commercial fishery). Samples may be collected from one location on all occasions or from several geographic locations and may be obtained strictly by recapturing individuals, by resighting them without physical capture, by recovering dead animals, or through

some combination of these mechanisms. Different models have been developed for all of these possible scenarios (see Amstrup et al., 2003; Williams et al., 2002; Seber, 1982, for overviews).

The second consideration in the analysis of capture-recapture data is the assumed behaviour of the individuals and how this may affect their probabilities of capture, survival etc. The primary distinction is whether the population of interest is closed or open. A population is closed if no individuals enter into (either by birth or immigration) or depart from the population (either by death or emigration) between the capture occasions. A population is open if individuals enter or depart so that the composition of the population, in terms of the exact identity of the individuals, changes over time. In general, the statistical models for analysing data from closed populations are simpler because there can be only one reason that an individual was not observed on a capture occasion – it must have been present and was simply not captured. When modelling an open population it is possible that an individual was not captured on a specific occasion because it had not yet entered the population or had left the population, either permanently or temporarily. Statistical models for analysing data from open populations must model at least some of the processes by which animals enter and/or leave the population depending on what assumptions can be made about the animals' behaviour and what data is considered as fixed.

Models must also account for the differences among individuals alive on each capture occasion and the changes in these individuals over time. Survival probabilities, capture probabilities, and other parameters of interest (e.g., fecundity) can depend on a variety of factors. Changes in the environment may increase or decrease the chances of survival of all individuals from one capture occasion to the next. Further differences between the individuals, such as age, sex, body mass or breeding status, may affect the chances of survival or capture at any given time. In some experiments, the process of capture itself may affect the behaviour of individuals, and these effects need to be considered in fitting statistical models. Individuals captured on one occasion may avoid traps in future, decreasing their probability of being captured on subsequent occasions (trap-shyness) or may learn to seek out traps if there is a reward for being captured (trap-happiness).

Differing objectives may also determine which aspects of the population dynamics need to be modelled and which can be ignored. Some capture-recapture studies aim to estimate the size of the population at a single instant or how population size changes over time. This requires models which make assumptions about the entire population, including the

individuals that were never captured during the experiments. Other studies may focus on the rates of entry to and exit from the population and the factors that affect these rates, and need only consider the individuals captured on one occasion or more.

1.2 Summary of Projects

The first project of my thesis (Chapter 2) addresses a problem in modelling data from two-sample capture-recapture experiments of closed populations. This type of experiment is commonly employed in monitoring salmon populations, either to estimate the number of smolts (young salmon) leaving the freshwater spawning grounds in one year or the number of adults returning upstream to breed. Fish are captured at two trapping locations along their migration route. Those captured at the first trap (the upstream trap for smolts and the downstream trap for adults) are marked so that they can be identified if caught again, and are returned to the river to continue their migration. At the second trap, a new sample of individuals is captured and the numbers of marked and unmarked individuals are recorded.

These migrations can last for very long periods of time so that trapping is conducted each day for several weeks or even months. During this time, the sampling probabilities or behaviour of the fish can change considerably and failure to account for this may produce biased estimates of the population size or its uncertainty. To account for the variations over time, the analysis is often conducted by stratifying the data – conceptually fitting separate models with independent parameters for each day or week of the experiment. While stratification does allow the changes over time to be modelled, the resulting models may have large numbers of parameters, and the data collected are often too sparse to produce precise estimates of the population size in each stratum (day or week) separately. Moreover, this strategy does not account for the temporal structure of the data: one would expect the number of fish migrating past the traps in one stratum to be similar to the numbers in adjacent strata and less similar to the numbers in much earlier or later strata. Models that reduce the number of parameters have been developed by imposing parametric assumptions on the time fish take to move between the trapping locations (Schwarz and Dempson, 1994) and by hierarchical Bayesian modelling (Mantyniemi and Romakkaniemi, 2002), but these still do not account for the temporal structure. My solution is to model the number of fish in each stratum as a smooth function of time, specifically using the Bayesian penalised spline model (Lang and Brezger, 2004). In section 2.4, I apply this model to analyse two data sets

collected from the migration of Atlantic salmon (*Salmo salar*) smolts along the Conne River in Newfoundland and from the migration of Chinook salmon (*Oncorhynchus tshawytscha*) smolts along the Trinity River in California. The new method provides estimates of the total number of migrating smolts that are more precise and less influenced by the presence of outliers than previous models.

My second project considers the problem of estimating the size of an open population from capture-recapture data when the sampling probability or survival probability of individuals depends on a time-dependent, continuous covariate like body mass. The challenge with this type of covariate is that its value can only be known for an individual on the occasion when that individual is captured. On the occasions when the individual is not captured the value of the covariate is unknown. Some analyses of capture-recapture data attempting to deal with the missing values have simply treated the covariate as a constant, considering the mean of the observed values for each individual as the fixed value on all occasions. Others have imputed the missing values by some method and then treated the imputed values as if they had actually been observed. Simulations that I have conducted show that both of these approaches may produce biased inference regarding the covariate's effect (Bonner, unpublished work). In previous work (Bonner and Schwarz, 2006), I developed a method for including such covariates as predictors of the individuals' capture and survival probabilities by modelling the distribution of the covariate to make inference about the missing values and then employing Bayesian methods to examine the covariate's effect. The project in Chapter 3 extends this work to estimate the population size on each capture occasion. The method is applied to study the dynamics over a 15 year period of a population of Soay sheep (*Ovis aries*) living on the Island of Hirta in the Scottish archipelago of St. Kilda. The results of this analysis show that body mass is an important predictor of the survival and capture probabilities, particularly for younger sheep. Simulations based on these results indicate that estimates of the population size may be biased if the effect of the covariate on the capture probabilities is ignored.

In Chapter 4, I develop a semi-parametric model to relate differences in individual survival probabilities to the effects of a continuous, time-dependent predictor. Again, this builds on the model of Bonner and Schwarz (2006). When dealing with continuous predictors of capture or survival probabilities, it is common to employ a generalised linear model framework which assumes that the effect of the predictor is linear on some transformed scale.

The specific transformation is almost always chosen for mathematical or computational convenience, not biological realism, and this choice often contains implicit assumptions about the covariate's effect that are not acknowledged. When studying complex biological relationships, the assumption of linearity may not adequately capture the predictor's effect, even if higher order polynomial terms are included in the linear predictor. Instead, I model the linear predictor as a smooth but flexible non-linear function of the continuous covariate by implementing the adaptive Bayesian spline model of Biller (2000). Simulations show that the new model is able to capture complex relationships between a covariate and the survival probability that cannot be fit by simple polynomial models. The method is applied to data obtained from the Dutch Constant-Effort-Sites bird banding study between 1994 and 2003 to examine the effect of individual reed warbler's (*Acrocephalus scirpaceus*) body condition on their probability of surviving from one year to the next. Details on the differences between the penalised and adaptive spline approach are discussed in section 1.3.3.

1.3 Common Methodology

1.3.1 Bayesian Methods

All three of these projects employ the Bayesian approach to statistical inference. In frequentist methods of inference, uncertainty about model parameters is quantified by how much the parameter estimates would vary if many data sets, all generated from the same model with the same true parameter values, were analysed in the same way. Bayesian inference quantifies uncertainty by assigning probability distributions to the parameters. Given a model defined through the likelihood function, $L(\boldsymbol{\theta}|\boldsymbol{X})$, inference is derived from the distribution of the parameters, $\boldsymbol{\theta}$, conditional on the observed data, $\boldsymbol{X}$. This distribution is called the posterior distribution of the parameters and its density will be denoted by $\pi(\boldsymbol{\theta}|\boldsymbol{X})$. Applying Bayes' rule to reverse the conditioning in the likelihood, the posterior density is computed as:

$$\pi(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{L(\boldsymbol{\theta}|\boldsymbol{X})\pi(\boldsymbol{\theta})}{\int L(\boldsymbol{\theta}|\boldsymbol{X})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

where $\pi(\boldsymbol{\theta})$ is the marginal density of the distribution of the parameters before any data is collected, called the prior distribution. The denominator in this expression is unique given $L(\boldsymbol{\theta}|\boldsymbol{X})$ and $\pi(\boldsymbol{\theta})$, and so the posterior distribution can be defined by the simpler

proportionality:

$$\pi(\boldsymbol{\theta}|\boldsymbol{X}) \propto L(\boldsymbol{\theta}|\boldsymbol{X})\pi(\boldsymbol{\theta}),$$

the product of the likelihood function and the prior density. In practise, inference is usually described in terms of summary statistics of the posterior distribution including the most probable values of the parameters (the mean, median or mode of the distribution) and regions of the parameter space with high posterior probability (credible intervals and regions) (see, for example, Lee, 2004; Gelman et al., 2003, for a thorough introduction to Bayesian inference).

One of the challenges, and most controversial components of Bayesian inference, is the selection of the prior distribution. Subjective Bayesian statisticians believe that the prior distribution should encode information about the parameters that is available before the data are collected, either from previous experiments or from expert knowledge on the system of interest. The prior distribution is chosen to concentrate high probability in regions of the parameter space believed more plausible before collecting the data. Bayes' rule can then be viewed as an explicit formula for combining these prior beliefs with new information in the observed data to obtain new beliefs described by the posterior distribution. If the information in the data is very strong then the prior and posterior distribution may be very different; if the information is weak then the prior and posterior may be very similar.

An alternative approach is to select prior distributions that are vague or non-informative so that inference is objective insofar as the shape of the posterior is determined almost exclusively by the data. Common choices for non-informative distributions are flat densities (which are improper if the parameter's value is unbounded and admissible only if the resulting posterior still integrates to 1), distributions with large variances and families of reference distributions, like Jeffrey's priors, which are invariant to certain transformations of the parameters. However, there is no single concept of a non-informative distribution and many formulations have been considered, both in general and for specific models (see for example Gelman, 2006).

A great advantage of Bayesian methods is that the prior distribution can be specified to incorporate extra structure in the model parameters that is not contained in the likelihood function. In most capture-recapture experiments, it is not realistic to believe that the capture probabilities are exactly the same on every capture occasion, but these values are related and will usually be very similar. If there is no information to differentiate between

the capture probabilities on each occasion then the parameters are exchangeable. Let $\boldsymbol{p} = (p_1, \ldots, p_T)$ denote the vector of capture probabilities for a capture-recapture experiment with T capture occasions. The elements of $\boldsymbol{p}$ are exchangeable if the joint distribution of $\boldsymbol{p}$ is independent of the ordering of the elements (Gelman et al., 2003, pg. 121). This structure can be included in a Bayesian model by defining a hierarchical prior distribution for the set capture probabilities. Instead of defining an independent prior for the capture probability on each capture occasion, the hierarchical prior models the set of capture probabilities as a random sample from a further distribution. This distribution is called the hyper-prior and its parameters the hyper-parameters. Hyper-parameters may be assigned fixed values or they may be modelled further. For example, a simple Bayesian model would assign independent prior distributions to the capture probabilities such that:

$$\text{logit}(p_j) \sim N(\mu_j, \sigma_j^2), \quad j = 1, \ldots, T$$

where μ_j and σ_j^2 are assigned fixed values, not necessarily the same for all j. A hierarchical model might instead model the capture probabilities such that:

$$\text{logit}(p_j) \sim N(\mu, \sigma^2), \quad j = 1, \ldots, T$$

where the hyper-parameters, μ and σ^2, are treated as unknown but equal for all j. These hyper-parameters may then be assigned fixed hyper-priors, for example:

$$\begin{aligned} \mu &\sim N(0, \tau^2) \\ 1/\sigma^2 &\sim \text{Gamma}(\alpha, \beta) \end{aligned}$$

or may themselves be modelled further if the experiment is repeated in several years or at several different locations. Models of this type are referred to as hierarchical Bayesian models because of the hierarchical structure of the prior distribution (see Gelman et al., 2003, ch. 5).

The gain from modelling the parameters in this way is that information about one parameter in the related set is included in inference about the other parameters in the set. For example, in a capture-recapture study of an open population, the data contains no information about the capture probability on the first capture occasion, p_1. However, if the hierarchical prior above is used then inference about the μ and σ^2 is obtained through

$p_2, \ldots, p_T$ from the data on the other capture occasions. This information can then be used to learn about the capture probability on the first capture occasion and other important quantities, like the initial population size and the rate of recruitment between the first and second capture occasion. Similarly, if the sample size on day j is small and the sample size on day k is large, then the information from day k will be used in estimating p_j through the estimation of the common hyper-parameters μ and σ^2. Hierarchical modelling plays an important role in all three projects in this thesis.

Another complication in Bayesian inference lies in computing the summary statistics for the posterior distribution. In simple problems, the posterior distribution has a simple, recognisable form that is easy to interpret. However, in most practical applications the posterior distribution is too complex for summary statistics to be derived analytically, especially when the model contains a large number of parameters. One solution is to approximate the posterior distribution by a simpler form. Under regularity conditions, the posterior will converge to a normal distribution as the sample size is increased, and so a normal approximation to the posterior may be appropriate (Gelman et al., 2003, pg. 101). A second solution is to sample from the posterior distribution and then approximate the summary statistics by the equivalent sample statistics (e.g., approximating the true posterior mean by the average of the sampled values).

Many different methods exist for sampling from general distributions, but the most commonly implemented in Bayesian applications is Markov chain Monte Carlo (MCMC) sampling. An MCMC algorithm works by defining the transition probabilities of a Markov chain so that starting from any initial values for the parameters, the distribution of the i^{th} realisation of the chain converges to the distribution of interest as i increases. In practise, realisations from the chain are simulated for a sufficiently long burn-in period until the distribution has (approximately) converged and the values generated on the subsequent iterations are treated as realisations from the distribution of interest.

Samples from the posterior distributions in Chapters 2 and 3 are obtained by MCMC sampling implemented in the `OpenBUGS` software package (Thomas et al., 2006). This software makes use of two specific MCMC algorithms, the Gibbs sampler and the Metropolis-Hastings (MH) algorithm, and their variants. The Gibbs sampler generates realisations from the posterior distribution by drawing values from the distribution of each parameter conditional on the data and all other parameters in the model (the full conditional distribution). On each iteration of the Markov chain, a new value for each parameter is generated from

its full conditional on the most recently sampled values for the other parameters (Gelfand, 2000; Tierney, 1994; Gelfand and Smith, 1990). If full conditionals cannot be computed for all parameters, or can be computed but cannot be sampled easily, then the MH algorithm can be used instead. Similar to the Gibb's sampler, the MH algorithm generates new values for each parameter on each iteration of the chain by sampling conditional on the most recent values for the remaining parameters. Rather than sampling from the full conditional distribution of a parameter, a value is generated from a specified proposal distribution, with density $Q(\theta)$, which may depend on the current value of the parameter, the remaining parameter values and the data. The value drawn from the proposal distribution is then accepted as the new value of the parameter with probability:

$$A = \min\left(1, \frac{\pi(\theta'|X) \cdot Q(\theta)}{\pi(\theta|X) \cdot Q(\theta')}\right) \tag{1.1}$$

where θ denotes the current value of the parameter and θ' the proposed value. If the proposal is not accepted, then the previous value of the parameter is retained and the algorithm continues with sampling the remaining parameters. More details of the MH algorithm are provided in many books on Bayesian inference and statistical computing (see Lee, 2004; Gelman et al., 2003; Chen et al., 2000; Gilks et al., 1996; Chib and Greenberg, 1995).

While the MH algorithm guarantees that the parameter values from subsequent iterations of the Markov chain converge in distribution to the posterior (under simple regularity conditions), many iterations may be required to reach convergence. Moreover, the values sampled on adjacent iterations are almost always dependent and the auto-correlation is often be very large. This means that many more samples would have to be collected to obtain a specified level of precision in estimating the posterior summary statistics than would be expected if the samples were independent. Some theoretical results exist concerning the time to convergence and autocorrelation for simple models, but not for complex chains with many parameters. To monitor the convergence of the chains generated with the MH algorithm, I have implemented the Gelman-Rubin-Brooks (GRB) convergence diagnostic (Brooks and Gelman, 1998,Gelman et al., 2003, pg. 296). The GRB diagnostic is computed by running several Markov chains in parallel starting from widely spaced initial values, and then comparing the distribution of the samples obtained from each chain. In particular, the GRB diagnostic compares some measure of the dispersion of the distribution within each chain (the variance or distance between the upper and lower $\alpha/2$ percentiles) with the dispersion

of the distribution of the values from all chains. At convergence, the dispersion within and between the chains should be the same. In practise, the GRB statistic estimates the ratio of the dispersion and adequate convergence is achieved when the ratio is close to 1. Gelman et al. (2003, pg. 297) recommends values of 1.1 or smaller.

In Chapter 4, the posterior distribution is a mixture of distributions with different dimensions formed from a set of candidate models with different numbers of parameters. Samples from this distribution cannot be obtained with either the Gibbs sampler or the standard MH algorithm described above. Instead, a sample is obtained from the posterior distribution via the reversible-jump Markov chain Monte Carlo (RJMCMC) algorithm, which extends the MH algorithm to the multi-dimensional parameter space (Green, 1995). Like the MH algorithm, each iteration of the RJMCMC algorithm involves proposing a new state for the parameters conditional on the current parameter values and accepting or rejecting the new values. The difference is that the proposal distribution selected for the RJMCMC algorithm must allow for the chain to move between subsets of the parameter space with different dimensions. This is usually accomplished in two steps by first selecting a model (perhaps by randomly adding or removing a predictor variable in the current model) and then proposing new parameter values for the selected model. The probability with which the proposal is accepted must also be modified to account for the possible change in dimension between the current and proposed parameter vectors. The new acceptance probability is:

$$A = \min\left(1, \frac{\pi(\theta'|X)\cdot Q(\theta,u)}{\pi(\theta|X)\cdot Q(\theta',u')}\mathcal{J}((\theta,u),(\theta',u'))\right) \tag{1.2}$$

where u and u' are random variables required to produce a bijection between the parameter spaces and $\mathcal{J}((\theta,u),(\theta',u'))$ is the Jacobian of this transformation. The acceptance probability was first derived by Green (1995); a simplified explanation is given in Waagepetersen and Sorensen (2001) and more details are available in recent books on Bayesian analysis or statistical computation (e.g. Chen et al., 2000). When implementing an MCMC algorithm for a complex model, different methods can be combined for updating the parameters on each iteration. If the full conditional distributions can be computed for some parameters, then these parameters can be updated by Gibbs sampling while the remaining parameters are updated in MH or RJMCMC steps.

1.3.2 Model Comparison and Assessing Model Fit

In Chapters 2 and 3, several models with similar structure are tested in application to the real data sets. I have chosen to compare these models with the deviance information criterion (DIC) introduced by Spiegelhalter et al. (2002). Given a sample of K vectors from the posterior distribution of the model, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$, Spiegelhalter et al. (2002) define a measure of the effective number of parameters in the model as:

$$p_D = \frac{1}{K}\sum_{k=1}^{K} D(\boldsymbol{\theta}_k) - D(\hat{\boldsymbol{\theta}}),$$

where $D(\boldsymbol{\theta}) = -2 \cdot L(\boldsymbol{\theta}|\boldsymbol{X})$ is the deviance function and $\hat{\boldsymbol{\theta}}$ is an estimate of the parameter values, usually taken to be the posterior mean. By itself, p_D provides a measure of complexity that is interpreted as the number of unique, estimable parameters in the model. This is particularly useful for understanding the structure of a hierarchical Bayesian model in which the number of parameters depends on the variance of the prior. Continuing the example of modelling capture probabilities on each capture occasion as independent draws from a normal distribution with unknown mean and variance, on the logistic scale, the number of unique capture probabilities depends on the magnitude of this variance. If the variance is very close to zero, then the capture probabilities will all be effectively the same and there only one parameter to estimate. If there were no other parameters in the model then p_D would be close to 1. As the variance gets larger the capture probabilities become less dependent on each other and p_D would increase.

Given this measure of the complexity of the model, the DIC is defined in analogy to classical model comparison criterion, like the AIC, by adding twice the effective number of parameters to the deviance function. This produces the criterion:

$$\text{DIC} = D(\hat{\boldsymbol{\theta}}) + 2p_D. \tag{1.3}$$

Heuristically, the DIC penalises the likelihood for models which have better fit to the data but many more parameters. A model with increased complexity (as measured by p_D) will have a lower DIC value than a simpler model only if the improvement of the fit to the data (as measured by the deviance at $\hat{\boldsymbol{\theta}}$) outweighs (twice) the extra count of parameters. In practise, Spiegelhalter et al. (2002) suggest that differences in the DIC of 3 or more be

considered as strong evidence in favour of one model, while differences of 2 or less may be attributable to Monte Carlo error and should not be interpreted in favour of the model with lower DIC.

The primary advantage of the DIC over other Bayesian methods for model comparison and selection (e.g. Bayes factors or posterior model probabilities) is ease of computation. The DIC is computed directly from the sample of realisations from the posterior obtained through MCMC. The method also provides a direct measure of each model's complexity and there is no requirement that the different models be nested. A disadvantage is that the DIC does not provide a basis for multi-model inference by model averaging. This leads to difficulties when several models perform equally well because the DIC cannot be used to combine the separate inferences from these models into a single inference. Model averaging could be performed through RJMCMC (Green, 1995), posterior model weights (Raftery et al., 1997) or the Bayes Information Criterion for each model (Schwarz, 1978). These procedures require more complex computations than the DIC, and have not been explored in Chapters 2 and 3. In Chapter 4, the adaptive spline model is fit via RJMCMC, which effectively produces inference averaged over a set of models with different numbers of predictor functions (splines with differing numbers of knots).

Along with model selection, tools are also needed to check that a single, selected model is adequate. In Chapters 2 and 3 I have chosen to assess the fit of the selected models by the method of posterior predictive p-values (also called Bayesian p-values) (Meng, 1994; Gelman et al., 1996; Gelman et al., 2003, pg. 157-177). The philosophy underlying this method is that data simulated from the model will resemble the observed data if and only if the model fits the data well. Let $D(\boldsymbol{X}, \boldsymbol{\theta})$ denote some measure of the discrepancy between the data and the model given the parameters $\boldsymbol{\theta}$. The p-value associated with this discrepancy is (Gelman et al., 2003, pg. 162):

$$p = \int \int 1[D(\boldsymbol{X'}, \boldsymbol{\theta}) > D(\boldsymbol{X}, \boldsymbol{\theta})] f(X'|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}|X)\, dX'\, d\boldsymbol{\theta},$$

the probability under the posterior predictive density, $f(X'|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|X)$, that simulated data, $\boldsymbol{X'}$, will produce a larger discrepancy than the observed data. In practise, the integral cannot be computed analytically and is approximated by simulation. Several sets of parameters, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$, are generated. A new set of data is computed for each set of parameters, denoted

by $\boldsymbol{X}'_k$ for $\boldsymbol{\theta}_k$, and the p-value is approximated by:

$$p = \frac{1}{K}\sum_{k=1}^{K} 1[D(\boldsymbol{X}'_k, \boldsymbol{\theta}_k) > D(\boldsymbol{X}, \boldsymbol{\theta}_k)].$$

There is no evidence for poor fit of the model if p is near .5.

The essential concept of the Bayesian p-value is very similar to that of a classical goodness-of-fit (GOF) test. In the classical GOF framework, a test statistic is computed from the observed data and compared to its sampling distribution under the assumption that the true parameters are fixed and equal to the computed estimates. The p-value is the probability that a new data set from the same model would produce test statistic more extreme than observed. There are two key differences between classical GOF and the Bayesian p-value. First, the Bayesian p-value does not consider a single, fixed set of parameters and instead integrates the probability over the entire space weighted according to the posterior. Second, the discrepancy function may depend on both the data and the parameter values – it need not be a test statistic in the classical sense. Moreover, as shown in the simulations of Chapter 2 the distribution of the Bayesian p-value is not necessarily uniform over the interval $(0, 1)$ for repeated data sets simulated from the true model. This has important implications for diagnosing lack of fit.

Allowing the discrepancy measure to depend on both data and parameters provides great flexibility in the types of function used to test a model's fit. Brooks et al. (2000) provide some specific guidelines for the use of Bayesian p-values in assessing mark-recapture models. In particular, they recommend measures of discrepancy based on the Freeman-Tukey statistic for comparing observed counts with their expected values:

$$\left(\sqrt{x} - \sqrt{E(X|\boldsymbol{\theta})}\right)^2$$

where x is the count of individuals with some outcome (capture history in the current context) and $E(X|\boldsymbol{\theta})$ its expected value. This measure assigns less weight to outcomes with small expected counts than other measures, like Pearson's χ^2, and provides more robust assessment when the expected counts of some outcomes are very close to 0. Note that the expected counts are functions of the parameter values and so this function would not provide a proper test statistic unless the parameter values were replaced with point estimates.

1.3.3 Splines

Splines are functions commonly employed in the problem of smoothing bivariate data – selecting a flexible function to describe the relationship in bivariate data with few assumptions beyond that the function must be continuous. The models developed in Chapters 2 and 4 both incorporate splines to describe the dependence of some quantity (specifically, daily population sizes in Chapter 2 and individual survival probabilities in Chapter 4) on an auxiliary variable (time in Chapter 2 and a continuous covariate in Chapter 4). Part of the appeal of fitting splines is that the model can be formulated as an extension of polynomial regression and so much of the theory and methods of linear models applies.

Suppose that we wish to model a response variable, y, as a function of a predictor, x, in the range $[\xi_l, \xi_u]$. The disadvantage of a straight polynomial regression that the fit of a polynomial is global: the value of the function over any (small) interval determines its value over the entire range of data (Schumaker, 1993, pg. 103). If the relationship between the two variables has a complicated shape with local features, then the polynomial model may require very high degree to adequately describe the data. Splines allow for local behaviour by introducing an extra set of predictor functions which are non-zero only over sub-intervals of the range of x. The set of basis functions for a spline of order q includes the first q polynomial basis functions, namely $\{1, \ldots, x^{q-1}\}$, as well as K truncated polynomial terms defined as:

$$(x - \xi_k)_+^{q-1} = \begin{cases} 0 & x < \xi_k \\ (x - \xi_k)^{q-1} & x \geq \xi_k \end{cases}$$

for some set of K points, $\xi_l = \xi_0 < \xi_1 < \ldots < \xi_K < \xi_{K+1} = \xi_u$, called the knots of the spline. In the interval between any adjacent pair of knots, $[\xi_{k-1}, \xi_k]$, the spline will be equal to a polynomial of order q. In the next interval, $[\xi_k, \xi_{k+1}]$, the exact form of the polynomial will be altered by the addition of some multiple of the term $(x - \xi_k)^{q-1}$. The spline will still be continuous and will have $q - 2$ continuous derivatives over the entire range, but the $q - 1^{st}$ derivative will be discontinuous with a jump at each knot. Cubic splines (order $q = 4$) are employed in both chapters so that the fitted curve is a sequence of connected cubic polynomials with continuous first and second derivatives and jumps in the third derivative at each knot (Schumaker, 1993, pg.108).

When the order of the spline model is fixed, the flexibility of the fitted curve is determined by two factors: 1) the number and placement of the knots and 2) the size of the jumps allowed

in the $(q-1)^{st}$ derivative at each knot point. Splines with larger numbers of knots and/or larger jumps in the derivative will produce a wider range of shapes but generally less smooth fits. This dichotomy leads to two competing approaches for fitting splines to bivariate data. Penalised spline methods begin by defining a large number of knots with fixed positions and then constraining the size of the jumps allowed at each knot to maintain smoothness (Eilers and Marx, 1996; Ruppert et al., 2003, pg.65). Adaptive or free-knot spline methods allow the larger jumps but include the placement of knots in fitting the curve so that knots are located only where necessary and the number of knots is kept as small as possible (Jupp, 1978; DiMatteo et al., 2001). In Chapter 2, I implement the Bayesian penalised spline method of Lang and Brezger (2004). The method developed in Chapter 4 employs the Bayesian adaptive spline model of Biller (2000).

In practise, computation with the basis of truncated polynomial functions can lead to numerical problems. The columns of the design matrix will be highly correlated if two knot points are very close to each other and the entries of the design matrix may be very large which makes matrix operations unstable. These problems can be avoided by working with an equivalent basis – a set of $q+K$ functions such that every spline can be written as a linear combination of the new set of functions and vice versa. A common choice is the basis of B-spline functions (De Boor, 1978; Schumaker, 1993; Ruppert et al., 2003). The advantages of B-splines over the truncated polynomial basis are: 1) that the basis functions are positive and sum to 1 at every point in $[\xi_l, \xi_u]$, 2) as a result, the values of the design matrix will always be between 0 and 1, and 3) each basis function is local in that it is positive only over a fixed sub-interval. This last point means that the design matrix has a banded structure which further simplifies the computation. The methods of both Chapters 2 and 4 make use of the B-spline basis instead of the truncated polynomial basis. Formulas for computing the B-spline basis functions are provided in De Boor (1978) and Schumaker (1993).

Chapter 2

Bayesian P-Spline Methods for Modelling Sparse Data from the Stratified-Petersen Experiment

2.1 Introduction

Stratification is a commonly employed in modelling capture-recapture data to avoid potential biases of abundance estimates that might be caused by heterogeneity in the capture probabilities among individuals in the population of interest. Instead of estimating the total abundance directly, the population is divided into groups of individuals, or strata, for which the capture probabilities can be assumed equal. Estimates are then computed separately for each stratum and summed to produce an estimate of the total population size. While this strategy does allow heterogeneity to be accounted for, the numbers of individuals marked and/or recaptured in some strata may be very small. This can lead to numerical problems in computing estimates for some strata or produce estimates with very low precision. My approach is to explicitly model the population size of each stratum using splines in order to share information among strata and increase the precision of the estimates of abundance in strata with sparse data.

While stratified capture-recapture models are used in studies of many different animals and stratification may be based on many factors (e.g., geographic location, age or sex), I will consider the specific case of the two-sample experiment with individuals stratified

temporally. These types of experiments are commonly used in fisheries research to monitor the number of salmon migrating along a river – either as juveniles moving from freshwater to the ocean or as adults returning to the spawning grounds to breed. At one location in the river individuals are trapped, marked in some way, and returned to the population. Farther along the river (downstream for juveniles and upstream for adults) a second trap collects a sample which will contain both marked and unmarked fish (see Figure 2.1). The proportion of marked fish recaptured provides information about the capture probability, which can be combined with the total number captured to estimate the population size.

This design is modified in some experiments so that trapping need be conducted at only one location in order to reduce the cost and effort required. Before the single trapping location a group of marked individuals is released into the river. Often, these are individuals that were captured earlier in the experiment. Samples are then captured from both the population of interest and the introduced group of marked individuals. Figure 2.2 shows how this emulates the second trapping location in the experiment with two traps. The difference in analysing the data from this experiment is that the population of interest does not normally include the marked individuals. If these individuals were previously captured then they have already been included in the estimate of population size and do not need to be counted again.

If it is reasonable to believe that the probability of being caught in the second trap is the same for all individuals, then no stratification is necessary. Assuming that the population is closed, i.e. that no individuals enter or leave the population between the two capture locations, and that individuals behave independently, the probability of capture at the second trap can be estimated by the proportion of marked individuals that were recaptured during the entire experiment. The total population size is then approximated by a Horvitz-Thompson type estimator, dividing the total number of individuals caught at the second trap by the estimated capture probability. The resulting estimator is commonly referred to as the Lincoln-Petersen index (Seber, 1982, pg. 59).

However, salmon migrations (also called runs) often last for several weeks and the capture probabilities may vary considerably over this time. For this reason, it is common to stratify the population by time, estimating capture probabilities and population sizes separately for each day, or perhaps week, of the experiment. If sufficient numbers of fish are marked in each stratum then the Lincoln-Petersen estimator, or a variation, can be applied to each stratum separately and the resulting estimates summed to produce an estimate of total

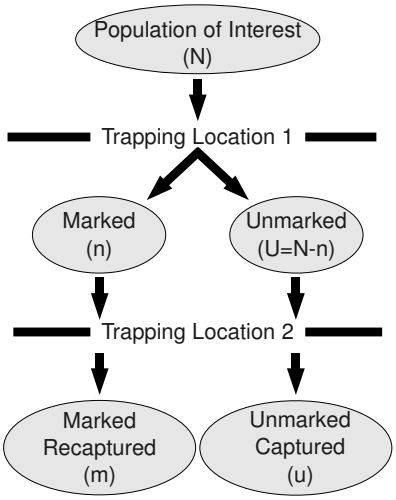

Figure 2.1: Schematic diagram of the standard two-sample capture-recapture experiment. A sample of size n from the population of size N is captured at the first trapping location, marked and released. At the second location samples are obtained from the populations of both marked and unmarked individuals. Notations for the number of individuals in each group are included in brackets.

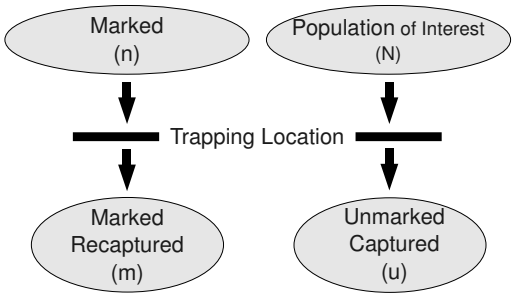

Figure 2.2: Schematic diagram of the modified two-sample capture-recapture experiment with only one trapping location. At the single location samples are captured from the populations of both marked and unmarked individuals. Notations for the number of individuals in each group are included in brackets. Marked individuals are, most often, excluded from the population of interest to avoid double counting.

abundance. Often there are some days in the experiment when very few fish are marked, making the estimated capture probability for this day, and hence the estimate of abundance, very imprecise (Seber, 1982, pg. 60).

A second challenge that arises in most experiments of this type is that fish marked at the first site on one day do not necessarily all pass the second site on the same day. Instead, some fish may migrate very quickly so that they pass the second trap on the same day that they are released at the first while other fish may move more slowly and don't pass the second site until several days after being marked and released. When this occurs, it is not possible to know exactly how many marked fish were present on each day at the second location, and so the Lincoln-Petersen estimates cannot be computed for each day directly. Instead, it is necessary to model the movement of the marked fish between the two sites in order to approximate the number of marked fish available on each day and then to estimate the capture probability Darroch (1961). This adds even more uncertainty to the daily estimates of population size.

Provided that unique marks are applied on each day it is possible to know when a recaptured fish was originally marked, and data from these experiments are commonly summarised by a matrix whose i, j entry indicates the number of fish marked on stratum i and recaptured in stratum j. This matrix will be diagonal if it is known that fish passing the first location in stratum i always pass the second location in stratum i and experiments generating such data will be referred to as diagonal experiments. For example, data from a diagonal experiment with three strata might be displayed as:

Marked	Recaptured		
10	1		
10		1	
10			1
Unmarked	100	100	100

indicating that in each stratum 10 individuals were marked at the first location and 1 was recaptured at the second location along with 100 more unmarked individuals. This type of data might arise if individuals are stratified by day and the traps are close enough to each other so that fish take at most a few hours to move from the first location to the second. The more general experiment in which individuals marked in stratum i may be recaptured outside of this stratum will be referred to as the non-diagonal stratified experiment. Sample data might be displayed as:

Marked	Recaptured		
10	1	1	1
10		1	1
10			1
Unmarked	100	100	100

indicating that 10 individuals were marked in each stratum at the first location, 1 individual from each was recaptured at the second location in each of the subsequent strata and a further 100 unmarked individuals were captured in each stratum at the second location. The empty cells indicate structural zeros in the data; when the strata are based on time it is not possible for individuals marked in one stratum to be recaptured in a previous stratum.

The standard methods for analysing stratified data treat the counts in each stratum as completely independent of counts in all other strata. This is too general for studies of migrating salmon and other temporally stratified mark-recapture data. While fluctuations in the counts from day-to-day will always occur, migrations tend to follow a fairly predictable pattern: few fish pass on the days early in the migration period, the numbers grow fairly steadily to one or two peaks in the middle of the migration and then drop back down at the end of the period. The result is that the abundance of fish on one day is strongly associated with the abundance on the neighbouring days. The objective of my work in this project is to model the daily abundance of fish passing the second site as a smooth curve in order to explicitly model this dependence. To capture this behaviour without making too many assumptions about the exact shape of the smooth curve, I model the daily abundance by fitting a spline. In particular, I apply the Bayesian P-spline model developed by Lang and Brezger (2004).

The format of this chapter is as follows. Sections 2.2 and 2.3 describe the development of the spline model for the cases of diagonal and non-diagonal stratified experiments, and provide results from simulation studies comparing the model with other estimators of abundance for stratified data. In section 2.4, the model is applied to two data sets, the first collected during the migration of juvenile Atlantic salmon (*Salmo salar*) through the Conne River in Newfoundland in 1987 and the second collected during the migration of juvenile Chinook salmon (*Oncorhynchus tshawytscha*) through the Trinity River in California in 2003. This is followed by a discussion of the results and some extensions of the model.

2.2 The Diagonal Stratified-Petersen

2.2.1 Introduction

In this section, I consider data from experiments in which individuals cannot change strata between the two sampling locations. This would be the case in estimating the total size of a salmon run if, for example, the two sampling locations were close enough to guarantee that fish always pass both sites in one day. Suppose that fish are marked and released for a total of s consecutive days. Let n_i denote the number of fish marked and released at the first site on day i, m_i the number of these fish that are recaptured at the second site, and u_i the number of unmarked fish captured at the second site. Let U_i denote the total number of unmarked individuals which pass the second site on day i. The objective is to estimate $U_1, \ldots, U_s$ from which the total population size can be computed as $N = \sum_{i=1}^{s}(n_i + U_i)$.

The standard assumptions for modelling data from this type of study are that the samples of marked and unmarked fish captured at the second location on each day form simple random samples from the sets of marked and unmarked fish available, with the same sampling probability for both. More formally, it is assumed that:

1) the population is closed (i.e., no fish enter or leave the population between the two sites)

2) marks are not lost between the two samples

3) all fish in one stratum have the same probability of being captured at the second site

4) and whether or not any individual is captured at the second site is independent of the capture of all other individuals.

Under these assumptions, the numbers of marked and unmarked fish captured at the second site on day i have independent binomial distributions:

$$m_i \sim \text{Binomial}(n_i, p_i) \text{ and } u_i \sim \text{Binomial}(U_i, p_i)$$

where p_i is the probability that any individual in the i^{th} stratum is captured at the second site. The likelihood for the two sets of parameters $\boldsymbol{p} = (p_1, \ldots, p_s)$ and $\boldsymbol{U} = (U_1, \ldots, U_s)$ is then constructed by multiplying the contributions from the marked and unmarked fish for

each day:

$$L(\boldsymbol{p}, \boldsymbol{U}|\boldsymbol{n}, \boldsymbol{m}, \boldsymbol{u}) = \prod_{i=1}^{s} \left[\binom{n_i}{m_i} p_i^{m_i} (1-p_i)^{n_i - m_i} \cdot \binom{U_i}{u_i} p_i^{u_i} (1-p_i)^{U_i - u_i} \right]. \tag{2.1}$$

Note that the model makes no assumptions regarding the capture and marking of individuals at the first site.

2.2.2 Previous Estimators of Abundance

Given this model, the simplest method for estimating the total population size is to estimate each U_i separately given n_i, m_i and u_i and then to sum these values. The intuitive estimator of p_i, and also maximum likelihood estimator (MLE), is the proportion of marked fish recaptured, $\hat{p_i} = m_i/n_i$. Equating the expected and observed number of unmarked fish captured on day i and substituting $\hat{p}_i$ then yields the estimator $\hat{U}_i = (n_i u_i)/m_i$ or equivalently $\hat{N}_i = (n_i \cdot (u_i + m_i))/m_i$, which is the Lincoln-Petersen estimator of abundance.

This estimator presents two problems in practice, particularly when the capture probabilities are small. First, the estimate does not exist when $m_i = 0$ since this entails division by 0. Second, the estimator is biased for all experiments – in fact, $E(\hat{N}_i) = \infty$ because there is always a non-zero probability that $m_i = 0$. The estimator commonly used instead is the Chapman estimator (Seber, 1982):

$$\tilde{N}_i = \frac{(n_i + 1)(u_i + m_i + 1)}{(m_i + 1)} - 1. \tag{2.2}$$

which essentially adds 1 to the counts of marked, recaptured and unmarked inviduals caught in order to avoid division by 0, and then subtracts 1 from the final estimate to account for the change. This estimator can always be computed and is known to be unbiased when $n_i + u_i + m_i \geq N_i$, a condition which ensures that at least one marked individual was recovered at the second site. In the case where $n_i + u_i + m_i < N_i$, Seber (1982, pg. 60) notes that the bias is negligible ($< .02$ 95% of the time) provided that $m_i \geq 7$. The variance of the Chapman estimator for a single strata can approximated by (Seber, 1982):

$$\widehat{V(\tilde{N}_i)} = \frac{(n_i + 1)(m_i + u_i + 1)(n_i - m_i)u_i}{(m_i + 1)^2(m_i + 2)}. \tag{2.3}$$

By independence, the variance of the estimator of total abundance, $\tilde{N} = \sum_{i=1}^{s} \tilde{N}_i$, is approximated by the sum of the variances for the individual strata, $\widehat{V(\tilde{N})} = \sum_{i=1}^{s} \widehat{V(\tilde{N}_i)}$.

Seber (1982) demonstrates that the relative standard error of the Chapman estimator for a single strata is approximately $1/\sqrt{E(m_i)}$. When $E(m_i)$ is very small, the relative standard error will be very large and the estimate of population size for stratum i will be highly variable. With $E(m_i) = 5$ the standard error of $\hat{N}_i$ is approximately $.45\hat{N}_i$ and with $E(m_i) = 1$ is approximately $\hat{N}_i$. The standard solution to this problem is to pool the data – i.e., to reduce the number of strata by adding n_i, m_i and u_i for several consecutive days, and then computing the Chapman estimates for each of the new strata. Complete pooling results if the analysis ignores the stratification entirely, summing the data over all days and computing an estimate of the total population size directly. Let $n = \sum_{i=1}^{s} n_i$, $m = \sum_{i=1}^{s} m_i$ and $u = \sum_{i=1}^{s} u_i$. The completely pooled Chapman estimator of the population size is:

$$\tilde{N}_{pool} = \frac{(n+1)(m+u+1)}{m+1} - 1 \tag{2.4}$$

and an estimate of the variance of this estimator is:

$$\widehat{V(\tilde{N}_{pool})} = \frac{(n+1)(m+u+1)(n-m)u}{(m+1)^2(m+2)}. \tag{2.5}$$

Key to the validity of this estimator is the variation in the capture probabilities over the t days. If the capture probabilities are constant, then $\tilde{N}_{pool}$ is unbiased for N and $\widehat{V(\tilde{N}_{pool})}$ will provide an accurate estimate of the uncertainty. However, if the capture probabilities vary over time the estimator may be biased and, often more importantly, $\widehat{V(\tilde{N}_{pool})}$ may severely underestimate the true variability of the estimator. Simulations I have performed indicate that when $E(p_i)$ is small the bias in $\widehat{V(\tilde{N}_{pool})}$ may be as much 40%. This is also observed in the simulations presented in section 2.2.4 below.

Bayesian hierarchical modelling presents an alternative to pooling that draws on the similarities between strata without assuming exact equality of the capture probabilities. Mantyniemi and Romakkaniemi (2002) developed a hierarchical model for non-diagonal data which I have simplified here for diagonal experiments. In the simplest, non-hierarchcial Bayesian model of the SP experiment, each p_i and U_i would be assigned independent prior distributions, and the posterior distribution for each depend would only on the data collected on day i. In the hierarchical model, the parameters from different days are linked by

assuming that $\boldsymbol{p}$ and $\boldsymbol{U}$ form random draws from some hyperpriors. Mantyniemi and Romakkaniemi (2002) suggest a normal model for the logit transformed capture probabilities and a multinomial prior for the daily run sizes where the parameters of these distributions, including the total population size, are assigned hyperpriors with fixed parameter values.

The advantage of the hierarchical model is that inference for the parameters in one stratum will depend on the data from all strata. In this way, data is shared among the strata, but the parameters are still allowed to vary day-to-day. The disadvantage of the hierarchical approach is that it makes no adjustment for the ordering of the data – the same amount of information is shared between days 1 and 2 as days 1 and 10 or days 1 and 100.

2.2.3 The Bayesian P-Spline Model

While the hierarchical model allows for data to be shared among strata, it does not consider the natural ordering of the data when strata are based on time. By assuming that each p_i and U_i form independent draws from their respective hyper-priors, the amount of information shared between (p_i, U_i) and (p_j, U_j) is the same regardless of whether $i-j$ is 1, 5 or $t-1$. In the case of temporal stratification, and particularly the application to salmon migrations, it seems intuitive that the number of fish passing the second trap will be more similar for days that are close together and less similar for days that are further apart. This can be achieved through several different extensions of the hierarchical Bayesian model, for example, by assuming that the correlation between U_i and U_j follows some decreasing function of $|i-j|$ or by defining an autoregressive model as in time series analysis. I have chosen to model U_j explicitly as a smooth curve using the Bayesian penalized spline (P-spline) method of Lang and Brezger (2004).

As discussed in the introduction, two factors control the smoothness of a spline: the number and locations of the knots and the variation in the coefficients of the basis function expansion. The classical P-spline method of Eilers and Marx (1996) approaches this dichotomy by fixing a large number of knot points and then penalizing the first or second order differences of the coefficients. In the original implementation, the spline curve is fit by minimizing a target function which adds the sum-of-squared residuals and a penalty term formed as the product of a smoothing parameter and the sum of the differences of the spline coefficients. Increasing the smoothing parameter places more weight on the penalty term and results in a smoother curve. Decreasing the smoothing parameter places more weight on the sum-of-squared residuals and produces a fit that comes closer to interpolating the

data.

Lang and Brezger (2004) develop a Bayesian formulation of the P-spline method which replaces the penalty term by a particular prior distribution for the coefficients. Let $B(x)_1, \ldots, B(x)_{K+q}$ denote the B-spline basis functions for a spline of order q with K knots and $b_1, \ldots, b_{K+q}$ the corresponding coefficients. To ensure that the population size in each strata remains positive, I model the log transformed daily counts and the fitted spline has the form:

$$\log(U_j) = \sum_{k=1}^{K+q} b_k B_k(j). \tag{2.6}$$

The prior distribution for the parameters $b_1, \ldots, b_{K+q}$ suggested by Lang and Brezger (2004) models the differences in these coefficients as either a first or second order random walk. Here I apply the second order prior such that:

$$b_{k+1} - b_k = (b_k - b_{k-1}) + \delta_k$$

where the δ_k are independent, normally distributed random variables with mean 0 and variance τ_U^2. The initial coefficients, b_1 and b_2, are assigned improper flat priors.

In the Bayesian P-spline approach, the variance parameter τ_U plays the same role as (though opposite in direction to) the smoothing parameter in the classical method. If τ_U is small, then the b_k will all be very similar and the spline will be smooth. If τ_U is large then the b_k may vary more widely and the spline will be more flexible. Rather than fixing the value of τ_U, this parameter is assigned a prior distribution which incorporates uncertainty in the amount of smoothing required. Following Lang and Brezger (2004), I define an inverse gamma prior, $1/\tau_U^2 \sim \Gamma(\alpha, \beta)$, with the parameters α and β chosen so that $E(\tau_U)$ is small, which *a priori* favours a smooth fit, but $V(\tau_U)$ is large, which allows a less smooth curves if the data requires. Lang and Brezger (2004) suggest $\alpha = 1$ with $\beta = .005$, $\beta = .0005$, or $\beta = .00005$. By default, I have chosen $\alpha = 1$ and $\beta = .0005$.

Although the spline model may reflect the trend in the daily population size, similar to a running mean, it is unlikely that U_j will exactly follow a smooth curve. If the deviations from a smooth curve are small then it seems reasonable that forcing U_j to be smooth will not have a large impact on the estimation of N. However, it was not clear how the estimator of the total population size would be affected if U_j deviated significantly from a smooth curve.

To assess this, I also consider an extension of the spline model with an added error term:

$$\log(U_j) = \sum_{k=1}^{K+q} b_k B_k(j) + \epsilon_j \tag{2.7}$$

where the ϵ_j are assumed independent and normally distributed with mean 0 and variance τ_e. The variance parameter τ_e was assigned the inverse gamma distribution $1/\tau_\epsilon^2 \sim \Gamma(1, .05)$.

Samples from the posterior distributions of these models were generated through MCMC sampling in the OpenBUGS software package (Thomas et al., 2006). BUGS code for the Bayesian P-spline models with and without the error term are included in Appendix A.1 and Appendix A.2. One complication in sampling from the posterior distribution of the Bayesian P-splines model was that the chains had very high autocorrelation and mixed very slowly. This was due to the strong associations between some parameters that prevented the chains from making large jumps in the parameter space. To ensure that the samples adequately covered the parameter space, several chains were run in parallel starting from dispersed initial values. Each chain was also thinned considerably to reduce the auto-correlation and avoid storing the output from many highly correlated realizations that provided little extra information about the shape of the posterior. Convergence was assessed by comparing trace plots for select parameters and by computing the BRG diagnostics.

Fit of the selected model was assessed by computing Bayesian p-values as described in section 1.3.2. P-values were generated for three different discrepancy measures assessing different components of the model:

$$\begin{aligned} D_1(\boldsymbol{X}, \boldsymbol{\theta}) &= -2 \cdot L(\boldsymbol{p}, \boldsymbol{U} | \boldsymbol{X}) \\ D_2(\boldsymbol{X}, \boldsymbol{\theta}) &= \textstyle\sum_{i=1}^{s} \left(\sqrt{m_i} - \sqrt{n_i p_i}\right)^2 \\ D_3(\boldsymbol{X}, \boldsymbol{\theta}) &= \textstyle\sum_{i=1}^{s} \left(\sqrt{u_i} - \sqrt{U_i p_i}\right)^2 . \end{aligned}$$

The first measure is simply the deviance and is intended to provide an overall assessment of fit, the second and third are the Freeman-Tukey statistics comparing the observed and expected number of marked individuals recaptured and comparing the observed and expected number of unmarked individuals captured at the second site.

2.2.4 Simulation Studies

I conducted two simulation studies to assess the performance of the Bayesian P-spline model and compare it with other methods of estimating the population size. In the first study, the daily run size exactly followed a smooth curve. Accuracy and precision of the estimators of total population size from the P-spline models were compared with 4 other methods of estimating the population size: summing daily Chapman estimates, complete pooling, fitting a simple Bayesian model with complete independence between days and fitting a hierarchical model of the capture probabilities and daily run size. The second simulation compared only the P-spline model with and without the extra error term when the true model deviated from a smooth curve. Parameters values chosen for the simulations mimicked the daily run size and capture probabilities estimated from fitting the P-spline model to the Conne River data (see section 2.4.1 for details).

My expectation for the first simulation was that the P-spline model would outperform the other methods for estimating the population size. I anticipated that all methods would produce unbiased estimates. However, the precision of the estimator from the P-spline model should have been improved by taking advantage of the knowledge that the run size was exactly smooth. I hoped that the P-spline model without the error term would also perform as well or better than the model with the extra error term in the second simulation even though the daily run sizes did not follow a smooth curve. My reasoning was that the errors from the strict P-spline model might cancel each other out so that even though the daily run size would be overestimated on some days and underestimated on others, the estimator of total population size would be unbiased. If this were true, it would not be necessary to fit the more complex model with the extra error term.

Simulation 1a: Run size follows a smooth curve

The first simulation compared the performance of the P-spline model and standard models when the true numbers of fish passing the second site each day, $N_1, \ldots, N_s$, exactly followed a smooth curve. The shape of the run for the Conne River data Atlantic salmon smolts was approximated by a curve formed from two quadratic segments on the log-scale with a single

maximum at 13.5 days. The exact formula was:

$$N_i = \begin{cases} 7100^{(1-(i-13.5)^2/13.5^2)} + 125 & i = 1, \ldots, 13 \\ 7100^{(1-(i-13.5)^2/20^2)} + 125 & i = 14, \ldots, 40 \end{cases} \tag{2.8}$$

with N_i rounded to the nearest whole number. The number of fish marked per day was chosen to be a fixed 100 ($n_i = 100$, $i = 1, \ldots, 40$), so that $U_i = N_i - 100$. The resulting number of unmarked fish passing the second site per day starts and ends at 25, and reaches a maximum of 7103 on day 14 (see Figure 2.3). The true total population size resulting for this model was 87 thousand. Capture probabilities were simulated independently for each day from a beta distribution with parameters $\alpha = 2.5$ and $\beta = 50$. The resulting mean and standard deviation were .05 and .03 with 5^{th} and 95^{th} percentiles of .01 and .10.

From these models of the run size and the capture probabilities, I generated 100 independent data sets in two steps. First, daily capture probabilities, $p_1, \ldots, p_{40}$ were independently sampled from the specified beta distribution. Counts of the marked and unmarked fish observed on each day at the second site were then simulated as independent binomial random variables $m_i \sim \text{Binomial}(100, p_i)$ and $u_i \sim \text{Binomial}(U_i, p_i)$.

For each data set, estimates of the total populations size were computed from 6 methods as follows:

1) Daily Chapman estimates

   - Daily estimates computed via equation (2.2) and then summed to produce an estimate of total population size.

2) Pooled estimator

   - Data pooled completely by summing the number of marks, recaptures, and unmarked individuals over all days of the experiment. An estimate of population size was then computed via equation (2.4).

3) Simple Bayesian model

   - Bayesian model fit with independent priors on each p_i and U_i:

$$\begin{aligned} \text{logit}(p_i) &\sim N(-2, 1.22) \\ \log(U_i) &\sim N(7.5, 4). \end{aligned}$$

4) Hierarchical Bayesian model:

- Bayesian model fit with hierarchical prior distributions for both p_i and U_i:

$$\begin{aligned}\text{logit}(p_i) &\sim N(\xi_p, \sigma_p^2), \text{ with } \xi_p \sim N(-2, 1.22) \text{ and } 1/\sigma_p^2 \sim \Gamma(.001, .001)\\ \log(U_i) &\sim N(\xi_U, \sigma_U^2), \text{ with } \xi_U \sim N(7.5, 4) \text{ and } 1/\sigma_U^2 \sim \Gamma(.001, .001).\end{aligned}$$

5) Bayesian P-spline model:

- Bayesian model fit with hierarchical prior for p_i, as in previous model, and Bayesian P-spline model of equation (2.6) for $\log(U_i)$. The second difference prior was selected for the coefficients of the spline:

$$b_1 \propto 1, \quad b_2 \propto 1, \quad b_k \sim N(2b_{k-1} - b_{k-2}, \sigma_U^2) \text{ for } k = 3, \ldots, K+4$$

and $1/\tau_U^2 \sim \Gamma(1, .0005)$.

6) Bayesian P-spline model with Error:

- Bayesian model fit with hierarchical prior for p_i, as in model 4, and Bayesian P-spline model with error of equation (2.9) for $\log(U_i)$. The prior for the spline coefficients, $b_1, \ldots, b_{K+q}$ was the same as in model 4. The variance of the error terms was assigned the inverse gamma prior $1/\tau_\epsilon^2 \sim \Gamma(1, .05)$.

The prior distributions for the simple and hierarchical Bayesian models were chosen to be somewhat informative about the model parameters. The prior median of U_i for the simple Bayesian model was 1808 with 5^{th} and 95^{th} percentiles 3 and 1.30×10^6. The prior median of p_i was .12 with 5^{th} and 95^{th} percentiles .02 and .50. Summaries of the posterior distribution of total abundance, $N = \sum_{i=1}^{40}(U_i + n_i)$, for the Bayesian models were computed from samples of U_i generated via MCMC in OpenBUGS (Thomas et al. (2006)). Each Markov chain was run for 100,000 iterations, discarding the first 20,000 iterations as burnin and retaining every 5^{th} of the remaining iterations for computing summary statistics. Posterior means were chosen as point estimates of the total population size and posterior standard deviations as measures of uncertainty.

Note that the data generating model does not exactly match the models for analysis. In particular, the capture probabilities were generated from a beta distribution but the 4 fitted Bayesian models assume that the capture probabilities follow a normal distribution

on the logit scale. However, the normal distribution on the logit scale can be made to fit very closely to the chosen beta distribution. If $X \sim N(-3.15, .65)$ then the distribution function of $e^X/(1+e^x)$ differs from the distribution function of the beta distribution with parameters $\alpha = 2.5$ and $\beta = 50$ by at most .03. Because of this, the difference between the genrating and fitted models will not affect the models' performance.

Performance of the 6 methods was assessed in terms of the bias and mean squared error (MSE) of the estimator of the total population size, along with the accuracy of the associated estimator of uncertainty. For any one of the above methods, let $\hat{N}_k$ denote the point estimate of N computed from the k^{th} simulated data set and $\hat{S}_k$ the associated estimate of uncertainty. Let N^* and S^* denote the observed mean and standard deviation of $\hat{N}_k$ across the 100 simulated data sets: $N^* = \sum_{k=1}^{100} \hat{N}_k/100$ and $S^* = \sum_{k=1}^{100}(\hat{N}_k - N^*)^2/99$. Accuracy of the estimate of total population size was assessed by the percent bias:

$$\%\text{Bias}(\hat{N}) = 100\frac{(N^* - N)}{N}.$$

MSE of the estimator was computed as usual, but converted to a percentage of the square of the true population size:

$$\%\text{MSE}(\hat{N}) = 100\frac{\sum_{k=1}^{100}(\hat{N}_k - N)^2}{N^2}$$

This produced a unitless quantity that could be averaged across data sets even if the population size was not constant, as in the second simulation. Accuracy of the estimate of error was assessed through the percent bias:

$$\%\text{Bias}(\hat{S}) = 100\frac{(\sum_{k=1}^{100}\hat{S}_k/100 - S^*)}{S^*}.$$

For the daily Chapman estimators, $\hat{S}_k$ was computed as square root of the sum of the daily variance provided in equation (2.3); for the pooled estimator, $\hat{S}_k$ was computed from the square root of the variance in equation (2.5); and for the Bayesian models, $\hat{S}_k$ was taken to be the posterior standard deviation of the total population size. The posterior predictive p-values and DIC were also computed for the 3 Bayesian models.

Table 2.1 presents the results for all 6 methods. The pooled Chapman estimator, P-spline model, and P-spline with error all provided nearly unbiased estimates of N – the %Bias for all

3 estimators being within 1 standard error of 0. The summed daily Chapman estimates had a slight negative bias, while the posterior mean from the simple and hierarchical Bayesian model had significant positive biases. The simple and hierarchical Bayesian models also produced the highest %MSE. The estimated %MSE for the posterior mean of the simple Bayesian model was approximately 1.5 times that of the %MSE of both the Daily and Pooled Chapman methods, which, in turn, were almost twice the %MSE of both the P-spline model with and without error.

As expected, the standard error of the pooled Chapman estimator significantly underestimated its true variability. In contrast, the posterior standard deviation from the simple Bayesian model severely overestimated the variability of this estimator. Posterior standard deviations for both the P-spline and P-spline without and with error provided very accurate measures of the variability of the estimator of total population size.

Table 2.2 presents summary statistics for the DIC and the Bayesian p-values for the 4 Bayesian models. Not surprisingly, the P-spline model has the lowest measure of complexity. By taking advantage of the temporal structure, the effective number of parameters is reduced by 1/3 in comparison with the hierarchical model. Introducing the extra error term into the P-spline model did increase p_D slightly, but the value was still much lower than for the hierarchical model. The P-spline model without error also produced the lowest DIC when averaged over all 100 datasets and was selected (i.e. had the lowest DIC) for 79% of the simulated datasets. The P-spline model with error had the lowest DIC value for the remaining 21%. On average, the mean difference in DIC between these two models was less than 1.1, suggesting very little difference in the models' fit. Indeed, the absolute difference in the estimate of population size for the two P-spline models was less than 4300 (5% of the total population size) for all 100 data sets.

One surprising result was that the Bayesian p-values were not uniformly distributed on $(0, 1)$ for any model or discrepancy measure. Given that the true daily run size followed a smooth curve, I expected the p-values be uniformly distributed for all models. Instead, the means of the p-value were always close to .5 but the 5^{th} and 95^{th} percentiles were all near .2 and .8, not .05 and .95 as I had expected.

Simulation 1b: U_j follows a smooth curve plus noise

In reality, the true daily population sizes will never conform exactly to a smooth curve; fluctuations will always occur as the result of changes in the environment and the natural

behaviour of the migrating fish. To examine the effects of deviations from the smooth curve on the performance of the Bayesian P-spline method, I conducted a second simulation study in which random noise was introduced into the model of the run size. Daily numbers of fish passing the second trap were generated from the same model as in the first simulation study, except that random Gaussian noise with standard deviation of .5 was added to the logarithm of the daily run sizes. That is:

$$N_i = \begin{cases} 7100^{(1-(i-13.5)^2/13.5^2+\epsilon_i)} + 125 & i = 1, \ldots, 13 \\ 7100^{(1-(i-13.5)^2/20^2+\epsilon_i)} + 125 & i = 14, \ldots, 40 \end{cases} \tag{2.9}$$

where each $\epsilon_i \sim N(0, .5^2)$ independently, and N_i is rounded to the nearest whole number.

In this study, 30 data sets were generated from each of 30 different realizations of $N_1, \ldots, N_{40}$. The total population size for the 30 realizations ranged between 73 and 130 thousand, with mean 99 thousand. Capture probabilities and the numbers of marked and unmarked fish captured at the second trap for each data set given the true counts were simulated exactly as in the first study. I fit both the Bayesian P-spline models without and with the error term to each data set and compared the results with the same statistics used in the previous section.

Table 2.3 presents the summary statistics comparing the estimators of population size from the P-spline models without and with the error term. As in Simulation 1a, both models produced nearly unbiased estimates of the population size. However, the posterior standard deviation from the P-spline model without error underestimated the variability in the population size by just over 10% on average, while the model with the error extra error still produced a very accurate estimate of the uncertainty. The actual variability in the estimates of the daily population size, averaged over the 30 realizations, was smallest for the Bayesian P-spline model with error.

Table 2.4 presents the DIC and Bayesian p-values for the two models. In contrast to the previous study, the P-spline model with the extra error term produced DIC values that were, on average, 10 units lower than the DIC values for the model without error and lowest for more than 90% of the simulated data sets. The Bayesian p-values provide further evidence that the strict P-spline model could not fit the simulated data adequately. The means of all three p-values were below .5, and the 5^{th} and 95^{th} percentiles were not centred on the unit interval.

Summary

In the first simulation study, summing the daily Chapman estimates underestimated the true population size by slightly more than 5%. The reason for this is that the daily estimator is not unbiased when the number of individuals marked is very small. Pooling the data over all days avoids the small counts and produces an estimator that is almost unbiased. However, this estimator completely ignores the variation in the capture probabilities and the resulting estimate of uncertainty was severely, negatively biased. This is very important for wildlife managers who may be led to make strong decisions on what action to take under the false belief that they have a very precise estimate of the population size.

The poor performance of the simple Bayesian model had two root causes. Because the model considered data from each day independently the estimate of population size was highly influenced by outliers in the data. In particular, the model tended to severely overestimate the population size on days when the number of recaptured individuals was low, which lead to the overall positive bias. When data is sparse the results are also highly dependent on the choice of prior distributions. This was in fact the reason that the third Bayesian p-value was consistently above .5 for this model. The prior distribution for the capture probabilities placed significant mass on values above .5, far higher than the true values. When data was simulated from the posterior-predictive distribution the numbers of unmarked individuals captured at the second site almost always was much higher than the observed number which has specific influence on the 3^{rd} discrepancy measure. Further to this, the choice of priors with large variances inflated the posterior standard deviation of the total population size which biased the estimator of uncertainty. Selecting more informative prior distributions did decrease the bias and also corrected the Bayesian p-value (results not shown), but it is not clear that strongly informative prior distributions can be provided for real experiments.

The hierarchical model remedies these problems somewhat by sharing information among the days. This reduced the effect of outliers, decreased the bias of the estimator of population size and also lessened the sensitivity to the prior distribution. However, the hierarchical model treated all days equally without regard for the order of the data. In these simulations, and in most real experiments, the data was most sparse at the start and end of the migration period. The tendency of the hierarchical model was to pull the posterior distribution for the population size on these days up toward the posterior distributions in the middle of the

run, when more individuals were captured and data was much richer. The result was that the daily population size was often overestimated at the start and end of the run, creating a positive bias overall.

Estimators of the total population size from both the P-spline models without and with the error term out-performed the other methods when the daily run size followed a smooth curve. However, the model without the extra error term could not adapt to the noise in the second simulation. The model over-smoothed the data and the DIC and Bayesian p-values both indicated that the fit was not adequate. My conclusion is that the P-spline model with error provides a compromise between complete exchangeability and exact smoothness that performs almost as well as the strict P-spline model when the true run size is smooth but is likely to provide better fit to real data.

Method	N^*	%Bias(SE)	% MSE	Mean SE	Obs. SD	% Bias(SE)
Daily Chapman	81.2	−7.1(1.2)	1.9	10.5	10.3	2.0(3.2)
Pooled Chapman	85.7	−1.9(1.4)	1.8	5.9	11.8	−49.6(0.9)
Simple Bayes.	97.1	11.0(1.3)	2.9	15.5	11.4	35.4(4.6)
Hierarchical Bayes.	93.8	7.3(1.3)	2.3	12.3	11.6	5.6(3.8)
P-Spline	88.0	0.7(1.0)	1.0	8.8	8.6	1.6(1.7)
P-Spline w Error	88.9	1.7(1.0)	1.1	9.1	8.9	2.2(1.7)

Table 2.1: Comparison of the estimators of abundance and their uncertainty for Simulation 1a. Estimation of abundance is assessed according to the %Bias and %MSE of the point estimators. Values provided are the mean estimate of N (scaled by 10^3), the %Bias (with standard error in parentheses), and the %MSE. The true population size in the simulations was 87 thousand. Estimation of the uncertainty is assessed according to the %Bias of the point estimator of the standard error. Values provided are the mean of the uncertainty estimates from the 100 simulations and the observed standard deviation of the estimator of N (both scaled by 10^3), and the %Bias of the uncertainty estimate (with its standard error in parentheses).

	DIC			Bayesian P-values		
	Mean DIC	Mean pD	% Lowest	1	2	3
Simple Bayes.	459.6	69.9	0	.51(.46,.57)	.45(.19,.73)	.61(.53,.70)
Hierarchical Bayes.	455.4	61.5	0	.46(.37,.56)	.41(.17,.76)	.53(.42,.65)
P-Spline	435.1	41.1	79	.44(.17,.68)	.48(.08,.86)	.46(.20,.71)
P-Spline w Error	436.2	44.7	21	.50(.29,.68)	.50(.14,.84)	.49(.23,.71)

Table 2.2: DIC and Bayesian p-values for the Bayesian models applied in Simulation 1a. The left half of the table presents the mean values of DIC and p_D over the 100 simulated data sets. Also provided is the % of the data sets for which each method produced the lowest DIC value. The right half of the table presents the mean Bayesian p-value for each of the three discrepancy measures. The 5^{th} and 95^{th} percentiles of the Bayesian p-values are provided in parentheses.

Method	%Bias(SE)	% MSE	Mean SE	Obs. SD	% Bias(SE)
P-Spline	1.4(0.4)	1.5	10.2	11.8	−11.7(0.6)
P-Spline w error	0.8(0.4)	1.3	10.8	11.1	−0.7(0.6)

Table 2.3: Comparison of the estimators of abundance and their uncertainty for Simulation 1b. Estimation of abundance is assessed according to the %Bias and %MSE of the point estimators. Values provided are the mean estimate of N (scaled by 10^3), the %Bias (with standard error in parentheses), and the %MSE. Estimation of the uncertainty is assessed according to the %Bias of the point estimator of the standard error. Values provided are the mean of the uncertainty estimates from the 100 simulations and the observed standard deviation of the estimator of N (both scaled by 10^3), and the %Bias of the uncertainty estimate (with its standard error in parentheses).

	DIC			Bayesian P-values		
	Mean DIC	Mean pD	% Lowest	1	2	3
P-Spline	457.5	42.4	6.8	.11(.00,.36)	.25(.01,.64)	.41(.14,.66)
P-Spline w error	445.6	50.0	91.4	.42(.27,.56)	.45(.12,.79)	.51(.27,.72)

Table 2.4: DIC and Bayesian p-values for the Bayesian models applied in Simulation 1b. The left half of the table presents the mean values of DIC and p_D over the 300 simulated data sets. Also provided is the % of the data sets for which each method produced the lowest DIC value. The right half of the table presents the mean Bayesian p-value for each of the four discrepancy measures. The 5^{th} and 95^{th} percentiles of the Bayesian p-values are provided in parentheses.

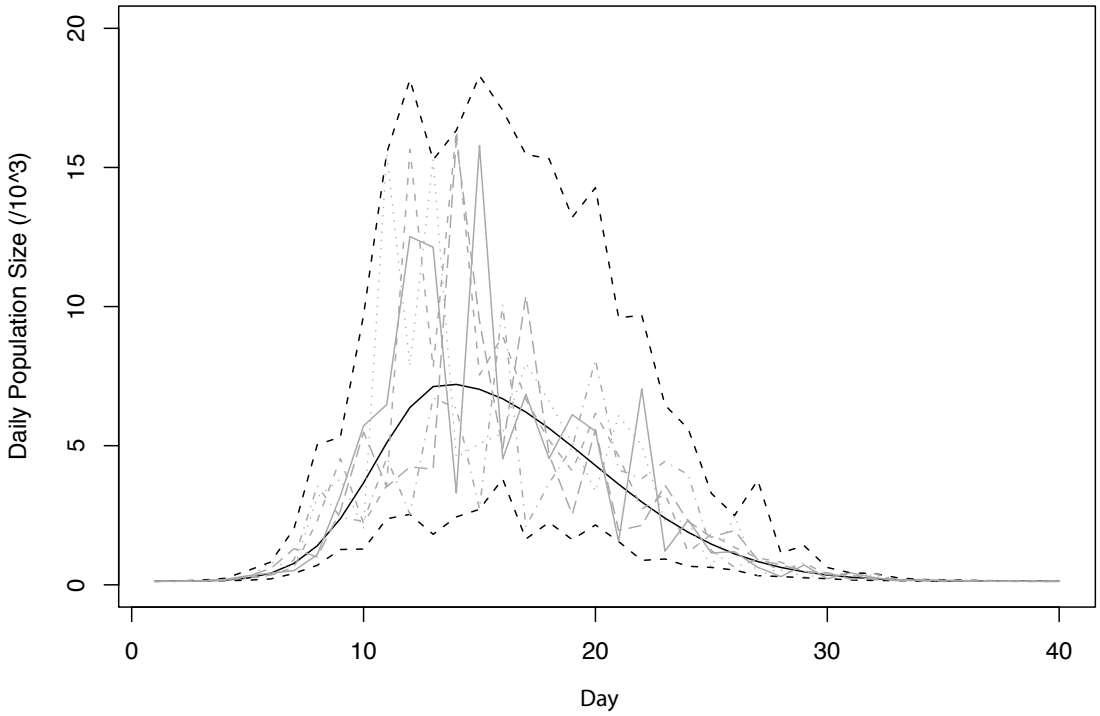

Figure 2.3: Number of fish passing the second trap on each day in Simulations 1a and 1b. The solid black line indicates the number of fish passing on each day in Simulation 1a. The grey lines represent separate realizations from Simulation 1b, obtained by adding Gaussian error to the values of Simulation 1a (on the logit scale). The dashed black lines indicate the daily maximum and minimum numbers of fish passing on each day over the 300 data sets in simulation 1b.

2.3 Non-Diagonal Data from Two-Sample Studies

2.3.1 Introduction

In many fisheries studies, it is not reasonable to assume that all of the fish marked at the first site on one day will pass the second site on the same day. Some fish may move between the sites more slowly than others so that fish marked on one day are recovered for several days afterward. The result is that the numbers of marked fish passing the second site on each day cannot be known exactly and capture probabilities cannot be estimated simply by comparing the daily counts of marked and recovered fish. Instead, more complex models are needed to produce estimates of the number of marked fish in each stratum at the second site, from which the capture probabilities can be estimated.

The general design for a stratified, two-sample experiment does not require that the number of strata be the same in both samples. In fisheries monitoring programs, trapping is often conducted for longer at the second site to allow the fish marked during the final days of capture at the first site time to move to the second site. Let s denote the number of strata in the first sample, indexed by i, and t the number strata in the second sample, indexed by j. Let n_i denote the number of individuals marked from stratum i, m_{ij} denote the number of individuals marked in stratum i which are recaptured in stratum j and u_j the number of unmarked individuals captured in stratum j of the second sample. In fisheries work, it is usually reasonable to assume that fish do not move backward from site 2 to site 1, so that individuals marked on day i may only be recovered at site 2 on days $i, i+1, i+2, \ldots, t$. The resulting matrix of recoveries, the $s \times t$ matrix with entries m_{ij} denoted by $\boldsymbol{M}$, will be upper triangular.

The standard model for non-diagonal data assumes that the probability of belonging to stratum j at the second site is the same for all individuals marked in stratum i at the first site. In the fisheries context, this means that all fish released at the first site on day i have the same probability of passing the second site on day j, for each $j = i, i+1, \ldots, t$. Let θ_{ij} denote the probability that an individual marked and released in stratum i at the first site moves to stratum j at the second site. Assuming independence between movement and capture at the second site, the probability that an individual is marked in stratum i and recaptured in stratum j is $\theta_{ij}p_j$. The probability that an individual marked in stratum i is never recovered is $1 - \sum_{j=i}^{t} \theta_{ij}p_j$. The recoveries of the individuals marked in stratum i are

then modelled according to the multinomial distribution:

$$(m_{i1}, \ldots, m_{it}, n_i - m_{i\cdot}) \sim \text{Multinomial}\left(n_i, \left(\theta_{i1}p_1, \ldots, \theta_{it}p_t, 1 - \sum_{j=i}^{t} \theta_{i1}p_j\right)\right) \tag{2.10}$$

where $m_{i\cdot} = \sum_{j=i}^{t} m_{ij}$ denotes the total number of individuals marked in stratum i and ever recovered. As in the model for the diagonal experiment, the model conditions on the number of unmarked individuals alive in each stratum at the second site, U_j, so that:

$$u_j \sim \text{Binomial}(U_j, p_j). \tag{2.11}$$

No assumptions are made concerning the movement of unmarked individuals among the strata of the two samples. The full likelihood is again formed by multiplying the contributions for the marked and unmarked individuals over all days of the experiment yielding:

$$\begin{aligned} L(\boldsymbol{p}, \boldsymbol{\Theta}, \boldsymbol{U} | \boldsymbol{n}, \boldsymbol{M}, \boldsymbol{u}) = \; & \prod_{i=1}^{s} \left[\frac{n_i!}{m_{i1}! \cdots m_{it}! (n_i - m_{i\cdot})!} \left(\prod_{j=i}^{t} (\theta_{ij} p_j)_{ij}^{m} \right) \left(1 - \sum_{j=1}^{t} \theta_{ij} p_j \right)^{(n_i - m_{i\cdot})} \right] \\ & \cdot \prod_{j=1}^{t} p_j^{u_j} (1 - p_j)^{U_j - u_j} \end{aligned} \tag{2.12}$$

with $\boldsymbol{\Theta}$ denoting the $s \times t$ matrix of transition probabilities.

2.3.2 Previous Estimators of Abundance

Several methods have been developed to compute estimates of population size based on this model of the two-sample study. Schaeffer (1951) developed an ad hoc approach to estimation that is credited as the first treatment of the model, though (Schwarz et al., 2008) recently showed that the resulting estimators perform no better than pooling the data and ignoring the stratification completely. Darroch (1961) applied the method of maximum likelihood deriving explicit MLEs for the special case with $s = t$, and showing that explicit formula did not exist otherwise. For cases with $s < t$ or $s > t$ Darroch recommended pooling strata in order to reproduce the case $s = t$, and then applying the explicit formula. Methods for computing MLEs and their standard errors in all cases based on numerical optimization of the likelihood were later developed by Plante et al. (1998). Other estimators of the total population size have been produced by the method of moments (Chapman and Junge, 1956) and by least squares estimation (Banneheka et al., 1997). The advantage of these

approaches over maximum likelihood methods is that estimates can, in theory, be computed analytically regardless of the relationship between s and t without the need for lengthy numerical optimization. However, the resulting estimators do not satisfy the asymptotic optimality properties of MLEs.

A problem that faces all of these methods is that the estimate of abundance can only be computed if the matrix of recoveries, $\boldsymbol{M}$, has full rank. If this is not the case, then unique solutions to the estimating equations will not exist, and the different solutions may produce very different estimates of abundance. For example, imagine that a simple study was conducted with two strata in each sample and the following data was observed:

$n_1 = 10$	$m_{11} = 1$	$m_{11} = 1$
$n_2 = 10$	$m_{21} = 1$	$m_{22} = 1$
	$u_1 = 10$	$u_2 = 100.$

Applying the method of moments, equating the expected values of m_{ij} and U_j with their observed values, produces the set of six estimating equations:

$$\begin{aligned} &10\theta_{ij}p_j = 1, \quad i, j \in \{1, 2\} \\ &U_1 p_1 = 10 \\ &U_2 p_2 = 100 \end{aligned}$$

One solution to these equations is given by: $\theta_{11} = \theta_{21} = 1/10$, $p_1 = 1$, $\theta_{12} = \theta_{22} = 9/10$, $p_2 = 1/9$, $U_1 = 10$ and $U_2 = 900$. The resulting estimate of total abundance is 930. A second solution is given by: $\theta_{11} = \theta_{21} = 1/10$, $p_1 = 1/9$, $\theta_{12} = \theta_{22} = 1/10$, $p_2 = 1$, $U_1 = 90$ and $U_2 = 100$. This solution produces an abundance estimate of 210.

While it is unlikely that the $\boldsymbol{M}$ will be rank deficient when many fish are marked and captured in each stratum, this happens frequently with sparse data sets either because no fish are recaptured for one stratum (i.e., a row of 0s) or because the same pattern of recoveries occurs in two different strata (i.e., linear dependence). The essence of the problem is that the movement probabilities cannot all be distinguished when data is sparse or the matrix of recoveries does not have full rank. In the example presented, it is impossible to tell if most of the marked individuals moved into the first or second stratum of the second sample. As a result, it is not possible to estimate the total number of marked individuals in each stratum, capture probabilities, or the numbers of unmarked individuals in each stratum, and

no sensible estimate of the abundance can be produced. Even if the matrix of recoveries does have full rank, it is often the case that some linear combination of rows or columns is close to **0** so that the resulting estimate of abundance will have very low precision.

As in the case of sparse diagonal data sets, the most common solution in problems with sparse data is to pool adjacent strata to reduce the number of parameters needed in the model and increase the counts in each cell of the matrix of recoveries. Another solution that has been proposed is to define a parametric model for the movement probabilities which depends on a smaller number of parameters. Schwarz and Dempson (1994) considered the specific application to studies of salmon migrations and modelled the time each marked fish takes to move between the two sites as a log-normal random variable with separate mean and variance parameters, μ_i and σ_i^2, for each day of marking. The probability that a fish marked on day i passes the second site on day $i+k$ is then given by:

$$\theta_{ij} = \begin{cases} \Phi\left(\frac{\log(d)-\mu_i}{\sigma_i}\right) & k = 0 \\ \Phi\left(\frac{\log(k+1+d)-\mu_i}{\sigma_i}\right) - \Phi\left(\frac{\log(k+d)-\mu_i}{\sigma_i}\right) & k = 1, \ldots, t-i \end{cases} \tag{2.13}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. The value d is an adjustment which Schwarz and Dempson (1994) set to .5 to account for marked fish being released throughout each day rather than at a single time point. This value should be set to 0 if all fish marked in one day were released at the same time. Schwarz and Dempson (1994) numerically optimized the likelihood in equation (2.12) with each θ_{ij} replaced by the expression in equation (2.13) to compute MLEs of μ_i, σ_i, p_j and U_j and of the total population size. When $t > 2$, this model reduces the number of parameters required to model the movements of marked fish from $s(2t-s+1)/2$ (assuming forward movement only) to $2s$.

Mantyniemi and Romakkaniemi (2002) developed a Bayesian implementation which extended this model to the hierarchical framework. Instead of defining independent prior distributions for each parameter, the elements of the vectors $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_s)$, $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_s)$ and $\boldsymbol{p} = (p_1, \ldots, p_t)$ were modelled as random draws from hyper-prior distributions with unknown hyper-parameters. Mantyniemi and Romakkaniemi (2002) specifically selected normal hyper-priors for each of μ_i, $\log(\sigma_i)$ and $\text{logit}(p_j)$, these transformations being standard to ensure that $\sigma_i > 0$ and $p_j \in (0,1)$. The prior distribution for the vector of daily run sizes, $\boldsymbol{U} = (U_1, \ldots, U_t)$, was defined through a two-step process. First, the marginal

distribution of the total number of unmarked fish was assigned the improper Jeffrey's prior $\pi(U) \propto 1/U$. The daily numbers of unmarked individuals were then modelled conditionally as:

$$\boldsymbol{U} \sim \text{Multinomial}(U, (\rho_1, \ldots, \rho_t))$$

with the hyper-parameter $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_t)$ assigned a Dirichlet prior.

Both the models of Schwarz and Dempson (1994) and Mantyniemi and Romakkaniemi (2002) provide solutions that compromise between the fully stratified and completely pooled estimators. Information is shared among strata, so that the data is not viewed as completely separate, but the estimated capture probabilities are still allowed to vary. In the model of Schwarz and Dempson (1994) the sharing of information was produced through the explicit parametric assumptions regarding the travel times, which introduced dependence among the number of fish arriving in neighbouring strata at the second trapping location. The model of Mantyniemi and Romakkaniemi (2002) introduced further dependence by assuming that the parameters from separate strata are similar, but not equal, using the data to determine the degree of similarity (quantified by the variance of the hyper-priors).

2.3.3 The Bayesian P-spline Model

The model I develop for non-diagonal data from temporally stratified two-sample studies incorporates the Bayesian P-spline method into the models of Schwarz and Dempson (1994) and Mantyniemi and Romakkaniemi (2002). I begin by constructing a simple Bayesian implementation of the model with independent prior distributions for the parameters of the travel time distribution, μ_i and σ_i in equation (2.13), and for both p_j and U_j, $i = 1, \ldots, s$, $j = 1, \ldots, t$. Following the work of Mantyniemi and Romakkaniemi (2002), I propose the prior distributions:

$$\mu_i \sim N(\xi_\mu, \tau_\mu^2) \text{ and } \log(\sigma_i) \sim N(\xi_\sigma, \tau_\sigma^2)$$

for the parameters of the travel times distributions. Note that the prior distribution for σ_i is not a standard prior density, the more common choices being the conjugate inverse-gamma prior or the Jeffrey's prior $\pi(\sigma_i^2) \propto 1/\sigma_i^2$ (Lee, 2004). However, the parameters of the log-normal distribution can be chosen to make its distribution function match that of the inverse-gamma very closely, and this formulation is easily extended to more the more complex models. The prior distributions selected for p_j and U_j are the same as those chosen for the diagonal model in section 2.2.3. For the simple model, the values of the parameters

of the prior distributions (e.g., ξ_μ and τ^2_μ) are fixed a priori and assumed to be the same for all i or j.

Next, I extend the simple model to a fully hierarchical model by specifying hyper-prior distributions for the parameters of the prior distributions, rather than fixing these values. In my simple Bayesian model, the prior distributions all have the form:

$$\eta_i^X \sim N(\mu_X, \sigma_X)$$

where X is one of μ, σ, p or U, and $\eta_i^\mu = \mu_i$, $\eta_i^\sigma = \log(\sigma_i)$, $\eta_i^p = \text{logit}(p)$ and $\eta_i^U = \log(U_i)$. Here i is used to index the strata at either site for notational convenience. The idea is that all four sets of parameters have the exactly the same probability structure after transformation. In my formulation of the hierarchical model the prior parameters are assigned the hyper-prior distributions:

$$\mu_X \sim N(0, \tau_X^2) \text{ and } 1/\sigma_X^2 \sim \Gamma(\alpha_X, \beta_X)$$

with τ_X^2, α_X and β_X chosen separately for each of $X = \mu, \sigma, p, U$. This formulation is similar to the hierarchical model of Mantyniemi and Romakkaniemi (2002) except for the choice of prior distribution for U_j. Rather than defining this prior through in two-steps modelling $U = \sum_{j=1}^t U_t$ and then $(U_1, \ldots, U_t)|U$, I have defined the prior for each U_j independently.

As with the hierarchical model of diagonal data, this model shares information among strata but does not account for the natural ordering in temporally stratified data. To account for this ordering, I again consider models based on the Bayesian P-spline model of Lang and Brezger (2004). Because the four sets of parameters all share the same probability structure in the hierarchical model, the P-spline model described in section 2.2.3 can be applied equally to any of $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{p}$ or $\boldsymbol{U}$. It is also possible to smooth more than one set of parameters at the same time. In this section I focus only on smoothing of $\boldsymbol{U}$ and discuss smoothing of the remaining sets of parameters in section 2.5.

Exactly as in section 2.2.3 I define:

$$\log(U_j) = \sum_{k=1}^{K+q} b_k B_k(j) \tag{2.14}$$

where $B_k(\cdot)$ is the k^{th} B-spline function of order q and b_k the associated coefficient. The

coefficients are again assigned the second difference prior:

$$b_{k+1}|b_k, b_{k-1} \sim N(b_k + (b_k - b_{k-1}), \tau_U^2) \tag{2.15}$$

for $k = 3, \ldots, K + q$, with improper flat priors for the initial two coefficients and inverse-gamma prior for the variance parameter, τ_U^2. I also consider the extended model:

$$\log(U_j) = \sum_{k=1}^{K+q} b_k B_k(j) + \epsilon_j \tag{2.16}$$

where $\epsilon_j \sim N(0, \tau_\epsilon^2)$ allows for random deviations from the spline model. The prior for the variance of the errors is chosen to be the same as the prior selected for τ_U^2, $1/\tau_\epsilon^2 \sim$ Gamma(α, β).

Inference for the Bayesian P-spline model was again based on numerical summaries of the posterior distribution computed from samples obtained via MCMC in OpenBUGS. Several chains were run in parallel and convergence was monitored by comparing trace plots and computing the Gelman-Rubin-Brooks diagnostics.

Model fit was assessed with the Bayesian p-values computed from the discrepancy measures defined in section 2.2 adjusted to account for the non-diagonal matrix of recoveries. The new discrepancy measures were:

$$\begin{aligned}
D_1(\boldsymbol{X}, \boldsymbol{\Theta}) &= -2 \cdot L(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{p}, \boldsymbol{U}|\boldsymbol{X}) \\
D_2(\boldsymbol{X}, \boldsymbol{\Theta}) &= \textstyle\sum_{j=1}^{t} \left(\sqrt{\sum_{i=1}^{s} m_{ij}} - \sqrt{\sum_{i=1}^{s} n_i \theta_{ij} p_j} \right)^2 \\
D_3(\boldsymbol{X}, \boldsymbol{\Theta}) &= \textstyle\sum_{i=1}^{t} \left(\sqrt{u_j} - \sqrt{U_j p_j} \right)^2 .
\end{aligned}$$

As in section 2.2, the first discrepancy provided an overall measure of fit based on the likelihood in equation (2.12), the second compared the observed and expected numbers of marked individuals recovered in each stratum, and the third compared the observed and expected numbers of unmarked individuals captured in each stratum at the second site. I also introduced a fourth discrepancy to compare the observed and expected counts in each cell of $\boldsymbol{M}$:

$$D_4(\boldsymbol{X}, \boldsymbol{\Theta}) = \sum_{i=1}^{s} \sum_{j=1}^{t} \left(\sqrt{m_{ij}} - \sqrt{n_i \theta_{ij} p_j} \right)^2 .$$

Model selection was again based on the DIC presented in equation (1.3). However, when

I initially computed the DIC for simulated, non-diagonal data I found that the values of p_D produced by the simple Bayesian model with independent priors were often negative. Spiegelhalter et al. (2002) noted that this could occur if the marginal posterior distributions of some parameters were very skewed so that posterior means were a poor measure of the centre of these distributions. Plotting the marginal densities of the parameters of the simple Bayesian model showed that the posterior distributions of the capture probabilities were highly skewed and that the posterior means and medians differed considerably. Replacing $\hat{\boldsymbol{\theta}}$ in the DIC with the vector of marginal posterior medians produced values of p_D for the simple Bayesian model that were both positive and consistent with the other models. For consistency, DIC values for all Bayesian models were computed with marginal posterior medians as the point estimates of all parameters. The effect on the DIC of the remaining models was negligible.

2.3.4 Simulation Study

I again assessed the performance of the Bayesian P-spline models by analysing simulated data sets. In section 2.2.4, I conducted simulations for diagonal experiments with the daily run size modelled both exactly by a smooth curve (Simulation 1a) and also by a smooth curve plus noise on the log scale (Simulation 1b). Comparing the results in these two situations allowed me to assess how the smoothness of the true population size affected the performance of the P-spline estimators. The results showed that including the extra error term in the Bayesian P-spline model made little difference when the true counts exactly followed a smooth curve, but improved estimation when the true counts included added noise. In reality, the true daily population size will never exactly follow a smooth curve and so I only present simulations from the smooth curve with noise for non-diagonal experiments.

As in section 2.2.4, I designed the simulations to mimic the results of the analysis of the Conne River data (details in section 2.4.1). The the new simulation study assumed that fish trapping was conducted for 40 days at the first site, with exactly 100 individuals marked per day, and 45 days the second site. Ten realizations of the vector of daily run sizes, $\boldsymbol{N} = (N_1, \ldots, N_{40})$, were randomly generated from the model in equation (2.9), and for each of these realizations 30 data sets were simulated as follows. First, I sampled the parameters of the log-normal travel time distribution and capture probabilities independently from the

distributions:

$$\begin{aligned} \mu_i &\sim N(.5, 2.5) && i = 1, \ldots, 40 \\ \log(\sigma_i) &\sim N(-1.0, .5) && i = 1, \ldots, 40 \\ p_j &\sim \text{Beta}(2.5, 50) && j = 1, \ldots, 45. \end{aligned}$$

I then computed the entries of the matrix of transition probabilities, $\boldsymbol{\Theta}$, according to equation (2.13) with $d = 0$, and simulated the matrix of recoveries, $\boldsymbol{M}$, by applying the multinomial model in equation (2.10) with $n_i = 100$. I generated the numbers of unmarked individuals present in each stratum at the second site by first generating the intermediate values:

$$(U^*_{ii}, \ldots, U^*_{it}, U^*_{i-}) \sim \text{Multinomial}\left(N_i - n_i, (\theta_{ii}, \ldots, \theta_{it}, 1 - \sum_{j=i}^{t} \theta_{it})\right)$$

which simulate the movements of the unmarked individuals by assuming that all fish, marked or unmarked, have the same transition probabilities and then summing across days, $U_j = \sum_{i=1}^{\min(i,t)} U^*_{ij}$. The variable U^*_{i-} represents the number of fish passing the first site on day i that were not marked and that arrived at the second site after the end of the study. Although this violates the assumption of closure, estimates should still be unbiased for the total population size, $N = \sum_{i=1}^{40} N_i$ (Seber, 1982, pg. 72). Finally, the numbers of unmarked fish captured at the second site on each day, $u_1, \ldots, u_{45}$ were simulated according to the independent binomial model in equation (2.11).

Estimates of the total population size for each of the 300 data sets were computed with 5 different methods, namely: the pooled Chapman estimator, the simple Bayesian model, the fully hierarchical model, the Bayesian P-spline model and the Bayesian P-spline with error. Daily Chapman estimates cannot be computed because the exact numbers of marked individuals in each stratum at the second site are unknown. The following prior distributions were selected for the parameters of each of the Bayesian models:

1) Simple Bayesian model:

$$\begin{aligned} \mu_i &\sim N(0, 4) \\ \log(\sigma_i) &\sim N(0, 4) \\ \text{logit}(p_j) &\sim N(-2, 1.22) \\ \log(U_j) &\sim N(7.5, 4) \end{aligned}$$

2) Fully hierarchical Bayesian model:

$$\mu_i \sim N(\xi_\mu, \sigma_\mu^2), \text{ with } \xi_\mu \sim N(0,4) \text{ and } 1/\sigma_\mu^2 \sim \Gamma(.001,.001)$$
$$\log(\sigma_i) \sim N(\xi_\sigma, \sigma_\sigma^2), \text{ with } \xi_\sigma \sim N(0,4) \text{ and } 1/\sigma_\sigma^2 \sim \Gamma(.001,.001)$$
$$\text{logit}(p_j) \sim N(\xi_p, \sigma_p^2), \text{ with } \xi_p \sim N(-2,1.22) \text{ and } 1/\sigma_p^2 \sim \Gamma(.001,.001)$$
$$\log(U_j) \sim N(\xi_U, \sigma_U^2), \text{ with } \xi_U \sim N(7.5,4) \text{ and } 1/\sigma_U^2 \sim \Gamma(.001,.001).$$

3) Bayesian P-spline model:
Prior distributions for the parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$ and $\boldsymbol{p}$ were the same as in 2). The vector $\boldsymbol{U}$ was modelled according to equation (2.14) with the second difference prior of equation (2.15) for the coefficients of the spline and $1/\tau_U^2 \sim \Gamma(1,.05)$.

4) Bayesian P-spline model with Error:
Prior distributions for the parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$ and $\boldsymbol{p}$ were the same as in 2). The vector $\boldsymbol{U}$ was modelled according to equation (2.16) with the second difference prior for the coefficients of the spline, $1/\tau_U^2 \sim \Gamma(1,.05)$ and $1/\tau_\epsilon^2 \sim \Gamma(1,.05)$.

As in section 2.2.4, I selected the parameters of the prior distributions and hyper-priors for the simple and hierarchical models to produce weakly informative priors. For the four Bayesian models, posterior summary statistics were computed from random samples generated via MCMC in OpenBUGS (Thomas et al. (2006)). Each chain was run for 100,000 iterations, the first 20,000 iterations were discarded as burn-in, and the remaining iterations were thinned by a factor of 5.

Figure 2.4 illustrates ten sample realizations of the daily numbers of unmarked fish passing the second site. Although the mean daily population size over all 300 datasets formed a smooth curve, the daily values for one realization varied considerably from one day to the next. Total population size for the 10 realizations of $\boldsymbol{N}$ ranged from 26,280 to 36,890.

Table 2.5 presents the statistics comparing the estimators of total population size from the 5 methods. As in the earlier simulations, the estimator of total population size computed from the pooled data was almost unbiased, but its standard error greatly underestimated the estimator's true variability. Although the mean standard error for the pooled estimator was lower than the posterior standard deviation for any of the Bayesian models, its true variability was in fact higher than the variability of the posterior mean for the simple Bayesian model and both the Bayesian P-spline models.

The estimator of population size from the simple Bayesian model had a slight negative bias, in contrast to the previous simulations in which it showed a positive bias. However, the overwhelming problem with this method remained; the standard deviation of the posterior distribution greatly overestimated the true variability of the estimator. Results from the hierarchical Bayesian model were consistent with the previous simulations, indicating that the posterior mean slightly overestimated the true population size, on average, and that the posterior standard deviation overestimated the true variability in the estimator. The true variation in the posterior mean for the hierarchical Bayesian model was almost 1.4 times that of the other Bayesian models.

When the true daily population size deviated from the smooth curve in the simulations of the previous section, the Bayesian P-spline model with error performed better than the model without error. In particular, the model with error produced more accurate estimates of the uncertainty in the total population size. In contrast to this, the two models produced virtually identical results in the current simulation study. The %Bias, %MSE, mean posterior standard deviation of the population size and true variability in the posterior mean were all very similar (within .02 for all 4 statistics). The largest difference between the models was in the %Bias of the posterior standard deviation as an estimator of uncertainty, but the magnitude of this difference (4%) is highly unlikely to be important for making wildlife management decisions. Note that the %MSEs for both models were much lower than for all other models – approximately 2/3 the %MSE for the Pooled Chapman estimator and simple Bayesian model, and less than half the %MSE for the hierarchical Bayesian model.

Table 2.6 presents summary statistics for the DIC and the Bayesian p-values for the 4 Bayesian models. The differences in the DIC values provide very strong evidence for the Bayesian P-spline models over the simple or hierarchical models. The difference between the DIC for the P-spline models, on average, is marginal according to the guidelines of Spiegelhalter et al. (2002) and does not provide strong evidence for one over the other. Note that the complexity of the model with error was slightly greater, as would be expected. My interpretation is that the P-spline model with the added error term does provide better fit to the data (lower deviance) on average, but the improvement is marginal in comparison to the extra complexity of the model. Out of the 300 data sets, the P-spline model with error provided the lowest DIC value for almost 72%. Together, the Bayesian P-spline models with and without error produced the lowest DIC values for just under 96% of the simulated data sets.

The mean Bayesian p-values for the simple Bayesian model were all above .5 and the intervals formed by the 5^{th} and 95^{th} percentiles were skewed toward 1. As in section 2.2.4, these values do not indicate poor model fit, but rather result from the non-informative prior overwhelming the information in the data. Average p-values for the remaining models were all between .4 and .6 with the exception of the values associated with the first and third discrepancy measures for the P-spline model without error. This suggests that although the estimates of total population size differ negligibly between the spline models without and with the error term, the estimates of daily population size from the model with error fit better with the data. This is not surprising given that the true run size does not exactly follow a smooth curve. It is again clear that the p-values for all discrepancy measures do not follow uniform distributions on the unit interval. In general, the distributions place more of their mass closer to .5 than would be expected under the uniform. This means that the .05 or .1 cut off rule for assessing lack of fit would be more conservative rules than intended, rejecting the fit of the true model less often than would be expected by chance under the true model.

Method	%Bias(SE)	% MSE	Mean SE	Obs. SD	% Bias(SE)
Pooled Chapman	−1.4(0.8)	1.9	7.2	14.5	−50.0(0.5)
Simple Bayes.	−5.6(0.6)	1.7	15.7	11.6	39.5(2.7)
Hierarchical Bayes.	8.2(0.9)	2.9	15.4	15.1	4.7(2.5)
P-Spline	0.4(0.6)	1.2	11.2	11.4	−1.0(1.1)
P-Spline w error	0.5(0.6)	1.1	11.4	11.1	3.2(1.1)

Table 2.5: Comparison of the estimators of abundance and their uncertainty for Simulation 2. Estimation of abundance is assessed according to the %Bias and %MSE of the point estimators. Values provided are the mean estimate of N (scaled by 10^3), the %Bias (with standard error in parentheses), and the %MSE. Estimation of the uncertainty is assessed according to the %Bias of the point estimator of the standard error. Values provided are the mean of the uncertainty estimates from the 300 simulated data sets and the observed standard deviation of the estimator of N (both scaled by 10^3), and the %Bias of the uncertainty estimate (with its standard error in parentheses).

	DIC			Bayesian P-values			
	Mean DIC	Mean pD	% Lowest	1	2	3	4
Simple Bayes.	732.8	115.0	3.2	0.63(0.51,0.74)	0.57(0.36,0.76)	0.64(0.43,0.82)	0.63(0.52,0.72)
Hierarchical Bayes.	720.5	92.4	0.6	0.48(0.36,0.62)	0.44(0.19,0.67)	0.43(0.18,0.72)	0.47(0.33,.062)
Bayes. P-Spline	708.6	71.9	24.3	0.34(0.16,0.54)	0.51(0.13,0.85)	0.38(0.10,0.69)	0.45(0.15,.071)
BPS w error	706.0	75.5	71.6	0.43(0.29,0.59)	0.55(0.21,0.85)	0.43(0.16,0.72)	(0.49,0.22,0.73)

Table 2.6: DIC and Bayesian p-values for the Bayesian models applied in Simulation 2. The left half of the table presents the mean values of DIC and p_D over the 300 simulated data sets. Also provided is the % of the data sets for which each method produced the lowest DIC value. The right half of the table presents the mean Bayesian p-value for each of the four discrepancy measures. The 5^{th} and 95^{th} percentiles of the Bayesian p-values are provided in parentheses.

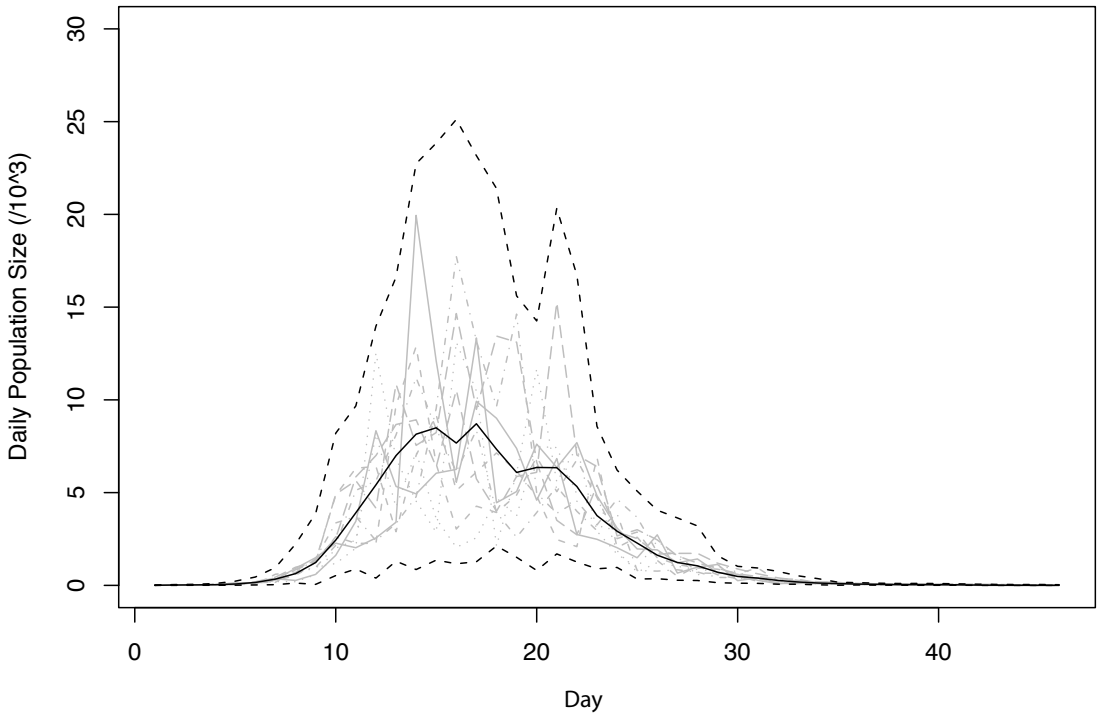

Figure 2.4: Number of unmarked fish passing the second trap on each day in Simulations 2. Each grey lines represents one of the numbers of unmarked fish. The solid black line indicates the daily mean and the dashed black lines the daily maximum and minimum numbers of fish passing on each day over the 300 simulated data sets.

2.4 Applications

2.4.1 Conne River Atlantic Salmon

The data set I have analysed in this first application is a modified version of data originally collected during a study Atlantic salmon smolts migrating along the Conne River, Newfoundland, in 1987. In the study, smolts were trapped at two sites along the river each day for a period of 1.5 months between April 26 and June 10. A total of 4975 smolts were captured and tagged at the first site of which 998 (20.1%) were recovered at the second site. A further 13,363 unmarked fish were captured at the second site.

The full data set was analysed previously by both Schwarz and Dempson (1994) and Mantyniemi and Romakkaniemi (2002). Schwarz and Dempson (1994) estimated the total population size (including both marked and unmarked individuals) to be 74.5 thousand fish (95% CI=62.5,76.7 thousand), though the analysis ignored data from a few days at both the beginning and end of the study when the numbers of marked fish were too low to produce estimates of the capture probabilities and population size. Mantyniemi and Romakkaniemi (2002) were able to analyse the entire data set with the hierarchical Bayesian implementation of the Schwarz-Dempson model and estimated the total run size at 74.5 thousand smolts (95% CI=68.7,80.8).

Although Schwarz and Dempson (1994) found it necessary to exclude the data collected from a few days at the start and the end of the migration, large numbers of smolts were marked and recovered on most days of the study and data sparsity was not a severe issue. Note that the analysis of Mantyniemi and Romakkaniemi (2002) including all days produced exactly the same point estimate of the population size, suggesting that a negligible number of fish passed on the days excluded by Schwarz and Dempson. My method is designed specifically to address the problems of modelling sparse data, and so I modified the data set by reducing the numbers of marked and unmarked fish captured on each day at the second site. To do this, I generated new values for the numbers of marked and unmarked fish captured at the second site, m^*_{ij} and u^*_j, as:

$$m^*_{ij} \sim \text{Binomial}(m_{ij}, .2)$$
$$u^*_j \sim \text{Binomial}(u_j, .2)$$

where m_{ij} and u_j are the original data values. In effect, this subsampling reduced the capture

probabilities at the second site on each day of the study by a factor of 5, but maintained the other relationships in the data. The new data set comprised the same number of releases at the first site, but the total numbers of marked and unmarked smolts captured at the second site were reduced to 183 and 2697 respectively. The raw proportion of marked smolts recovered for the modified data set was .04. Table 2.7 provides the subsampled data set.

Estimates of the population size were produced from the completely pooled Chapman estimator and the same four Bayesian models considered in the simulations of section 2.3.4. Inference for the Bayesian models was again based on samples from the posterior distribution generated via MCMC sampling implemented in OpenBUGS. For each model, 5 chains were run for 500,000 iterations each, the first 200,000 iterations of each chain were discarded as burn-in and values from every 50^{th} of the remaining iterations were retained for inference. Dispersed initial values for the 5 chains were generated by running a single chain for 100,000 iterations, computing the mean and variance matrix for the final 50,000 iterations, thinned every 10 iterations, and then simulating 5 independent multivariate-normal values with that mean and an inflated variance matrix. Convergence of each model was confirmed by computing the GRB diagnostics. Figures 2.5 and 2.6 depict summary statistics for the posterior distributions of the daily numbers of unmarked fish and the capture probabilites for each of the Bayesian models. Table 2.8 summarizes the posterior distributions of total population size.

The pooled estimate of the total run size was 77.9 thousand fish with 95% CI (72.5,83.4). Although the point estimate is very close to the estimates produced from the full data, the simulations in section 2.3.4 suggest that the interval estimate greatly underestimated the true variability. The 95% CI was in fact narrower than the interval estimates produced by both Schwarz and Dempson (1994) and Mantyniemi and Romakkaniemi (2002) for the full data set, even though the amount of information had been greatly reduced.

Estimates of the daily numbers of unmarked fish passing the second location, U_j, from the simple Bayesian model increased over the initial days of the run, peaked at day 14 and declined thereafter with the exception of a slight increase on day 17 to 18. Points estimates of p_j did not show a strong trend and varied considerably over time. However, the inference regarding these parameters was very imprecise. The 95% CIs for many values of p_j ranged from near .00 to above .40 and the average width of these CIs was .34. The uncertainty in p_j also lead to high uncertainty in U_j. At the peak of the run, the 95% CIs for the number

of unmarked smolts extended from below 5 thousand to above 15 thousand. The estimate of total population size from the simple model was 58.4 thousand, which was well below the estimate from the full data. The absolute width of the 95% was 41.0 thousand and the width relative to the point estimate was 69.1%.

The most obvious changes in the inference between the simple and hierarchical Bayesian models concerned the estimation of the capture probabilities. Posterior estimates of p_j for the hierarchical model were lower on average, were much less variable among days, and had much higher precision. These changes in the estimation of p_j also had corresponding effects on the estimation of U_j. Most obviously, the estimates of $p_{10}, \ldots, p_{13}$ were much lower for the hierarchical model and this lead to a corresponding increase in the estimates of $U_{10}, \ldots, U_{13}$, creating an earlier, and more extended, increase in the run size. The estimated total population size from the hierarchical model was 81.5 thousand; the absolute and relative widths of the 95% CI were both approximately half those of the simple Bayesian model.

Figure 2.5C shows that point estimates of U_j produced from the Bayesian P-spline model effectively smoothed the point estimates from the hierarchical model. The result was a smooth, unimodal curve that increased steadily from the start of the study, peaked on days 13 and 14, and then decreased to the end of the study. Throughout all 46 days, the smoothed estimates were much more precise than those produced by the hierarchical model. Introducing the error term into the P-spline model, Figure 2.5D, reintroduced some of the daily variation seen in the hierarchical model, including the increase from U_{17} to U_{18}, but the magnitude of these variations was diminished. As expected, the precision of the estimate for each U_j was lower for the P-spline model with error than for the model without, but this had a minimal effect on the precision of the estimate of total population size. The estimates of total population size for the Bayesian P-spline models with and without error were 68.6 thousand and 71.2 thousand and the 95% CIs from the two models were very similar, both absolutely and relative to the estimated run size. These intervals were .8 times as wide as the 95% CI for the hierarchical Bayesian model.

Table 2.9 compares the DIC and Bayesian p-values for the 4 Bayesian models. The lowest DIC was produced by the Bayesian P-spline model with the added error term, followed in order by the P-spline model without the error, the hierarchical model and the simple Bayesian model. The difference in DIC between the P-spline models with and without the error term was on the border of significance suggested by Spiegelhalter et al. (2002) and

provides weak evidence that the model with the error better fits the daily run size. The remaining DIC values provided strong evidence in favour of the P-spline smoothing of $\boldsymbol{U}$.

Bayesian p-values for the hierarchical model and both Bayesian P-spline models were all greater than .20 and less than .66. Based on the results of the simulation, I believe that these values provide no evidence of lack of fit. Note that the Bayesian p-values based on discrepancy measures 2 and 3 for the simple Bayesian model were well below the values obtained in the simulation. Further, the value of p_D for the simple Bayesian model was lower than the value for the hierarchical model suggesting that the simple model had fewer effective parameters. These results will be discussed further in section 2.5.

Figure 2.7 provides traceplots and plots of the estimated auto-correlation function of the 5 MCMC chains for the Bayesian P-spline model with error for selected values of $\log(U_j)$. These plots demonstrate that the 5 parallel chains converged very quickly from the initial values to a common distribution and that the correlation between iterations in the same chain were very low, after thinning. In fact, the correlation between the values on successive iterations (lag of 1) retained for inference was negligible suggesting that the chains could have been run for fewer iterations with less thinning to obtain the same level of precision in the posterior summary statistics. Table 2.10 summarises the GRB diagnostics for both the posterior mean and 97.5 percentile of $\log(U_j)$. After the selected burn-in period, 200,000 iterations, the maximum GRB values were 1.01 and 1.04 for the mean and 97.5 percentile respectively. This confirmed that the 5 chains had converged to similar distributions by this point. GRB diagnostic values were similar for the remaining parameters.

Date	Number	Recoveries on Days Following Tagging										Total	Number
	Tagged	0	1	2	3	4	5	6	7	8	9	Recoveries	Untagged
26/04/1987	8	0	0	0	0	2	0	0	0	0	0	2	0
27/04/1987	5	0	0	0	0	0	0	0	0	0	0	0	2
28/04/1987	6	0	0	0	0	0	0	0	0	0	0	0	1
29/04/1987	17	0	0	2	1	1	0	0	0	0	0	4	2
30/04/1987	66	0	1	0	2	3	2	0	0	0	0	8	39
01/05/1987	193	0	1	7	7	2	2	0	0	0	0	19	226
02/05/1987	90	0	2	0	0	0	2	0	0	0	0	4	75
03/05/1987	260	0	0	14	6	1	1	1	0	1	0	24	129
04/05/1987	368	0	9	46	4	2	1	0	3	0	1	66	120
05/05/1987	506	0	38	33	11	0	1	3	0	0	0	86	380
06/05/1987	317	1	27	26	3	1	4	0	0	0	0	62	921
07/05/1987	43	0	4	3	0	2	0	0	0	0	0	9	1005
08/05/1987	259	1	42	5	2	0	0	0	0	0	0	50	1181
09/05/1987	259	1	32	27	1	0	0	0	0	0	0	61	1087
10/05/1987	249	1	85	3	1	0	0	0	0	0	0	90	1108
11/05/1987	250	3	21	19	2	0	0	0	0	0	0	45	1685
12/05/1987	298	42	16	11	9	1	0	0	0	0	0	79	671
13/05/1987	250	1	7	25	6	4	0	0	0	0	0	43	1766
14/05/1987	193	0	9	18	8	0	0	0	0	0	0	35	636
15/05/1987	207	0	17	21	2	0	0	0	0	0	0	40	483
16/05/1987	175	0	18	10	1	0	0	0	0	0	0	29	170
17/05/1987	141	0	12	14	7	1	1	0	0	0	0	35	269
18/05/1987	155	0	1	19	13	6	2	0	0	0	0	41	212
19/05/1987	123	0	5	22	5	0	0	0	1	0	0	33	260

Continued on following page.

Date	Number	Recoveries on Days Following Tagging										Total	Number
	Tagged	0	1	2	3	4	5	6	7	8	9	Recoveries	Untagged
Continued from previous page.													
20/05/1987	128	0	6	17	2	1	0	0	0	0	0	26	154
21/05/1987	72	0	11	9	2	0	0	0	0	0	0	22	145
22/05/1987	57	0	6	8	0	1	0	0	0	0	0	15	99
23/05/1987	49	0	4	2	1	0	0	0	0	0	0	7	58
24/05/1987	57	14	2	1	0	0	0	0	0	0	0	17	74
25/05/1987	18	0	3	0	0	0	0	0	0	0	0	3	40
26/05/1987	20	0	3	4	0	0	0	0	0	0	0	7	50
27/05/1987	16	0	3	0	0	0	0	0	0	0	0	3	59
28/05/1987	15	0	0	2	0	0	0	0	1	0	0	3	40
29/05/1987	10	0	1	0	1	0	0	0	0	0	0	2	9
30/05/1987	13	0	0	2	0	0	0	0	0	0	0	2	14
31/05/1987	8	0	3	1	0	0	0	0	0	0	0	4	13
01/06/1987	2	0	1	0	0	0	0	0	0	0	0	1	22
02/06/1987	23	0	6	0	0	0	0	0	0	0	0	6	24
03/06/1987	20	0	2	0	0	0	0	0	0	0	0	2	33
04/06/1987	10	0	4	1	0	0	0	0	0	0	0	5	19
05/06/1987	10	3	1	0	0	0	0	0	0	0	0	4	12
06/06/1987	5	0	2	0	0	1	0	0	0	0	0	3	7
07/06/1987	2	0	0	0	0	0	0	0	0	0	0	0	4
08/06/1987	2	0	1	0	0	0	0	0	0	0	0	1	0
09/06/1987	0	0	0	0	0	0	0	0	0	0	0	0	0
10/06/1987	0	0	0	0	0	0	0	0	0	0	0	0	59
Total	4975											183	2697

Table 2.7: Data for the analysis of the 1987 run of Atlantic salmon along the Conne River. This data is sub-sampled from the data originally collected in the experiment.

	Mean	95% HPD	Width	Rel. Width
Simple Bayes.	54.4	(37.2,78.3)	41.0	75.4
Hierarchcial Bayes.	77.5	(62.2,94.8)	32.5	42.0
Bayes. P-spline	68.6	(56.9,81.1)	24.3	35.4
Bayes. P-spline + Error	71.2	(58.3,84.1)	25.8	36.2

Table 2.8: Abundance estimates from the four models of the subsampled Conne River Atlantic Salmon smolt data.

			Bayesian P-values			
	DIC	p_D	1	2	3	4
Simple Bayes.	761.2	89.4	0.97	0.03	0.05	0.55
Hierarchcial Bayes.	673.4	90.6	0.47	0.20	0.24	0.48
Bayes. P-spline	667.1	82.8	0.66	0.23	0.37	0.50
Bayes. P-spline + Error	664.2	83.8	0.55	0.26	0.33	0.47

Table 2.9: DIC and Bayesian p-values for the Bayesian models applied to the subsampled Conne River Atlantic Salmon smolt data. The left half of the table presents the DIC and p_D for each of the four models. The right half of the table presents the observed Bayesian p-value for each of the four selected discrepancy measures.

	50000	100000	200000	500000
Mean	1.01(1.00,1.04)	1.00(1.00,1.03)	1.00(1.00,1.01)	1.00(1.00,1.00)
97.5% Percentile	1.02(1.00,1.10)	1.01(1.00,1.07)	1.01(1.00,1.04)	1.00(1.00,1.01)

Table 2.10: Summary of the Gelman-Rubin-Brooks diagnostics for assessing convergence of $\log(U_j)$. The diagnostics were computed from 5 parallel chains of 500,000 iterations for each of the 46 days. The table reports the mean(minimum,maximum) value of the diagnostic over the 46 days at 50, 100, 200, and 500 thousand iterations for both the posterior mean (top row) and the 97.5% percentile (bottom row).

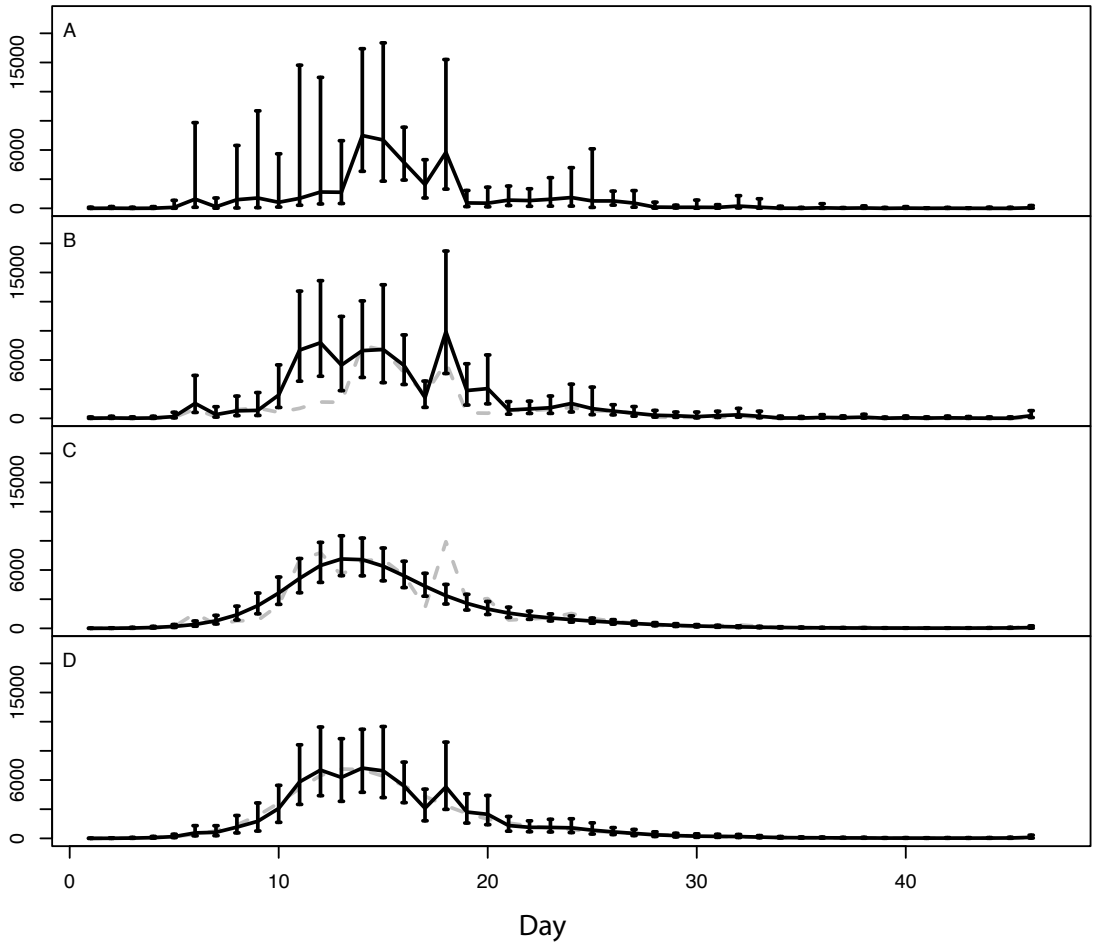

Figure 2.5: Estimated daily abundance of unmarked fish for the subsampled Conne River Atlantic Salmon data. The four panels illustrate the results from the 4 models fit to the data: A) simple Bayesian model, B) hierarchical Bayesian model, C) Bayesian P-spline model, and D) Bayesian P-spline model with error. In each plot, the solid black line connects the daily posterior median abundance and the error bars indicate the extents of the daily 95% CIs. The dashed grey lines in panels B, C and D indicate the daily medians from the preceding models, A, B and C respectively.

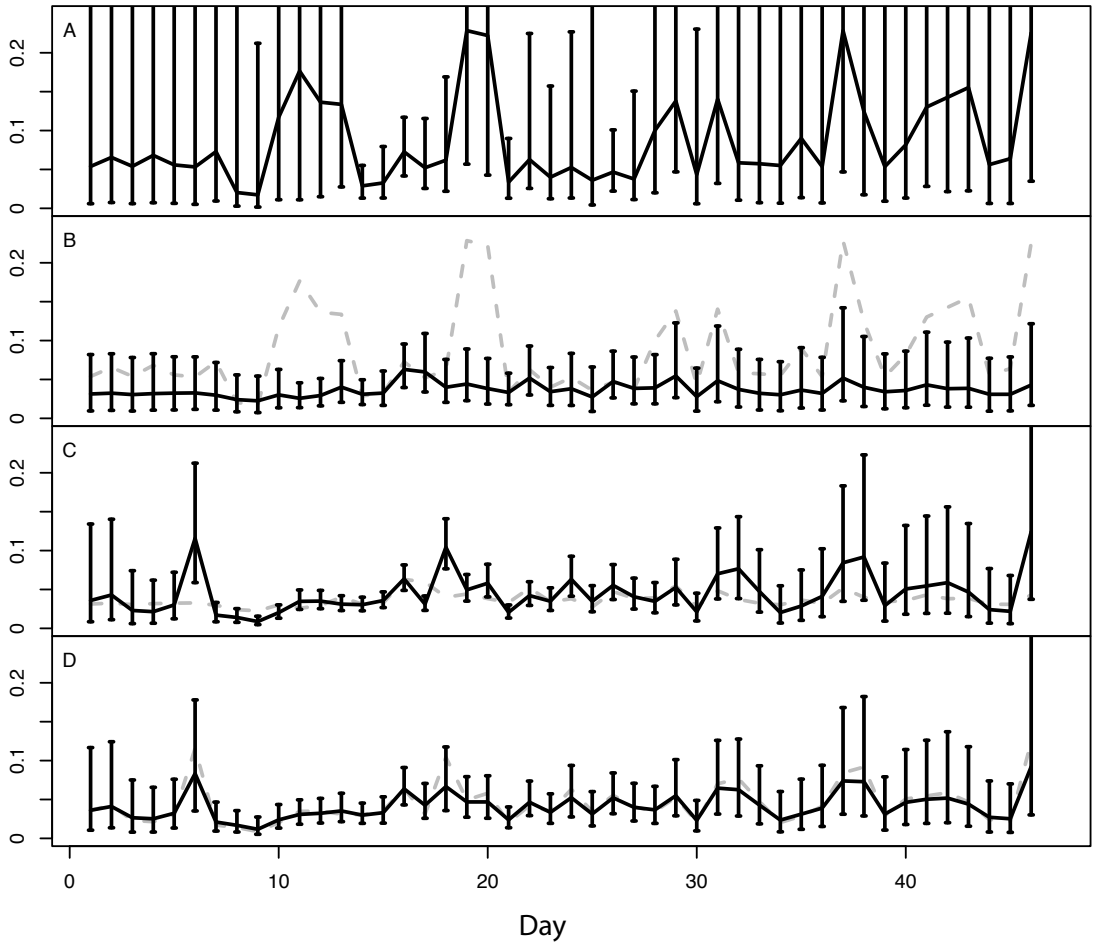

Figure 2.6: Estimated daily capture probabilities for the subsampled Conne River Atlantic Salmon data. The four panels illustrate the results from the 4 models fit to the data: A) simple Bayesian model, B) hierarchical Bayesian model, C) Bayesian P-spline model, and D) Bayesian P-spline model with error. In each plot, the solid black line connects the daily posterior median capture probabilities and the error bars indicate the extents of the daily 95% CIs. The dashed grey lines in panels B, C and D indicate the daily medians from the preceding models, A, B and C respectively.

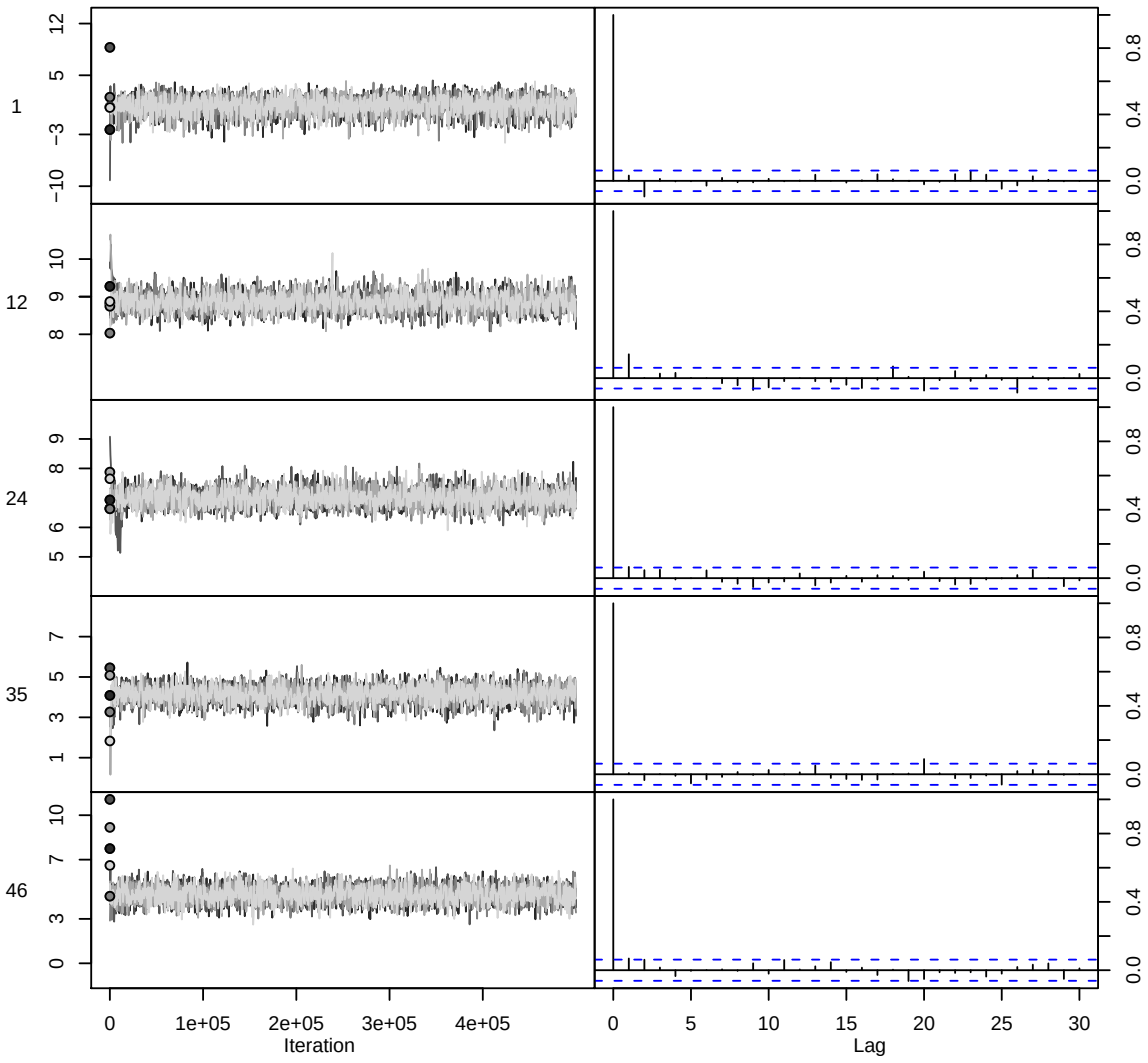

Figure 2.7: Traceplots and auto-correlation profiles for selected values of $\log(U_j)$. Traces for all 5 chains are represented in the left-hand plots for 5 equally spaced days, $j = 1, 12, 24, 35, 46$. The points at iteration 0 represent the randomly generated initial values for each of the chains. The right-hand plots illustrate the auto-correlation of the first chain at lags from 0 up to 30 iterations.

2.4.2 Trinity River Chinook Salmon

A second data set to which I have applied the Bayesian P-spline model was collected from an experiment monitoring the migration of Chinook salmon smolts along the Trinity River in Northern California. This data was collected as part of the Trinity River Restoration Program, a long term project aimed at restoring and monitoring fish habitat, and was provided by the Hoopla Valley Tribal Fisheries Department. My analysis focuses on estimating the number of migrating smolts from data collected at one trapping location, near Junction City, in 2003. In this year, smolts were trapped at the site for 38 consecutive weeks. Along with the wild smolts migrating downstream, smolts were released from a hatchery above Junction City in the 15^{th} and 31^{st} weeks of trapping.

The challenge in analysing this data set was not the sparsity of the data. A total of 50,660 salmon were marked over the 38 week period and 2442 of these were recaptured, an average of 64 (.05%) per week. Moreover, the data was almost diagonal; only 27 smolts were recovered outside of the week in which they were marked and these were ignored during the analysis. Summing the weekly Chapman estimates of the run size produced an estimate of 16.2 million smolts migrating over the entire period. However, there was one week in the data set, week 33, when the capture probability was apparently much lower than in the other weeks. In week 33, 2880 salmon were marked and only 8 (.003%) recovered. The proportion of marked fish recovered in this week was 6 times lower than the lowest recapture rate in the remaining 37 weeks, and the Chapman estimate of population size in this week, 11.1 million fish, was more than 2/3 of the estimate of the total population size.

It is clear that the numbers of smolts marked and recaptured in week 33 are not reliable. Either one of these numbers has been recorded incorrectly (e.g., the number of recaptures should be 80 not 8), or something happened during the experiment that greatly reduced the recapture rate (e.g., a large proportion of the marked fish were killed as they were transported to the release site). I have chosen to delete both the number of individuals marked and recaptured in week 33, but this leaves no information about the catchability in this week without assuming some relationship to the data in other weeks. The challenge in this analysis was to compute a reliable estimate of the total population size without any direct information regarding the capture probability in week 33.

In my analysis I fit two different models to the Trinity River Chinook data after deleting the numbers of smolts marked and recaptured in week 33: the fully hierarchical model and

the Bayesian P-spline model with error. The fully hierarchical model was exactly the same as that described in section 2.2 with the same prior distributions as chosen in the simulation of 2.2.4. The run size could not be expected to be smooth over all 38 weeks because of the two introductions of hatchery fish and so the Bayesian P-spline model had to be modified to allow for jumps in the run size. To accommodate the additions before weeks 15 and weeks 31 the spline was broken into 3 segments: the first modelled the run over weeks 1 to 14 with 3 equally spaced knots, the second modelled weeks 15 to 31 also with 3 equally spaced knots, and the third weeks 32 to 38 with only 1 knot. Prior distributions for the coefficients of each segment of the spline were defined separately, but the hyper-parameters were forced to be the same in order to achieve a constant degree of smoothnes. Explicitly, the prior distributions were:

$$\begin{aligned}
&\pi(b_1) \propto 1, \quad \pi(b_2) \propto 1, \quad b_{k+1} - b_k \ N(0, \tau_U^2) \quad k = 3, \ldots, 7 \\
&\pi(b_8) \propto 1, \quad \pi(b_9) \propto 1, \quad b_{k+1} - b_k \ N(0, \tau_U^2) \quad k = 10, \ldots, 14 \\
&\pi(b_{15}) \propto 1, \quad \pi(b_{16}) \propto 1, \quad b_{k+1} - b_k \ N(0, \tau_U^2) \quad k = 17, 18, 19
\end{aligned}$$

where τ_U^2 was the same for all 3 segments of the spline and was modelled with the same inverse gamma prior as in section 2.2.4.

Models were again fit via MCMC sampling in OpenBUGS, though only 1 chain of 500,000 iterations with a burn-in of 100,000 was run for each model. Figure 2.8 summarizes the posterior distributions of the capture probabilities and numbers of unmarked fish passing per week for both models. The plot of the capture probabilities shows that in week 33, the capture probability estimated from the hierarchical model shrank back toward the hierarchical mean (on the logistic scale) estimated from the remaining weeks. However, in the adjacent weeks the posterior mean capture probabilities were all lower than the hierarchical mean and there was a clear decreasing trend from weeks 29 to 35. Although the number of unmarked fish captured decreased from week 32 onward, the posterior mean run size for the hierarchical model drops from week 32 to 33, rises to week 34, and then drops again for the remaining weeks. Jumps in the estimated daily population size in weeks 15 and 32 result from the addition of the hatchery fish.

The Bayesian P-spline model applies the same hierarchical prior to the capture probabilities, but the shrinkage to the hierarchical mean in week 33 did not occur. Instead, the daily run sizes were smoothed by the spline such that the posterior means actually increased slightly from week 32 to 33 and then decreased steadily from week 33 onward. The posterior

mean capture probability in week 33 was well below the hierarchical mean to accommodate the trend in run size, but this estimate seemed to fit better with the pattern of capture probabilities observed on the neighbouring weeks.

Inference for both the capture probabilities and the run sizes for the remaining weeks were almost identical between the two models. Estimates of the total population size for the two model were 6.1 million (95% CI=5.6,6.7) for the hierarchical model and 6.3 million (95% CI=5.9,6.8) for the Bayesian P-spline model. However, the DIC for the P-spline model (686.7) was substantially lower than for the fully hierarchical model (693.8), mostly because the estimate of the effective number of parameters for the P-spline model ($p_D = 71.8$) was smaller than that of the hierarchical model (75.2). Goodness of fit was not examined for these models.

Week	Number	Recoveries			Total	Total
	Tagged	0	1	2	Recoveries	Catch
1	0	0	0	0	0	4135
2	1465	32	19	0	51	10449
3	1145	120	0	0	120	2188
4	229	25	0	0	25	669
5	20	0	0	0	0	294
6	183	17	0	0	17	699
7	707	74	0	0	74	969
8	632	93	0	0	93	974
9	1372	62	0	0	62	2398
10	283	7	3	0	10	467
11	647	31	0	0	31	895
12	275	11	0	0	11	422
13	277	12	1	0	13	406
14	335	13	1	0	14	524
15	4000	212	0	0	212	39917
16	4000	55	0	0	55	17588
17	2890	114	1	0	115	7861
18	3119	197	0	1	198	6891
19	2479	80	0	0	80	3571

Continued on following page.

Week	Number	Recoveries			Total	Total
	Tagged	0	1	2	Recoveries	Catch
Continued from previous page.						
20	1292	71	0	0	71	1715
21	2336	153	0	0	153	4220
22	2544	155	0	0	155	5014
23	2341	275	0	0	275	3288
24	1013	101	0	0	101	1301
25	729	65	1	0	66	989
26	333	44	0	0	44	444
27	269	33	0	0	33	339
28	77	7	0	0	7	107
29	62	9	0	0	9	79
30	26	3	0	0	3	35
31	20	1	0	0	1	26
32	4791	188	0	0	188	35131
33	2880	8	0	0	8	34530
34	3993	74	0	0	74	14832
35	1761	27	0	0	27	3617
36	1535	28	0	0	28	1819
37	485	14	0	0	14	632
38	115	4	0	0	4	115
Total	50,660				2442	209,550

Table 2.11: Data for the analysis of the 2003 run of Chinook salmon along the Trinity River. The columns indicate how many fish were tagged, how many of these were recovered and the total catch for each week. Smolts were recovered in the same week as tagging (lag 0) or 1 or 2 weeks after tagging.

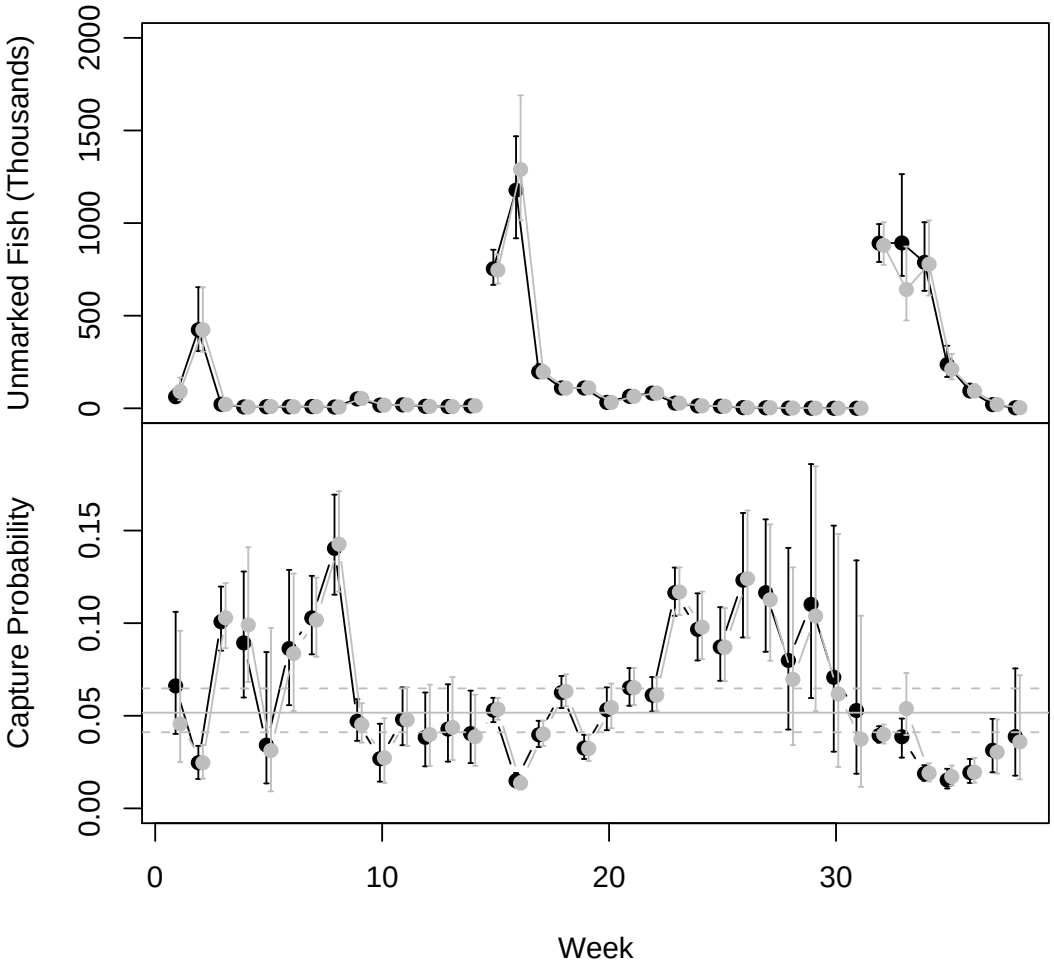

Figure 2.8: Summaries of the posterior distributions of the daily unmarked population size and the capture probabilities for the Trinity River Chinook salmon data. Grey lines and symbols represent the results from the fully hierarchical model. Black lines and symbols represent the results from the Bayesian p-spline model with error. For each model, the points indicate the posterior means and the error bars the extents of the 95% CIs. The solid, grey horizontal line in the second panel indicates the posterior mean of the hierarchical mean of the capture probabilities from the fully hierarchical model, transformed to the natural scale.

2.5 Discussion

The objective of this project was to develop a method for smoothing estimates of the populations size from temporally stratified capture-recapture data in order to share information between strata, reduce the effective number of parameters and produce more precise estimates when the data are sparse. I have done this by explicitly modelling the run size as a smooth curve using the Bayesian P-spline model proposed by Lang and Brezger (2004). In the Bayesian P-spline approach, the smoothness of the fitted curve is ensured by specifying a prior distribution on the coefficients of the basis functions which favours small changes in the spline at each knot and hence a smooth fit. The simulations I have presented demonstrate that the strict spline model produces better estimates of the total population size when the daily population size exactly follows a smooth curve, but I found it necessary to introduce an extra error term into the model when the population size deviated from the smooth curve. This was the preferred model in application to the Conne River Atlantic Salmon smolt data.

The previous method which is most similar to the Bayesian P-spline model is the hierarchical Bayesian model of Mantyniemi and Romakkaniemi (2002). In this approach, a Dirichlet prior distribution is specified for the proportions of unmarked individuals passing the second location in each stratum. In theory, the parameters of this distribution could be chosen to favour specific shapes for the run; e.g, if one specified parameters for the Dirichlet distribution that follow a parabolic curve then the expected prior daily run sizes would follow a similar shape. However, these parameters are fixed in the model of Mantyniemi and Romakkaniemi (2002) and there are many questions that would need to be answered in specifying the prior: How fast is the run size expected to increase/decrease? Are the increase and decrease symmetrical? Exactly where should the peak occur? Is one peak correct or should there be multiple peaks? The spline model avoids these questions by assuming only that the curve is smooth without imposing any further restrictions on its shape.

Although I have focused this chapter on smoothing the run size between strata, exactly the same methodology can be applied to the other sets of parameters in the model. For example, if the time that fish take to move between the two locations is a function of an environmental quantity that is unknown but thought to change smoothly over time (perhaps water flow rate) then one might hypothesize that the daily mean log travel-times follow a

smooth curve over time. Several models combining independent Bayesian P-splines for different combinations of the model parameters (e.g., smoothing only the capture probabilities or smoothing the run size as well as both the mean and standard deviation of the log-travel time) were tested on simulated data and also fit to the modified Conne River data. These models did not provide as good a balance between the number of parameters and fit to the data as measured by the DIC (results not shown).

One unexpected result of combining smoothing for different parameters was that any model that simultaneously smoothed both the run size and the capture probabilities produced a very poor fit to the Conne River data. The reason for this is that smoothing the run size and smoothing the capture probabilities are essentially opposite operations when the observed numbers of unmarked fish captured in each strata vary non-smoothly. If the true run size follows a smooth curve, then the capture probabilities must fluctuate in order to allow for the variation in the counts of unmarked fish and vice versa. Imposing smooth models on both sets of parameters results in a model that cannot adjust to the variation and fits neither the run size nor the capture probabilities properly. As a result, any model combining the P-spline for both the run sizes and capture probabilities produced a much higher DIC value than any other model, even though the effective number of parameters was reduced.

An approach that might improve the fit of these models is to incorporate the effects of covariates on the capture probabilities, parameters of the travel time distribution, or possibly the daily population size. For example, water flow rate is known to be a predictor of the capture probabilities in some experiments on Pacific salmon. The effect of this variable could be included in the linear predictor in order to explain some of the variation in the capture probabilities about the fitted spline and reduce the unexplained variation. However, such information was not available for either the Conne River or Trinity River data sets.

One extension of the model which I explored briefly removed the assumption of log-normally distributed travel times for the marked fish in each strata. The assumption of log-normality was first imposed by Schwarz and Dempson (1994) in order to reduce the number of parameters in the stratified-Petersen model and avoid problems with sparse data. My hope was that smoothing the run size might reduce the effective number of parameters sufficiently that this assumption could be relaxed. Allowing the transition probabilities, $\{\theta_{ij}\}$, to vary completely independently would require a very large number of parameters, and so I decided to apply a similar smoothing method to share information between the

travel times in adjacent strata at the first location (e.g., so that $\theta_{i,i+j}$ would be similar to $\theta_{i+1,i+1+j}$). The main challenge is that the number of transition probabilities for each stratum at the first location vary between strata (there are $t-i$ possibly non-zero values of θ_{ij} for stratum i) but must sum to 1 in each strata. This makes it difficult to relate the vectors of transition probabilities between different strata.

The solution I have considered is to assume that all fish marked at the first location must pass the second location in a fixed and known number of days, d, and that trapping is continued at the second location for at least this number of days after marking ends ($t-s \geq d$). This forces the number of (possibly) non-zero transition probabilities to be the same for all strata at the first location. I then implement the methods of Pawlowsky-Glahn and Egozcue (2006) for component data to map the vectors of transition probabilities for each stratum into a $d-1$ dimensional space where each element is unbounded and the sum-to-one constraint is enforced by the inverse mapping. Finally, I smooth the transformed vectors in the new space by modelling the differences in these vectors between adjacent strata as random draws from a multivariate normal distribution with mean zero vector and variance matrix selected to favour small changes. I have tested this approach in some simple models with limited simulations and the results are promising, but more work needs to be done. The challenges that still need resolving include choosing an appropriate transformation to the $d-1$ dimensional space and selecting the structure of the variance matrix. For example, one might assume a model for the covariances (perhaps AR(1) or exponential decreases) so that smoothing of the transition probabilities occurs within as well as among the strata.

One aspect of the approach I have taken that I find somewhat unsatisfactory is that the Bayesian P-spline models the number of unmarked individuals in each stratum, not the total population size. The total population in each stratum is the quantity of real interest and would be more likely to follow a smooth curve than the numbers of unmarked individuals which depend on the protocol for capture and marking at the first location. The problem with modelling the total in each stratum directly is that this quantity does not appear as a parameter in the model likelihood. Although this might be an issue in some experiments with high capture probabilities, I believe that modelling the numbers of unmarked individuals will be adequate for most applications. The number of marked individuals is usually a very small proportion of the total population size so that if the total follow a smooth curve then the numbers of unmarked individuals will also fall very close to a smooth curve. If the differences are large, then this can be accommodated in the P-spline

model with the extra error term. Another possible solution is discussed in Chapter 5.

There were two results in the analysis of the Conne River data that I found strange and that I believe require further discussion, both concerning the simple Bayesian model. First, the p_D value was lower for the simple Bayesian model than for the hierarchical model, suggesting that the simple model had fewer unique parameters. The reason for this is that p_D quantifies the number of estimable parameters in a model and the thinned Conne River data was too sparse for all of the parameters to be estimable without further assumptions. The second odd result was that the Bayesian P-values for the 2^{nd} and 4^{th} discrepancy measures suggested poor model fit, though the P-values for the hierarchical and P-spline models, which can be viewed as sub-models, did not. As discussed in section 2.2.4, this is caused by the influence of the prior distribution in the simple model. Figure 2.6 shows that the marginal posterior distributions of the capture probabilities for the simple model placed considerable mass on values well above what is known to be realistic. When parameters were simulated from this part of the posterior distribution and new data were generated from these values, the discrepancy for the observed data was higher than the discrepancy for the simulated data, resulting in the small p-value. Repeating the analysis with more informative prior distributions restricting the capture probabilities to smaller values produced Bayesian P-values that were closer to .5, but it may not be reasonable to assume this level of prior knowledge in practice. My conclusion is that the simple model does not provide enough structure for estimation from sparse data sets.

A second issue with the Bayesian P-values was that the values produced in the simulation studies did not follow a uniform distribution, even under the true model. This makes it difficult to interpret the P-values obtained from a single data set. From the simulations it seems that P-values outside of $(.2, .8)$ may be indicative of poor fit, but it is not clear if these bounds are fixed or depend on the parameters of the true model.

In conclusion, I believe that the Bayesian P-spline model provides an important improvement for estimating population size from temporally stratified data. By taking advantage of the temporal structure, the model produces more precise estimates of the population size than standard methods. In particular, I believe that the P-spline model with the extra error term will be applicable in a wide range of fisheries monitoring studies.

Chapter 3

Continuous, Time-Dependent, Individual Covariates in the Jolly-Seber Model

3.1 Introduction

This chapter and the next consider problems in the analysis of capture-recapture data from open populations – populations which change over the duration of the experiment as individuals enter through birth and immigration and depart through death and emigration. In particular, both chapters consider challenges in the analysis when the probabilities of capture or survival depend on covariates like body mass or fitness that are continuous and vary over time as well as among individuals.

There are two models which provide the basis for studying open populations from capture-recapture data, the Jolly-Seber (JS) and Cormack-Jolly-Seber (CJS) models (Cormack, 1964; Jolly, 1965; Seber, 1965). The key difference between the two is that the CJS model makes no assumptions about the behaviour of the individuals in the population that are never captured and remain unmarked for the entire experiment. Instead, the likelihood function models only the behaviour of each individual after it is first captured and marked. This provides inference about the survival rates during the experiment, but not about the population size. To make inference about the population size, the JS model adds the assumption that all individuals have the same probability of capture on each occasion,

regardless of whether they are marked or not.

Imagine an experiment comprising T capture occasions spaced far enough apart in time that individuals may enter or leave the population over the course of the experiment. Data for each of the individuals captured during the study can be summarized in a capture history, a vector of T 1's and 0's indicating whether or not the individual was captured on each occasion. The CJS model essentially assumes that the set of marked individuals captured on an occasion and the set which survive to the next occasion both form simple, random samples from the set of all marked individuals alive (full assumptions can be found in many sources including Seber (1965, pg. 197), Williams et al. (2002, pg. 422) and Pollock and Alpizar-Jara (2005, pg. 42)). Given these assumptions, a probability can be assigned to each capture history that is defined in terms of two sets of parameters, the T capture probabilities:

$$p_t = P(\text{an individual alive on occasion } t \text{ is captured}), \quad t = 1, \ldots, T$$

and the $T - 1$ survival probabilities:

$$\phi_t = P(\text{an individual alive on occasion } t \text{ survives to } t + 1), \quad t = 1, \ldots, T - 1.$$

Because the assumptions of the CJS model concern only the marked individuals, the probability assigned to each history considers only the events after the individual was first captured, marked and released. For example, an individual captured on occasions 2 and 4 of an experiment with 5 occasions would have the capture history 01010 which is assigned the probability:

$$\phi_2(1 - p_3)\phi_3 p_4((1 - \phi_4) + \phi_4(1 - p_5)).$$

This models the survival of the individual from occasion 2 to 3, the failure to capture it on occasion 3, etc. The full likelihood is constructed by multiplying these contributions for all individuals captured on at least one occasion, and maximum likelihood or Bayesian methods can be applied to make inference about the probabilities of survival (Seber, 1965; Poole, 2002).

To make inference about the population size on each occasion it is necessary to estimate the number of individuals, both marked and unmarked, that were alive but not captured. This requires further assumptions regarding the behaviour of the unmarked individuals alive

on each occasion. The JS model extends the CJS model by assuming that all individuals share the same probability of capture, whether or not they have been previously captured and marked. A new component is then added to the likelihood which models the number of individuals captured for the first time on each occasion of the experiment. In the classical methods originally provided by Jolly (1965) and (Seber, 1965), inferences about population's size are obtained in two steps. First, the CJS likelihood is maximized to produce estimates of the capture probabilities, $\hat{p}_2, \ldots, \hat{p}_T$ (note that p_1 is not estimable). The number of individuals alive on occasion t is then estimated by the ratio estimator:

$$\hat{N}_t^{JS} = \frac{n_t}{\hat{p}_t}$$

where n_t is the total number of individuals captured on occasion t, both marked and unmarked. Bayesian methods have also been developed (e.g. Royle et al., 2007).

The key assumption for estimating population size is that all individuals alive on one occasion have the same probability of capture. Although this might be a reasonable approximation in some populations, there are many factors that might affect an individual's catchability. This heterogeneity has long been known to induce bias in the estimates of population size (Gilbert, 1973; Carothers, 1973). In general, there have been two approaches to account for natural variation in the capture and survival probabilities. The first is to consider all differences as the results of unexplainable, random heterogeneity. Hwang and Chao (1995) derived a large sample bias correction for the specific case in which capture probabilities vary randomly among individuals, but with additive effects for variation over time. Pledger and Efford (1998) introduced a simulation based method for a simplified model involving individual capture probabilities, but assuming that all parameters, including the population size, are constant over time. Attempts have also been made to explicitly model individual capture probabilities as random effects with a parametric distribution (Pledger et al., 2003; Otis et al., 1978), but these methods have seen limited application. Large amounts of data are necessary to estimate separate parameters for each individual. Furthermore, Link and Barker (2005) demonstrated that two different distributions for the random effects can fit the data equally well but produce very difference estimates of the populations size.

The approach which has been applied more widely models the capture and survival probabilities as functions of explanatory variables. Lebreton et al. (1992) constructed a

general method for including predictors of the parameters of the CJS model by considering the capture and survival probabilities in a generalized linear modelling (GLM) framework. This was extended to the JS model and estimation of population size by McDonald and Amstrup (2001). As in the original JS model, inference for the population size on each occasion is obtained in two steps. First, the extended CJS likelihood of Lebreton et al. (1992) is maximized to produce estimates of the capture probabilities for each individual on each occasion after it is first captured. The number of individuals alive on occasion t is then estimated by the Horvitz-Thompson (HT) estimator:

$$\hat{N}_t^{HT} = \sum_{i=1}^{n_t} \frac{1}{\hat{p}_{it}}$$

where i indexes only the n_t individuals captured on occasion t and $\hat{p}_t$ is the estimated capture probability for the i^{th} individual. Note that this is exactly the JS estimate of population size if $\hat{p}_{it} = \hat{p}_t$ for all individuals.

Part of the appeal of the GLM framework was that it unified the models for many different types of variation. Explanatory variables included in the model could be discrete or continuous, individual or environmental (i.e., common to all living individuals) and constant or time-varying, provided that their values were known on all occasions after each individual was first captured. However, there are some explanatory variables which vary both among individuals and over time and which can only be known for an individual on the occasions that it was actually captured.

If the covariate is discrete then the multi-state model can be applied (Arnason, 1973; Schwarz et al., 1993; Brownie et al., 1993). This model was originally developed to account for differences in catchability when individuals could move randomly among different geographic locations. By assuming that movements follow the Markovian property, it is possible to model the location of each individual even when it is not captured. The likelihood contribution for each individual is then formed by summing the CJS likelihood contributions, with location specific capture and survival probabilities, over all possible sets of locations, weighted according to the transition probabilities. Inference about the population size may again be obtained with either classical maximum likelihood or Bayesian methods (Schwarz et al., 1993; Dupuis and Schwarz, 2007).

Bonner and Schwarz (2006) developed a method to incorporate continuous covariates as predictors in the CJS model. As in the multi-state model, it is assumed that the covariate

follows the Markovian property, but now with a continuous kernel. The resulting likelihood function requires integration over the unknown covariate values for each individual, which makes it very difficult to compute for any specific set of parameter values. Instead, inference was conducted through Bayesian methods via Markov chain Monte Carlo (MCMC) sampling.

The objective of the current project is to extend this model to the estimation of population size. The method I develop is similar to the two step procedure of both the standard JS model and the method of McDonald and Amstrup (2001). In the first step, the model of Bonner and Schwarz (2006) is applied to determine the effect of the covariate on the probabilities of capture and survival for each individual. In particular, MCMC sampling is used to generate a sample from the posterior distribution of these parameters. Inference about the population size is then made by computing a modified HT estimator for each realization, and treating these values as a sample from the posterior distribution of the population size.

The structure of this chapter is as follows. In section 3.2, I review the extended CJS model of Bonner and Schwarz (2006) and then describe the method for estimating population size. In section 3.3, I provide justification for this method as a fully Bayesian procedure. Sections 3.4 and 3.5 present an application of the method to study the dynamics of a population of Soay sheep and the results of a subsequent simulation study. The final section provides some comments on the method and discusses future work.

3.2 Methods

3.2.1 Notation

Data

T	Number of capture occasions (indexed by t)
n	Number of individuals captured at least one time (indexed by i)
$\boldsymbol{\omega_i}$	Capture history for captured individual i
$\boldsymbol{\Omega}$	$n \times T$ matrix of capture histories for the captured individuals
$\boldsymbol{z_i}$	Vector of covariate values for individual i (partially observed)
$\boldsymbol{Z}$	$n \times T$ matrix of covariates for the captured individuals

Parameters

Population Size:

M_t Number of marked individuals alive on occasion t

U_t Number of unmarked individuals alive on occasion t

N_t Total number of individuals alive on occasion t ($N_t = M_t + U_t$)

B_t Number of individuals recruited into the population between occasions t and $t+1$ which are alive on occasion $t+1$. B_0 is the number of individuals recruited before the first occasion which are still alive on the first occasion.

S Size of the super-population, the set of individuals ever available for capture during the study ($S = \sum_{t=0}^{T-1} B_t$)

Capture and Survival Probabilities:

p_{it} Probability that individual i is captured on occasion t given that it is alive

β_{0t}, β_1 Coefficients of the linear predictor of $\log(p_{it}/(1-p_{it}))$ as a function of z_{it}

ϕ_{it} Probability that individual i survives to occasion $t+1$ given that it was alive on occasion t

γ_{0t}, γ_1 Coefficients of the linear predictor of the $\log(\phi_{it}/(1-\phi_{it}))$ as a function of z_{it}

χ_{it} Probability that individual i is not captured after occasion t

Distribution of Covariates:

z_∞ Asymptotic mean value of the covariate

r Growth rate parameter

σ Standard deviation of $z_{it}|z_{i,t-1}$

3.2.2 Cormack-Jolly-Seber Model with a Continuous Covariate

The first step in estimating the population size is to determine the effect of the covariate on both the capture and survival probabilities. This is done by restricting the model to consider only the events after each individual is first captured, marked, and released. The restricted model is commonly referred to as the Cormack-Jolly-Seber model. Suppose that a total of n individuals are each captured one or more times over T capture occasions. When an individual is first captured, it is marked with a unique mark so that it can be identified if captured again. Every time an individual is captured the value of the continuous covariate

thought to affect the capture or survival probability is recorded without error. The data from this experiment are summarized by a pair of vectors for each individual. The first vector is the capture history, denoted $\boldsymbol{\omega}_i$ for individual i, a string of T indicator variables such that $\omega_{it} = 1$ if individual i was captured on occasion t and 0 otherwise. Two useful summary statistics are $a_i = \min\{t : \omega_{it} = 1\}$ and $b_i = \max\{t : \omega_{it} = 1\}$, the first and last occasions that individual i is captured. The second vector, $\boldsymbol{z}_i$, records the values of the covariate for individual i. For simplicity, I consider the case when z_{it} is a scalar, though the method should also apply when z_{it} is itself a vector quantity. It is assumed that z_{it} is known if individual i is captured on occasion t and is missing otherwise. The $n \times T$ data matrices with (i, t) entries ω_{it} and z_{it} summarizing all data from the experiment will be denoted by Ω and $\boldsymbol{Z}$.

The assumptions of the extended model are almost identical to those of the original CJS model. It is assumed that the sampling periods are instantaneous (so that the population does not change during a capture occasion), that emigration is permanent, that events in one individual's capture history are independent of the events in the capture histories of all other individuals, that tags are not lost between capture occasions, and that marks are not missed or misread when individuals are captured. The only difference is that the assumption of homogeneous capture and survival probabilities is relaxed so that these probabilities vary among the individuals alive on each occasion as functions of the covariate. In particular, it is assumed that the probability that individual i is captured on occasion t, p_{it}, and the probability that it survives to the next occasion, ϕ_{it}, can be modelled as:

$$\begin{aligned} \text{logit}(p_{it}) &= \beta_{0t} + \beta_1 z_{it} \\ \text{logit}(\phi_{it}) &= \gamma_{0t} + \gamma_1 z_{it} \end{aligned}$$

such that the effect of the covariate is linear on the logistic scale with time dependent intercept and constant slope. Models with different structures, e.g. constant intercept or quadratic effects, could be fit and tested against each other, but this is not my focus. Losses on capture have been ignored but could be incorporated by modelling the probability that a captured individual is not returned to the population exactly as in Seber (1965) and Jolly (1965).

One way to make inference about the capture and survival probabilities in spite of the missing covariate data is to model the distribution of the unknown values of the covariate.

The model developed in Bonner and Schwarz (2006) assumes that changes in the value of the covariate between two capture occasions are identically distributed for all individuals. Specifically, I defined the diffusion based model:

$$z_{i,t+1}|z_{it} \sim N(z_{it} + \mu_t, \sigma^2) \tag{3.1}$$

allowing the mean, μ_t, to vary over time but assuming a constant variance, σ^2. In Bonner and Schwarz (2006) the model was applied to study the effect of body mass on the survival rates of meadow voles (*Microtus pennsylvanicus*) and the population was restricted to adult voles only. In this case, it was reasonable to assume that all individuals had reached their adult body mass and that changes over time were the effect of external pressures affecting all individuals in the same way. This justified modelling the change in a vole's mass independently of its current mass. In the application to Soay sheep that I will consider, individuals of all ages are included in the population of interest and so it is necessary to adapt the model so that younger, smaller sheep have, on average, larger changes in body mass than older, bigger sheep.

The new model I use is based on the Ludwig van Bertalanffy (LVB) model of growth. This model derives from the assumption that the rate of change of a quantity, in continuous time, is proportional to the difference between the current value and an asymptotic value, z_∞, described mathematically by the differential equation:

$$\frac{dz}{dt} = r(z_\infty - z).$$

The parameter r is the constant of proportionality which determines the exact rate of change. The analytic solution to this equation is:

$$z_t = z_\infty(1 - e^{-r(t-t_0)}) \tag{3.2}$$

where t_0 is the constant of integration needed to produce a unique solution. For $r > 0$ and $z_\infty > 0$ the differential equation produces a growth-like trajectory which starts at $z_0 = z_\infty(1-e^{rt_0})$ at time 0 and increases asymptotically toward z_∞ at a constantly decreasing rate.

To make use of this function in the extended CJS model, equation (3.2) needs to be adapted to discrete time. Moreover, the age of individuals cannot usually be known exactly

and so it is not possible to model the covariate as a function of absolute time since birth. Instead, the model will be defined in terms of the changes in the covariate between capture occasions conditional on the value when the individual is first captured. Suppose that the value of the covariate is known for individual i on occasion t. From equation (3.2), the ratio between z_{it} and the covariate's asymptotic value is:

$$\frac{z_{it}}{z_\infty} = (1 - e^{-r(t-t_0)})$$

which implies that:

$$e^{-r(t-t_0)} = 1 - \frac{z_{it}}{z_\infty}.$$

Substituting into the expression for $z_{i,t+1}$ yields:

$$z_{i,t+1} = z_\infty \left(1 - e^{-r(t+1-t_0)}\right) = z_\infty \left(1 - e^{-r}\left(1 - \frac{z_{it}}{z_\infty}\right)\right) = z_\infty(1 - e^{-r}) + z_{it}e^{-r}.$$

Quinn and Deriso (1999, pg. 156) suggest taking this to be the expected value of the covariate on occasion $t+1$ conditional on the value at occasion t and further assuming additive errors that are normally distributed with mean 0 and constant variance σ^2. The new model is:

$$z_{i,t+1}|z_{it} \sim N(z_\infty(1 - e^{-r}) + z_{it}e^{-r}, \sigma^2) \tag{3.3}$$

and the density of this distribution will be denoted by $f(z_{i,t+1}|z_{it})$. The joint density of the covariate values for individual i conditional on the value observed when it is first captured, z_{i,a_i}, is simply the product $f(\boldsymbol{z}_i|z_{i,a_i}) = \prod_{t=a_i+1}^{T} f(z_{it}|z_{i,t-1})$. Note that conditioning on the value of the covariate the first capture has entirely removed the parameter t_0 so that $f(\cdot|\cdot)$ depends obn only two parameters, z_∞ and r. The time since birth, t, is also removed from the model. Non-constant time between capture occasions can be handled easily by multiplying both the rate of growth, r, and the error variance, σ^2, in equation (3.3) by Δ_t, the time between capture occasions t and $t+1$.

The likelihood function for the CJS model including a continuous covariate is most easily formed by starting with the complete data likelihood as if there were no missing data. Suppose that the value of the covariate could in fact be known on every occasion after an individual was first captured and not only the occasions when it actually was captured. The

complete data likelihood contribution for individual i would then be:

$$L_i(\boldsymbol{\beta}_0, \beta_1, \boldsymbol{\gamma}_0, \gamma_1, z_\infty, r|\boldsymbol{\omega}_i, \boldsymbol{z}_i) = \prod_{t=a_i+1}^{b_i} \left(\phi_{i,t-1} p_{it}^{\omega_{it}} (1-p_{it})^{(1-\omega_{it})}\right) \chi_{ib_i} \cdot f(\boldsymbol{z}_i|z_{i,a_i}) \quad (3.4)$$

where χ_{ib_i} is the probability that individual i is not observed after occasion b_i. This quantity is defined by the recursive relationship $\chi_{it} = (1-\phi_{it}) + \phi_{it}(1-p_{it})\chi_{i,t+1}$ with $\chi_{iT} = 1$. Adopting the convention that empty products are equal to 1, $L_i(\boldsymbol{\beta}_0, \beta_1, \boldsymbol{\gamma}_0, \gamma_1, z_\infty, r|\boldsymbol{\omega}_i, \boldsymbol{z}_i) = \chi_{ib_i}$ if $a_i = b_i$ and $L_i(\boldsymbol{\beta}_0, \beta_1, \boldsymbol{\gamma}_0, \gamma_1, z_\infty|\boldsymbol{\omega}_i, \boldsymbol{z}_i) = 1$ if $a_i = T$. As in the CJS model, the likelihood contribution conditions on the individual's first release so that only the events in the capture history after a_i are modelled, and not the initial capture on occasion a_i. The complete data likelihood is formed by multiplying these contributions for each individual. The observed data likelihood function could then be computed by integrating with respect to each of the missing covariate values; however, these integrals do not have analytic solutions and would be difficult to approximate numerically each time the likelihood function needed to be evaluated. Instead, the integration can be performed implicitly by forming the posterior distribution from the complete data likelihood and then treating the missing covariate values in the same way as the other unknowns in the model, simulating new values on each iteration of the MCMC sampling algorithm.

The final step in constructing the posterior distribution is to define a prior for the model parameters. For the vectors of time-dependent intercept terms, $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$, I adopt a hierarchical approach modelling the elements of each vector as random draws from a common normal distribution on the logistic scale and then assigning vague, conjugate hyper-priors. The intention is to encode the belief that the intercept terms on different occasions should be similar without making strong assumptions about their exact values. The full distribution for $\boldsymbol{\beta}_0$ is defined by:

$$\begin{aligned} \beta_{0t} &\sim N(\mu_\beta, \tau_\beta^2), \quad t = 1, \ldots, T \\ \mu_\beta &\sim N(0, 1000) \\ 1/\tau_\beta^2 &\sim \text{Gamma}(.001, .001) \end{aligned}$$

and exactly similar for $\boldsymbol{\gamma}_{0t}$. The remaining parameters have been assigned vague priors such

that:

$$\begin{gathered}\beta_1 \sim N(0, 1000) \\ \gamma_1 \sim N(0, 1000) \\ z_\infty \sim U(0, 100) \\ r \sim U(0, 10) \\ \text{and } 1/\sigma^2 \sim \text{Gamma}(.001, .001).\end{gathered}$$

All values are assumed independent a priori, except for the dependence introduced by the hierarchical structures on $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$.

The posterior distribution for the full set of parameters in the extended CJS model, $\boldsymbol{\theta} = \{\boldsymbol{\beta}_0, \beta_1, \boldsymbol{\gamma}_0, \gamma_1, z_\infty\}$, is then defined by the product of the joint prior density and the complete data likelihood above. Inference about these parameters is obtained by generating D random realizations from the posterior, $\tilde{\boldsymbol{\theta}}^1, \ldots, \tilde{\boldsymbol{\theta}}^D$, via MCMC sampling. Simulation was performed in OpenBUGS (Thomas et al., 2006).

3.2.3 Estimating Population Size

The second step in my method is to simulate values for the population size on each capture occasion for each of the sets of parameters sampled above. This is done by separately generating values for the number of marked and unmarked individuals alive on occasion t for each $\tilde{\boldsymbol{\theta}}^d$ and summing the two to obtain a value for the total population size, $\tilde{N}_t^d$. I then derive inference about the population size by treating the resulting values, $\tilde{N}_t^1, \ldots, \tilde{N}_t^D$, as a sample of size D from the posterior distribution of N_t. The posterior mean of N_t is approximated by $\bar{N}_t = \frac{1}{D}\sum_{d=1}^{D} \tilde{N}_t^d$, the variance by $\sum_{d=1}^{D}(\tilde{N}_t^d - \bar{N}_t)^2/(D-1)$ and quantiles by the order statistics of $\tilde{N}_t^d$.

To simulate the number of marked individuals alive on occasion t, M_t, I define the vector of indicator variables $\boldsymbol{S}_t = (s_{1t}, \ldots, s_{nt})$ such that $s_{it} = 1$ if individual i is alive on occasion t and was previously marked and 0 otherwise. By definition, $s_{it} = 0$ for $t \leq a_i$ and $s_{it} = 1$ for $a_i < t \leq b_i$. The remaining values of s_{it} can then generated sequentially as conditional Bernoulli random variables with:

$$P(s_{it} = 1 | s_{i,t-1}) = \begin{cases} \frac{\phi_{i,t-1}(1-p_{it})\chi_{it}}{\chi_{i,t-1}} & s_{i,t-1} = 1 \\ 0 & s_{i,t-1} = 0 \end{cases}$$

where p_{it}, ϕ_{it} and χ_{it} are computed given $\boldsymbol{\theta}_t^d$ and the completed matrix of covariates for the

d^{th} iteration of the Markov chain. This produces a complete vector of indicator variables, S_t^d, from which I compute $\tilde{M}_t^d = \sum_{i=1}^n s_{it}^d$ as an estimate of M_t.

An estimate for the number of unmarked individuals alive on occasion t, U_t, is generated through a modification of the HT estimator. Let $A_t = \{i : a_i = t\}$ denote the set of individuals first captured on occasion t. Given $\boldsymbol{\theta}_t^d$ it is possible to compute the capture probability on occasion t for each individual in A_t, denoted $\hat{p}_{it}^d$. The HT estimate of U_t would then be:

$$\hat{U}_t^d = \sum_{i \in A_t} \frac{1}{\hat{p}_{it}^d}.$$

However, as shown in the following section, direct use of $\hat{U}_t^d$ would underestimate the uncertainty in U_t. Instead, I replace $1/\hat{p}_{it}^d$ with a randomly generated value:

$$\tilde{U}_{it} \sim \text{Neg. Bin.}(1, \hat{p}_{it}^d)$$

for each $i \in A_t$ and define:

$$\tilde{U}_t^d = \sum_{i \in A_t} \tilde{U}_{it}^d.$$

Applying the procedure for each realization of $\boldsymbol{\theta}$ produces the set of values $\tilde{U}_t^1, \ldots, \tilde{U}_t^D$ for each t. The value $\tilde{N}_t^d$ is then defined as $\tilde{N}_t^d = \tilde{M}_t^d + \tilde{U}_t^d$. Note that $E(\tilde{U}_{it}^d | \tilde{\boldsymbol{\theta}}^d) = 1/\hat{p}_{it}^d$ so that $E(\tilde{U}_t^d | \tilde{\boldsymbol{\theta}}^d) = \hat{U}_t^d$, but the sample variance of $\tilde{U}^1, \ldots, \tilde{U}^D$ is greater than the sample variance of $\hat{U}^1, \ldots, \hat{U}^D$.

Given these simulated values it is also possible to construct an estimate of the number of individuals recruited on each occasion from each iteration of the Markov chain. The method is similar to that used by Seber (1965) and Jolly (1965). First, I estimate the number of marked and unmarked individuals alive on occasion t that were also alive on occasion $t-1$ and then I subtract these values from $\tilde{N}_t^d$ to estimate the number of new individuals in the population. For the d^{th} iteration of the Markov chain, an estimate of the number of individuals alive on both occasions and marked on occasion t is $M_t^{*d} = \sum_{i=1}^n 1(t-1 \geq a_i) \cdot s_{it}^d$. An estimate of the number of individuals alive on both occasions and unmarked on occasion t is $U_t^{*d} = \sum_{i \in A_{t-1}} U_{it}^{*d}$ where $U_{it}^{*d} \sim \text{Binomial}(\tilde{U}_{it-1} - 1, \phi_{i,t-1})$ models the number of individuals similar to individual i that were alive but not captured on occasion $t-1$ and that survived to occasion t. It is necessary to subtract one individual from the number of trials in the binomial because individual i itself has already been included as a marked individual

in M^{*d}_t. My estimate of the number of recruits on occasion t is then $\tilde{B}_{t-1} = \tilde{N}_t - M^{*d}_t - U^{*d}_t$ for $t = 2, \ldots, T$. Note that it is possible that $\tilde{B}_{t-1} < 0$, yielding a negative estimate of recruitment. On occasion 1, the estimated number of recruits is simply $\tilde{B}_0 = \tilde{N}^d_t$. The super-population size, the number alive and available on at least one capture occasion, can then be estimated by $\tilde{S}^d = \sum_{t=0}^{T-1} B_t$. Simulation from the posterior distribution and computation of the population size on each occasion, the number of recruits, and the super-population size were all performed in OpenBUGS (Thomas et al., 2006). BUGS code is included in Appendix B.

3.2.4 Goodness of Fit

To assess the fit of the LVB growth model, I propose a Bayesian p-value comparing the observed and expected covariate values. The exact discrepancy measure I have selected is the mean of standardized errors:

$$D(\boldsymbol{Z}, \boldsymbol{\theta}) = \frac{1}{n^{recap}} \sum_{i=1}^{n} \sum_{\{t:t>a_i, \omega_{it}=1\}} \left(\frac{z_{it} - E(z_{it}|\boldsymbol{\theta}, z_{it_i^-})}{\sigma\sqrt{t - t_i^-}} \right)^2 \tag{3.5}$$

where $n^{recap} = \sum_{i=1}^{n} \sum_{t=a_i+1}^{T} \omega_{it}$ denotes the total number of recaptures over all individuals and $t_i^- = \max\{s = 1, \ldots, t : \omega_{is} = 1\}$ the last occasion prior to t that individual i was captured. Because the LVB growth model specifies the conditional behaviour of the covariate, expected values can only be computed for the occasions after the each individual is first captured, limiting the sum to the occasions when each individual was recaptured. Division by $\sqrt{t - t_i^-}$ is required to standardize the residuals whose variance is proportional to the time since last capture. Division by σ is not a necessity, but converts the discrepancy measure into a scaleless quantity. As discussed in section 1.3.2, the Bayesian p-value is computed by simulating new data, $(\boldsymbol{\Omega}'_d, \boldsymbol{Z}'_d)$, for each of the D sets of parameters sampled from the posterior and computing the discrepancies for both the observed and simulated data, $D(Z, \tilde{\boldsymbol{\theta}}^d)$ and $D(\boldsymbol{Z}'_d, \tilde{\boldsymbol{\theta}}^d)$. The Bayesian p-value is the proportion of the K sets of parameters for which $D(Z, \tilde{\boldsymbol{\theta}}^d)$ is less than $D(\boldsymbol{Z}'_d, \tilde{\boldsymbol{\theta}}^d)$.

3.3 Justification

To justify this method of making inference about the population size, I will demonstrate that the two step procedure actually produces a valid sample from the joint posterior distribution of N_t and $\boldsymbol{\theta}$ for a specific choice of prior on $\boldsymbol{U} = (U_1, \ldots, U_t)$. First note that the joint posterior density of the entire set of parameters, $\boldsymbol{U} \cup \boldsymbol{\theta}$, can be decomposed as the product of the conditional posterior distribution of $\boldsymbol{U}$ given the data and $\boldsymbol{\theta}$ and the marginal posterior distribution of $\boldsymbol{\theta}$:

$$\pi(\boldsymbol{U}, \boldsymbol{\theta} | \boldsymbol{\Omega}, \boldsymbol{Z}) = \pi(\boldsymbol{U} | \boldsymbol{\theta}, \boldsymbol{\Omega}, \boldsymbol{Z}) \cdot \pi(\boldsymbol{\theta} | \boldsymbol{\Omega}, \boldsymbol{Z}).$$

The implication of this result is that samples can be obtained from the joint posterior distribution in two steps, first simulating from $\boldsymbol{\theta} | \boldsymbol{\Omega}, \boldsymbol{Z}$ and then drawing values from $\boldsymbol{U} | \boldsymbol{\theta}, \boldsymbol{\Omega}, \boldsymbol{Z}$. These steps are equivalent to the two steps in my method. Two things remain to be shown:

1) that $\pi(\boldsymbol{\theta} | \boldsymbol{\Omega}, \boldsymbol{Z})$ depends only on the events in each individual's capture history after its first release

2) that $\tilde{\boldsymbol{U}}^d = (\tilde{U}_1^d, \ldots, \tilde{U}_T^d)$ can be considered a random realization from $\pi(\boldsymbol{U} | \tilde{\boldsymbol{\theta}}^d, \boldsymbol{\Omega}, \boldsymbol{Z})$.

I will begin by proving these claims for the case where the capture and survival probabilities for each individual are restricted to a finite set of K values. The results will then be extended to the case where the capture and survival probabilities may take any value in a continuous range.

Suppose that the individuals alive on each occasion can be assigned to K different groups and that all individuals in group k on occasion t have the same probabilities of being captured and of surviving to the next occasion. The complication of unequal numbers of groups on different occasions can be removed by allowing empty groups so that K can be assumed the same for all t. Let p_{kt} and ϕ_{kt} denote the probabilities of capture and survival for group k on occasion t, U_{kt} the number of unmarked individuals alive in group k, and u_{kt} the number of these individuals that are captured. As in the continuous case, $\boldsymbol{\omega}_i$ will denote the capture history for individual i but now $\boldsymbol{z}_i$ will denote the vector of group memberships such that $z_{it} = k$ if individual i belongs to group k on occasion t. The full data are still defined by the two matrices $\boldsymbol{\Omega}$ and $\boldsymbol{Z}$. Note that u_{kt} is simply a summary statistic which can be formally defined as $u_{kt} = \sum_{i \in A_t} 1(z_{it} = k)$.

Following the development of Seber (1965) for the JS model (but ignoring losses on capture) the full likelihood can be formed as the product of two components, the first modelling the capture of unmarked individuals on each occasion and the second modelling the subsequent events for these individuals. Taking U_{kt} to be fixed and assuming independence among individuals, the number of unmarked individuals captured on occasion t will have a binomial distribution with parameters U_{kt} and p_{kt}. The resulting likelihood contribution is:

$$L_{1kt} = \binom{U_{kt}}{u_{kt}} p_{kt}^{u_{kt}} (1 - p_{kt})^{U_{kt} - u_{kt}}$$

and the first component of the likelihood is the product of these terms over all groups and occasions, $L_1 = \prod_{t=1}^{T} \prod_{k=1}^{K} L_{1kt}$. The exact form of the second component of the likelihood is not needed in the following theory and so it will simply be denoted by L_2. If each individual's group membership is fixed in time then L_2 is simply the product of standard CJS likelihoods defined separately for each group; if individuals are allowed to move among the groups then L_2 would represent the likelihood of a multi-state CJS model (Arnason, 1973; Schwarz et al., 1993; Brownie et al., 1993); and in the problem with a continuous covariate L_2 is exactly the likelihood defined in equation (3.4).

The only prior distribution that needs to be specified for the following theory is the prior on the number of unmarked individuals alive on each occasion, $\pi(\boldsymbol{U})$. The standard non-informative prior distribution for the number of trials in a binomial distribution is the improper Jeffrey's prior $\pi(N) \propto 1/N$ (King and Brooks, 2001). This suggests that the prior for $\boldsymbol{U}$ be specified by assuming that $\pi(U_{kt}) \propto 1/U_{kt}$ independently for each k and t. The difficulty that arises with this choice is that the prior is not defined when $U_{kt} = 0$ and the resulting posterior distribution is not determined if $u_{kt} = 0$ for any k and t. My solution to this problem is to approximate the standard prior by the, still improper, distribution:

$$\pi(U_{kt}) \propto \begin{cases} 1/\pi_0 & U_{kt} = 0 \\ 1/U_{kt} & U_{kt} > 0 \end{cases} \tag{3.6}$$

for some $\pi_0 << 1$.

One final assumption required in the following theory is that the probability of capture is bounded below by some known value greater than 0. That is, $\exists p_{\min} > 0$ such that $p_{kt} > p_{\min}$ $\forall k = 1, \ldots, K$, $\forall t = 1, \ldots, T$.

My first claim is that L_1, the component of the likelihood modelling when individuals are first captured, contributes no information about the capture probabilities, the survival probabilities or any parameters modelling the distribution of the individuals among the K groups. Intuitively, this seems fairly clear. It is not possible to know if an individual captured for the first time on occasion t was present on previous occasions, so that its capture provides no information about survival before occasion t. Further, the parameters U_{kt} and p_{kt} are completely confounded when considering only L_1. Capturing only a few individuals from group k on occasion t may indicate either that there were few individuals in group k or that the capture probability was small.

Claim 1. *The marginal posterior distribution of $\boldsymbol{\theta}$ does not depend on L_1 in that $\forall \epsilon_1 > 0$ $\exists \pi_0$ dependent only on $p_{\min}$, K, and T such that:*

$$\frac{L_2\pi(\boldsymbol{\theta})}{\int L_2\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \cdot (1-\epsilon_1) < \pi(\boldsymbol{\theta}|\boldsymbol{\Omega}, \boldsymbol{Z}) < \frac{L_2\pi(\boldsymbol{\theta})}{\int L_2\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \cdot (1+\epsilon_1)\,.$$

Lemma 1.1. *If $u_{kt} > 0$ then the marginal posterior distribution of $\boldsymbol{\theta}$ is independent of L_{1kt}.*

Proof. The full posterior distribution for $(\boldsymbol{U}, \boldsymbol{\theta})$ is defined by the proportionality:

$$\pi(\boldsymbol{U}, \boldsymbol{\theta}|\boldsymbol{\Omega}, \boldsymbol{Z}) \propto L_1 \cdot L_2 \cdot \prod_{k=1}^{K} \pi(U_k)\pi(\boldsymbol{\theta}) = L_2\pi(\boldsymbol{\theta}) \cdot \prod_{t=1}^{T}\prod_{k=1}^{K} L_{1kt}\pi(U_{kt}). \tag{3.7}$$

To compute the marginal posterior distribution of $\boldsymbol{\theta}$ one needs to sum this expression over all possible values of $\boldsymbol{U}$ and then normalize to obtain a proper distribution. That is:

$$\begin{aligned} \pi(\boldsymbol{\theta}|\boldsymbol{\Omega}, \boldsymbol{Z}) &\propto L_2\pi(\boldsymbol{\theta}) \cdot \textstyle\sum_{U_{11}=u_{11}}^{\infty} \cdots \sum_{U_{KT}=u_{KT}}^{\infty} \left[\prod_{t=1}^{T}\prod_{k=1}^{K} L_{1kt}\pi(U_{kt})\right] \\ &= L_2\pi(\boldsymbol{\theta}) \cdot \textstyle\prod_{t=1}^{T}\prod_{k=1}^{K} \left[\sum_{U_{kt}=u_{kt}}^{\infty} L_{1kt}\pi(U_{kt})\right]. \end{aligned}$$

Then:

$$\begin{aligned} \textstyle\sum_{U_{kt}=u_{kt}}^{\infty} L_{1kt}\pi(U_{kt}) &= \textstyle\sum_{U_{kt}=u_{kt}}^{\infty} \binom{U_{kt}}{u_{kt}} p_{kt}^{u_{kt}}(1-p_{kt})^{U_{kt}-u_{kt}} \cdot \frac{1}{U_{kt}} \\ &= \textstyle\sum_{U_{kt}=u_{kt}}^{\infty} \frac{1}{u_{kt}}\binom{U_{kt}-1}{u_{kt}-1} p_{kt}^{u_{kt}}(1-p_{kt})^{U_{kt}-u_{kt}} \\ &= \frac{1}{u_{kt}} \end{aligned}$$

by noting that $\binom{U_{kt}-1}{u_{kt}-1} p_{kt}^{u_{kt}}(1-p_{kt})^{U_{kt}-u_{kt}}$ is the probability mass function (pmf) of a negative binomial random variable with parameters u_{kt} and p_{kt} and the summation is over the entire

range of this variable.

Finally, u_{kt} belongs to the observed data which implies that:

$$\pi(\boldsymbol{\theta}|\boldsymbol{\Omega}, \boldsymbol{Z}) \quad \propto \quad L_2\pi(\boldsymbol{\theta}) \cdot \prod_{\{(l,s):l=1,\ldots,K;s=1\ldots,T\}\setminus(k,t)} \left[\sum_{U_{ls}=u_{ls}}^{\infty} L_{1ls}\pi(U_{ls})\right]$$

which is independent of L_{1kt}. □

Applying this result for all k and t such that $u_{kt} > 0$, the marginal posterior distribution for $\boldsymbol{\theta}$ can be defined by the proportionality:

$$\pi(\boldsymbol{\theta}|\boldsymbol{\Omega}, \boldsymbol{Z}) \quad \propto \quad L_2\pi(\boldsymbol{\theta}) \cdot \prod_{\{k,t:u_{kt}=0\}} \left[\sum_{U_{kt}=0}^{\infty} L_{1kt}\pi(U_{kt})\right]. \tag{3.8}$$

The same strategy cannot be applied for k and t such that $u_{kt} = 0$ because the cancellation resulting in the negative binomial pmf does not occur when $u_{kt} = 0$, and hence $\binom{U_{kt}}{u_{kt}} = 1$ for all values of U_{kt}. Instead, I show that the effect of L_{1kt} is negligible for small enough π_0.

Lemma 1.2. *The sequence:*

$$\sum_{U=1}^{\infty} \frac{(1-p)^U}{U}$$

converges for any $p \in (0, 1)$.

Proof. The ratio of the $U + 1^{st}$ and U^{th} terms of the sequence is:

$$\frac{\left(\frac{(1-p)^{U+1}}{U+1}\right)}{\left(\frac{(1-p)^U}{U}\right)} = (1-p)\frac{U}{U+1}$$

which tends to $(1 - p) < 1$ as $U \to \infty$ for any $p \in (0, 1)$. Hence, by the ratio test, the sequence is absolutely convergent. Since all terms are positive, it is convergent. □

Let $c_1 = \sum_{U=1}^{\infty} \frac{(1-p_{\min})^U}{U}$.

Proof of Claim 1. Suppose that $u_{kt} = 0$. Then:

$$\begin{aligned} \sum_{U_{kt}=0}^{\infty} L_{1k}\pi(U_{kt}) &= \tfrac{1}{\pi_0}\tbinom{0}{0}p_{kt}^0(1-p_{kt})^0 + \sum_{U=1}^{\infty}\tbinom{U}{0}p_{kt}^0(1-p_{kt})^U\tfrac{1}{U} \\ &= \tfrac{1}{\pi_0} + \sum_{U=1}^{\infty}\tfrac{(1-p_{kt})^U}{U}. \end{aligned}$$

Substituting this expression into equation (3.8), the normalized marginal posterior distribution of $\boldsymbol{\theta}$ is given by:

$$\pi(\boldsymbol{\theta}|\boldsymbol{\Omega},\boldsymbol{Z}) = \frac{L_2\pi(\boldsymbol{\theta})\cdot\prod_{\{k,t:u_{kt}=0\}}\left(\frac{1}{\pi_0}+\sum_{U=1}^{\infty}\frac{(1-p_{kt})^U}{U}\right)}{\int L_2\pi(\boldsymbol{\theta})\cdot\prod_{\{k,t:u_{kt}=0\}}\left(\frac{1}{\pi_0}+\sum_{U=1}^{\infty}\frac{(1-p_{kt})^U}{U}\right)d\boldsymbol{\theta}}.$$

Noting that:

$$\frac{1}{\pi_0} < \left(\frac{1}{\pi_0}+\sum_{U=1}^{\infty}\frac{(1-p_{kt})^U}{U}\right) < \frac{1}{\pi_0}+c_1$$

implies:

i)

$$\begin{aligned}\pi(\boldsymbol{\theta}|\boldsymbol{\Omega},\boldsymbol{Z}) &> \frac{L_2\pi(\boldsymbol{\theta})\cdot\left(\frac{1}{\pi_0}\right)^{K_{0\cdot}}}{\int L_2\pi(\boldsymbol{\theta})\cdot\left(\frac{1}{\pi_0}+c_1\right)^{K_{0\cdot}}d\boldsymbol{\theta}}\\ &= \frac{L_2\pi(\boldsymbol{\theta})}{\int L_2\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}}\cdot(1+\pi_0c_1)^{-K_{0\cdot}}\\ &> \frac{L_2\pi(\boldsymbol{\theta})}{\int L_2\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}}\cdot(1+\pi_0c_1)^{-KT}\end{aligned}$$

ii)

$$\begin{aligned}\pi(\boldsymbol{\theta}|\boldsymbol{\Omega},\boldsymbol{Z}) &< \frac{L_2\pi(\boldsymbol{\theta})\cdot\left(\frac{1}{\pi_0}+c_1\right)^{K_{0\cdot}}}{\int L_2\pi(\boldsymbol{\theta})\cdot\left(\frac{1}{\pi_0}\right)^{K_0}d\boldsymbol{\theta}}\\ &= \frac{L_2\pi(\boldsymbol{\theta})}{\int L_2\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}}\cdot(1+\pi_0c_1)^{K_{0\cdot}}\\ &< \frac{L_2\pi(\boldsymbol{\theta})}{\int L_2\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}}\cdot(1+\pi_0c_1)^{KT}\end{aligned}$$

where $K_{0t}=\sum_{k=1}^{K}1(u_{kt}=0)<K$ is the number of groups in which no unmarked individuals were captured on occasion t and $K_{0\cdot}=\sum_{t=1}^{T}K_{0t}<KT$. Choosing $\pi_0<\frac{(1-\epsilon_1)^{-1/(KT)}-1}{c_1}$ implies $(1+\pi_0c_1)^{-KT}>(1-\epsilon_1)$ and hence that:

$$\pi(\boldsymbol{\theta}|\boldsymbol{\Omega},\boldsymbol{Z}) > \frac{L_2\pi(\boldsymbol{\theta})}{\int L_2\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}}\cdot(1-\epsilon_1)\,.$$

Choosing $\pi_0<\frac{(1+\epsilon_1)^{1/(KT)}-1}{c_1}$ implies $(1+\pi_0c_1)^{KT}<(1+\epsilon_1)$ and hence that:

$$\pi(\boldsymbol{\theta}|\boldsymbol{\Omega},\boldsymbol{Z}) < \frac{L_2\pi(\boldsymbol{\theta})}{\int L_2\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}}\cdot(1+\epsilon_1)\,.$$

It follows that

$$\frac{L_2\pi(\boldsymbol{\theta})}{\int L_2\pi(\boldsymbol{\theta})\,d\,\boldsymbol{\theta}}\cdot(1-\epsilon_1) < \pi(\boldsymbol{\theta}|\boldsymbol{\Omega},\boldsymbol{Z}) < \frac{L_2\pi(\boldsymbol{\theta})}{\int L_2\pi(\boldsymbol{\theta})\,d\,\boldsymbol{\theta}}\cdot(1+\epsilon_1)\,.$$

for any:

$$\pi_0 < \min\left(\frac{(1-\epsilon_1)^{-1/(KT)}-1}{c_1}, \frac{(1+\epsilon_1)^{1/(KT)}-1}{c_1}\right).$$

□

The result of Claim 1 is that for π_0 small enough, the marginal posterior of $\boldsymbol{\theta}$ depends only on the second component of the likelihood, L_2, and the prior distribution for this set of parameters. Inference for the capture probabilities, the survival probabilities and the parameters modelling the distribution of individuals among the K groups depends only on the data from the individuals after each has been first captured, marked and released. How many unmarked individuals were captured in each group on each occasion does not contribute any information about these parameters.

My second claim is that inference concerning the number of unmarked individuals alive on occasion t depends only on the groups in which $u_{kt} > 0$. Note that U_{kt} appears only in L_{1kt} which depends on the data only through the summary statistic u_{kt} and on the parameter p_{kt}. This implies that $\boldsymbol{U}$ depends only on $\boldsymbol{p} = \{p_{kt}; k = 1,\ldots,K, t = 1,\ldots,T\}$ and $\boldsymbol{u} = \{u_{kt}; k = 1,\ldots,K, t = 1,\ldots,T\}$ so that:

$$\pi(\boldsymbol{U}|\boldsymbol{\theta},\boldsymbol{\Omega},\boldsymbol{Z}) = \pi(\boldsymbol{U}|\boldsymbol{p},\boldsymbol{u}).$$

Claim 2. *For any $\epsilon_2 > 0$ and $\epsilon_3 > 0$ $\exists$ $\pi_0 > 0$ dependent only on $p_{\min}$ and K such that:*

1)

$$\sum_{\{k:u_{kt}>0\}}\frac{u_{kt}}{p_{kt}} < E(U_t|\boldsymbol{p},\boldsymbol{u}) < \sum_{\{k:u_{kt}>0\}}\frac{u_{kt}}{p_{kt}} + \epsilon_2$$

2)

$$\sum_{\{k:u_{kt}>0\}}\frac{u_{kt}(1-p_{kt})}{p_{kt}^2} < Var(U_t|\boldsymbol{p},\boldsymbol{u}) < \sum_{\{k:u_{kt}>0\}}\frac{u_{kt}(1-p_{kt})}{p_{kt}^2} + \epsilon_3.$$

for all $t = 1,\ldots,T$ simultaneously.

Lemma 2.1. *If $u_{kt} > 0$ then $U_{kt}|u_{kt},p_{kt} \sim$ Neg. Bin.(u_{kt},p_{kt}).*

Proof. Note that U_{kt} must be greater than u_{kt} so that the posterior probability for any $U_{kt} < u_{kt}$ is zero. For $U_{kt} \geq u_{kt}$ the posterior probability is defined by:

$$\begin{aligned}
\pi(U_{kt}|u_{kt}, p_{kt}) &\propto L_{1kt} \cdot \pi(U_{kt}) \\
&= \tbinom{U_{kt}}{u_{kt}} p_{kt}^{u_{kt}} (1-p_{kt})^{U_{kt}-u_{kt}} \cdot \tfrac{1}{U_{kt}} \\
&= \tfrac{1}{u_{kt}} \tbinom{U_{kt}-1}{u_{kt}-1} p_{kt}^{u_{kt}} (1-p_{kt})^{U_{kt}-u_{kt}} \\
&\propto \tbinom{U_{kt}-1}{u_{kt}-1} p_{kt}^{u_{kt}} (1-p_{kt})^{U_{kt}-u_{kt}}
\end{aligned}$$

which is the pmf of a negative binomial random variable with parameters u_{kt} and p_{kt}. □

It follows from this lemma that $E(U_{kt}|u_{kt}, p_{kt}) = u_{kt}/p_{kt}$ and $Var(U_{kt}|u_{kt}, p_{kt}) = u_{kt}(1 - p_{kt})/p_{kt}^2$.

Lemma 2.2. *If $u_{kt} = 0$ then $\exists \pi_0 > 0$ dependent only on $p_{\min}$ and K such that $E(U_{kt}|u_{kt} = 0, p_{kt}) < \epsilon_2/K$.*

Proof. By definition:

$$\begin{aligned}
E(U_{kt}|u_{kt}=0, p_{kt}) &= \textstyle\sum_{U_{kt}=0}^{\infty} U_{kt} \pi(U_{kt}|u_{kt}=0, p_{kt}) \\
&= \frac{\sum_{U_{kt}=0}^{\infty} U_{kt} \binom{U_{kt}}{0} (1-p_{kt})^{U_{kt}} \pi(U_{kt})}{\sum_{U_{kt}=0}^{\infty} \binom{U_{kt}}{0} (1-p_{kt})^{U_{kt}} \pi(U_{kt})} \\
&= \frac{0 + \sum_{U_{kt}=1}^{\infty} U_{kt} (1-p_{kt})^{U_{kt}} \cdot U_{kt}^{-1}}{\frac{1}{\pi_0} + \sum_{U_{kt}=1}^{\infty} \frac{(1-p_{kt})^{U_{kt}}}{U_{kt}}} \\
&< \frac{\left(\frac{1}{p_{kt}} - 1\right)}{\frac{1}{\pi_0} + c_1} \\
&< \pi_0 \left(\frac{1}{p_{kt}} - 1\right) \\
&< \pi_0 \left(\frac{1}{p_{\min}} - 1\right).
\end{aligned}$$

Thus, $E(U_{kt}|u_{kt} = 0, p_{kt}) < \epsilon_2/K$ for any $\pi_0 < \frac{\epsilon_2 p_{\min}}{K(1-p_{\min})}$. □

Lemma 2.3. *The sequence:*

$$\sum_{U=1}^{\infty} U(1-p)^U$$

converges for any $p \in (0, 1)$.

Proof. The ratio of the $U + 1^{st}$ and U^{th} terms of the sequence is:

$$\frac{(U+1)(1-p)^{U+1}}{U(1-p)^U} = (1-p)\frac{U+1}{U}$$

which again tends to $(1-p) < 1$ as $U \to \infty$ for any $p \in (0,1)$. Hence, by the ratio test, the sequence is absolutely convergent. Since all terms are positive, it is convergent. □

Let $c_2 = \sum_{U=1}^{\infty} U(1-p_{\min})^U$.

Lemma 2.4. *If $u_{kt} = 0$ then $\exists \pi_0 > 0$ dependent only on $p_{\min}$ and K such that $Var(U_{kt}|p_{kt}, u_{kt}) < \epsilon_3/K$.*

Proof. By definition:

$$\begin{aligned} Var(U_{kt}|u_{kt}=0,p_{kt}) &= \textstyle\sum_{U_{kt}=0}^{\infty} U_{kt}^2 \pi(U_{kt}|u_{kt}=0,p_{kt}) - E(U_{kt}|u_{kt}=0,p_{kt})^2 \\ &< \frac{\sum_{U_{kt}=0}^{\infty} U_{kt}^2 \binom{U_{kt}}{0}(1-p_{kt})^{U_{kt}}\pi(U_{kt})}{\sum_{U_{kt}=0}^{\infty}\binom{U_{kt}}{0}(1-p_{kt})^{U_{kt}}\pi(U_{kt})} \\ &= \frac{0+\sum_{U_{kt}=1}^{\infty} U_{kt}(1-p_{kt})^{U_{kt}}}{\frac{1}{\pi_0}+\sum_{U_{kt}=1}^{\infty}\frac{(1-p_{kt})^{U_{kt}}}{U_{kt}}} \\ &< \frac{\sum_{U=1}^{\infty} U(1-p_{\min})^U}{\frac{1}{\pi_0}} \\ &= \pi_0 c_2. \end{aligned}$$

It follows that $Var(U_{kt}|p_{kt}, u_{kt}) < \epsilon_3/K$ for any $\pi_0 < \frac{\epsilon_3}{c_2 K}$. □

Proof of Claim 2. By independence of $U_{1t}, \ldots, U_{Kt}$ it follows that:

$$E(U_t|\boldsymbol{p},\boldsymbol{u}) > \sum_{\{k:u_{kt}>0\}} \frac{u_{kt}}{p_{kt}}$$

and

$$\mathrm{Var}(U_t|\boldsymbol{p},\boldsymbol{u}) > \sum_{\{k:u_{kt}>0\}} \frac{u_{kt}(1-p_{kt})}{p_{kt}^2}$$

Suppose now that π_0 is chosen so that $\pi_0 < \min\left(\frac{\epsilon_2 p_{\min}}{K(1-p_{\min})}, \frac{\epsilon_3}{c_2 K}\right)$. Combining the results of Lemmas 2.1 and 2.2 yields:

$$\begin{aligned} E(U_t|\boldsymbol{p},\boldsymbol{u}) &= \textstyle\sum_{k=1}^{K} E(U_{kt}|u_{kt},p_{kt}) \\ &< \textstyle\sum_{\{k,t:u_{kt}>0\}} \frac{u_{kt}}{p_{kt}} + \sum_{\{k,t:u_{kt}=0\}} \frac{\epsilon_2}{K} \\ &< \textstyle\sum_{\{k,t:u_{kt}>0\}} \frac{u_{kt}}{p_{kt}} + \epsilon_2. \end{aligned}$$

Combining the results of Lemmas 2.1 and 2.3 yields:

$$\begin{aligned} Var(U_t|\boldsymbol{p},\boldsymbol{u}) &= \sum_{k=1}^{K} Var(U_{kt}|u_{kt},p_{kt}) \\ &< \sum_{\{k:u_{kt}>0\}} \frac{u_{kt}(1-p_{kt})}{p_{kt}^2} + \sum_{\{k:u_{kt}=0\}} \frac{\epsilon_2}{K} \\ &< \sum_{\{k:u_{kt}>0\}} \frac{u_{kt}(1-p_{kt})}{p_{kt}^2} + \epsilon_2. \end{aligned}$$

□

The consequence of this result is that groups for which $u_{kt} = 0$ on occasion t can be ignored in inference concerning $U_t = \sum_{k=1}^{K} U_{kt}$. To compute the posterior expected value or variance of U_t one need only sum the posterior expected values or variances of U_{kt} for k such that $u_{kt} > 0$. Lemma 2.1 further indicates that a sample from the conditional posterior distribution of $U_t|\boldsymbol{\theta},\boldsymbol{\Omega},\boldsymbol{Z}$ can be generated by simulating values $\tilde{U}_{kt} \sim$ Neg. Bin.(u_{kt},p_{kt}) for each k such that $u_{kt} > 0$ and summing these values. Computing the HT estimator for each t, $\hat{U}_t = \sum_{k=1}^{K} 1/p_{kt}$, is equivalent to summing the conditional posterior expected value of U_{kt} given $\boldsymbol{\theta}$ for each $k = 1,\ldots,K$. These values would provide appropriate inference about the posterior mean of U_t, but treating these values as a sample from the posterior distribution of U_t would ignore the uncertainty in U_t given $\boldsymbol{\theta}$ and underestimate the posterior variance, as mentioned in the previous section.

Together, Claims 1 and 2 justify the two step procedure to generate a sample of size D from the full posterior distribution of $(\boldsymbol{U},\boldsymbol{\theta})$ when the capture and survival probabilities vary across K distinct groups in the population. To extend this result to the case where the capture and survival probabilities are functions of a continuous covariate, I simply partition the range of the covariate into a sequence of small, disjoint intervals and then apply the above results. Although a covariate like body mass may be continuous in reality, in practice any variable is measured on a discrete scale. A partition of intervals may then be defined so that each interval covers at most one observable value of the covariate. Individuals with the same value of the covariate are considered as a single group, and the capture and survival probabilities for this group are computed at the observable value. If the covariate is measured on a fine enough scale then it is highly improbable that any two individuals will ever share the same value, as we would expect for a continuous covariate. In this case, each

group will contain either 0 or 1 individuals and the posterior mean of U_t given $\boldsymbol{\theta}$ will be:

$$E(U_t|\boldsymbol{\theta}, \boldsymbol{\Omega}, \boldsymbol{Z}) = \sum_{i=1}^{N} \frac{1(\omega_{it} = 1)}{p_{it}} = \sum_{i \in A_t} \frac{1}{p_{it}}$$

which is in fact the HT estimator of U_t if the individual capture probabilities were known. A sample from the posterior distribution is generated by simulating $\tilde{U}_{it} \sim \text{Neg. Bin.}(1, p_{it})$ for each $i \in A_t$ (equivalently $\tilde{U}_{it} - 1 \sim \text{Geometric}(p_{it})$) and setting:

$$\tilde{U}_t = \sum_{it} \tilde{U}_{it}.$$

The value $\tilde{U}_{it} - 1$ can be considered a guess at the number of unmarked individuals similar to individual i that were alive but not captured on occasion t. Simulation of the number of marked individuals alive on each occasion extends directly from the probability model.

In practice, the value of π_0 need never be specified. No matter the values of K, $p_{\min}$ and T a value of π_0 can be chosen so that Claims 1 and 2 are satisfied as closely as desired. Moreover, the accuracy of the approximation only improves as π_0 decreases to 0. Let U_{zt} denote the number of individuals present on occasion t with covariate values in the interval $[z, z + \Delta]$. My interpretation is that the inference from the two stage procedure is the same as inference given a prior distribution for each U_{zt} that is negligibly different from $1/U_{zt}$ but still yields a proper posterior distribution.

3.4 Application

In this section, I apply my method to study the dynamics of a population of Soay sheep (*Ovis aries*) living on the Scottish Island of Hirta. Hirta is a small island of only 6.3 km^2 located 200 km off the west coast of mainland Scotland in the St. Kilda Archipelago. The sheep were originally introduced by early settlers of the archipelago to an even smaller island, Soay, though exactly when and by whom is not clear. In the early twentieth century, the population of the islands was reduced by disease and emigration, and in 1930 the remaining inhabitants were evacuated. At this time, the sheep were removed from the islands except for a small population that was transferred from Soay to Hirta (Catchpole et al., 2000). The population is ideal for study because it is geographically closed, has no natural predators or competitors, and the sheep are easily located so that capture probabilities are very high.

The sheep on Hirta have been studied a great deal and much is known about the population's dynamics. Because there is no predation and no competition for food with other species, the population goes through cycles in which it grows quickly when conditions are favourable and then experiences sudden crashes when conditions worsen (Coulson et al., 2001). Further studies have examined the effects of sex, body mass, horn type, coat type and environmental variables (including winter severity and spring precipitation levels) on the survival of the sheep (Catchpole et al., 2000; King et al., 2006). The objective of my analysis is to incorporate the effects of body mass on the capture and survival probabilities when estimating the population size.

Sheep in one area of the island, Village Bay, have been captured during three censuses conducted in April, August and October of each year since 1986. Newly captured sheep are tagged with uniquely numbered ear tags and several variables are recorded including coat and horn type and body mass. Most of the sheep are captured as lambs (during their first year of life) and so their year of birth and age in every following year is known. Searches for dead sheep are also conducted during each census which provides exact information on the time of death for some individuals. However, the current analysis considers only the summer (August) captures of sheep ignoring the captures in the other seasons and the recovery of dead individuals. Previous studies have shown that the survival of males and females differ considerably and are affected by different factors, and the analysis is further restricted to the sub-population of female sheep.

The original data contained records of 2393 sheep captured in at least one summer of the 15 years between 1986 and 2000, of which 1208 (50%) were female. The majority of the female sheep, 612 (51%), were captured only one time, largely because of the high mortality for lambs. One complication in the analysis was that sheep were weighed on only slightly more than half of the times they were captured. My method requires that the covariate be observed each time an individual was captured and so captures without body mass being recorded were removed from the data. The result is that the estimated capture probabilities should be interpreted as the probability that a sheep was captured and weighed. This will not bias the estimates of the capture or survival probabilities or population size, but the precision of all estimates will be reduced simply because there are less data. The data set in my analysis comprised records for 833 sheep of which 488 (59%) were captured and weighed on only one occasion, 150 (17%) on two occasions and 205 (25%) on 3 or more occasions. An average of 56 individuals were captured and weighed for the first time in each of the 15

years, with counts ranging from 21 in 1986 to 96 in 1987.

Previous analysis of the Soay sheep at Village Bay has shown a strong link between age and survival. Catchpole et al. (2000) and Catchpole et al. (2008) categorised the female sheep into 4 age classes: lambs (year of birth), yearlings (second year of life), adults (ages 3-7), seniors (8 years or older). In my analysis, I fit four different models considering the possible effects of age and body mass:

Model 1: no effect of age or body mass (original Jolly-Seber model) $(p(t), \phi(t))$

Model 2: effect of age only $(p(t \times a), \phi(t \times a))$

Model 3: effect of body mass only $(p(t \times w), \phi(t \times w))$

Model 4: effect of both age and body mass $(p(t \times w \times a), \phi(t \times w \times a))$.

The structure of Model 4 included age and time dependent intercepts and age dependent slopes in the linear predictors of both the capture and survival probabilities. The hierarchical prior was applied separately to the time dependent intercept terms within each age class and no associations were included among the age classes. Estimates of $\boldsymbol{\beta}_0$ and β_1 could not be produced separately for the lambs because it was impossible for lambs to be previously captured and marked. To estimate the capture probabilities for lambs, required for the estimation of the total population size, it was assumed that the parameters $\boldsymbol{\beta}_0$ and β_1 were equal for the lambs and the yearlings. The remaining three models are obtained by placing restrictions on the parameters in Model 4. Model 3 is derived by assuming that the intercepts and slopes are the same for all age classes, Model 2 by assuming that the slopes are uniformly equal to 0 and Model 1 by assuming both that the intercepts are equal across age classes and the slopes are uniformly 0.

Samples from the posterior distributions of all four models were generated through MCMC sampling implemented in OpenBUGS. Markov chains were run for 500,000 iterations total for each model; realizations from the first 100,000 iterations were discarded as burn-in and the remaining realizations were thinned every 40 iterations to yield a sample of size 10,000 from the posterior.

Table 3.1 presents the DIC for each of the four models. The DIC decreased monotonically as the complexity of the models increases. The smallest DIC value was produced by Model 4 including the effects of both age and body mass. DIC values for the Models 2 and 3

Model	DIC	p_D
1) $p(t), \phi(t)$	3937.9	21.4
2) $p(t \times w), \phi(t \times w)$	3744.2	29.3
3) $p(t \times a), \phi(t \times a)$	3697.0	59.7
4) $p(t \times w \times a), \phi(t \times w \times a)$	3612.1	66.1

Table 3.1: DIC selection criterion for the four models fit to the female Soay sheep data.

including only age and body mass were greater by 85 and 132 respectively providing very strong evidence that both are important predictors of capture and/or survival.

The estimated survival and capture probabilities for the best fitting model are plotted as functions of time for fixed values of the covariate in Figure 3.1 and as functions of the covariate for fixed time in Figure 3.2. Posterior summaries for the model parameters are given in Table 3.2. Capture probabilities decreased with the age of the sheep but are fairly consistent over time. The main exception was in 1989 when the capture probability was lower for all ages of sheep. Capture probabilities for the seniors varied more than for the other age classes. There was some suggestion that the capture probability of the lambs and yearlings decreased with increasing body mass, but the 95% CIs of β_1 covered 0 for all age classes.

Survival probabilities for the seniors and adults were also fairly close to constant over time. However, the survival probabilities for the lambs varied considerably. In particular, the survival probabilities were very low in the periods 1988-1989, 1991-1992, 1994-1995 and 1998-1999. Survival probabilities for the yearlings were very low in 1988-1989 and lower in 1998-1999 than in the other years. Higher body mass had a positive effect on the survival probabilities of all ages of sheep. This effect was strongest for the lambs and yearlings, and weakest for the adults and seniors.

Posterior summary statistics for the number of female sheep alive and the number recruited in each year are plotted in Figure 3.3. The estimated population size was very low in 1986, but I believe that this resulted from a data anomaly. The data contains no direct information about the capture probabilities in this year and all inference was derived from the relationship to the other terms in the hierarchical model. However, a much smaller proportion of the sheep captured in 1986 were weighed than in other years. Removing the captures when sheep weren't weighed and imposing the hierarchical model has overestimated the capture probability in this year which has lead to a severe underestimate of the population size. This also caused the number of births in 1987 to be overestimated. Beyond

this, the population of female sheep seems to have had an overall increase during the 15 year period with cycles of decline followed by rebound in 1989, 1992 and perhaps in 1999, though data from more recent years would be needed to confirm a rebound. The number of female lambs born each year was fairly steady from 1989 onward with the exception of decreases in 1989, 1992 and possibly 2000. The posterior mean super-population size was 954 with 95% CI (876,1105).

The results of Catchpole et al. (2000) show very similar patterns of survival by age and time. Estimated survival probabilities were highest for the adults and yearlings and lowest for lambs. Estimates for all ages were lower than usual in the periods 1988-1989, 1991-1992 and 1994-1995, with the lambs being affected most and the adults least. Estimates of the size of the female population also showed similar trends with low numbers and few births in 1986, 1989 and 1992.

Values of the discrepancies computed to assess the fit of the LVB growth model are plotted in Figure 3.4 for 1000 equally spaced samples from the posterior of the selected model. The Bayesian p-value, .75, was close enough to .5 that it does not provide substantial evidence against the LVB growth model, but I was concerned again by the fact that the discrepancy was most often larger for the simulated data than for the observed data. Further examination showed that this was because the distribution of the errors in body mass for the sheep was not exactly Gaussian. Instead, the distribution of the errors was slightly more peaked and there were a few large values far in the tails of the distribution. These observations increase the value of σ so that the errors were, on average, slightly larger in most of the simulated data sets. It is possible that a distribution with heavier tails, like a t-distribution, would provide a better fit, but I am confident that this would have little effect on the estimated capture probabilities or population size.

I had hoped to further demonstrate the utility of the Bayesian p-value by repeating the analysis with the distribution of the covariate given by the diffusion model of equation (3.1) in place of the LVB growth model in equation (3.3). However, the fit of this model was so poor that it was not even possible to generate a sample from the posterior distribution. The MCMC sampler had clearly not converged after 50,000 iterations, and trace plots for some parameters showed an almost linear increasing or decreasing trend over this period. In particular, the value of μ^β for the yearlings decreased almost monotonically and was below -15.0 by the final iteration, an impossibly small value on the logistic scale. The reason for this was that the model did not allow for the differential growth of young and old sheep.

The estimated mean change in body mass was close to 0 kg (between -3.0 and 1.5) for all years. If this were true then we'd expect to have captured yearlings whose body mass was the same as when they were captured as lambs, but no such individuals were observed. The only explanation is that the capture probability for yearlings and lambs of this size is 0, forcing down the value of β_{0t} in all years. The further result was an impossibly high estimate of the population size; the mean of the super-population size over the 50,000 iterations was more than 19,000. This clearly demonstrates the importance of choosing an adequate model of the covariate and checking its fit.

Parameter		Lambs	Yearlings	Adults	Seniors
z_∞	24.43(24.10,24.81)				
r	0.49(0.45, 0.52)				
σ	1.93(1.85, 2.01)				
μ_β			3.16(-0.44, 7.14)	0.60(-0.97,2.08)	-0.50(-3.72,2.81)
τ_β			0.64(0.08, 1.21)	0.63(0.34,0.98)	0.73(0.14,1.30)
β_1			-0.12(-0.33, 0.07)	0.00(-0.06,0.06)	0.02(-0.12,0.14)
μ_γ		-5.23(-7.10,-3.48)	-4.02(-8.43,-0.21)	-0.95(-2.98,1.08)	-2.76(-5.88,0.22)
τ_γ		1.79(0.96, 2.76)	1.74(0.60, 3.29)	0.55(0.06,1.11)	0.49(0.04,1.33)
γ_1		0.41(0.29, 0.53)	0.35(0.14, 0.61)	0.13(0.05,0.22)	0.14(0.01,0.27)

Table 3.2: Posterior summary statistics for parameters of Model 4 including the effects of both age and body mass on the capture and survival probabilities. Statistics provided are the posterior mean and 95% CI in parentheses. Parameters of the covariate distribution are common to all sheep. The remaining parameters vary by age class.

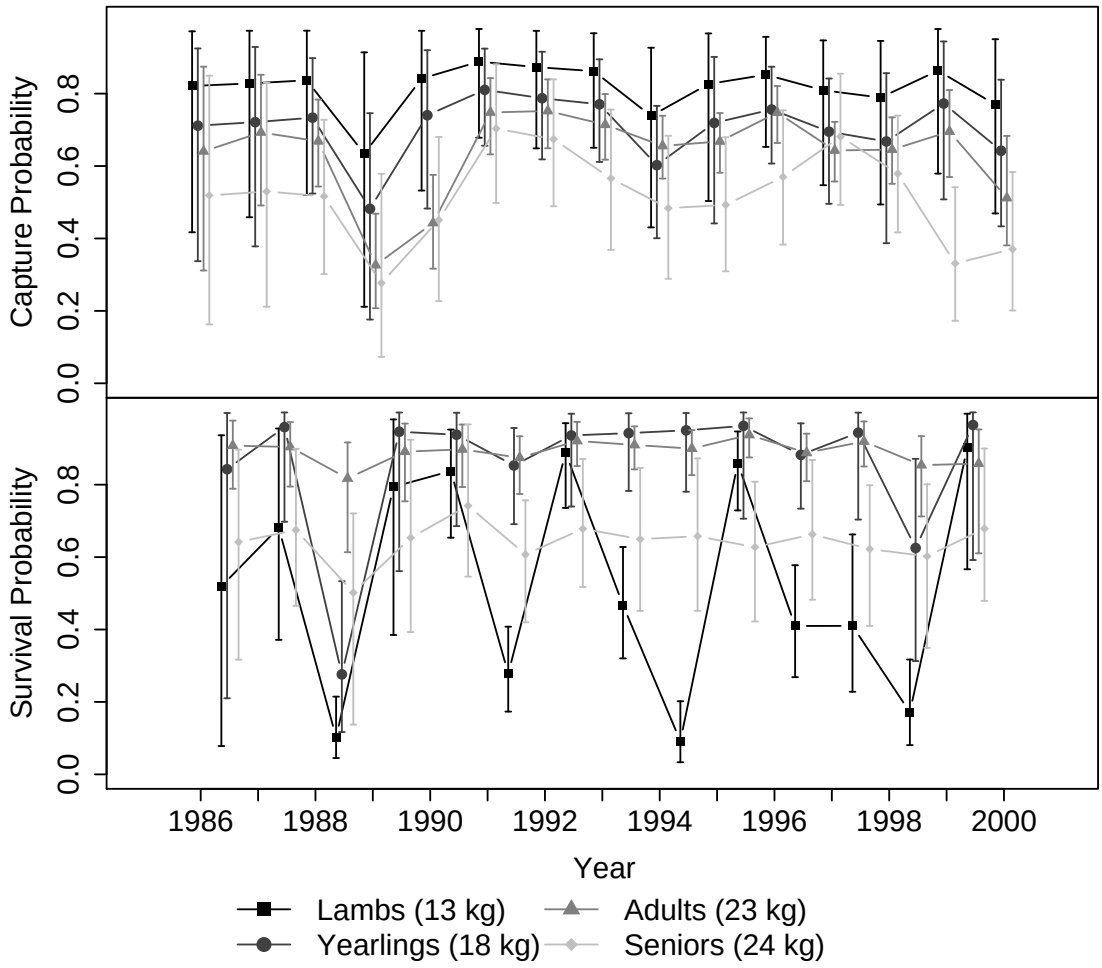

Figure 3.1: Effect of time on the capture and survival probabilities of the female sheep. The plotted values summarize the posterior distribution of the of the capture probability (top) and survival probability (bottom) of sheep with the average observed body mass for each age class, as indicated in the legend. For each age $\times$ year the point represents the posterior mean and the error bars indicate the extents of the 95% CIs.

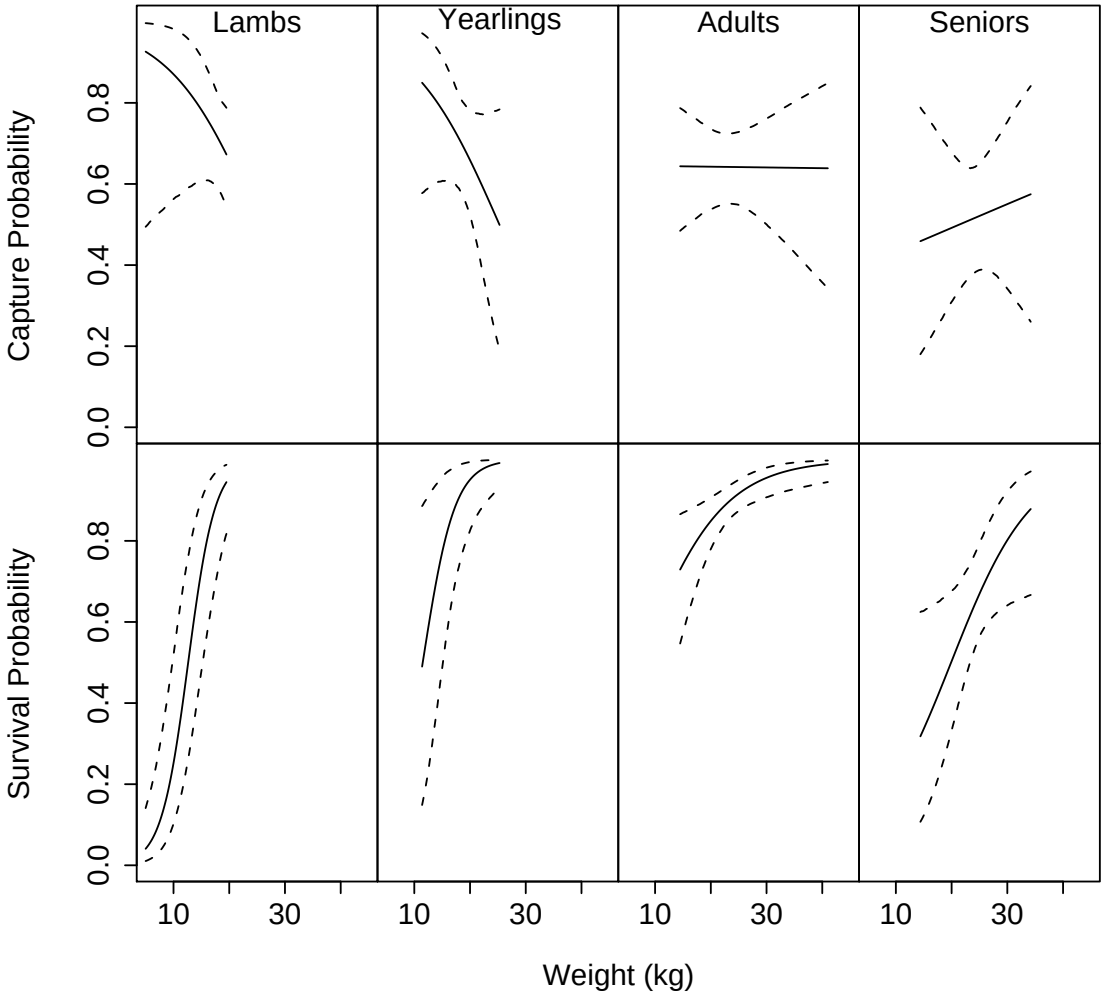

Figure 3.2: Effect of body mass on the capture and survival probabilities of the female sheep. The solid lines indicate the posterior mean capture probability (top) and survival probability (bottom) as a function of body mass, with the intercept parameters equal to the hierarchical mean for each age group. The dotted lines indicate the extents of the pointwise 95% CIs. For each age class, the posterior distributions are only summarized over the range of observed covariate values.

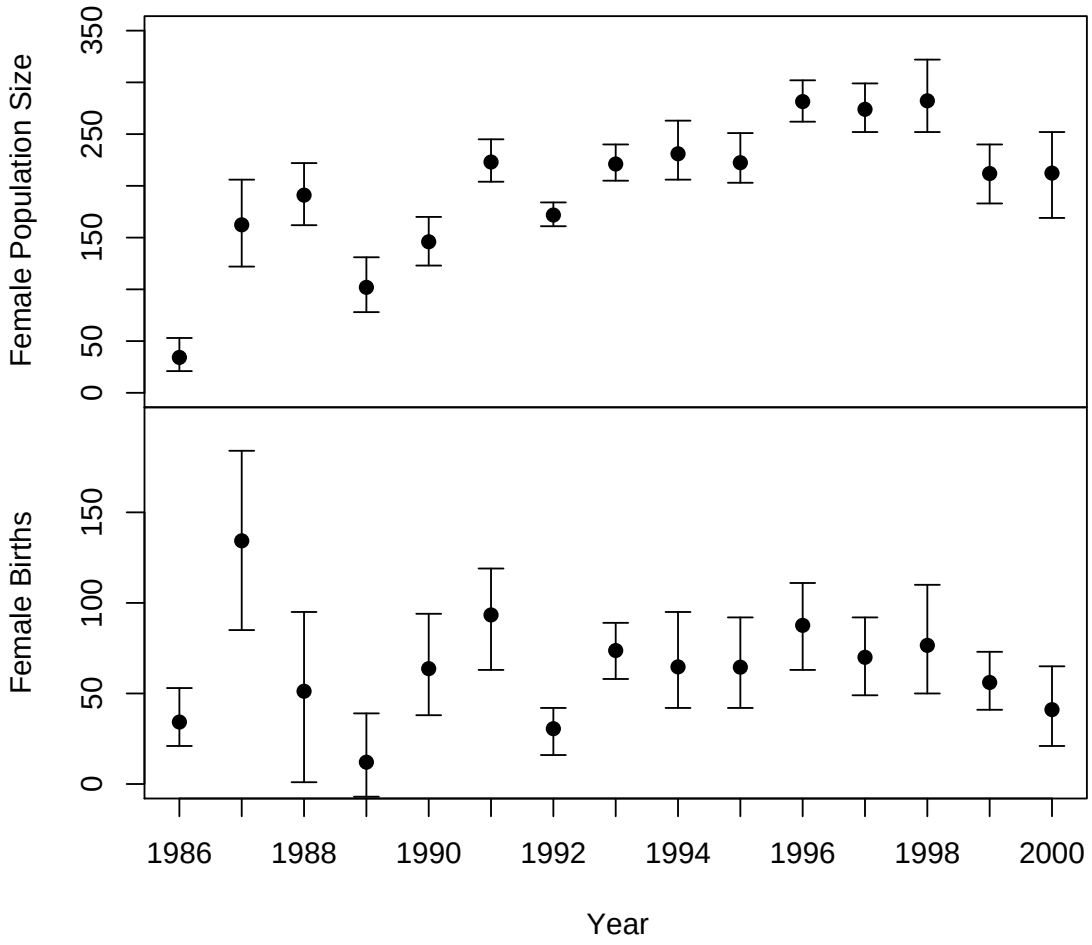

Figure 3.3: Estimated female population size and number of female births by year. Points in the top plot indicate the posterior mean population size for each year with 95% CIs represented by the error bars. Points in the bottom plot indicate the posterior mean number of births per year with 95% CIs represented by the error bars.

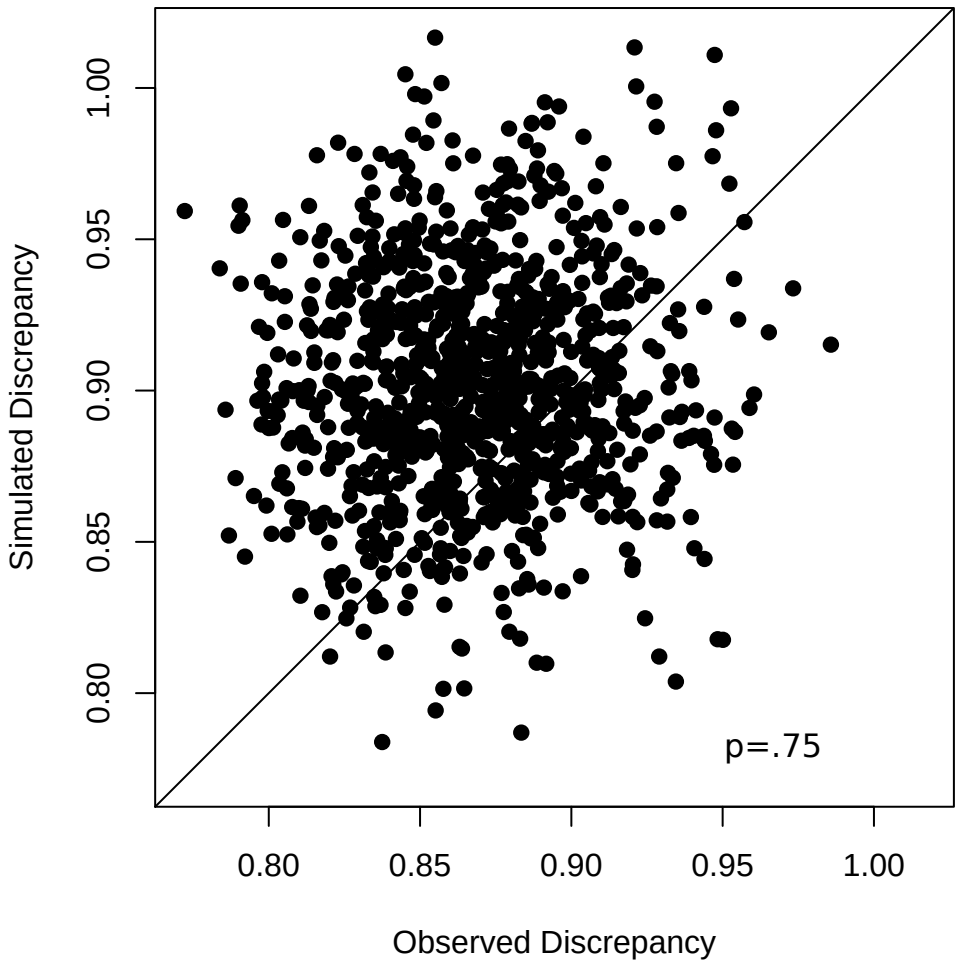

Figure 3.4: Observed and simulated discrepancies for computing the Bayesian p-value assessing the model of the covariate. The Bayesian p-value is the proportion of times the simulated discrepancy is greater than the observed discrepancy (i.e., above the plotted line).

3.5 Simulation

The performance of the new method was assessed through a simulation study based on the results of the Soay sheep analysis. To simplify the models in the simulation, the age structure was ignored. Instead, data was generated under two scenarios, the first based on the parameter estimates for the yearling sheep and the second based on the estimates for the adults. All simulated data sets were analysed with both the standard JS model and my extended model incorporating the covariate. When the capture probabilities were affected by the covariate, the standard JS model overestimated the population size. When the capture probabilities did not depend on the covariate, the standard model produced estimates that were both slightly more accurate and precise than my extended model.

For each of the two scenarios, I generated 100 data sets comprising capture histories for a population of 1000 individuals over 15 capture occasions. Data for each individual were generated by first simulating the time of recruitment and the values of the covariate, and then simulating the survival and capture events. Time of recruitment was randomly selected from a period of twenty time units, starting 5 units before the first capture occasion to establish a population of adults. The value of the covariate at the time of recruitment was simulated from a normal distribution with mean 11.6 and standard deviation 2.5, the observed mean and standard deviation of body mass of the Soay sheep lambs. Values of the remaining parameters in each of the two scenarios are provided in Tables 3.3 and 3.4. In the first simulation, based on the yearling sheep, the covariate had a weak, negative effect on the capture probability and a stronger, positive effect on the survival probability. Survival probabilities also varied widely over time for the same value of the covariate. In the second scenario, based on the adults, the covariate had no effect on the capture probability and a much smaller effect on the survival probability, and there was less variation in the capture probabilities over time.

Each data set was analysed with both a Bayesian implementation of the standard JS model assuming identical capture and survival probabilities for all individuals alive on each occasion and my model allowing for the effect of the covariate on both sets of parameters. All models were fit using MCMC simulation in OpenBUGS to generate samples from the posterior distribution. Markov chains were run for 50,000 iterations in total with the first 10,000 discarded as burn-in. The remaining 40,000 samples were thinned every 8 iterations to yield 5000 samples from the posterior distribution. Accuracy of the models was compared by

the bias of the posterior mean as a point estimate of selected parameters and the proportion of times the 95% posterior CIs for these parameters covered the true values. Precision was compared by the width of the 95% CIs.

Results for the first simulation scenario are presented in Table 3.3. Posterior means for my model including the covariate's effect provided approximately unbiased estimates of all parameters, except for the variance of the survival probability intercept terms, τ^{ϕ}. The posterior mean did overestimate the super-population size slightly, but the bias, less than 1%, is small enough to be ignored in practice. The 95% CIs for most parameters were very narrow and covered the true parameter values for at least 95% of the data sets. The model ignoring the covariate overestimated the super-population size by more than 7% and the 95% CI covered the true population size for only 15 of 100 data sets.

In the second scenario, the estimate of the super-population size from the model ignoring the covariate was approximately unbiased (see Table 3.4). This occurred despite the facts that the model did not account for the effect of the covariate on the survival probability and the posterior mean of $\boldsymbol{\beta}_0$ was biased by more than 20%. In contrast, the estimate of the super-population size from the model incorporating the covariate had a positive bias greater than 1.5%. The 95% credible interval for the super-population size was also marginally wider for the modelling including the covariate effect. Estimates for the remaining parameters were approximately unbiased, though the coverage of the 95% CIs was less than 95% for 6 of the 10 selected parameters.

The first simulation shows the potential for biased estimates when the effect of the covariate on the capture probability is ignored. Further simulations with even stronger effects of the covariate on the capture probabilities produced even larger biases (results not shown). When the capture probabilities are the same for all individuals, the standard JS model produced estimates that were both more accurate and more precise, even though the model ignored the effect on the survival probabilities. However, the loss of accuracy and precision from including the covariate when it did not affect the capture probability was much less than the loss from ignoring the significant effect.

		Jolly-Seber Model with Covariate					Jolly-Seber Model				
		Posterior Mean		95% CI			Posterior Mean		95% CI		
Parameter	Truth	Mean(SE)	% Bias(SE)	Width	% Width	% Coverage	Mean(SE)	% Bias(SE)	Width	% Width	% Coverage
S	1000.0	1004.8(1.5)	0.5(0.1)	83.1	8	99	1074.0(2.8)	7.4(0.3)	128.6	13	15
μ^p	3.5	3.5(0.0)	−1.1(1.4)	1.9	55	96	1.1(0.0)	-	0.7	-	-
τ^p	0.6	0.6(0.0)	4.0(2.3)	0.6	95	97					
β_1^p	−0.1	−0.1(0.0)	1.4(1.9)	0.1	74	95					
μ^ϕ	−3.5	−3.4(0.0)	2.7(1.4)	2.4	70	95	2.6(0.0)	-	2.2	-	-
τ^ϕ	1.7	1.9(0.0)	11.6(2.6)	2.5	72	95					
β_1^ϕ	0.3	0.3(0.0)	0.9(0.4)	0.1	19	98					
z_∞	24.5	24.5(0.0)	0.0(0.0)	0.5	2	98					
r_z	0.5	0.5(0.0)	0.0(0.3)	0.1	12	97					
σ_z	2.0	2.0(0.0)	−0.2(0.1)	0.1	6	99					

Table 3.3: Results of the simulation based on the parameter estimates of the yearling sheep. The left half of the table presents the results for the extended JS model including the effect of the covariate, the right for the standard JS model ignoring the covariate. Performance of each method was assessed by the bias of the posterior mean, and the width and % coverage of the 95% credible interval.

		Jolly-Seber Model with Covariate					Jolly-Seber Model			
		Posterior Mean		95% CI			Posterior Mean		95% CI	
Parameter	Truth	Mean(SE)	% Bias(SE)	Width	% Width	% Coverage	Mean(SE)	% Bias(SE)	Width	% Width	% Coverage
S	1000.0	1015.5(4.6)	1.6(0.5)	238.2	24	99	1005.5(3.2)	0.5(0.3)	200.3	20	100
μ^p	0.5	0.5(0.1)	−4.4(10.4)	2.0	403	93	0.4(0.0)	−20.4(3.7)	0.7	145	88
τ^p	0.6	0.6(0.0)	0.7(2.8)	0.6	102	95					
β_1^p	0.0	0.0(0.0)	0.1(0.2)	0.1	-	91					
μ^ϕ	−1.0	−1.0(0.0)	1.6(2.8)	1.1	111	95	1.1(0.0)	-	0.7	-	-
τ^ϕ	0.5	0.5(0.0)	6.6(3.2)	0.7	137	93					
β_1^ϕ	0.1	0.6(0.0)	−0.7(1.2)	0.0	45	96					
z_∞	24.5	24.5(0.0)	−0.2(0.1)	0.8	3	93					
r_z	0.5	0.5(0.0)	0.3(0.5)	0.1	17	93					
σ_z	2.0	2.0(0.0)	−0.4(0.2)	0.2	9	91					

Table 3.4: Results of the simulation based on the parameter estimates of the adult sheep. The left half of the table presents the results for the extended JS model including the effect of the covariate, the right for the standard JS model ignoring the covariate. Performance of each method was assessed by the bias of the posterior mean, and the width and % coverage of the 95% credible interval.

3.6 Discussion

The objective of this project was to incorporate the effects of continuous covariates on individual capture and survival probabilities when estimating the size of an open population from capture-recapture data. My solution extends the Bayesian implementation of the CJS model that by Bonner and Schwarz (2006). Inference about the population size is conducted in two stages. First, the Bayesian CJS model is fit via MCMC to obtain a sample from the posterior distribution of the parameters governing the capture and survival probabilities. A modification of the HT estimator is then applied to each of the sampled parameter sets to obtain a sample from the posterior distribution of the number of unmarked individuals, and hence the total population size, alive on each occasion.

There are two concerns I have with the new method, both relating to the prior distribution for the number of unmarked individuals. The first is that the number of unmarked individuals alive on occasion t, U_t, is not a quantity of direct importance, its value being dependent on the effort put into capture. The true parameter of interest is the total population size, N_t, and this is the parameter to which the prior should be assigned. However, N_t is not a parameter in the likelihood function and so it cannot be assigned a prior distribution directly. My second concern is that the analyst cannot assign a prior distribution to U_t other than that defined in section 3.3. In most (if not all) capture-recapture studies the researchers will have at least some, and possibly considerable, prior information about the population size. Unfortunately, Claim 1 is only true for the specified non-informative prior distribution in equation (3.6), and so the two stage approach cannot work with any other prior.

I believe that both of these problems can be solved by the method of Bayesian melding (Poole and Raftery, 2000). Bayesian melding was developed for the analysis of complex deterministic models in which data provides estimates for a set input parameters which are fed into the model to produce a set of output parameters. In some situations, the researchers may have prior beliefs about both the inputs and outputs of the model. Incorporating the prior on the inputs is straightforward, but the standard Bayesian formula does not allow for specification of a prior for the outputs. Bayesian melding provides a means to define prior distributions for the two sets of parameters simultaneously. A sample from the posterior is generated by first sampling values from the prior distribution of the inputs, computing the corresponding outputs from the deterministic model, and then re-weighting each realization

based on the prior distribution for the outputs and the value of the likelihood function. This could be applied to my method by considering $N_t = U_t + M_t$ as a deterministic function of U_t and M_t, and would allow for the specification of a prior for each N_t directly as well as for the input parameters $\boldsymbol{U}$, $\boldsymbol{\beta}_0$, β_1, $\boldsymbol{\gamma}_0$ and γ_1.

Although I developed this method specifically to allow the inclusion of continuous covariates in the JS model, it can easily be extended to many other capture-recapture models. The two claims in section 3.3 depend only on the first component of the likelihood, L_1 and the exact form of L_2, the component modelling the events in each individuals capture history subsequent to their first capture, is irrelevant. This means that the two-stage procedure can be directly applied to any model which considers the number of unmarked individuals captured on each occasion in the same way – as a random sample from a population of unmarked individuals of fixed but unknown size. For example, the method could be extended to models including recoveries of dead individuals as well as the capture of live individuals or to the robust design (Pollock, 1982).

An important model to which the method does not extend is the Schwarz-Arnason super-population model (Schwarz and Arnason, 1996). There are several disadvantages to conditioning on the number of unmarked individuals alive on each occasion as fixed values including that the estimate of recruitment on each occasion need not be positive and that constraints cannot be imposed on the numbers of recruits. Schwarz and Arnason (1996) extended the original JS model by explicitly modelling the number of individuals recruited on each occasion as a sample from the super-population, the hypothetical collection of individuals available to be captured at least one time during the experiment, and modelling the events in each capture history prior to the individuals first capture. This ensures that the number of births is non-negative on all occasions and allows the probabilities of recruitment to be constrained or modelled as functions of covariates, including environmental variables or even the population size itself. The difficulty in incorporating continuous covariates into the super-population model is in modelling the events in an individuals capture history before it is first captured. This would require more assumptions to model the unconditional distribution of the covariate and also more intensive computing to generate a sample from the posterior distribution. Incorporating continuous covariates into the Schwarz-Arnason model is an important topic for future research.

There is one important caveat with the application of my method to the Soay sheep. Although the trends in the estimated population size match well with Catchpole et al.

(2000), my estimates are in fact smaller in every year. This is due to the assumption that $\boldsymbol{\beta}_0$ and β_1 are the same for both the lambs and yearlings. The capture probabilities decline with body mass for the yearlings ($\beta_1 < 0$) and extrapolating to the lower body masses observed for the lambs results in significantly higher capture probabilities (.80 for the lambs versus .65, approximately, for the yearlings). From the original data including the captures when body mass was not recorded, I know that this overestimates the capture probabilities of the lambs, which leads to underestimation of the population size. Unfortunately, the capture probabilities for the lambs, and hence the total population size, cannot be estimated without some assumption relating the lambs to the other sheep. Alternatives would be to exclude the lambs from the estimate of population size or to model the intercept of the capture probability as a linear function of age. However, the analysis seems sufficient as an illustration of the method.

Chapter 4

Continuous Covariates in the Cormack-Jolly-Seber Model: A Bayesian Adaptive Spline Approach

4.1 Introduction

This chapter presents a second project concerning the use of continuous predictors of the capture and survival probabilities in capture-recapture models of open populations. As discussed in Chapter 3, predictors of the capture probabilities are often included to account for variations in individual catchability that might otherwise bias estimates of the population size. However, the effect of the explanatory variables on survival may also be of direct interest. For example, capture-recapture experiments have studied the effect of body mass on the survival of meadow voles (*Microtus pennsylvanicus*)(Nichols et al., 1992; Bonner and Schwarz, 2006), winter temperatures on the survival of grey herons (*Ardea cinerea*) (North and Morgan, 1979) and sex, horn type and birth weight on the survival probabilities for the population of Soay sheep (*Ovis aries*) considered in the last chapter (Catchpole et al., 2000; King et al., 2006). A common assumption in most models is that the effect of a continuous covariate is linear on some scale (usually logistic) so that the estimated survival probabilities either increase or decrease monotonically with the covariate. The objective of this project

was to develop a flexible method that would make as few assumptions as possible about the nature of these relationships.

In some biological systems the effect of a continuous covariate on survival may be very complex. Suppose that it had not been possible to categorize the Soay sheep by age class in the application of the previous chapter. Instead of modelling the effect of body mass on the survival probability of each class separately, it would have been necessary to define a single function modelling this relationship. The resulting function would have several peaks with high probabilities of survival at the optimum body mass for each age class and low values in between. Some covariates may also be influenced by stabilizing selection so that the survival probability peaks at one specific value and decreases at both higher and lower values. Fitting models assuming a monotone relationship would not capture either of these behaviours and would lead to incorrect conclusions about the covariate's effect.

One solution to this problem is to partition the values of the covariate into a number of discrete categories and allow the survival probability to vary independently in each category. In Nichols et al. (1992), for example, the body mass of the meadow voles is divided into 4 distinct classes. The disadvantages of this approach is that it introduces discontinuities into the function relating the covariate to body mass, many categories may be required to describe complex relationships and the definition of the categories is usually subjective, rather than biologically based. Another method is to introduce higher order polynomial terms to allow for non-linearities. While this may be adequate for simpler relationships, many higher order terms would be required if the relationship has local behaviours, as in modelling the effect of body mass on the survival probability of all ages of sheep simultaneously.

My solution is to model the linear predictor of the survival probabilities as a spline function. As discussed in Chapter 1, there are two general approaches to fitting splines and ensuring a proper balance between smoothness and complexity. In Chapter 2, penalized splines were applied to model the daily run size of migrating salmon populations as a smooth function of time. In this chapter, I have chose an adaptive spline approach. In particular, I have implemented the adaptive Bayesian B-spline method of Biller (2000).

The challenge of fitting an adaptive spline can be viewed as a large variable selection problem. Each function in the spline basis contributes a new variable to the model, and the goal is to select the set (or sets) of basis functions that produce the best balance between fit and complexity. Bayesian inference using the reversible jump Markov chain Monte Carlo (RJMCMC) algorithm of Green (1995) provides an ideal method for exploring the space of

spline models and selecting the best sets of basis functions. The method of Biller (2000) provides an particular implementation of the RJMCMC algorithm for sampling from a specific set of candidate models formed by a finite set of knots over the range of the covariate. Inference is then obtained from the summary statistics of the sampled models.

The basic approach of my method is very similar to that of Gimenez et al. (2006). In this work, splines were used to flexibly model the effect of body mass on the survival probability of a population of sociable weavers (*Philetairus socius*) in South Africa captured over a 15 year period. There are two key differences between this project and my work. First, Gimenez et al. (2006) implemented a penalized spline approach similar to that used in Chapter 2. Their model also assumes that body mass is a constant, averaging all values observed for an individual to obtain a single covariate, whereas my method allows the value of the covariate to change over time as in Bonner and Schwarz (2006) and the model of the previous chapter.

Section 2 of this chapter provides further details of adaptive splines and the method for fitting the survival probabilities as a function of a continuous covariate. Here I adapt the method of Biller (2000) to allow for the unobserved values of the covariate when individuals are not captured and the estimation of the remaining parameters in the model. Section 3 examines a simulation study where the survival probability is known to have local dependence on the covariate. Fit of the spline model is compared with the fit of a simpler cubic polynomial model. In section 4, I apply the method to study the relationship between the survival and condition of reed warblers (*Acrocephalus scirpaceus*) captured as part of the Dutch Constant Effort Sites (CES) ringing program. The final section discusses advantages and disadvantages of the method and provides some suggestions for its use in future capture-recapture studies.

4.2 Methods

4.2.1 Notation

Observed Data & Latent Variables:
T =number of capture occasions (indexed by t).
z_{it}=value of time-varying covariate for individual i at time t (*partially observed*).

Parameters:

ϕ_{it} =probability that individual i is available for capture at occasion $t+1$ given that it was available for capture at occasion t

p_t =probability that any individual available for capture at occasion t is captured

$s(z)$ =spline function modelling dependence of ϕ_{it} on z_{it}

κ =number of potential knot locations

K =number of knots in a single realization of the adaptive spline

ξ_k =location of the k^{th} knot

ξ_l, ξ_u=lower and upper boundary knots

$\boldsymbol{\beta}$ =vector of coefficients of $s(z)$

μ_t =population mean change in covariate between t and $t+1$

σ^2 =population variance of change in covariate between adjacent capture occasions

4.2.2 Cormack-Jolly-Seber Model with Individual Covariates

The basis for the new method I have developed is the extended CJS model of Bonner and Schwarz (2006). As discussed in Chapter 3, this model allows for the CJS capture and survival probabilities to be modelled as functions of a continuous covariate that can only be observed when an individual is captured. To account for the missing values of the covariate, their distribution is modelled conditional on the observed values for each individual. Bayesian methods are then applied to make inference about the covariate's effect.

In both Bonner and Schwarz (2006) and Chapter 3 it was assumed that the capture and survival probabilities were linear functions of the covariate on the logistic scale. That is:

$$\text{logit}(p_{it}) = \beta_{0t} + \beta_1 z_{it}$$

and:

$$\text{logit}(\phi_{it}) = \gamma_{0t} + \gamma_1 z_{it}$$

where and z_{it}, p_{it} and ϕ_{it} are the value of the covariate, the capture probability and the survival probability for individual i on capture occasion t. The new model allows $\text{logit}(\phi_{it})$ to be a non-linear function of the covariate. In particular, the model specifies that $\text{logit}(\phi_{it}) = s(z_{it})$ where $s(z)$ is a spline function fit through the adaptive Bayesian B-spline approach. For simplicity, I will focus only on the effect of the covariate on survival and assume that

the capture probabilities are dependent only on time (i.e., $p_{it} = p_t$ for all i). Otherwise, the likelihood contribution for each individual is exactly as given in equation (3.4).

4.2.3 Adaptive Bayesian B-Spline Model

The adaptive spline fitting paradigm considers the number of knots, K, and their locations, $\xi_1, \ldots, \xi_K$ as unknown values that must be estimated along with the coefficients of the spline. Suppose that we wished to model the mean of some response as a function of the covariate, z, which ranges between ξ_l and ξ_u, called the boundary knots in spline terminology. The idea of Biller (2000) is to select a large number,κ, of potential knot locations between the boundary knots and then select from models with knots at these points. The model space is then taken to be the restricted set of cubic spline functions with between 1 and κ knots at these locations. There are $\binom{\kappa}{K}$ possible splines with K knots and hence $\sum_{K=0}^{\kappa}\binom{\kappa}{K}$ candidate models in total.

In essence, standard Bayesian methods are applied to make inference about the models in this space. A prior distribution is defined for the entire set of models and then combined with information from the data, in the form of the likelihood, to generate the posterior distribution from which inference can be obtained. The main challenge is that even though there is a finite set of candidate models, the number of possible configurations for the knots is too great to compute posterior probabilities for each model separately, either numerically or analytically. Moreover, the standard sampling techniques of Bayesian inference, like the Metropolis-Hastings algorithm, cannot be used to sample from the posterior distribution because the space contains models of varying dimension (i.e., varying numbers of basis functions). Instead, Biller (2000) provides a reversible jump MCMC (RJMCMC) algorithm to explore the model set and generate a sample of realizations from which point and interval estimates can be computed.

Let $\{B_1^{(q)}(z), ..., B_{q+K}^{(q)}(z)\}$ denote the set of B-spline basis functions for a spline of order q with K knots and $\boldsymbol{b} = (b_1, \ldots, b_{q+K})$ the corresponding vector of coefficients so that:

$$s(z) = \sum_{j=1}^{q+K} b_j B_j^{(q)}(z).$$

To define a prior distribution over the entire model space, it is necessary to assign a density to each of the possible sets of knots and the coefficients of the resulting spline functions.

Following Biller (2000) I construct the joint prior density, $\pi(\boldsymbol{b}, \boldsymbol{\xi}, K)$, as the product of three densities: the prior probability on the number of knots, $\pi(K)$, the probability of the knot locations given their number, $\pi(\boldsymbol{\xi}|K)$, and the density of the coefficients given both the number of knots and their locations, $\pi(\boldsymbol{b}|\boldsymbol{\xi}, K)$. First, the number of knots is assigned a truncated Poisson(λ) distribution so that:

$$\pi(K) \propto e^{-\lambda}\lambda^{K}(K!)^{-1}$$

for each $K = 1, \ldots, \kappa$. Next, each of the $\binom{\kappa}{K}$ possible configurations of K knots are assigned equal probability conditional on K so that:

$$\pi(\boldsymbol{\xi}|K) = \binom{\kappa}{K}^{-1}.$$

Finally, independent, diffuse normal priors with mean 0 and large variance τ^2 are specified for the coefficients of the spline, $\boldsymbol{b}$.

Each iteration of the RJMCMC algorithm developed in Biller (2000) includes three separate steps involving the proposal and then acceptance or rejection of a new spline function. The first step in each iteration allows a change in the dimension of $s(z)$ by adding or deleting one knot (equivalently, adding or deleting one B-spline basis function from the set of predictor variables). Starting from a spline with K knots the number of knots in the proposal, K', is increased to $K + 1$ with probability:

$$P(K+1|K) = \begin{cases} 0 & K = \kappa \\ .5 & 0 < K < \kappa \\ 1 & K = 0 \end{cases}$$

and decreased to $K - 1$ with probability $P(K-1|K) = 1 - P(K+1|K)$. If $K' = K + 1$, the location of the new knot is chosen from the $\kappa - K$ currently unoccupied locations with uniform probability. The coefficient for the new basis function is then set equal to the weighted average $ub_{j^-} + (1-u)b_{j^+}$ in which b_{j^-} and b_{j^+} represent the coefficients from the neighbouring basis functions and $u \sim U(0,1)$. Minor, deterministic adjustments are also required for the coefficients b_{j-} and b_{j^+} as their associated basis functions are changed by the addition of the new knot. Let $\boldsymbol{\xi'}$ and $\boldsymbol{b'}$ denote the proposed vectors of knot locations and coefficients. Because of the change in dimension, the probability of accepting the

proposed spline must be computed from the RJMCMC acceptance ratio in equation (1.2) with the proposal density $Q(\theta') = P(K+1|K)/(\kappa-K)$, the density of the current parameter values $Q(\theta) = P(K|K+1)/(K+1)$ and the Jacobian $\mathcal{J}((\theta,u),(\theta',u')) = |b_{j^-} - b_{j^+}|$. If $K' = K-1$, then one knot is deleted in a similar way. The knot to be deleted is randomly selected with uniform probability, the coefficients of the proposal are computed and the new spline is accepted according to the RJMCMC acceptance ratio with $Q(\theta') = P(K-1|K)/K$, $Q(\theta) = P(K|K-1)/(\kappa-K-1)$ and $\mathcal{J}((\theta,u),(\theta',u')) = 1/(|b_{j^-} - b_{j^+}|)$.

The second step of each iteration proposes a limited adjustment to the location of one knot. For the j^{th} knot in the spline, a neighbourhood is defined which contains all of the empty knot locations between the previous knot, ξ_{j-1}, and the subsequent knot, ξ_{j+1}. The knot is moveable if this neighbourhood contains at least one unoccupied location. To update the spline, one knot is selected from the set of moveable knots with uniform probability and moved to one of the empty knot locations in its own neighbourhood, again with uniform probability. The B-spline basis functions are recomputed given the new vector of knots, ξ', but the vector of coefficients is not changed. The dimension of the proposal is the same as that of the current model and so acceptance or rejection is performed in a simple MH step with acceptance probability given in equation (1.1). The purpose of this step is to allow the sampler to move through different models with the same number of knots without having to move first through models of differing dimension. Keeping the new location for the selected, moveable knot within its neighbourhood limits the amount of change possible on one iteration, but ensure that the proposal is accepted with high probability.

The third step in updating the spline model is to propose new values for the vector of coefficients, $\boldsymbol{b}$. The most common method for updating the coefficients of linear models is to perturb the current values by the addition of random, Gaussian noise. When this is done, the variance of the added noise needs to be tuned so that the deviations are not to large, which would cause the proposal to be accepted very rarely, or too small, which would cause the chain to mix very slowly. Biller (2000) notes that this method for proposing new coefficients cannot be used in the adaptive algorithm because the optimal variance would depend on the exact number of knots and their locations. Instead, the method of Gamerman (1997) is recommended. One step of the iteratively reweighted least squares algorithm is performed and the resulting estimates of the coefficients and their variance matrix are taken as the mean and variance of a multivariate normal distribution from which the proposal is generated. This step also does not alter the dimension of the model, and so the acceptance

probability is again computed from the standard MH ratio in equation (1.1).

Implementation of this algorithm for modelling the CJS survival probabilities as a function of a continuous covariate poses some difficulties because of the missing covariate values. First, the RJMCMC algorithm for sampling from the posterior distribution requires extra steps to impute the unobserved covariate values on each iteration. Further steps are also needed to update the remaining parameters in the model including the capture probabilities and the parameters specifying the distribution of the covariate. The second complication is in the choice of the boundary knots, ξ_l and ξ_U. Biller (2000) suggests setting these equal to the minimum and maximum values of the covariate and spacing the κ potential knot locations evenly between. This is not possible in my application because the minimum and maximum values may not have been observed. If ξ_l and ξ_u are set equal to the observed minimum and maximum, say $z_{\min}^{\text{obs}}$ and $z_{\max}^{\text{obs}}$, it is possible for some of the covariate values imputed on each MCMC iteration to lie outside of the boundary knots. Although the value of each B-spline basis function is constrained to $[0, 1]$ within $[\xi_l, \xi_u]$, the values are not constrained outside of this interval and this can lead to the same numerical problems that arise with the truncated polynomial basis. Instead, I recommend choosing ξ_l and ξ_u to enclose a wide range about the observed data (e.g. $\xi_l = z_{\min}^{\text{obs}} - (z_{\max}^{\text{obs}} - z_{\min}^{\text{obs}})$ and $\xi_u = z_{\max}^{\text{obs}} + (z_{\max}^{\text{obs}} - z_{\min}^{\text{obs}})$) but still spacing the potential knots equally between $z_{\min}^{\text{obs}}$ and $z_{\max}^{\text{obs}}$. This arrangement allows the imputed covariate values to lie outside of $[z_{\min}^{\text{obs}}, z_{\max}^{\text{obs}}]$ without concern for numerical instability, but constrains the spline to be equal to a cubic polynomial on the intervals $[\xi_l, z_{\min}^{\text{obs}}]$ and $[z_{\max}^{\text{obs}}, \xi_u]$ where there is, in fact, no observed data.

The final algorithm for generating a sample from the posterior distribution of all random variables has the following structure. First the algorithm is initialized by defining the boundary knots and selecting κ potential knot locations between $z_{\min}^{\text{obs}}$ and $z_{\max}^{\text{obs}}$. Initial values are then chosen for all parameters including the parameters of the spline, the parameters of the covariate distribution and the capture probabilities. Once the algorithm has been initialized, MCMC iterations are conducted in the following steps:

1) Simulating the missing covariate values.

2) Updating the parameters of the covariate distribution.

3) Updating $\text{logit}(\phi(z)) = s(z)$ by:

   a adding or deleting one knot at a randomly selected location,

b moving one randomly selected knot to an empty location in its neighbourhood, and

c updating the coefficients.

4) Updating the capture probabilities.

These iterations are repeated until the chain converges and a large sample of realizations from the posterior is generated. Because there is no single set of knots, an estimate of logit(ϕ_{it}) cannot be obtained simply by plugging estimates of the coefficients into equation the B-spline formula. Instead, the value of $s(z)$ at a specific point is approximated by averaging the value of the function across the sampled realizations. Precision of $s(z)$ is assessed with pointwise 95% highest posterior density (HPD) credible intervals. That is, for each value of z in $[\xi_l, \xi_u]$ I compute the shortest interval which covers 95% of the sampled values of $s(z)$. The RJMCMC algorithm was implemented in the `R` programming language (R Development Core Team, 2008).

4.3 Simulation Study

In my simulation study, capture histories for 500 individuals were generated from a CJS model with 3 capture occasions – the minimum required to estimate a survival probability. Covariate values for each individual were simulated from the diffusion based model:

$$z_{i,t+1}|z_{it} \sim N(z_{it} + \mu_t, \sigma^2)$$

with the initial distribution $z_{a_i} \sim N(0, 1)$ and parameters $\mu_1 = \mu_2 = 0.00$ and $\sigma^2 = 1.00$. Survival probabilities for each interval were computed from a bimodal function of the covariate with modes at $z = -1$ and $z = 1$. This function is plotted in Figure 4.1. Capture probabilities were $p_2 = p_3 = .85$. Recruitment times were assigned so that half of the individuals were first captured and marked on occasion 1 and half on occasion 2.

Three different models of the survival probability were fit to the simulated data to study the new method's performance. The first model fit a cubic polynomial to the logit of the survival probability. The second and third fit the adaptive spline model described above using two different prior distributions for the number of active knots: truncated Poisson with rate parameter $\lambda = 25$ and truncated Poisson with rate parameter $\lambda = 75$. Markov chains for all three analyses were run for 100,000 iterations. The initial 10,000 were discarded as burn-in, and every 10^{th} of the remaining 90,000 iterations was retained for inference.

Estimates of the survival probability as a function of the covariate for all 3 models are plotted along with their pointwise 95% HPD credible intervals in Figure 4.1. The cubic function clearly was too rigid to adjust to the local changes in the survival probability. Instead, the fitted curve decreased throughout the range of the covariate, and the 95% credible intervals failed to cover the true survival probability for much of the range.

In comparison, the spline fit using the Poisson(25) prior easily captured the bimodality of the survival probability. The pointwise 95% credible intervals covered the true function completely, though they were between 2 and 3 times as wide as the credible intervals for the cubic fit. A trace plot and histogram of the number of knots in the spline for each MCMC iteration and a plot indicating the proportion of times each potential knot location was occupied are shown in Figure 4.2. The number of knots appeared to converge very quickly to a stable distribution which placed 95% of the posterior probability on models with between 5 and 14 knots. The posterior median was 10 knots. Knot locations that were occupied most often occupied were centred near the largest mode, although most knot locations throughout the range of the covariate were occupied on approximately 1% of the iterations.

The posterior distribution of the spline fit using the Poisson(75) prior assigned probability to more complex models with many more knots. Ninety-five percent of the posterior probability was assigned to models with between 14 and 35 knots, and the posterior median was 23 knots. Most knot locations throughout the range of the covariate were occupied on approximately 2% of the iterations.

The result of using the Poisson(75) prior was that more knots were included in the realization of the spline on each MCMC iteration and so the estimated survival probability as a function of the covariate was less smooth. In fact, the new estimate contained considerable noise and the most frequently occupied knot locations were associated with an anomalous local change in the survival probability at z=1.8. This spike in the survival probability was caused by a chance grouping of individuals all of which were captured on one occasion with covariate values near 1.8 and failed to survive until the next capture occasion. Examination of the true survival status (available from the simulated data) versus the covariate showed exactly the same result. Note that even though the estimated survival probability was far from the truth at some points, the 95% credible intervals still covered the truth at all points.

Posterior summary statistics for the remaining parameters are provided in Table 4.1. Changing the model of the survival probabilities had negligible affect on these parameters,

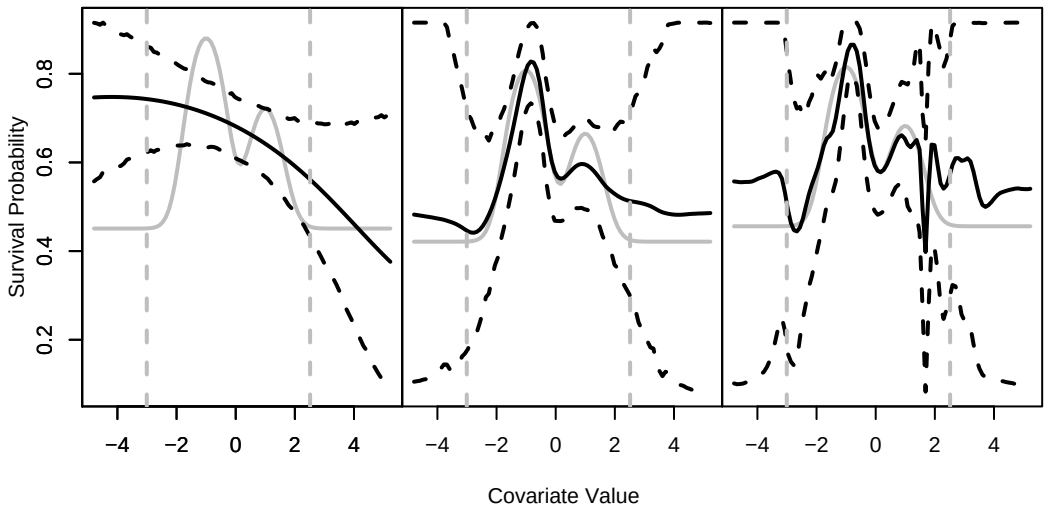

Figure 4.1: Estimated survival probability as a function of the covariate for the simulated data. The left plot illustrates the estimated function assuming a cubic fit, the centre plot using the adaptive spline method with a Poisson(25) prior on the number of knots, and the right plot using the adaptive spline method with a Poisson(75) prior on the number of knots. In each plot the solid grey line indicates the true function, solid black line the pointwise posterior mean fit, and dotted black lines the bounds of the pointwise posterior 95% credible interval. The vertical dotted grey lines indicate the 2.5 and 97.5 percentiles of the simulated covariate values.

and the posterior means and 95% credible intervals produced by all three models were remarkably similar. The reason for this was that many individuals were captured on at least 2 of the 3 capture occasions. Of the 500 individuals, 300 were captured on at least 2 occasions. With this many recaptures, good estimates of the capture probabilities could be obtained from direct comparison of the capture histories and the observed covariate values allowed accurate estimation of μ_1, μ_2, and σ^2, without any adjustment for the covariate's effect on the survival probability.

4.4 Example

Data for the study of reed warblers (*Acrocephalus scirpaceus*) was obtained from the Dutch Constant Effort Sites (CES) banding project (Speek, 2006). In the CES project, volunteer ringers capture birds on 12 day-long visits between April and August to each of 38 sites

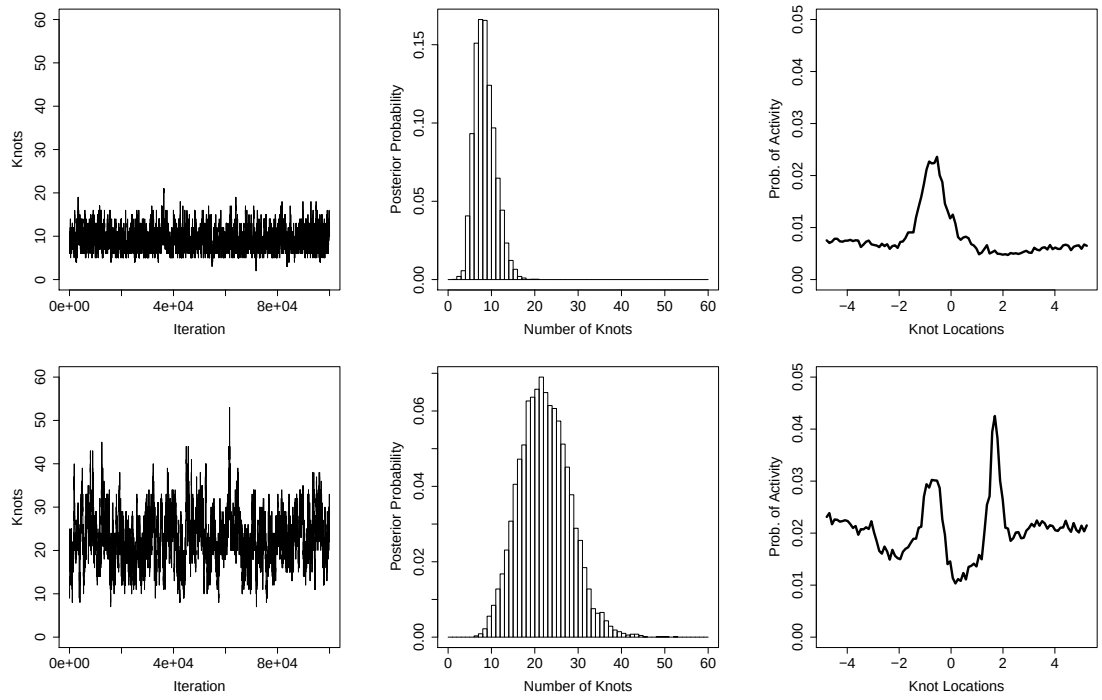

Figure 4.2: Number and locations of knots in the splines fit to the survival probability for the simulated data set. The upper row shows the results using the Poisson(25) prior distribution on the number of knots and the lower row using the Poisson(75) prior. The plots illustrate, from left to right, the number of knots in the spline on each MCMC iteration, the proportion of times different numbers of knots occurred and the proportion times each potential knot location was occupied.

Parameter	True Value	Cubic Polynomial	Poisson(25) Prior	Poisson(75) Prior
p_1	0.85	0.85(0.77,0.91)	0.84(0.76,0.90)	0.84(0.76,0.90)
p_2	0.85	0.84(0.73,0.96)	0.82(0.72,0.93)	0.80(0.70,0.90)
μ_1	0.00	-0.06(-0.23,0.10)	-0.06(-0.23,0.10)	-0.06(-0.23,0.10)
μ_2	0.00	0.04(-0.09,0.17)	0.04(-0.09,0.17)	0.04(-0.09,0.17)
σ	1.00	1.03(0.96,1.10)	1.03(0.96,1.10)	1.03(0.96,1.10)

Table 4.1: Estimates of capture probabilities and parameters of the covariate distribution for the simulated data. Results on the left are from the model assuming a cubic relationship between survival and the covariate and on the right are from the spline models. Values given are the posterior means with 95% HPD credible interval in parentheses.

in Holland. Ringers optionally record demographic and biometric characteristics of the captured birds including age, sex, body mass, wing length and tarsus length. The program was initiated in 1994 and data was available for 10 years up to 2003.

This analysis applies my extension of the CJS model to the final 5 years of data using a measure of the birds' condition as a predictor of survival. Each year of the study was considered as a single capture occasion. Multiple captures of the same bird in one year were combined by collapsing records into a single capture indicator and averaging the biometrics measurements. The condition measure for a single bird in a single year was defined as the ratio of its average observed body mass to its average observed wing length. After removing outliers, the measurements of body mass ranged from 9.9 g to 17.5 g and wing length from 62 mm to 71 mm. The range of the condition measure was .15 g/mm to .24 g/mm.

The CES database contains records of approximately 300,000 captures of 25,000 reed warblers captured between 1999 and 2003. The majority of these birds were observed only once, which was taken as evidence of large numbers of transients in the population, birds that pass through the study sites as they migrate to other breeding grounds. To avoid heterogeneity in the survival probability resulting from the immigration of these transient birds, I used an ad hoc method restricting the data set to birds captured 2 or more times – even if the 2 captures occurred in the same year – and conditioning the each capture history on the bird's second release. Observations of juvenile birds were removed from the data because it was believed that the probability of survival, and its relation to condition, was likely to differ between juveniles and adults. It was also necessary to remove many individuals for whom the condition measure could not be computed in any year of the study because of missing data on body mass or wing length. The final data set for this analysis contained capture histories of 592 birds, with 111 captured in 2 or more years.

As in the simulation study, three different models of the relationship between the survival probability and body condition were tested: a cubic model on the logit scale, and two adaptive spline models with different priors on the number of knots. For the spline models, $\kappa = 100$ potential knot locations were equally spaced between the minimum and maximum observed condition values and the boundary knots were located at .05 and .34 g/mm. The prior distributions on the number of knots, K, were truncated Poisson with $\lambda = 25$ in the first run and truncated Poisson with $\lambda = 75$ in the second. In all three models, I introduced a time dependent intercept term in the linear predictor of $\text{logit}(\phi(z))$ to allow for some change in the survival probabilities over time. The resulting models allowed $\text{logit}(\phi(z))$

to move up or down on each occasion but assumed that the relative effects of any two values of the covariate were constant. In the standard CJS model, the final capture and survival probabilities are completely confounded and only their product is estimable. I found that although $\phi_4(z)$ and p_5 were not completely confounded in my model because of the dependence on z, separate estimates had very low precision. Thus, I only estimated their product $\phi_4(z)p_5$. Body condition was modelled according to diffusion based density in equation (3.1) because the data set was restricted to adult warblers and no systematic growth was expected.

Figure 4.3 compares the survival probabilities estimated from the three different models. Here the fit of the curves was very similar for all three models. Some local effects did appear in the spline estimates, but these occurred at extreme values of body condition where few birds were observed and come at the cost of much lower precision. Indeed, the peak in the posterior mean survival probability at .22 g/mm arose from two birds captured with covariate values near this point. The credible intervals at this point range from approximately .10 to .90 indicating extreme uncertainty and the peak disappeared when the capture histories for these two birds were ignored (results not shown). Figure 4.4 plots the number of knots on each iteration of the MCMC algorithm for the two spline models. As in the simulation, the posterior distribution arising from the truncated Poisson(75) prior placed higher probability on models with more knots, which decreased the smoothness and precision of the estimated survival probabilities as a function of body condition. The 95% credible intervals for all three models overlapped at all points.

Posterior summary statistics for the remaining parameters and a point estimate of survival at the median value of body condition, .17 g/mm, are provided in Table 4.2. Also included are the results of fitting a Bayesian implementation of the CJS model with no effect of the covariate. The results were very similar for all four models. The estimates suggest no significant change in the capture probability over time, though there was a slight decrease in the survival probability at the median condition. Credible intervals of $\phi_t(z)$ were wider for the spline models than for the cubic model, but this does not affect the remaining parameters. The estimates of μ_t are all close to 0 indicating that there was no distinct increase or decrease in the birds' condition over any period. The posterior mean of σ was .012 g/mm for all models. If a birds wing length were fixed at 62(71) , this would translate to a standard deviation in mass between of .74(.85) g between capture occasions.

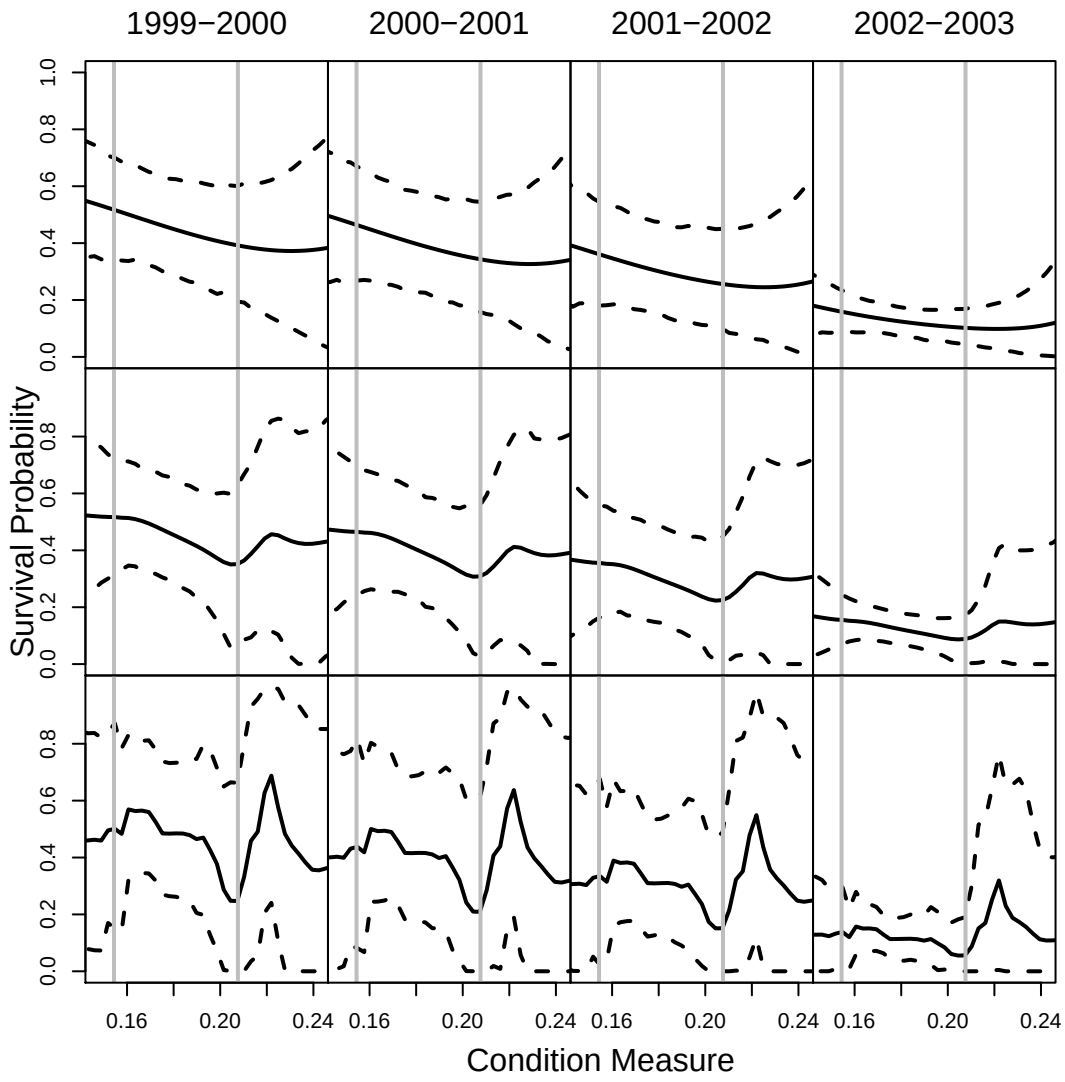

Figure 4.3: Estimated survival probabilities as a function of condition for the reed warbler data. Estimates from the cubic model are shown in the top row, from the spline model with truncated Poisson(25) prior on the number of knots in the middle, and from the spline model with truncated Poisson(75) prior on the bottom. Solid black lines indicate the pointwise posterior mean and dashed lines the bounds of the posterior pointwise 95% credible intervals. In each plot, the vertical grey dashed line indicate the 2.5-th and 97.5-th percentiles of the observed condition values. The plots for 2002-2003 actually estimate the product $\phi_4(z)p_5$ because of the weak identifiability of these parameters separately.

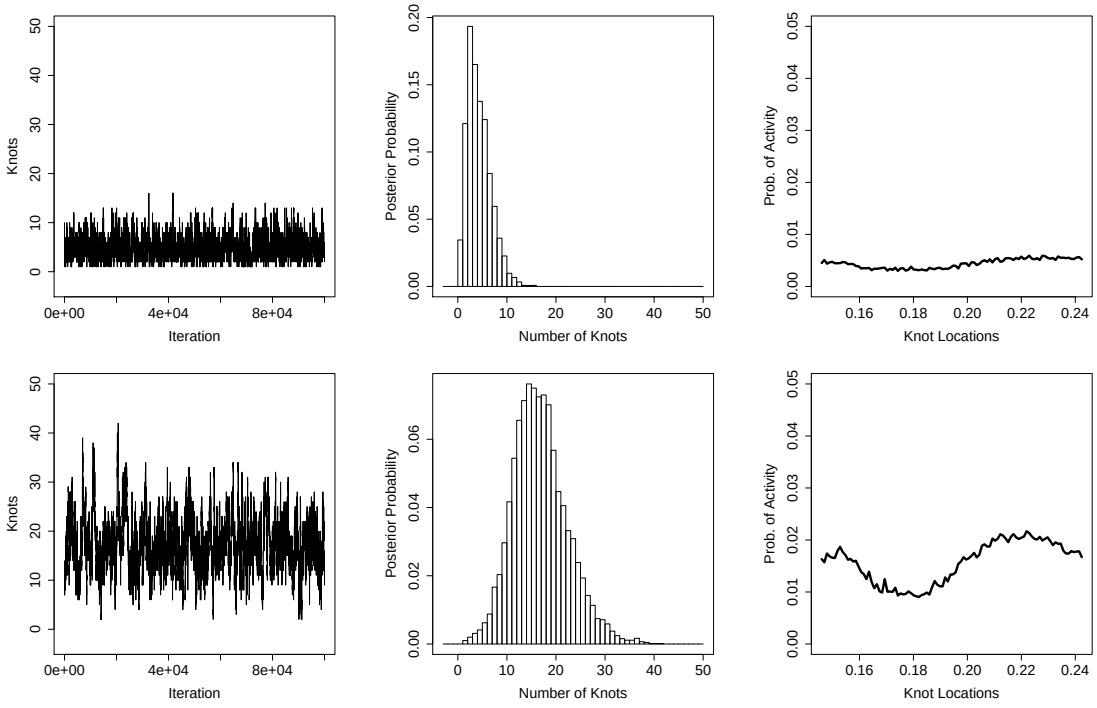

Figure 4.4: Number and locations of knots in the splines fit to the survival probabilities for the reed warbler data. The upper row shows results using the Poisson(25) prior distribution on the number of knots and the lower using the Poisson(75) prior. The plots illustrate, from left to right, the number of knots in the spline on each MCMC iteration, the proportion of times different numbers of knots occurred and the proportion times each potential knot location was occupied.

Quantity	CJS Model	Cubic Polynomial	Poisson(25) Prior	Poisson(75) Prior
$\phi_1(0.17)$	0.48(0.33,0.67)	0.47(0.32, 0.64)	0.48(0.32, 0.68)	0.52(0.32, 0.78)
$\phi_2(0.17)$	0.41(0.27,0.63)	0.42(0.25, 0.60)	0.43(0.26, 0.65)	0.45(0.23, 0.71)
$\phi_3(0.17)$	0.33(0.19,0.55)	0.32(0.17, 0.49)	0.32(0.17, 0.50)	0.34(0.14, 0.58)
$\phi_4(0.17)p_5$	0.13(0.09,0.19)	0.13(0.08, 0.19)	0.14(0.08, 0.19)	0.13(0.06, 0.20)
p_2	0.45(0.29,0.62)	0.46(0.29, 0.64)	0.45(0.29, 0.63)	0.42(0.26, 0.60)
p_3	0.46(0.28,0.66)	0.46(0.28, 0.67)	0.45(0.27, 0.66)	0.43(0.24, 0.65)
p_4	0.46(0.25,0.71)	0.48(0.26, 0.73)	0.48(0.26, 0.72)	0.44(0.23, 0.70)
μ_1	NA	0.000(-0.004,0.004)	0.000(-0.004,0.004)	0.000(-0.004,0.004)
μ_2	NA	0.000(-0.005,0.004)	-0.001(-0.006,0.004)	-0.001(-0.006,0.004)
μ_3	NA	-0.002(-0.007,0.004)	-0.002(-0.007,0.003)	-0.002(-0.007,0.002)
μ_4	NA	-0.002(-0.010,0.007)	-0.001(-0.010,0.007)	-0.003(-0.011,0.005)
σ	NA	0.012(0.010,0.014)	0.012(0.010,0.014)	0.012(0.010,0.014)

Table 4.2: Estimates of capture probabilities and parameters of the covariate distribution for the reed warbler data. Results on the left are from the CJS model assuming no effect of the covariate on survival, then from a model assuming a cubic relationship between survival and the covariate on the logistic scale, from the spline model with Poisson(25) prior on the number of knots, and on the right are from the spline model with Poisson(75) prior on the number of knots. Posterior means with 95% credible intervals are provided for each parameter. Estimated survival probabilities are computed at the median observed value of condition, .17 g/mm. Note that p_5 is almost indistinguishable from $\phi_4(z)$ and their product was estimated as a single parameter.

4.5 Discussion

To my knowledge, only one other lead author has suggested the use of splines to allow more flexibility in modelling a covariate's effect in capture recapture methods. Gimenez, Covas, Brown and Anderson (2006) incorporated splines to model the effect of the southern oscillation index (SOI) on the survival of snow petrels (*Pagodroma nivea*) living in Terre Adèlie, and Gimenez, Crainiceanu, Barbraud, Jenouvrier and Morgan (2006) to model the effect of body mass on the survival of sociable weavers in South Africa. Both applications employ Bayesian inference but differ from my method in two respects. First, there are no missing data in the covariate: SOI is an environmental variable that can be observed regardless of the capture of individual birds, and the body mass for the sociable weavers is defined by a single, static covariate computed by averaging the observations for each bird. The second difference is that both applications make use of penalized splines with fixed number of knots and fixed knot locations, rather than adaptive (free-knot) splines.

Another application of spline methods in analysing capture-recapture data is given by Fewster and Patenaude (2008). The objective was to estimate the distribution of times that southern right whales (*Eubalaena australis*) spend on their breeding grounds in the Auckland Islands by analysing data from multiple sightings of individual whales. Penalized cubic splines were fit to the density function of the residence times to ensure a smooth result without make strong assumptions about the shape of the distribution. In contrast to my method, and the work of Gimenez et al, Fewster applies classical likelihood methods to fit the spline function.

The advantage of adaptive spline methods over penalized spline methods is that by considering the number and location of knots as unknowns, knots can be located only where required to improve fit. One consequence is that the degree of smoothing can vary across the spline by clustering knots where the curve changes most rapidly and placing no knots where the curve is most smooth. The posterior distribution computed from the simulated data set with the Poisson(25) prior assigns low probability of being occupied to most knot locations, except for those in the area of the largest mode which are 2 to 3 times more likely to be occupied. Denison et al. (1998) provides some examples of adaptive spline fits to even more rapidly changing functions with jump discontinuities. In contrast, a penalized spline fit has an upper limit to the flexibility which is determined by the spacing of the fixed knot points. Of course, the added flexibility also increases the potential for overfitting as seen

with the Poisson(75) prior in both the simulation study and the analysis of the reed warbler data.

The Bayesian approach to adaptive spline fitting also provide a natural way to incorporate uncertainty concerning the location of the knots. Classical adaptive splines methods seek a single, best fitting set of knots which are then taken to be fixed. This is akin to selecting a single set of predictor variables in a multiple regression problem and poses two problems. First, different fitting criteria may lead to different models. Second, treating the selected set of knots (predictor variables) as fixed ignores one component of variation when computing the estimates of uncertainty. Bayesian posterior model probabilities naturally rank the different knot configurations, and model averaging using these probabilities to weight the different splines provides a clear way to aggregate inference from the different models. The posterior mean over all configurations of knots will favour the features of the most probable models, but also include some features of the less probable configurations down-weighted accordingly. Credible intervals computed from the entire set of models also accounts for uncertainty in the locations of the knots and will generally be wider than those computed from a single set of knots (Hoeting et al., 1999).

The primary difficulty with the Bayesian adaptive spline method of Biller (2000) lies in selecting the joint prior distribution on the set of models indexed by $\boldsymbol{b}$, $\boldsymbol{\xi}$, and K. As is evident from the examples in this paper, the choice of prior is a very important determinant of the smoothness of the fitted spline. A prior distribution that places too much mass on simple models risks ignoring important aspects of the data, while a prior that favours complex models risks overfitting. Apart from the prior distribution there is no penalty for the model complexity, and overfitting is a serious concern. In simple smoothing of a scatterplot it is possible to choose the priors subjectively and then plot the fitted curve over-top of the raw data to assess the fit. The difficulty in this application is that neither the covariate nor the response are completely observed, so the fit of the spline cannot be visualized directly.

My recommendation is that several prior distributions be selected and the resulting curves compared to see how the fit changes and whether the changes are biologically plausible. For the reed warbler data, the obvious difference between the cubic fit to the survival probability and the spline fits with Poisson(25) and Poisson(75) priors is the peak at .22 g/mm. The size of the peak in the last set of curves in Figure 4.3 is striking, but the point where this occurs is well beyond the 97.5%-ile of the observed covariate values, where

there are little data, and the 95% credible intervals at this point are very wide. Further analysis revealed that the peak was due to 2 birds and disappeared when these individuals were ignored. Moreover, there is no biological reason to believe in an increase in the survival probability at this point. Despite the size of the peak, it seems clear that there is no evidence of a jump in the survival probability at .22 g/mm.

A second subjective decision that must be made in applying my method is the choice of boundary knots, ξ_l and ξ_u. In theory, this choice should have little effect on the fitted curve as long as $[\xi_l, \xi_u]$ encloses the observed data. The challenge is that if $[\xi_l, \xi_u]$ is very wide then the distance from the lower boundary to the first internal knot or from the last knot to the upper boundary will be large. This will lead to small values in the design matrix and the numerical algorithms may become unstable. Conversely, if $[\xi_l, \xi_u]$ is too narrow then it is possible for imputed values of the covariate to lie far outside this range and similar problems can occur. To assess the impact of this choice, I repeated the analysis of the simulated data with several values for of ξ_l and ξ_u and found no effect on the fitted survival probability. I also encountered no numerical problems using my default choices for the boundary knots.

One source of confusion with the adaptive-spline method might be the apparent discrepancy between the prior and posterior distribution on the number of knots. In both the simulation and example, the posterior distribution concentrates its mass on much simpler models than the prior distribution. This might seem contradictory, but exactly the same behaviour appears in the original examples of Biller (2000). In his discussion of another paper on Bayesian adaptive splines, Holmes (2002) explains that the apparent discrepancy is a result of the Bartlett-Lindley paradox. Ignoring the prior distribution on $\boldsymbol{\xi}$ and K, the vague multivariate-normal prior on the spline coefficients, $\pi(\boldsymbol{b}|\boldsymbol{\xi}, K)$, induces a prior predictive distribution for the data whose variance increases with the dimension of $\boldsymbol{b}$. As a result, any observed data has lower prior probability under more complex models and the prior distribution actually places less and less mass on models of higher and higher complexity.

The key in the adaptive spline model is that the prior on $\boldsymbol{b}|\boldsymbol{\xi}, K$ which favours simple models is partially offset by the Poisson prior on K which assigns very little mass to these models – when λ is large enough. The resulting prior distribution is a balance that assigns its mass to models simpler than those favoured by $\pi(K)$ alone. In essence, there is no discrepancy; rather, the prior on the set of models has to be interpreted through the full joint distribution, $\pi(\boldsymbol{b}, \boldsymbol{\xi}, K)$, and not simply the marginal prior $\pi(K)$.

Another caution with my approach, and with methods incorporating time-varying covariates in general, is the amount of data needed to provide adequate estimates of the survival probabilities. The final data set for the reed warblers contained 592 animals, but only 111 were captured on two or more occasions. Further, the condition measure was recorded two or more times for only 77 birds. This provides very little information regarding how the covariate changes over time and how differences in condition might affect the birds' survival. Fitting a Bayesian implementation of the standard CJS model to the data (assuming time-dependent survival and capture probabilities and ignoring any effect of the covariate) yields credible intervals for the survival probabilities with widths between .34 and .36. In light of this uncertainty when ignoring the covariate, it seems unlikely that any model will be able to detect an effect of condition on survival.

Based on my experience, including time-dependent covariates in the CJS model requires more data (i.e. capture of more individuals at given capture and survival rates) than models assuming homogeneity of individuals, or including environmental and fixed individual predictors. Using splines to model the survival probability as a function of the covariate will require even more data. Whereas a parametric curve borrows information from all values of the covariate to estimate the survival probability at any single point, the spline only uses information from a local neighbourhood of covariate values. The result is that if few birds are observed with values in a given range of the covariate then the survival probabilities estimated from the spline model will be highly variable in that range, though the estimates from a parametric model may still be very precise.

I did not examine the goodness-of-fit of the extended CJS model, but this could again be done by computing Bayesian p-values. Given that the method specifically addresses the issue of allowing the survival probabilities to vary as a function of the continuous covriate, the most important component to check is the model of the survival probabilities. However, it is difficult to define a measure of discrepancy that assesses the fit of the survival probabilities alone. For example, the discrepancy:

$$D_1(\boldsymbol{X}, \boldsymbol{\Theta}) = \sum_{i,t:\omega_{it}=1} \left(1(\omega_{i,t+1} = 1) - \sqrt{\phi_{it} p_{t+1}}\right)^2$$

is the Tukey-Freedman statistic comparing each $\omega_{i,t+1}$ with its expected value, given that the individual was alive and captured on the previous occasion ($\omega_{it} = 1$). This measure depends on the adequacy of both the survival probabilities and the capture probabilities.

A discrepancy measure that depends on the capture probabilities alone can be defined by considering only the capture occasions when each individual is known to be alive. Let A_t denote the set of individuals known to be alive on occasion t because $\omega_{it_1} = 1$ for some $t_1 < t$ and $\omega_{it_2} = 1$ for some $t_2 > t$. The discrepancy:

$$D_2(\boldsymbol{X}, \boldsymbol{\Theta}) = \sum_{t=2}^{T-1} \left(\sqrt{\sum_{i \in A_t} 1(\omega_{it} = 1)} - \sqrt{\sum_{i \in A_t} p_t} \right)^2$$

compares the observed and expected numbers of individuals from this set captured on occasion t, and depends only on the parameter p_t. One approach to assessing the adequacy of the model of the survival probabilities would be to compute a Bayesian p-value based on $D_2(\boldsymbol{X}, \boldsymbol{\Theta})$ to assess the adequacy of the modelled capture probabilities, and then compute a second Bayesian p-value based on $D_1(\boldsymbol{X}, \boldsymbol{\Theta})$. If the p-value computed from $D_1(\boldsymbol{X}, \boldsymbol{\Theta})$ indicated poor fit but the p-value computed from $D_2(\boldsymbol{X}, \boldsymbol{\Theta})$ did not, then this would suggest that the model of the survival probabilities needs to be more complicated possibly including further covariates or group effects.

As a final note, I address the removal of transient birds from the CES reed warbler data. The majority of individuals captured were never recaptured and it is likely that many of these individuals were passing through the capture sites while migrating to other locations. The apparent survival probability of these individuals would be 0. Pradel et al. (1997) developed a model to account for transient individuals in the standard CJS model and compared it to the *ad hoc* method of conditioning on second release that I employed. They found that the *ad hoc* method produced unbiased estimates of the survival probabilities and was almost as efficient as the more complicated model when capture probabilities are high. With capture probabilities of .4 they found relative efficiency greater than .8.

A side-effect of conditioning on the second release in my application may be filtering of the values of the covariate. Resident individuals with values that equate to low survival probabilities will have less chance of being recaptured and more chance to be mistaken for transients and removed from the analysis. This should not bias the estimates of survival probability, but will decrease the precision of the estimates where these individuals are removed. The analysis is not intended to be an exhaustive examination of the reed warbler data, and how to deal with transients properly remains an open question.

Chapter 5

Conclusion

The three projects in this thesis all consider problems in modelling capture-recapture data from populations in which the parameters (like catchability and survival) vary among individuals. In Chapter 2, the variation arises because individuals captured on different days may have different capture probabilities and may take different amounts of time to move between the two trapping locations. The challenge was to allow for these differences without introducing large numbers of extra parameters that would be difficult to estimate from sparse data. In Chapters 3 and 4, the capture and/or survival probabilities vary as functions of a continuous covariate that changes over time. The challenge in Chapter 3 was to estimate the population size while accounting for this variability. In Chapter 4 my objective was to develop a model that allowed for the variation but made minimal assumptions about the relationship between the covariate and the individual survival rates. Table 5.1 provides a brief summary of the experimental design, my contributions, and the related work for each of these projects. I will conclude by comparing and contrasting some of the similar methods used in the different projects, discussing some common problems and suggesting some directions for future work.

Chapters 2 and 4 both made use of splines to smooth the effect of a predictor, but did so in very different ways. In Chapter 2, I chose a penalized spline method which fixes the number of knots and their locations before the analysis, and I implemented an adaptive approach that includes placement of the knots as part of fitting the model in Chapter 4. One advantage of the adaptive spline method over the penalized method is that the smoothness of the curve could vary across the range of the predictor variable. In the simulation of Section 4.3, knots clustered at the highest mode where the true curve changed most quickly providing

more flexibility at this point. Penalized splines can also display varying smoothness if the penalty – the prior on the coefficients in the Bayesian framework – is allowed to change over the range of the data. Lang and Brezger (2004) suggest an extension of their model that weights the prior variance for each coefficient with random terms to allow different penalties (one could even smooth these weights so the smoothness changed smoothly). The clear disadvantages of the adaptive method are the difficulty in specifying the prior on the model space (number of knots) and the computation needed to sample from the posterior distribution. The shape of the fitted curve in both the simulation and application of Chapter 4 depended greatly on the prior distribution for the number of knots, and priors favouring large numbers of knots generated anomalous results. The RJMCMC algorithm was also more difficult to implement and required more computing time than the MCMC algorithms for fitting the penalized spline model.

Although I prefer the penalized spline approach based on my experience in these two projects, I think that there is a correspondence between the two methods. Suppose that I fit a Bayesian P-spline with many knots but specify the prior on the differences in the coefficients as a mixture of two distributions: a point mass at 0 with high probability and otherwise a normal distribution as in Lang and Brezger (2004). This is equivalent to specifying, a priori, that most knots have no effect on the shape of the spline but allowing non-zero effects when needed to fit the data, which is exactly the philosophy of the adaptive spline method. I believe that unifying the two approaches would provide insight into the best method for fitting splines and choosing the prior distributions, and I intend to explore this in future research.

In Chapters 2 and 3, I compared different models of the same data with the DIC whereas in Chapter 4 model selection (choosing the set of knots) was performed by RJMCMC. The biggest advantage of the DIC is the simplicity of the computation, requiring nothing further than the output of the MCMC runs for each model. It selected the true model for the great majority of data sets in the simulation of Chapter 1 and p_D provided a useful summary for comparing the complexity of the hierarchical models. DIC would be difficult to use if the model space were very large, e.g. many different predictors with possible interactions, as this would require MCMC sampling separately for each model. The method also provides no basis for model averaging and this seems like a severe disadvantage.

On the other hand, RJMCMC combines sampling from the posterior distributions of the possible models and model selection in one procedure so that only one chain need be run

regardless of the number of models being considered. The algorithm also naturally produces results averaged over the different models. My main concern with this method is that it seems impossible to know whether the chain has adequately explored the model space when many candidate models never occurred in the realizations of the chain. I did test several chains starting from different initial values when modelling the reed warbler data in section 4.4, and all chains converged to the same area of the sample space. But, many configurations of knots were never realized, and it is impossible to be certain that no better model existed. Bayesian model selection, particularly for complex, hierarchical models, seems like an area that requires more development.

One concern I have with the methods of both Chapters 2 and 3 is that the quantity of interest, the total population size, does not occur in the likelihood and cannot be modelled directly. Instead, my methods model the number of unmarked individuals alive in each strata or on each occasion and treats estimates of the total population size as derived parameters. The same problem occurs in any model that conditions on the number of unmarked individuals as fixed but unknown values. One solution for both projects is to model the number of unmarked individuals further, but this requires extra assumptions and introduces more parameters to estimate. In Chapter 2, I could have applied the P-spline to smooth the population size in each strata at the first location prior to marking rather than modelling the number of unmarked fish at the second location. However, I would then have had to model the capture and marking of fish at the first trapping location and also make assumptions about the movements of the unmarked fish between the two locations. In Chapter 3, I could have based my method on the model of Schwarz and Arnason (1996), but this would have required modelling the value of the covariate prior to each individual's first capture instead of conditioning on the value when each individual was first captured.

As discussed in section 3.5, Bayesian melding may offer an alternative solution. In effect, Bayesian melding provides a technique for resampling the realizations from the posterior distribution of a Bayesian model to accommodate prior beliefs about quantities that don't occur in the model likelihood but are produced as derived output. This may provide a way to incorporate prior information about the total population size in both models while still conditioning on the number of unmarked individuals as fixed values in the likelihood. For either model, a sample would be generated from the posterior distribution initially assuming a non-informative prior on the number of unmarked individuals and imposing no constraints on the total population size. The realizations in the sample would then be

reweighted according to the beliefs about the total population size. In applying Bayesian melding to smoothing data from the Stratified-Petersen experiment, realizations from the posterior that produced a smooth fit to the total run size as a function of time would be weighted more heavily than realizations that produced a very rough fit. In the JS model with a continuous covariate, the realizations would be weighted according to the researchers own prior beliefs about the population size on each capture occasion. The challenge I foresee is that if the prior distribution on the numbers of unmarked individuals is not consistent with the prior distribution on the total population size then the weights will be small for most realizations sampled from the posterior the method will be very inefficient. However, this method would have general implications in capture-recapture analysis beyond the models presented in my projects and is certainly a topic for future research.

Another important question is how the inferences about the population size in Chapter 3 and survival probabilities in Chapter 4 might be affected if distribution of the covariate for the uncaptured individuals is very different from the distribution for the individuals captured at least one time. Two possible scenarios can occur: 1) there is a segment of the population of interest that has very low, but positive, capture probabilties and 2) there is a segment of the population of interest that is not catchable. In the first scenario, the inference about the population size or survival probability will be unbiased, but will have very low precision. Suppose in the method of Chapter 3, for example, that there is a single individual in the population at time t that has a very, very small capture probability. If this individual is captured then the inverse of its capture probability will be very large and the estimated population size will be very large. If it is not captured, then it will not contribute to the estimate of population size at all, and the estimate will be much lower. The posterior mean will provide adequate inference about the population size on average, but will be highly variable. The effect is similar to the problem of estimating the weight of a group of elephants from the weight of a single elephant via the Horvitz-Thompson estimator as described by Basu (1971).

Situations where some individuals have zero capture probability can arise in fisheries experiments if the capture equipment is size selective. For example, trawl nets will only capture individuals of a certain length or higher. Assuming that the capture probability was constant, as in Chapter 4, or a linear function of the covariate on the logistic scale, as in Chapter 3, would completely ignore this change in the capture probability. The resulting estimates of the survival probability would not apply to the individuals with zero

capture probability and estimates of the population size would not include these individuals. However, if these individuals are truly not catchable then it is not possible to learn about this segment of the population. My view is that this is not a problem with the methods, but rather a problem in defining the population of interest. If one is truly interested in learning about the entire population, then the capture mechanism has to sample from the entire population. If this is not the case, then inference must be restricted to only the catchable segment of the population.

How to account for individual variation has long been an important question in modelling data from capture-recapture experiments. Successive models have allowed for more complex types of variation but have also required more sophisticated methods to estimate parameters. Early models of the stratified, two-sample, closed population experiment completely separated the estimation of capture probabilities and population size for each strata (Darroch, 1961). The models produced closed form MLEs in some cases, but the number of parameters was too high for estimates to be computed from sparse data sets. Schwarz and Dempson (1994) reduced the number of parameters by modelling the transition probabilities as functions of log-normal travel times and Mantyniemi and Romakkaniemi (2002) further added a Bayesian hierarchical structure to share information about the parameters in different strata. Fitting these models requires numerical optimization of the likelihood or algorithms to sample from the posterior distribution that cannot be done by hand.

Lebreton et al. (1992) developed a general framework for modelling the capture and survival probabilities of individuals in an open population as functions of known predictor variables, and McDonald and Amstrup (2001) extended these methods to the estimation of population size. The models of Schwarz et al. (1993) and Brownie et al. (1993) further allowed for categorical predictors that changed over time, and Bonner and Schwarz (2006) developed a Bayesian method for modelling the effects of continuous, time-dependent covariates. The three projects in this thesis build on these methods by explicitly modelling the temporal structure of data from stratified-Petersen experiments, by adding estimation of the population size to the model of Bonner and Schwarz (2006), and by allowing more flexibility in the relationship between a continuous predictor and the survival probabilities. Bayesian methods provide a convenient framework to make inference from the complex models that result.

	Population Structure, Source of Variation	Objective	Data Collection	Contributions	Related Literature
1	Closed, Temporal stratification	Total population size	Paired release-recapture	Penalized spline smoothing of daily population size	Schwarz and Dempson (1994), Mantyniemi and Romakkaniemi (2002), Lang and Brezger (2004)
2	Open, Cts. covariate (z_{it})	Population size on each occasion	Multiple mark-release-recapture	Dependence of capture/survival on cts. covariate; Modelling covariate with LVB growth model; Bayesian adaptation of Horvitz-Thompson estimator	Lebreton et al. (1992), McDonald and Amstrup (2001), Bonner and Schwarz (2006)
3	Open, Cts. covariate (z_{it})	Survival probability	Multiple mark-release-recapture	Semiparametric fitting of survival prob. as function of z_{it}; Bayesian adaptive spline fitting	Lebreton et al. (1992), Bonner and Schwarz (2006), Gimenez et al. (2006)

Table 5.1: Summary of contributions for each project. The first columns describe the structure of the population, the primary objective of the analysis, and the data collection method. The final two columns summarize the contributions in this thesis and related works.

Bibliography

Amstrup, S. C., McDonald, T. L. and Manly, B. J. F., editors (2003). *Handbook of Capture-Recapture Analysis*. Princeton University Press, Princeton, NJ.

Arnason, A. N. (1973). The estimation of population size, migration rates, and survival in a stratified population. *Research in Population Ecology* **15**, 1–8.

Banneheka, S. G., Routledge, R. D. and Schwarz, C. J. (1997). Stratfied two-sample tag-recovery census of closed populations. *Biometrics* **53**, 1212–1224.

Basu, D. (1971). An essay on the logical foundations of survey sampling, part 1. In Godambe, V. P. and Sprott, D. A., editors, *Foundations of Statistical Inference*, pages 203–242. Holt, Rinehart and Winston, Toronto.

Biller, C. (2000). Adaptive Bayesian regression splines in semiparametric generalized linear models. *Journal of Computational and Graphical Statistics* **9**, 122–140.

Bonner, S. J. and Schwarz, C. J. (2006). An extension of the Cormack-Jolly-Seber model for continuous covariates with application to *Microtus pennsylvanicus*. *Biometrics* **62**, 142–149.

Brooks, S. P., Catchpole, E. A. and Morgan, B. J. T. (2000). Bayesian animal survival estimation. *Statistical Science* **15**, 357–376.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**, 434–455.

Brownie, C., Hines, J. E., Nichols, J. D., Pollock, K. H. and Hestbeck, J. B. (1993). Capture-recapture studies for multiple strata including non-markovian transitions. *Biometrics* **49**, 1173–1187.

Carothers, A. D. (1973). The effects of unequal catchability on Jolly-Seber estimates. *Biometrics* **29**, 79–100.

Catchpole, E. A., Morgan, B. J. T., Coulson, T. N., Freeman, S. N. and Albon, S. D. (2000). Factors influencing soay sheep survival. *Applied Statistics* **49**, 453–472.

Catchpole, E. A., Morgan, B. J. T. and Tavecchia, G. (2008). A new method for analysing discrete life history data with missing covariate values. *Journal of the Royal Statistical Soceity, Series B* **70**, 445–460.

Chapman, D. G. and Junge, Jr, C. O. (1956). The estimation of the size of a stratified animal population. *The Annals of Mathematical Statistics* **27**, 375–389.

Chen, M.-H., Shao, Q.-M. and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer, New York.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician* **49**, 327–335.

Cormack, R. M. (1964). Estimates of survival from the sighting of marked animals. *Biometrika* **51**, 429–438.

Coulson, T., Catchpole, E. A., Albon, S. D., Morgan, B. J. T., Pemberton, J. M., Clutton-Brock, T. H., Crawley, M. J. and Grenfell, B. T. (2001). Age, sex, density, winter weather, and population crashes in soay sheep. *Science* **292**, 1528–1531.

Darroch, J. N. (1961). The two-sample capture-recapture census when tagging and sampling are stratified. *Biometrika* **48**, 241–260.

De Boor, C. (1978). *A Practical Guide to Splines*. Springer-Verlag, New York.

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society, Series B* **60**, 333–350.

DiMatteo, I., Genovese, C. R. and Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* **88**, 1055–1071.

Dupuis, J. A. and Schwarz, C. J. (2007). A Bayesian approach to the multistate Jolly-Seber capture-recapture model. *Biometrics* **63**, 1015–1022.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89–121.

Fewster, R. M. and Patenaude, N. J. (2008). Cubic splines for estimating the distribution of residence time using individual resightings data. In Thomson, D. L., Cooch, E. G. and Conroy, M. J., editors, *Modeling Demographic Processes in Marked Populations*, volume 3 of *Environmental and Ecological Statistics*, pages 393–416, Berlin. Springer.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing* **7**, 57–68.

Gelfand, A. E. (2000). Gibbs sampling. *Journal of the American Statistical Association* **95**, 1300–1304.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 515–533.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman and Hall, New York, 2nd edition.

Gelman, A., Meng, X.-L. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6**, 733–807.

Gilbert, R. O. (1973). Approximations of the bias in the Jolly-Seber capture-recapture model. *Biometrics* **29**, 501–526.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). Introducing Markov chain Monte Carlo. In Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., editors, *Markov chain Monte Carlo in Practice*, chapter 1, pages 1–19. Chapman and Hall, New York.

Gimenez, O., Covas, R., Brown, C. R. and Anderson, M. D. (2006). Nonparametric estimation of natural selection on a quantitative trait using mark-recapture data. *Evolution* **60**, 460–466.

Gimenez, O., Crainiceanu, C., Barbraud, C., Jenouvrier, S. and Morgan, B. J. T. (2006). Semiparametric regression in capture-recapture modelling. *Biometrics* **62**, 691–698.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–7332.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science* **14**, 382–417.

Holmberg, J., Norman, B. and Arzoumanian, Z. (2008). Robust, comparable population metrics through collaborative photo-monitoring of whale sharks rhincodon typus. *Ecological Applications* **18**, 222–233.

Holmes, C. (2002). Discussion of: Spline adaptation in extended linear models. *Statistical Science* **17**, 22–24.

Hwang, W.-D. and Chao, A. (1995). Quantifying the effects of unequal catchabilities on Jolly-Seber estimators via sample coverage. *Biometrics* **51**, 128–141.

Jolly, G. M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika* **52**, 225–247.

Jupp, D. L. B. (1978). Approximation to data by splines with free knots. *Siam Journal of Numerical Analysis* **15**, 328–343.

King, R. and Brooks, S. P. (2001). On the bayesian analysis of population size. *Biometrika* **88**, 317–336.

King, R., Brooks, S. P., Morgan, B. J. T. and Coulson, T. (2006). Factors influencing soay sheep survival: A Bayesian analysis. *Biometrics* **62**, 211–220.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* **13**, 183–212.

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

Lee, P. M. (2004). *Bayesian Statistics: an introduction*. Hodder Arnold, London, UK, third edition.

Link, W. A. and Barker, R. J. (2005). Modeling association among demogaphic parameters in analysis of open population capture-recapture data. *Biometrics* **61**, 46–54.

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

Mantyniemi, S. and Romakkaniemi, A. (2002). Bayesian mark-recapture estimation with an application to a salmonid smolt population. *Canadian Journal of Fisheries and Aquatic Science* **59**, 1748–58.

McDonald, T. L. and Amstrup, S. C. (2001). Estimation of population size using open capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics* **6**, 206–220.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics* **22**, 1142–1160.

Nichols, J. D., Sauer, J. R., Pollock, K. H. and Hestbeck, J. B. (1992). Estimating transition probabilities for stage-based population projection matrices using capture-recapture data. *Ecology* **73**, 306–312.

North, P. M. and Morgan, B. J. T. (1979). Modelling heron survial using weather data. *Biometrics* **35**, 667–681.

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

Pawlowsky-Glahn, V. and Egozcue, J. (2006). Compositional data and their analysis: an introduction. *Geological Society, London, Special Publications* **264**, 1–10.

Plante, N., Rivest, L.-P. and Tremblay, G. (1998). Stratified catpure-recapture estimation ofthe size of a closed population. *Biometrics* **54**, 47–60.

Pledger, S. and Efford, M. (1998). Correction of bias due to heterogeneous capture probability in capture- recapture studies of open populations. *Biometrics* **54**, 888–898.

Pledger, S., Pollock, K. H. and Norris, J. L. (2003). Open capture-recapture models with heterogeneity: I. Cormack-Jolly-Seber model. *Biometrics* **59**, 786–794.

Pollock, K. H. (1982). A capture-recapture design robust to unequal probability of capture. *Journal of Wildlife Management* **46**, 752–757.

Pollock, K. H. and Alpizar-Jara, R. (2005). Classical open-population capture-recapture models. In Amstrup, S. C., McDonald, T. L. and Manly, B. J. F., editors, *Handbook of Capture-Recapture Analysis*, chapter 3, pages 36–57. Princeton University Press, Princeton, NJ.

Poole, D. (2002). Bayesian estimation of survival from mark-recapture data. *Journal of Agricultural, Biological, and Environmental Statistics* **7**, 264–276.

Poole, D. and Raftery, A. E. (2000). Inference for deterministic simulation models: The Bayesian melding approach. *Journal of the American Statistical Association* **95**, 1244–1255.

Pradel, R., Hines, J. E., Lebreton, J.-D. and Nichols, J. D. (1997). Capture-recapture survival models taking account of transients. *Biometrics* **53**, 60–72.

Quinn, II, T. J. and Deriso, R. B. (1999). *Quantitative Fish Dynamics*. Oxford University Press, Inc., New York.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**, 179–191.

Royle, J. A., Dorazio, R. M. and Link, W. A. (2007). Analyis of multinomial models with unknown index using data augmentation. *Journal of Computational and Graphical Statistics* **16**, 67–85.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge, UK.

Schaeffer, M. B. (1951). Estimation of the size of animal populations by marking experiments. *U.S. Fish and Wildlife Services Bulletin* **69**, 191–203.

Schumaker, L. L. (1993). *Spline Functions: Basic Theory*. John Wiley & Sons, Inc., New York, 2nd edition.

Schwarz, C. J. and Arnason, A. N. (1996). A general methodology for the analysis of capture-recapture experiments in open populations. *Biometrics* **52**, 860–873.

Schwarz, C. J., Arnason, A. N. and Kirby, C. W. (2008). The siren song of the Schaefer estimator – no better than a pooled Petersen. Unpublished work.

Schwarz, C. J. and Dempson, J. B. (1994). Mark-recapture estimation of a salmon smolt population. *Biometrics* **50**, 98–108.

Schwarz, C. J., Schweigert, J. F. and Arnason, A. N. (1993). Estimating migration rates using tag-recovery data. *Biometrics* **49**, 177–193.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Seber, G. A. F. (1965). A note on the multiple recapture census. *Biometrika* **52**, 249–259.

Seber, G. A. F. (1982). *The Estimation of Animal Abundance and Related Parameters*. Macmillan, New York, 2nd edition.

Speek, H. G. (2006). CES handleiding. Technical report. `http://www.vogeltrekstation.nl/ces_handleiding.htm`.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* **64**, 583–640.

Thomas, A., O'Hara, B., Ligges, U. and Stutz, S. (2006). Making BUGS open. *R News* **6**, 12–17.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics* **22**, 1701–1762.

Waagepetersen, R. and Sorensen, D. (2001). A tutorial on reversible jump MCMC with a view toward applications in QTL-mapping. *International Statistical Review* **69**, 49–61.

Williams, B. K., Nichols, J. D. and Conroy, M. J. (2002). *Analysis and Management of Animal Populations*. Academic Press, San Diego.

Appendix A

BUGS Code for the Models of Chapter 2

A.1 Diagonal Bayesian P-spline Model with Error

```
model diagonal{
    ##### Prior distributions ######
    ## Run size and capture probabilities
    for(i in 1:s){
          etaU[i] <- inprod(Z[i,],bU[]) + eU[i]
          eU[i] ~ dnorm(0,taueU)
          etaP[i] ~ dnorm(xiP,tauP)
    }

    ##### Hyperpriors #####
    ## Run size
    bU[1] ~ dflat()
    bU[2] ~ dflat()

    for(i in 3:K){
          xiU[i-2] <- 2*bU[i-1] - bU[i-2]
          bU[i] ~ dnorm(xiU[i-2],tauU)
    }
    tauU ~ dgamma(1,.0005)
    sigmaU <- 1/sqrt(tauU)

    taueU ~ dgamma(1,.0005)
```

```
    sigmaeU <- 1/sqrt(taueU)

    ## Capture probabilities
    xiP ~ dnorm(-2,.666)
    tauP ~ dgamma(.001,.001)
    sigmaP <- 1/sqrt(tauP)

    ##### Prior Distributions #####
    ## Capture probabilities and run size
    for(i in 1:s){
        logit(p[i]) <- etaP[i]
        U[i] <- round(exp(etaU[i]))
    }


    ##### Likelihood contributions #####
    for(i in 1:s){
          ## Marked fish
          m[i] ~ dbin(p[i],n[i])

          ## Unmarked fish
          u[i] ~ dbin(p[i],U[i])
    }

    ##### Derived Parameters #####
    ## Total number of unmarked fish
    Utot <- sum(U[1:s])

    ## Total population size
    N <- sum(n[1:s]) + Utot
}
```

A.2 Non-diagonal Bayesian P-spline Model with Error

```
model non-diagonal{
    ##### Prior distributions ######
    ## Run size and capture probabilities
    for(j in 1:t){
          etaU[j] <- inprod(Z[j,],bU[]) + eU[j]
          eU[j] ~ dnorm(0,taueU)
```

```
      etaP[j] ~ dnorm(xiP,tauP)
}

## Mean and sd of log travel-times
for(i in 1:s){
      etaMu[i] ~ dnorm(xiMu,tauMu)
      etaSd[i] ~ dnorm(xiSd,tauSd)
}

##### Hyperpriors #####
## Run size
bU[1] ~ dflat()
bU[2] ~ dflat()

for(i in 3:K){
      xiU[i-2] <- 2*bU[i-1] - bU[i-2]
      bU[i] ~ dnorm(xiU[i-2],tauU)
}
tauU ~ dgamma(1,.0005)
sigmaU <- 1/sqrt(tauU)

taueU ~ dgamma(1,.05)
sigmaeU <- 1/sqrt(taueU)

## Capture probabilities
xiP ~ dnorm(-2,.666)
tauP ~ dgamma(.001,.001)
sigmaP <- 1/sqrt(tauP)

## Mean and sd of log travel times
xiMu ~ dnorm(0,.0625)
tauMu ~ dgamma(.001,.001)
sigmaMu <- 1/sqrt(tauMu)

xiSd ~ dnorm(0,.0625)
tauSd ~ dgamma(.001,.001)
sigmaSd <- 1/sqrt(tauSd)

##### Compute derived parameters #####
## Mean and sd of log travel times
for(i in 1:s){
    mu[i] <- etaMu[i]
```

```
    log(sd[i]) <- etaSd[i]
}

## Capture probabilities and run size
for(j in 1:t){
    logit(p[j]) <- etaP[j]
    U[j] <- round(exp(etaU[j]))
}

## Transition probabilities
for(i in 1:s){
    # Probability of transition in 0 days (T<.5)
    Theta[i,i] <- phi((log(1)-mu[i])/sd[i])


    for(j in (i+1):t){
    # Probability of transition in j days (j-1<T<j)
        Theta[i,j] <- phi((log(j-i+1)-mu[i])/sd[i])
                - phi((log(j-i)-mu[i])/sd[i])
    }

    Theta[i,t+1] <- 1-sum(Theta[i,i:t])
}

##### Likelihood contributions #####
## Marked fish
for(i in 1:s){
      # Compute cell probabilities
      for(j in i:t){
              Pmarked[i,j] <- Theta[i,j] * p[j]
          }
      Pmarked[i,t+1] <- 1- sum(Pmarked[i,i:t])

      # Likelihood contribution
      m[i,i:(t+1)] ~ dmulti(Pmarked[i,i:(t+1)],n[i])
}

## Unmarked and total number of fish
for(j in 1:t){
      # Likelihood contribution
      u[j] ~ dbin(p[j],U[j])
}
```

```
    ##### Derived Parameters #####
    ## Total number of unmarked fish
    Utot <- sum(U[1:s])

    ## Total population size
    N <- sum(n[1:s]) + Utot
}
```

Appendix B

BUGS Code for the Models of Chapter 3

```
model JSCov{
    #### Likelihood #####
    for(i in 1:n){

        ## Conditional model of covariate.
        for(t in (a[i]+1):ncap){
              mu.z[i,t] <- z.inf*(1-exp(-z.rate)) + Z[i,t-1]*exp(-z.rate)
              Z[i,t] ~ dnorm(mu.z[i,t],tau)
        }

        ## Compute capture prob
        for(t in a[i]:ncap){
              logit(p[i,t]) <- bp0[t] + bp1 * Z[i,t]
              q[i,t] <- 1-p[i,t]
        }

        ## Compute survival prob
        for(t in a[i]:ncap){
              logit(phi[i,t]) <- bphi0[t] + bphi1 * Z[i,t]
        }
```

```
## Compute chi term.
chi[i,ncap] <- 1
for(t in 1:(ncap-b[i])){
      chi[i,ncap-t] <- (1-phi[i,ncap-t]) +
            phi[i,ncap-t] * (1-p[i,ncap-t+1]) * chi[i,ncap-t+1]
}

## Model captures between first and last observation.
for(t in (a[i]+1):b[i]){
      W[i,t] ~ dbern(p[i,t])
}

## Model survival between first and last observations.
for(t in a[i]:(b[i]-1)){
      dummy1[i,t] <- 1
      dummy1[i,t] ~ dbern(phi[i,t])
}

## Model last capture.
dummy2[i] <- 1
dummy2[i] ~ dbern(chi[i,b[i]])

## Compute contributions to estimates of number of
## unmarked individuals.
Ui.tmp[i] ~ dnegbin(p[i,a[i]],1)
Ui[i,a[i]] <- Ui.tmp[i] + 1

## Compute contribution to number of unmarked individuals
## alive on both occasions t-1 and t.
## (Work around to avoid binomial of size 0)
Ui.tmp.tmp[i] <- max(1,Ui.tmp[i])
PhiUi.tmp[i] ~ dbin(phi[i,a[i]],Ui.tmp.tmp[i])
PhiUi[i,a[i]+1] <- PhiUi.tmp[i] * step(Ui.tmp[i]-1)

## Compute probability that an individual was alive and
## marked on each occasion after it was last captured and
## simulate survival indicator.
for(t in (b[i]+1):ncap){
      PM[i,t] <- (equals(t,a[i]+1) + Mi[i,t-1]) *
                (1-(1-phi[i,t-1])/chi[i,t-1])

      Mi[i,t] ~ dbern(PM[i,t])
```

```
  }
}

##### Compute derived parameters #####
## 1) Population size

for(t in 1:ncap){
      M[t] <- sum(Mi[,t])
      U[t] <- sum(Ui[,t])
      N[t] <- M[t] + U[t]
}

## 2) Births
B[1] <- N[1]

for(t in 2:ncap){
      ## Estimate number of individuals captured on occasion t
      ## or previously that are alive on both occasions t and t+1

      for(i in 1:n){
          PhiMi[i,t] <- (equals(a[i],t-1) + equals(Mi[i,t-1],1)) *
                        equals(Mi[i,t],1)
      }
      PhiM[t] <- sum(PhiMi[,t])

      ## Simulte the number of individuals not captured on
      ## occasion t-1 or previously that were alive on both
      ## occasions t-1 and t
      PhiU[t] <- sum(PhiUi[,t])

      B[t] <- N[t] - (PhiM[t] + PhiU[t])
}

## 3) Super-population size
Nsuper <- sum(B[])

##### Prior distributions #####

## Prior distributions for LVB parameters
z.inf ~ dunif(0,100)
z.rate ~ dunif(0,10)
tau ~ dgamma(.001,.001)
```

```
      sigma <- 1/sqrt(tau)

      ## Prior distributions for capture coefficients.
      mu.bp ~ dnorm(0,.0001)
      tau.bp ~ dgamma(.01,.01)
      sigma.bp <- 1/sqrt(tau.bp)

      for(t in 1:ncap){
            bp0[t] ~ dnorm(mu.bp,tau.bp)
      }
      bp1 ~ dnorm(0,.0001)

      ## Prior distributions for survival coefficients.
      mu.bphi ~ dnorm(0,.0001)
      tau.bphi ~ dgamma(.01,.01)
      sigma.bphi <- 1/sqrt(tau.bphi)

      for(t in 1:ncap){
            bphi0[t] ~ dnorm(mu.bphi,tau.bphi)
      }
      bphi1 ~ dnorm(0,.0001)
}
```