Bayesian analysis of mark-recapture data with travel-time-dependent survival probabilities

Saman MUTHUKUMARANA, Carl J. SCHWARZ and Tim B. SWARTZ

Key words and phrases: Bayesian analysis; latent variables; mark-recapture; salmon; WinBUGS. 

MSC 2000: Primary 62P10; secondary 62F15.

Abstract: The authors extend the classical Cormack-Jolly-Seber mark-recapture model to account for both temporal and spatial movement through a series of markers (e.g. dams). Survival rates are modelled as a function of (possibly) unobserved travel times. Because of the complex nature of the observed likelihood, a Bayesian approach is used based on the complete data likelihood, and integration of the posterior is accomplished through Markov chain Monte Carlo methods. The model is tested against simulated data and is also applied to actual salmon data arising from the Columbia river system. The methodology has been developed for use by the Pacific Ocean Shelf Tracking (POST) project.

Analyse bayesienne de données de capture-recapture a l’aide de probabilités de survie dépendantes

Résumé : The authors extend the classical Cormack-Jolly-Seber mark-recapture model to account for both temporal and spatial movement through a series of markers (e.g. dams). Survival rates are modelled as a function of (possibly) unobserved travel times. Because of the complex nature of the observed likelihood, a Bayesian approach is used based on the complete data likelihood, and integration of the posterior is accomplished through Markov chain Monte Carlo methods. The model is tested against simulated data and is also applied to actual salmon data arising from the Columbia river system. The methodology has been developed for use by the Pacific Ocean Shelf Tracking (POST) project.

1. INTRODUCTION

The Pacific Ocean Shelf Tracking (POST) project (http://www.postcoml.org) is part of the Census of Marine Life study (http://www.coml.org). In the POST project, acoustic transmitters are surgically implanted into fish (e.g. salmon smolt). The fish are then tracked during their migration by a series of listening lines along the ocean shelf. These listening lines record the acoustic-identification of the fish that pass near the receivers and their times of passage. A complicating factor in the analysis of the POST data is that sometimes fish do not pass sufficiently close to the receiver, and hence are not detected. Data are downloaded from the listening lines and are stored in a database that can be queried by researchers.

The POST project is a combination of two types of mark-recapture experiments. In the first type of mark-recapture experiment (Lebreton, Burnham, Clobert & Anderson 1992), animals are initially marked, and then a subset of these animals are recaptured at yearly intervals (for example). In this type of experiment, there is interest in the temporal dimension of survival. For example, one may be interested in the survival rates of a species from year to year. In the POST project,
we are also interested in the temporal dimension of survival and this information is captured by measuring the passage of time between listening lines.

In the second type of mark-recapture experiment, marked fish are released, and are detected as they swim past landmarks (p. 25 of Burnham, Anderson, White, Brownie & Pollock 1987). In this type of experiment, there is interest in the spatial dimension of survival. For example, one may be interested in survival rates between particular dams. In the POST project, the listening lines are placed in fixed locations which may correspond to interest in the spatial dimension of survival.

This paper considers methods to combine both the temporal and the spatial dimensions of the problem into a single mark-recapture model. We provide a generalization of the Cormack-Jolly-Seber (CJS) model (Cormack 1964; Jolly 1965; Seber 1965). The CJS model assumes that survival probabilities between listening lines are homogeneous amongst all animals. In our generalization, the travel times of individual animals between listening lines differ amongst animals and survival may be a function of travel time. Cowen & Schwarz (2005) considered a similar problem, but assumed that survival rates between listening lines are independent of travel time. We do not make this restrictive assumption. We model survival probabilities as a function of travel times. Pollock, Bunck, Winterstein & Chen (1995) considered Kaplan-Meier estimation in a context that allows for differential survival and the possibility of relocation of some animals. We allow for non-detection as in standard mark-recapture models.

A standard likelihood approach is difficult because of the presence of multidimensional integrals. A natural way to approach this problem is via a Bayesian framework through the use of latent (unobservable) variables. A Bayesian approach via simulation avoids the maximization of likelihoods which may be problematic in high-dimensional problems. Latent variables arise in the experiment when marked fish go unobserved at listening lines. A complete data likelihood is easy to construct as it treats the latent variables as though they are observable. We then “integrate” over the complete data likelihood by obtaining a Markov chain Monte Carlo (MCMC) sample from the posterior. Brooks, Catchpole & Morgan (2000) provide a review of Bayesian methods in mark-recapture experiments.

In Section 2, we provide a detailed development of the Bayesian model. The complete data likelihood is derived where survival probabilities depend on travel times. When the distances between listening lines vary greatly, the dependence structure is clearly important. Prior distributions are then defined on the model parameters. Computation is discussed in Section 3. As the posterior distribution is complex and high-dimensional, we obtain posterior summary statistics which describe key features in the study. In particular, posterior expectations are approximated through MCMC methods using WinBUGS software (Spiegelhalter, Thomas & Best 2003). Unlike the CJS model, it is shown that nonidentifiability is not a problem for the proposed Bayesian model. In Section 4, we discuss the topics of model selection and model adequacy. In Section 5, we provide some examples and demonstrate the reliability of the approach via simulation. We also provide sensitivity analyses with respect to some of the model assumptions. We conclude with a short discussion in Section 6.

2. MODEL DEVELOPMENT

Consider a population of $n$ fish where each fish is implanted with an acoustic transmitter. Without loss of generality, assume that all fish are released at location $j = 0$, and that listening lines are set up at locations $j = 1, \ldots, m$. The observed data for the experiment consist of $(\omega, T^{\text{obs}})$ where $\omega = \{\omega_{ij}\}$ is the detection history such that

$$
\omega_{ij} = \begin{cases} 
1 & \text{if the } i\text{-th fish is detected at location } j \\
0 & \text{if the } i\text{-th fish is not detected at location } j
\end{cases}
$$

and $\omega_{i0} = 1$. The vector $T^{\text{obs}} = \{T_{ij}\}$ corresponds to observed cumulative travel times such that $T_{ij}$ is the time required for the $i$-th fish to travel from the point of release to location $j$. When
a fish is not detected, then there is no observed cumulative travel time. We refer to the missing or latent cumulative travel times as $T^{\text{mis}}$ and let $T = (T^{\text{obs}}, T^{\text{mis}})$. Therefore $T$ represents the complete cumulative travel times. Note that when a fish has died (and is therefore not detected), we still imagine that there is a cumulative travel time associated with the fish. The value is missing but it represents the cumulative travel time that the fish would have taken had it been alive.

Associated with $(\omega, T)$ are the quantities $(S^{\text{obs}}, t)$ where $S^{\text{obs}} = \{S_{ij}\}$ is a function of the detection history data $\omega$ and $t = \{t_{ij}\}$ is a function of the complete cumulative travel times $T$. The variable $S_{ij}$ denotes the survival status of the $i$-th fish at location $j$ where $S_{ij} = 1$ (0) indicates that the $i$-th fish is alive (dead) at location $j$. Since fish are released alive, we have $S_{i0} = 1$. Note that whereas the entire vector $\omega$ is observed, some of the entries $S_{ij}$ are latent. This is due to the fact that an undetected fish may be either alive or dead. As an example, consider the observed data $(\omega_{i0}, \ldots, \omega_{i5}) = (1, 0, 0, 1, 0, 0)$. In this case, $(S_{i0}, \ldots, S_{i3}) = (1, 1, 1, 1)$ but $S_{i4}$ and $S_{i5}$ are latent. We supplement the observed $S^{\text{obs}}$ with the missing or latent $S^{\text{mis}}$ to give the complete survival history $S = (S^{\text{obs}}, S^{\text{mis}})$. The variable $t_{ij} = T_{ij} - T_{i,j-1}$ denotes the interval travel time for the $i$-th fish from location $j-1$ to $j$. Because some of the $T_{ij}$ may be missing, this implies that some of the $t_{ij}$ may be missing. In fact, there are at least as many missing $t_{ij}$’s as there are missing $T_{ij}$’s. As an example, consider $(T_{i0}, T_{i1}, T_{i2}, T_{i3}, T_{i4}, T_{i5}) = (0, x, NA, NA, y, z)$ where NA denotes “Not Available”. Then $(t_{i0}, t_{i1}, t_{i2}, t_{i3}, t_{i4}, t_{i5}) = (0, x, NA, NA, NA, z - y)$. Therefore, the vector $t$ consists of both observed and latent data. Missing data issues have been previously considered in mark-recapture experiments. For example, Bonner & Schwarz (2006) showed how the classical CJS model can be extended for time-dependent individual covariates which form a set of missing values when animals are unobserved. Dupuis (1995) used directed graphs for the Bayesian analysis of mark-recapture experiments of the first type (Lebreton et al. 1992).

We now describe the two primary parameters of interest in the model. We let $p_j$ denote the probability of detection at the $j$-th location. As the acoustic transmitters are identical and the fish comprise a sample from an underlying population, one typically assumes that the probability $p_j$ does not depend on fish $i$. In some instances, it may be reasonable to assume a common probability of detection (i.e. $p_j = p$ for all locations) although the general case causes no additional difficulty.

The second parameter of interest concerns survival where $\phi_{ij}$ denotes the survival probability of the $i$-th fish when travelling from location $j-1$ to location $j$ given that the fish was alive at location $j-1$. In Cowen & Schwarz (2005), the modelling assumption $\phi_{ij} = \phi_j$ implies that survival probabilities are independent of travel times. In our paper, we consider $\phi_{ij} = f(t_{ij})$ where $f$ is a specified decreasing parametric function. Therefore, the longer that it takes a fish to travel between locations $j-1$ and $j$, the greater the chance that the fish does not survive. In our datasets, travel times are measured in days, and we define $\phi_{ij} = q_j^{t_{ij}}$ such that $q_j$ denotes the daily survival probability when travelling between locations $j-1$ and $j$. Our modelling assumption implies that survival is independent across days. Therefore, the proposed framework reduces the primary parameters of interest to $(p, q)$ where $p = \{p_j\}$ and $q = \{q_j\}$. Gimenez, Crainiceanu, Barbraud, Jenouvrier & Morgan (2006) consider a penalized spline approach when modelling survival probabilities in a semi-parametric fashion.

In Cowen & Schwarz (2005), an observed likelihood is obtained based on the observed data $(\omega, T^{\text{obs}})$. The observed likelihood is complex as it involves integrals with respect to the latent cumulative travel times $T^{\text{mis}}$. We take an approach based on the complete data likelihood as in van Deusen (2002). The complete data likelihood treats latent variables as though they are available, and is especially well suited to Bayesian analyses (as will be seen). An advantage of the complete data likelihood over the observed likelihood is that it has a much simpler form. In our approach, we develop the complete likelihood based on $(\omega, S, t)$.

In obtaining the complete data likelihood, let $[A | B]$ generically denote the density function or probability mass function (pmf) corresponding to $A$ given $B$. In addition, let $\omega_i = (\omega_{i0}, \ldots, \omega_{im})$, where
$S_i = (S_{i0}, \ldots, S_{im})$, and $T_i = (T_{i0}, \ldots, T_{im})$. Then the complete data likelihood is given by

$$[\omega, S, T] = \prod_{i=1}^{n} [\omega_i, S_i, T_i] \prod_{i=1}^{m} [\omega_i | S_i, T_i] [S_i, T_i] \prod_{i=1}^{m} [\omega_i | S_i, T_i] [S_i | T_i] [T_i]$$

(1)

where the independence of fish is assumed and the expressions in (1) are based on conditional probability. The complete data likelihood for the $i$-th fish is therefore the product of three terms; the conditional pmf of detection history given survival and travel time history, the conditional pmf of survival history given travel time history and the travel time density. We now derive expressions for each of the three terms in the product (1). The first term is given by

$$[\omega_i | S_i, T_i] = [\omega_i | S_i] = \prod_{j=1}^{m} [\omega_{ij} | S_{ij}] = \prod_{j=1}^{m} (p_{ij}^{\omega_{ij}} (1 - p_j)^{1 - \omega_{ij}})^{S_{ij}}$$

(2)

where the key assumption in (2) is that detection at location $j$ does not depend on other locations, and we note that when a fish dies (i.e. $S_{ij} = 0$), then detection is impossible and there is no contribution to the complete data likelihood. Now

$$[S_i | T_i] = [S_{im} | S_{i0}, \ldots, S_{i,m-1}, T_i] [S_{i,m-1} | S_{i0}, \ldots, S_{i,m-2}, T_i] \cdots [S_{i1} | S_{i0}, T_i]$$

$$= \prod_{i=1}^{m} [S_{ij} | S_{i,j-1}, T_{ij-1}, T_j]$$

$$= \prod_{j=1}^{m} (\phi^{S_{ij}}_j (1 - \phi_{ij})^{1 - S_{ij}})^{S_{ij-1}}$$

$$= \prod_{j=1}^{m} (q_{ij}^{t_{ij}} (1 - q_j^{t_{ij}})^{1 - S_{ij}})^{S_{ij-1}}$$

(3)

where $t_{ij} = T_{ij} - T_{ij-1}$ and there is no survival contribution to the likelihood when a fish has already died (i.e. $S_{ij-1} = 0$). Putting (1), (2) and (3) together, we have the complete data likelihood

$$[\omega, S, T] = \prod_{i=1}^{n} [T_i] \prod_{j=1}^{m} (p_{ij}^{\omega_{ij}} (1 - p_j)^{1 - \omega_{ij}})^{S_{ij}} (q_{ij}^{t_{ij}} (1 - q_j^{t_{ij}})^{1 - S_{ij}})^{S_{ij-1}}.$$
Diffuse distributions are appealing in that they allow the data to drive the inference. Referring to (4) and (5), we consider the prior density
\[ [p, q, \mu, \Sigma] = [p] [q] [\mu] [\Sigma] \] (6)
where prior independence is assumed. As the \(p\)'s and \(q\)'s are probabilities defined on the simplex, it is customary to assume Beta priors. Specifically, we assume independent detection probabilities \(p_j\) where
\[ [p_j] \propto p^a_j (1 - p_j)^{b - 1} \]
and independent daily survival probabilities \(q_j\) where
\[ [q_j] \propto q^a_j (1 - q_j)^{b - 1}. \]
The \(a\)'s and the \(b\)'s may be pre-specified based on one's subjective understanding of the listening devices and the daily survival rates. We impose a diffuse Normal \(m(0, \sigma_\mu I)\) prior for the mean log travel time distribution \([\mu]\) where \(\sigma_\mu\) is set large and the normal distribution is constrained according to \(\mu_1 \leq \cdots \leq \mu_m\). We set \(\Sigma^{-1} \sim \text{Wishart}((1/m)I, m)\). Having specified the complete data likelihood (see (4) and (5)) and the prior (6), the ingredients for a Bayesian analysis have been determined.

3. COMPUTATIONS

We re-express the complete data likelihood \([\omega, S, t]\) in (4) as \([X_{obs}, X_{mis} | p, q, \mu, \Sigma]\) to emphasize the dependency on the unknown parameters and to emphasize that \((\omega, S, t)\) consists of both observed and missing values. The Bayesian paradigm then gives the following expression for the posterior
\[ [p, q, \mu, \Sigma | X_{obs}] \propto [X_{obs} | p, q, \mu, \Sigma] [p, q, \mu, \Sigma] \]
\[ = \int [X_{obs}, X_{mis} | p, q, \mu, \Sigma] [p, q, \mu, \Sigma] dX_{mis}. \] (7)

In theory, the posterior density (7) provides a complete description of the uncertainty in the parameters defined in the mark-recapture experiment. However, the dimensionality and the complexity of (7) is such that it is impossible to gain any meaningful insight. Alternatively, we consider the following expression
\[ [p, q, \mu, \Sigma, X_{mis} | X_{obs}] \propto [p, q, \mu, \Sigma, X_{obs}, X_{mis}] \]
\[ \propto [X_{obs}, X_{mis} | p, q, \mu, \Sigma] [p, q, \mu, \Sigma] \] (8)
where the last expression in (8) is the product of the complete data likelihood and the prior density which are familiar and simple forms.

Therefore, if we are able to sample variates \((p, q, \mu, \Sigma, X_{mis})\) from (8), then we can use the sampled components \((p, q, \mu, \Sigma)\) as realizations from the posterior distribution. However, sampling directly from (8) is a difficult/impossible task, and instead, a Markov chain is constructed which has the posterior as its stationary distribution. Fortunately, this is easily implemented using WinBUGS software (Spiegelhalter, Thomas & Best 2003). In WinBUGS, the user needs only to specify the form of the complete data likelihood, the prior and the observed data. WinBUGS then produces an appropriate Markov chain. The user may then proceed with the Markov chain output as seen fit. For example, output may be averaged to provide estimates of posterior expectations and marginal posterior densities may be approximated using density estimation techniques. We may even obtain posterior expectations of latent variables. Note that whereas classical methods (e.g. estimation and testing) often rely on asymptotic distributions of statistics, a sample from the
posterior is a sample from the distribution of interest. An overview of the use of MCMC methods is provided in the edited text by Gilks, Richardson & Spiegelhalter (1996). Carlin & Louis (2000) and Gelman, Carlin, Stern & Rubin (2003) provide further information on MCMC and give modern accounts of the Bayesian approach to statistics. A detailed description of the WinBUGS code for the POST project is given in Muthukumarana (2007). There are different versions of WinBUGS and only the full version 1.4.1 was able to handle the complexity of our model. When a model is not too complex, WinBUGS makes use of the Gibbs sampling algorithm to generate a Markov chain. With our model, the full conditional distributions required for Gibbs sampling have non-standard forms. In this case, WinBUGS makes use of the Metropolis-Hastings algorithm where Markov chain output contains streams of duplicate values due to the acceptance/rejection step in the algorithm. We note that conditional on \( S_{ij} \) and \( S_{i,j-1} \), there are two Bernoulli terms in the complete data likelihood. More specifically, we can express (4) as

\[
[\omega, S, T] = \prod_{i=1}^{n} \left[ T_i \right] \prod_{j=1}^{m} (S_{ij}p_j)^{\omega_{ij}} (1 - S_{ij}p_j)^{1-\omega_{ij}} (S_i,j-1q_j^{1j})^{S_{ij}} (1 - S_{i,j-1}q_j^{1j})^{1-S_{ij}}.
\]

The recognition of this fact enables a simpler expression for the complete data likelihood and dramatically reduces the computational time. We also note that our model contains constrained distributions for the log\((T_{ij})\) and the \( \mu_j \) variables. As constraints are not a standard feature of WinBUGS, we overcome this difficulty through the use of indicator variables. Using this approach, it is not necessary to determine the norming constant for the constrained distribution.

WinBUGS coding can initially be difficult; we hope that our code provides a beginning template for future Bayesian analyses in mark-recapture. We also note that some pre-processing was required for the data considered in Section 5.2. It was necessary to extract \((\omega, S)\) from \(T_{obs}\) prior to running WinBUGS. R code developed for the pre-processing stage is provided in Muthukumarana (2007).

3.1 Nonidentifiability

It is well known that final survival and capture rates are confounded in the classical CJS model as a result of nonidentifiability. To understand the problem at a deeper level, we recall that the observed likelihood in the CJS model is not the same as the complete data likelihood (4). The observed CJS likelihood differs in that it does not contain the cumulative travel-time distributions \([T_i]\) appearing in (4). Secondly, the individual survival probabilities \(\phi_{ij} = q_j^{1j}\) in (4) are replaced with the simpler probabilities \(\phi_{ij} = \phi_j\). Finally, the observed likelihood (which is difficult to write down in the general case) may be derived from the complete data likelihood (4) by summing over cases that are not directly observed. For example, the term in the complete data likelihood corresponding to the unobservable case \((S_{im} = 1, \omega_{im} = 0)\) is added to the term in the complete data likelihood corresponding to the observable case \((S_{im} = 0, \omega_{im} = 0)\). As a result, the terms \(p_m\) and \(\phi_m\) only appear in the observed data likelihood as the product \(p_m\phi_m\). The implication is that the data only allow us to learn about the product \(p_m\phi_m\) and not about the individual parameters \(p_m\) and \(\phi_m\). In this case, we say that \(p_m\) and \(\phi_m\) are nonidentifiable.

In general, the typical consequences of nonidentifiability include ridges in the likelihood surface and multimodality which are problematic for estimation. In a Bayesian context, these sorts of problems may not be as problematic as in the classical context. In theory, all that one needs to do in a Bayesian analysis is integrate to obtain the required posterior summaries. However, from a practical perspective, nonidentifiability still may cause problems in a Bayesian analysis. For example, MCMC algorithms may have difficulty traversing parameter spaces with elongated likelihoods.

Swartz, Haitovsky, Vexler & Yang (2004) have demonstrated that the use of informative (ie. non-diffuse) priors may be effective in mitigating the effects of nonidentifiability in Bayesian models. In our application, we recommend the use of informative priors particularly for the detection
probabilities \( p_j \). There is often good prior knowledge concerning the capabilities of the listening lines and the acoustic transmitters.

Keeping in mind the above discussion, there is an appealing by-product of the modelling assumption \( \phi_{ij} = q_{ij}^t \) with respect to nonidentifiability. In the CJS model, if one replaces \( \phi_{ij} = \phi_j \) with \( \phi_{ij} = q_{ij}^t \), the nonidentifiability disappears because the product \( p_m q_{im}^t \) appearing in the likelihood differs over the fish \( i = 1, \ldots, n \). This result is somewhat paradoxical as the model with \( \phi_{ij} = q_{ij}^t \) is more complex than the traditional CJS model with \( \phi_{ij} = \phi_j \) yet we gain better insight regarding the individual parameters with the more complex model. In the model proposed in this paper (which is a generalization of the CJS model), it follows that there is no problem with nonidentifiability.

4. MODEL SELECTION AND MODEL ADEQUACY

The topics of model selection and model adequacy are of fundamental importance in applied statistics and these topics are becoming increasingly important with the consideration of more complex models. However, in Bayesian statistics there are a myriad of approaches that have been proposed for both model selection and assessing model adequacy, and it is fair to say that there is no consensus on the “correct” approach to either of these problems. In this section, we provide some general remarks on model selection and model adequacy, and we provide some concrete suggestions that are relevant to the problem at hand.

4.1 Model selection

A principled Bayesian approach for comparing a finite number of competing models involves the calculation of the posterior probabilities of the models. When equal prior probabilities are assigned to each of the models, then the posterior comparison of two models (\( i \) and \( j \)) reduces to a study of the Bayes factor

\[
B_{ij} = \frac{\int f_i(x | \theta_i) \pi_i(\theta_i) \, d\theta_i}{\int f_j(x | \theta_j) \pi_j(\theta_j) \, d\theta_j} \tag{9}
\]

where \( f_i(x | \theta_i) \) is the likelihood of model \( i \) with parameter \( \theta_i \) and \( \pi_i(\theta_i) \) is the prior density corresponding to parameter \( \theta_i \). When the Bayes factor \( B_{ij} \) is greater (smaller) than 1, this provides evidence for (against) model \( i \) relative to model \( j \). A major practical difficulty with the use of Bayes factors is the calculation of the Bayes factor \( B_{ij} \). The expression (9) can rarely be evaluated analytically and it is typical to attempt to approximate \( B_{ij} \). For example, a method that is often unstable involves the approximation of the numerator by averaging \( f_i(x | \theta_i^{(k)}) \) where \( \theta_i^{(k)} \) is the \( k \)-th iteration of \( \theta_i \) from the prior distribution. Naturally this approach presupposes a proper prior. Methods of approximation based on output from MCMC simulation have also been proposed. However, even these methods are fraught with difficulties. For example, we have experienced unstable estimation and overflow with the complex models proposed in this paper. Another serious problem with the use of Bayes factors is one of calibration when improper priors are used. An overview of Bayes factors is given by Kass & Raftery (1995).

Due to the practical difficulties with the use of Bayes factors in complex models, a number of alternative diagnostics have been proposed that are often viewed as approximations to Bayes factors. For example, the diagnostics AIC (Akaike 1973), BIC (Schwarz 1978) and DIC (Spiegelhalter, Best, Carlin & van der Linde 2002) have all received prominent attention in the literature.

Amongst the numerous model selection diagnostics, it appears that the DIC is the most widely used in WinBUGS applications. In fact, WinBUGS provides DIC values as an option in its Inference menu. Unfortunately, DIC is unavailable in WinBUGS with our model due to the complexity of the model where some of the stochastic nodes (eg. survival status \( S_{ij} \)) are discrete.
One of the main thrusts of our paper is that complex mark-recapture models can be analyzed fairly easily using WinBUGS software. Therefore we believe that it would be against the spirit of the paper to require the investigator to fit models using WinBUGS and then carry out model selection using some sophisticated procedure outside of WinBUGS. For example, the method of Laud & Ibrahim (1995) based on posterior predictive discrepancies requires variate generation from marginal distributions outside of WinBUGS. For this reason, we want a model selection approach that can be easily implemented within WinBUGS.

For model selection in our mark-recapture models, we suggest the use of the BIC diagnostic. The BIC diagnostic is a little more sophisticated than AIC as it takes sample size into account. Denote the complete data likelihood (4) by \( L(\theta) \) where \( \theta = (\omega, S, T) \) and let \( \theta^{(i)} \) represent the \( i \)-th realization of \( \theta \) from MCMC simulation. Then the BIC is approximated by

\[
BIC = p \log(n^*) - \frac{2}{N} \sum_{i=1}^{N} \log L(\theta^{(i)})
\]  

(10)

where \( p \) is the number of parameters in the model (including missing values), \( n^* \) is the number of observed data values and \( N \) is the number of MCMC simulations. A model with a smaller value of (10) is a preferred model. Note that \( p \log(n^*) \) may be viewed as a penalty term that takes the dimensionality of the model into account. Note also that the Bayesian formulation of BIC is different than the classical version which evaluates \( \log L \) at the maximum likelihood estimate rather than averaging \( \log L \) over the posterior. Again, an important feature is that (10) may be evaluated directly in WinBUGS by coding BIC. Some discussion of the use of BIC and other model selection diagnostics in Bayesian applications is given by Aitkin (1991) and DeSantis & Spezzaferri (1997).

4.2 Model adequacy

As problematic as model selection may be with complex Bayesian models, the assessment of model adequacy in complex Bayesian models is even more problematic. A possible explanation for this is that aposteriori testing of model adequacy is not a Bayesian construct and may be seen as violating the Bayesian paradigm. From the point of view of a subjective Bayesian purist, any uncertainty concerning a model ought to be expressed via prior opinion. For example, if an experimenter is unsure whether the sampling distribution of the data is normal or Student, then the uncertainty might be expressed via a mixture. In theory, if we are able to express uncertainty in a model (and this includes both the sampling model and the parameters given the sampling model), then there is no need to assess model adequacy as all possible models have been considered and our inferences are subjective. However, from a practical point of view, it is typically difficult/impossible to determine the space of possible sampling models and parameters, and to assign prior opinion to the space.

Therefore, what does the practical Bayesian do in the context of model assessment? An honest answer may be that the assessment of complex Bayesian models is not a routine activity. When Bayesian model assessment is considered, it appears that the prominent modern approaches are based on the posterior predictive distribution (Gelman, Meng & Stern 1996). These approaches rely on sampling future variates \( y \) from the posterior predictive density

\[
f(y \mid x) = \int f(y \mid \theta) \pi(\theta \mid x) \, d\theta
\]

(11)

where \( x \) is the observed data, \( f(y \mid \theta) \) is the sampling density for \( y \) and \( \pi(\theta \mid x) \) is the posterior density. In MCMC simulation, approximate sampling from (11) proceeds by sampling \( y_i \) from \( f(y \mid \theta^{(i)}) \) where \( \theta^{(i)} \) is the \( i \)-th realization of \( \theta \) from the Markov chain. Model assessment then involves a comparison of the future values \( y_i \) versus the observed \( x \). One such comparison involves the calculation of posterior predictive p-values (Meng 1994).
A major difficulty with posterior predictive methods concerns double use of the data. Specifically, the observed data $x$ is used both to fit the model giving rise to the posterior density $\pi(\theta \mid x)$ and then is used in the comparison of $y_i$ versus $x$. For this reason, some authors prefer a cross-validated approach (Gelfand, Dey & Chang 1992) where the data $x$ and then is used in the comparison of $y$ versus $x$ respectively, the observed data $x$ versus $y$ that is similar to the posterior predictive methods but instead samples “model variates” $y$ from the prior predictive density

$$f(y) = \int f(y \mid \theta) \pi(\theta) \, d\theta$$

(12)

where $\pi(\theta)$ is the prior density. This approach was advocated by Box (1980) before simulation methods were common. Note that generating from (12) presupposes proper priors which are required in WinBUGS.

It is not difficult to write R code to simulate $y_1, \ldots, y_N$ from the prior predictive density in (12). It is then a matter of deciding how to compare the $y_i$’s against the observed data $x$. We advocate simple comparisons that are of direct interest to the application. For example, one might compare the mean observed cumulative travel time $\bar{T}_2 = \sum_{i=1}^{n} T_{ij}/n$ at the second listening line to the histogram formed by the $N$ variates $\bar{T}_2$ obtained from the prior predictive simulation. Naturally, as the priors become more diffuse, it becomes less likely to find evidence of model inadequacy. We investigate the assessment of model adequacy on the Columbia river data in Section 5.2.

5. EXAMPLES

5.1 Simulated data

Various simulation studies were carried out. We report on one such simulation. A dataset corresponding to $n = 500$ fish with $m = 5$ listening lines was simulated using R code. Detection probabilities at each listening line were set to $p_j = p = 0.8$ while daily survival probabilities between listening lines were set to $q_j = q = 0.99$, $j = 1, \ldots, m$. The logarithms of the cumulative travel times between the listening lines were generated from the constrained multivariate normal distribution (5) with $\mu = (1, 2, 3, 4, 5)'$ and $\Sigma = (\sigma_{ij})$ where $\sigma_{ii} = 1.0$ and $\sigma_{ij} = 0.2$ for $i \neq j$. We then generated a survival history $S$ based on $S_{ij} \sim \text{Bernoulli}(q_{ij})$ and a detection history $\omega$ based on $\omega_{ij} \mid S_{ij} = 1 \sim \text{Bernoulli}(p_j)$. Having generated the data as described, the data were partitioned into its observed and missing components of $(\omega, S, T)$; this is necessary for the construction of the complete data likelihood (4). In the simulated data, there are $5(500) = 2500$ cumulative travel times $T_{ij}$ of which 858 are missing. Therefore we have considered a challenging test case with a large proportion of missing values. At the fifth listening line, a typical survival probability is $\phi_{55} = q_{5}^{\exp(T_{55})-\exp(T_{54})} = (0.99)^{\exp(5)-\exp(4)} \approx 0.39$.

For the analysis of the simulated data, we first consider the “full model” which contains all of the parameters described in the paper. Uniform prior distributions for the parameters $p_j$ and $q_j$ were assigned according to $p_j \sim \text{Beta}(1.0, 1.0)$ and $q_j \sim \text{Beta}(1.0, 1.0)$, $j = 1, \ldots, 5$ with independence across the distributions. The specified priors provide a good test of the robustness of the methods with respect to the priors as the corresponding prior means $E(p) = E(q) = 1/2$ are not close to the pre-set parameter values. The cumulative travel times are modelled as in (5) and the rest of the prior settings are given as described in Section 2. The model was fit using
WinBUGS software where 500 iterations were used for the burn-in period. The posterior estimates in Table 1 were based on 8000 iterations which required approximately 15 minutes of computation on a personal computer. We observe that the posterior means of the primary parameters $p_j$ and $q_j$ are close to the pre-set values. The posterior means of the secondary parameters $\mu$ and $\Sigma$ also appear in agreement with the pre-set values.

We now wish to investigate aspects of the model selection diagnostic BIC in (10). With simulated data, we can investigate the diagnostic since we know the true model from which the data were generated. For the full model considered above with the uniform priors, we obtained $BIC = 9640.3$. Knowing the way that the data were simulated, we also fit the “true model” with $\phi_{ij} = q_{ij}$ and $p_{ij} = p$ where independent uniform priors were assigned to $p$ and $q$. For the true model, we obtained $BIC = 9557.5$. Therefore the BIC diagnostic preferred the true model over the full model.

We next fit an even simpler model with $\phi_{ij} = \phi_j$ and $p_j$ which is analogous to the Cowen and Schwarz (2005) model as it does not consider survival as a function of travel time. In this case, $BIC = 9725.2$. Therefore the BIC diagnostic rightly suggests that the travel time assumption is important. In fact, the full model is preferred to the model analogous to Cowen and Schwarz (1995).

Finally, we consider the sensitivity of the analysis with respect to the travel time assumption (5). The analysis here is the same as in the full model analysis except that we assume the simpler travel time distribution $(T_{i1}, \ldots, T_{im}) \sim Normal_m(\mu, \Sigma)$ subject to the constraint $0 < T_{i1} \leq \cdots \leq T_{im}$. Under the simpler assumption, the posterior means of the $p_j$’s and $q_j$’s are comparable to the posterior means in the analysis of the full model and $BIC = 9763.3$. This suggests that even though the overall fit of the simpler model is not good (in terms of BIC), the precise shape of the distribution of the cumulative travel times is not a critical assumption in the estimation of the primary parameters of interest.

5.2 Columbia river data
The model was then fit to data obtained from the Columbia River system. From April 25/2001 to May 30/2001, \( n = 324 \) radio tagged chinook salmon were released from the Rock Island Dam. Data were recorded at listening lines established at the \( m = 3 \) dams downstream at Wanapum, Priest Rapids and Hanford Reach. For example, corresponding to fish \( j = 4 \), we have data values \( \omega_4 = (1, 1, 0, 0) \) and \( (T_{40}, T_{41}, T_{42}, T_{43}) = (0, 6.1, \text{NA}, \text{NA}) \). This implies that the fourth fish was released at the Rock Island Dam and then detected at Wanapum but was undetected at both Priest Rapids and Hanford Reach. The fish took 6.1 days to reach Wanapum from the Rock Island Dam. The data gives rise to the survival history \( S_4 = (1, 1, \text{NA}, \text{NA}) \) since we know that the fish survived up to Wanapum but it is unknown whether the fish survived up to Priest Rapids or Hanford Reach. The interdam distances are approximately 37.6 miles, 18.7 miles and 15.0 miles respectively. In the Columbia River data, there are \( 3(324) = 972 \) cumulative travel times \( T_{ij} \) of which 294 are missing. Cowen & Schwarz (2005) also studied this dataset in the context of radio failure. Here, we ignore radio failure, and therefore survival is a function of both actual survival and radio failure. In this example, we fit a model which allows for the possibility of varying detection probabilities \( p_j \) and varying daily survival probabilities \( q_j \) at each of the dams. Uniform priors were assigned to the detection and daily survival probabilities. The remaining priors are given as in Section 2. Although we have not done so, we remark that it is possible to enhance the model by stratifying the salmon according to their release date.

Estimates of the posterior means of the parameters are given in Table 2. These are based on a MCMC simulation using WinBUGS with a burn-in period of 500 iterations followed by 4000 iterations. We observe that the detection probabilities \( p_j \) are high and are similar across the dams. We note that the daily survival probabilities \( q_1, q_2 \) and \( q_3 \) decrease and this appears to make biological sense. With respect to the estimated travel-time parameter \( \mu \), we refer to Figure 1 which provides density plots of the observed travel times between dams. The average travel times between the three dams are 9.9, 3.6 and 1.9 days respectively. These values are roughly in agreement with MCMC estimates of the mean interval travel times \( (E(t_{ij}) \approx \exp(\mu_1 + \Sigma_{11}/2)) \) which are 10.3, 4.6 and 2.7 days respectively. Figure 2 provides an estimate of the posterior density of \( \mu_1 \) using a kernel smoother from WinBUGS. The plot suggests a nearly symmetric unimodal distribution as might be expected. We observe strong positive correlations in the \( \Sigma \) matrix; this is expected as we are modelling cumulative travel times.

It is instructive to look at some plots related to the MCMC simulation. A trace plot for \( \mu_1 \) is given in Figure 3. The trace plot appears to stabilize immediately and hence provides no indication of lack of convergence in the Markov chain. In Figure 4, an autocorrelation plot for \( \mu_1 \) is also provided. The autocorrelations appear to dampen quickly. This provides added evidence of the convergence of the Markov chain and also suggests that it may be appropriate to average Markov chain output as though the variates were independent. Similar plots were obtained for all of the parameters in the model. In addition to the diagnostics described, multiple chains were obtained to provide further assurance of the reliability of the methods. For example, the Brooks-Gelman-Rubin statistic (Brooks & Gelman 1997) gave no indication of lack of convergence.

As discussed in Section 3.1, nonidentifiability poses neither a theoretical nor a practical obstacle for the model proposed in this paper. To investigate the degree of confounding between the final capture rate \( p_3 \) and the final daily survival probability \( q_3 \), we calculate the posterior correlation between \( p_3 \) and \( q_3 \) using the output from the MCMC simulation. The posterior correlation is found to be 0.02 which indicates a lack of confounding between the final capture rate and the final daily survival probability.

In assessing model adequacy, we have mentioned that a model with diffuse priors will almost...
always be viewed as adequate. The reason for this is that a model with diffuse parameters gives rise to diffuse data, and observed data is unlikely to be seen as inconsistent when compared to diffuse data. To provide a more stringent test, we consider a modification of our model where subjective priors are introduced. Since we know apriori that the listening lines and acoustic devices are of high quality, we assign prior probabilities of detection according to independent \( p_j \sim \text{Beta}(19,1) \). We also know that daily survival probabilities are high and we therefore set independent \( q_j \sim \text{Beta}(18,2) \). The prior distribution for the log cumulative travel times is a constrained multivariate normal distribution with mean vector and variance covariance matrix given by the posterior means from the initial analysis. This distribution is roughly consistent with the data and is far less diffuse. With this model, the posterior means of the primary parameters are very close to estimates provided in Table 2. This suggests that the data are informative and are dominating the inferences. To assess model fit, we generated 20 datasets from the prior predictive density (12) according to the model described above, and compared the generated data with the observed data. Although there are many features of the data that could be checked, we focus on a study of detection history \( \omega \) and cumulative travel time \( T \). For \( \omega \), Figure 5 provides a histogram of the proportion of fish detected in the simulated datasets. For each simulation, the proportion detected is given by \( \sum_{i=1}^{324} \sum_{j=1}^{3} \omega_{ij} / (924) \). The proportion detected for the observed data is 0.70 which appears consistent with the model. For \( T \), there appears to be no limit on the number of features that one may check. For illustration we consider the travel time to the first dam. Figure 6 provides a boxplot for each generated dataset using the 324 total travel times \( T_{1i} \). The boxplot for the observed data is also included and appears to be consistent with the generated data.

To check whether the travel-time assumption is relevant to the Columbia river data, we fit a second model that is analogous to the model considered by Cowen and Schwarz (2005). We keep everything the same as in the initial model with the diffuse priors but let \( \phi_{ij} = \phi_j \) rather than...
\( \phi_{ij} = q_{ij}^{t_{ij}} \). For the Cowen and Schwarz (2005) model, we obtain BIC = 3988.8 which is much worse than the initial model with BIC = 2521.1. The extremely bad fit of the Cowen and Schwarz (2005) model may have been anticipated as Figure 1 suggests large travel time differences between the three dams.

6. DISCUSSION

The Bayesian framework provides a straightforward approach to dealing with the complex observed likelihood which requires integration over the unobservable travel times. Furthermore, the Bayesian approach provides a convenient way of estimating the correlation in travel times between sampling locations. This addresses an important biological question as to whether some fish are intrinsically faster, or whether travel times are independent random events.

A alternative approach might be based on the EM algorithm as used by van Deusen (2002). However, the computation of the expected log-likelihood is not straightforward and numerical methods would likely be needed. In our Bayesian approach, Markov chain Monte Carlo (MCMC) methods avoid the necessity of numerical integration.

Finally, our model is easily extended to allow for individual time-independent covariates such as initial body mass \( m_{ij} \) by modelling \( q_{ij} = f(m_{ij}) \) for some function \( f \). Work is underway to extend our formulation to allow for time-dependent individual covariates and to more fully investigate the choice of \( f \). The revised model must account for missing values both in the travel times and in the individual covariates.

It is not immediately clear how our model might be extended to a two-dimensional spatial setting. For example, listening lines may be set up in the woods to track the movements of animals. In this case, in addition to death and radio tag failure, lack of detection may be due to nonstandard travel paths. We consider this to be an open research problem.

ACKNOWLEDGEMENTS
The authors thank the editor and two anonymous referees for helpful comments that lead to an improvement in the paper. Muthukumarana was partially supported by a MITACS fellowship. Schwarz and Swartz were partially supported by Discovery grants through the Natural Sciences and Engineering Research Council of Canada (NSERC). The WinBUGS code and the R code used in this paper are available from the first author upon request.

REFERENCES


Figure 3: Trace plot for $\mu_1$ based on MCMC simulation in Example 5.2.


Figure 4: Autocorrelation plot for $\mu_1$ based on MCMC simulation in Example 5.2.


S. Muthukumarana (2007). “Bayesian modelling and software development for the POST project”, *MSc project, Department of Statistics and Actuarial Science, Simon Fraser University*.


Figure 5: Histogram of the proportion of fish detected for the observed data and the 20 generated datasets in Example 5.2.

Figure 6: Boxplots of the travel time $T_{i1}$ to the first dam for the observed data and the 20 generated datasets in Example 5.2.